ORIGINAL ARTICLE

QTL mapping in outbred half-sib families using Bayesian model selection

M Fang^{1,2,3}, J Liu^{1,3}, D Sun¹, Y Zhang¹, Q Zhang¹, Y Zhang¹ and S Zhang¹

¹Department of Animal Genetics and breeding, Key Laboratory of Animal Genetics and Breeding of Ministry of Agriculture, National Engineering Laboratory for Animal Breeding, College of Animal Science and Technology, China Agricultural University, Beijing, China and ²Department of Life Science, Heilongjiang Bayi Agriculture University, Daqing, China

In this article, we propose a model selection method, the Bayesian composite model space approach, to map quantitative trait loci (QTL) in a half-sib population for continuous and binary traits. In our method, the identity-by-descent-based variance component model is used. To demonstrate the performance of this model, the method was applied to map QTL underlying production traits on BTA6 in a Chinese half-sib dairy cattle population. A total of four QTLs were detected, whereas only one QTL was identified using

the traditional least square (LS) method. We also conducted two simulation experiments to validate the efficiency of our method. The results suggest that the proposed method based on a multiple-QTL model is efficient in mapping multiple QTL for an outbred half-sib population and is more powerful than the LS method based on a single-QTL model.

Heredity (2011) **107**, 265–276; doi:10.1038/hdy.2011.15; published online 13 April 2011

Keywords: Bayesian model selection; variance component model; Markov chain Monte Carlo algorithm; QTL mapping

Introduction

Paternal half-sib families are widely used as resource population for quantitative trait loci (QTL) mapping. Many methods have been proposed for such designs. Georges *et al.* (1995) developed a maximum likelihood method for a single-family design and implemented it to map QTL for milk production traits in the US Holstein population. A regression method of interval mapping proposed by Knott *et al.* (1996) has been commonly used to map QTL in half-sib families, particularly in dairy cattle (Spelman *et al.*, 1996; Zhang *et al.*, 1998; Heyen *et al.*, 1999; Velmala *et al.*, 1999; Nadesalingam *et al.*, 2001; Plante *et al.*, 2001; Ron *et al.*, 2001; Freyer *et al.*, 2002). Grignola *et al.* (1996a,b) proposed a restricted maximum likelihood method for half-sib families, which was implemented by a number of investigators (for example, Zhang *et al.*, 1998; Freyer *et al.*, 2002; Liu *et al.*, 2004). However, all of these studies are based on a single-QTL model, resulting in estimation bias due to the polygenic nature of quantitative traits. Zeng (1994) proposed a composite interval mapping approach, in which the effects of the markers that do not bracket the tested QTL are treated as a cofactor to absorb the effects of other QTL. This method was successfully used by De Koning *et al.* (2001) in a half-sib population. Although composite interval mapping can greatly improve the precision of

QTL position estimation, it is essentially a single-QTL model. Kao *et al.* (1999) developed a multiple interval mapping approach, in which multiple QTL are simultaneously included in the model and the suitable multiple-QTL model is determined by sequential stepwise selection. Sequential stepwise selection is not optimal because of the dynamic changes of the null hypothesis, such that its efficiency is significantly influenced by the quality of the data (Raftery *et al.*, 1997; Gelman *et al.*, 2004).

With the development of the Bayesian model selection approach, the issues in multiple-QTL mapping have been solved both for inbred line crosses (Satagopan and Yandell, 1996; Heath, 1997; Uimari and Hoeschele, 1997; Sillanpää and Arjas, 1998; Stephens and Fisch, 1998; Gaffney, 2001; Xu, 2003; Yi *et al.*, 2003, 2005) and for outbred populations (Meuwissen and Goddard, 2004; Liu *et al.*, 2007; Yi and Xu, 2000). The stochastic search variable selection algorithm was first developed by George and McCulloch (1993) and used by Meuwissen and Goddard (2004) to fine map multiple QTL in outbred half-sib families. This method can simultaneously handle multiple QTL and also keeps model dimensions fixed. However, all of these authors assume that each marker interval contains one QTL and that the QTL position is fixed at the middle of marker, so that it cannot be applied for dense markers because of increased computational burden and correlations among potential QTLs. Liu *et al.* (2007) used the reversible jump Markov chain Monte Carlo (RJMCMC; Green, 1995) method for multitrait mapping of multiple QTL. The two methods described above are based on variance component models and use the efficient Gibbs sampler to first sample all QTL substitution effects and then update QTL variance using the sampled substitution effects. Because the parameter

Correspondence: Dr D Sun and Professor Q Zhang, Department of Animal Genetics and breeding, China Agricultural University, 2 Yuanmingyuan West Road, Beijing 100193, China.
E-mails: sundx@cau.edu.cn and qzhang@cau.edu.cn

³These authors contributed equally to this work.

Received 19 November 2009; revised 24 April 2010; accepted 14 June 2010; published online 13 April 2011

to estimate is large (including QTL substitution effect and variance), it converges slowly and has poor mixing.

The potential concern with RJMCMC is that it often has poor mixing due its variable dimensions (Yi, 2004; Wang *et al.*, 2005; Liu *et al.*, 2007; Banerjee *et al.*, 2008) and thus is not easy to implement (Banerjee *et al.*, 2008). To overcome this shortcoming of the RJMCMC algorithm, Godsill (2001) conceived a Bayesian composite model space approach. His method keeps the model dimensions fixed so that it mixes rapidly, and was used by Yi (2004) in QTL mapping for an allelic substitution model. In our previous work, we extended the Bayesian composite model space approach to a variance component model (Fang *et al.*, 2009). This method also directly estimates variance components with the M-H algorithm, as Yi and Xu (2000) did, but the model dimension was unchanged. The efficiency of the method was illustrated with an outbred full-sib population and a series of simulated data. In dairy cattle, the half-sib families' design is often used for QTL mapping. It is of interest to apply this method to multiple-QTL mapping in dairy cattle. In the present study, we modified this method for half-sib families for continuous traits as well as binary traits. The method optimally updates the covariance effect, polygenic variance and residual variance, so that the computational speed and efficiency are greatly improved compared with our previous method. Finally, we applied this method to analyse real data and a series of simulated data to fully assess its performance.

Bayesian modelling of multiple QTLs

Bayesian composite model space method

In a half-sib design, the parents of the half-sib families are randomly sampled from a large outbred population that is in Hardy-Weinberg and linkage equilibrium. Let $\mathbf{y}$ be an $n \times 1$ phenotypic vector, where n is the number of phenotypic observations. If $\mathbf{y}$ is continuously distributed, the multiple-QTL model can be expressed as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{a}\boldsymbol{\Gamma} + \mathbf{g} + \mathbf{e}, \quad (1)$$

where $\boldsymbol{\Gamma} = (\gamma_1, \dots, \gamma_j, \dots, \gamma_L)'$ is the maximal QTL number, and γ_j is a binary variable indicating that the corresponding QTL effect is present ($\gamma_j = 1$) or absent ($\gamma_j = 0$) in the model; $\mathbf{a} = (\mathbf{a}_1, \dots, \mathbf{a}_j, \dots, \mathbf{a}_L)$ is an $n \times L$ matrix, and $\mathbf{a}_j$ is an $n \times 1$ random QTL effect vector and $\mathbf{a}_j \sim N(\mathbf{0}, \Theta_j \sigma_j^2)$, where σ_j^2 is the QTL variance and Θ_j is the identity-by-descent (IBD) matrix that can be inferred by the conditional expectation approach (Xu and Gessler, 1998); $\mathbf{Z}$ is an $n \times n$ identity design matrix relating records to each individual; $\boldsymbol{\beta}$ is a vector of covariate effects and $\mathbf{X}$ is the corresponding design matrix; $\mathbf{g}$ is denoted as an $n \times 1$ vector of random polygenic effect and $\mathbf{g} \sim N(\mathbf{0}, \mathbf{A}\sigma_A^2)$, where $\mathbf{A}$ is the additive relationship matrix and σ_A^2 is the polygenic additive variance; $\mathbf{e}$ is the vector of random error with the distribution $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2)$, where $\mathbf{I}$ is $n \times n$ identity matrix and σ_e^2 is the residual variance. We assume that neither QTL nor polygenes have dominant effects. Because $\gamma_j^2 = \gamma_j$, the variance component model can be expressed as

$$\text{Var}(\mathbf{y}) = \mathbf{V} = \sum_{j=1}^L \gamma_j \Theta_j \sigma_j^2 + \mathbf{A}\sigma_A^2 + \mathbf{I}\sigma_e^2. \quad (2)$$

Prior, likelihood and joint posterior

The prior of indicator variable γ_j follows a Bernoulli distribution and $p(\gamma_j = 1) = l_0/L$, where l_0 is the expected QTL number and $L = l_0 + l_0\sqrt{3}$ is the maximal QTL number (see the study by Yi *et al.*, 2005 for details); the prior for QTL variance, polygenic variance and residual variance is assumed to follow informative inverted χ^2 distributions with expressions:

$$p(\sigma_i^2 | s_i^2, \omega_i) \sim \chi^{-2}(\omega_i, s_i^2),$$

$$p(\sigma_A^2 | s_A^2, \omega_A) \sim \chi^{-2}(\omega_A, s_A^2),$$

and

$$p(\sigma_e^2 | s_e^2, \omega_e) \sim \chi^{-2}(\omega_e, s_e^2),$$

where s (the scaled parameter) and ω (the degree of freedom) are hyperparameters; the priors of covariate effects are uniformly distributed; and the QTL position follows uniform distribution across the region. We denote marker information $\mathbf{M}$, unknown variables $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma_1^2 \sim \sigma_j^2, \gamma_1 \sim \gamma_j, \sigma_A^2, \sigma_e^2\}$ and QTL position $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^L$. Accordingly, the joint posterior probability of unknowns can be expressed as

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda} | \mathbf{y}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}) \propto f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}) p(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad (3)$$

where

$$f(\mathbf{y} | \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}) \propto |\mathbf{V}|^{-1/2} \times \exp \left\{ -\frac{1}{2} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\}, \quad (4)$$

and

$$p(\boldsymbol{\theta}, \boldsymbol{\lambda}) = p(\boldsymbol{\beta}) p(\sigma_A^2) p(\sigma_e^2) \prod_{j=1}^L p(\sigma_j^2) p(\gamma_j) p(\lambda_j). \quad (5)$$

Posterior calculation

Updating model indicator

We introduced an indicator $\{\gamma_j\}_{j=1}^L$ in the variance components model, which is key to the new method. The fast M-H algorithm is used to update the indicator, which is faster than the Gibbs sampler because it can avoid calculating the denominator $f(\mathbf{y}, \gamma_j)$. Our algorithm first proposes a new value $\gamma_j = \kappa$, $\kappa \in \{0, 1\}$, on the basis of its prior probability, and then accepts with a probability equal to $\min(1, r)$, where

$$r = \frac{p(\gamma_j = \kappa) \cdot f(\mathbf{y} | \gamma_j = \kappa, \boldsymbol{\theta}_{-\phi_j}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M})}{p(\gamma_j = 1 - \kappa) \cdot f(\mathbf{y} | \gamma_j = 1 - \kappa, \boldsymbol{\theta}_{-\phi_j}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M})}. \quad (6)$$

Updating polygenic variance, residual variance and covariance effects

Joint updating strategy: To reduce the computational burden, we employed a joint updating strategy to

generate the posterior distribution of σ_A^2 , σ_e^2 and β via the efficient random walk Metropolis–Hastings (RWM–H; Metropolis *et al.*, 1953; Hastings, 1970) algorithm. Specifically, new values $\sigma_A^{2(*)}$, $\sigma_e^{2(*)}$ and $\beta^{(*)}$ are proposed, which are accepted with a probability equal to $\min(1, r)$ with

$$r = \frac{L_1}{L_2} \cdot \frac{p(\beta^{(*)})p(\sigma_A^{2(*)}|\sigma_A^2, \omega_A)p(\sigma_e^{2(*)}|\sigma_e^2, \omega_e)}{p(\beta^{(0)})p(\sigma_A^{2(0)}|\sigma_A^2, \omega_A)p(\sigma_e^{2(0)}|\sigma_e^2, \omega_e)} \cdot hr_\beta \cdot hr_A \cdot hr_e \quad (7)$$

where $L_1 = f(\mathbf{y}|\beta^{(*)}, \sigma_A^{2(*)}, \sigma_e^{2(*)}, \theta_{-(\beta, \sigma_A^2, \sigma_e^2)}, \lambda, \mathbf{A}, \mathbf{Z}, \mathbf{X}, \mathbf{M})$ and $L_2 = f(\mathbf{y}|\beta^{(0)}, \sigma_A^{2(0)}, \sigma_e^{2(0)}, \theta_{-(\beta, \sigma_A^2, \sigma_e^2)}, \lambda, \mathbf{A}, \mathbf{Z}, \mathbf{X}, \mathbf{M})$; hr_β , hr_A and hr_e are proposal ratios for β , σ_A^2 and σ_e^2 , respectively. The new proposal for the covariate effect $\beta^{(*)}$ is sampled from a multiple-dimension normal distribution with mean value $\beta^{(0)}$ and (co)variance $\mathbf{V}_t$ equal to the tuning parameter (Gelman *et al.*, 1995; Browne, 1998). As the proposal is symmetric $p(\beta^{(*)}|\beta^{(0)}, \mathbf{V}_t) = p(\beta^{(0)}|\beta^{(*)}, \mathbf{V}_t)$, the proposal ratio $hr_\beta = 1$. The new proposal variance $\sigma^{2(*)}$, including $\sigma_A^{2(*)}$ and $\sigma_e^{2(*)}$, is proposed from a scaled inverted χ^2 distribution with v degrees of freedom and a scaled parameter with the expectation of the current variance $p(\sigma^{2(*)}|\sigma^{2(0)}, v) \sim \text{Inv-}\chi^2(v, (v-2)\sigma^{2(0)}/v)$; see Browne, 1998; Fang *et al.*, 2009). Thus, the proposal ratio hr for variance component can be expressed as

$$\begin{aligned} hr &= \frac{p(\sigma^{2(0)}|\sigma^{2(*)}, v)}{p(\sigma^{2(*)}|\sigma^{2(0)}, v)} \\ &= \frac{\text{Inv} - \chi^2(v, (v-2)\sigma^{2(*)}/v)}{\text{Inv} - \chi^2(v, (v-2)\sigma^{2(0)}/v)} \\ &= \left(\frac{\sigma^{2(*)}}{\sigma^{2(0)}}\right)^{v+1} \cdot \exp\left\{\frac{(v-2)}{2} \left(\frac{\sigma^{2(0)}}{\sigma^{2(*)}} - \frac{\sigma^{2(*)}}{\sigma^{2(0)}}\right)\right\}. \end{aligned} \quad (8)$$

Optimising the tuning parameter: In RWM–H analyses, the values of the degree of freedom v and the (co)variance $\mathbf{V}_t$ may influence the performance of the RWM–H algorithm. We introduce a general way to ascertain them in order to ensure the optimal performance of RWM–H.

When a multiple-dimension normal distribution with variance $\mathbf{V}_\beta$ is simulated, the optimal tuning parameter of the RWM–H is suggested as $\mathbf{V}_t = (2.38/\sqrt{d})^2 \mathbf{V}_\beta$ (Gelman *et al.*, 1995), where d is the number of dimensions. However, in our joint sampling scheme, β is updated with the polygenic and the residual variance; hence, d equals the number of the covariates plus 2. In a Bayesian framework, the posterior distribution of β is normally distributed with variance $\mathbf{V}_\beta = (\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X})^{-1}$, where $\mathbf{V}$ is the unknown phenotypic (co)variance, which can be approximately estimated as $\mathbf{V} = \sigma_A^2 \mathbf{A} + \sigma_e^2 \mathbf{I}$, where σ_A^2 is genetic variance including polygenic and QTL variance. σ_A^2 and σ_e^2 can be determined by restricted maximum likelihood. If a binary trait is analysed, $\sigma_e^2 = 1$, and thus only σ_A^2 needs to be estimated. The degree of freedom is empirically set as $v = 60$, and the whole setting of v and $\mathbf{V}_t$ may lead the acceptance rate falling into the range of 10–40%. Such an acceptance rate range is optimal (Roberts and Rosenthal, 2001) because the

target distribution can be efficiently explored and the autocorrelation is reduced dramatically under such conditions.

The process of updating QTL variance and QTL position is described in Appendix A. In mapping QTL for a binary trait, we employ a data augmentation approach to generate the liability y , which has been illustrated at length by Yi and Xu (2000).

Applications

Real data analysis

The data from a Chinese Holstein population, with a daughter design, have been described by Chen *et al.* (2005). The population consisted of 26 bulls and their 2270 daughters, and the following five phenotypes of milk production traits were measured: milk yield (MY), fat yield (FY), protein yield (PY), fat percentage (FP) and protein percentage (PP). The bulls and daughters were genotyped at 14 marker loci on chromosome 6, covering a total distance of 55.7 cM. Marker names and their relative positions are shown in Figures 1 and 2. For each trait, the estimated breeding value (EBV) of each daughter was used as phenotypic data. The EBV was further standardised for convenience to choose the initial value (Jansen, 2003; Xu, 2003; Yi *et al.*, 2005), and the standardised EBV was subject to Bayesian analyses.

Because maternal genotypes are not available in a half-sib design and treated as missing data, they are randomly sampled from possible alleles with equal probability. The Bayes factor (BF) defined by Yi *et al.* (2005) was used to support the existence of a QTL, and the BF threshold of 3 supports a claim of significance (Kass and Raftery, 1995). To compare the new method with the least square (LS) method, the real data were also analysed with the software QTL Express (Seaton *et al.*, 2002).

In real data analyses, the hyperparameters $\omega_j = \omega_A = \omega_e = 3$ lead to a rather non-informative prior (Fang *et al.*, 2009); the expected QTL number $l_0 = 1$ and thus the maximum QTL number $L = l_0 + l_0 \sqrt{3} \approx 3$. Accordingly, the prior probability of the model indicator $p(\gamma_j = 1) = l_0/L = 0.33$; the prior distribution of QTL variance was truncated at 0.2 according to our prior information that the heritability explained by QTL is <20%; and the tuning parameter for QTL position was 2 cM. The MCMC ran for 21 000 rounds, the thinning interval was set to 10, and the burn-in period was 1000. As a result, the size of the posterior sample was 2000.

The profiles of the BF from the new method and the F -statistic of the LS method are plotted in Figure 1, and the profiles of posterior probability are presented in Figure 2. The profiles of F are the same as that in Chen *et al.* (2005) and only one QTL that affects FY was significant at the 0.01 level. However, with our method, four QTLs were detected, one affecting FY, one affecting FP and two affecting PP. Figure 1 shows that the peaks of the profiles of our method and the LS method almost overlap. The point estimates and highest posterior density regions of the QTL position from the two methods are listed in Table 1. The estimated heritabilities for the detected QTL are also given in Table 1; they are all about equal to 4%.

Figure 1 Profiles of the posterior probability (proposed method) and F -statistic (LS method) obtained from real data analysis. The black line and light grey line represent the BF score under the maximum QTL number $L = 3$ and $L = 2$, respectively, and the dark grey line represents the F -statistic from the LS method. The grey dashed horizontal line represents the threshold value of the LS method, and the black dashed horizontal line represents the standard threshold 3.2 of BF score. Figures **a–e** are the results of the five traits MY, FY, PY, FP and PP.

The acceptance rate of the M–H algorithm for the joint updating of the covariance effect, the polygenic variance and the residual variance was 14.3% for MY, 15.5% for FY, 14.2% for PY, 18.9% for FP and 17.5% for PP and varied from 10 to 40%. The histograms of posterior distributions are plotted in Supplementary Figures S1–S4 of the Supplementary Material, and a typical output is shown in Figure 3, demonstrating that posterior distributions of the covariance effect, polygenic variance and

residual variance are well explored. The results suggest that the new joint updating algorithm is more efficient.

To test the sensitivity for the maximum QTL number (or expected QTL number), we also set the expected QTL number $l_0 = 0.73$, leading to the maximum QTL number $L = l_0 + l_0\sqrt{3} = 2$ and thus the prior probability of the model indicator $p(\gamma_j = 1) = l_0/L = 0.37$. The profiles of the BF are also plotted in Figure 1 and are similar to that from $l_0 = 1$. The profiles of posterior probability are also

similar to that from $l_0 = 1$ (data not shown). The results show that the new method is not highly sensitive to L (or l_0).

Simulation studies

We conducted two simulation experiments to further evaluate the performance of the method. In the first experiment, we simulated 20 independent half-sib families, each with 100 individuals, so that the final sample size was 2000. The parents of each family were randomly sampled from a large outbred population in linkage equilibrium and Hardy–Weinberg equilibrium. Eleven evenly spaced markers, each having six alleles with equal frequency, covered a 100 cM chromosome region with an average marker interval of 10 cM. Two QTL were simulated at positions 15 cM and 75 cM. The additive variances of the two QTL were $\sigma_1^2 = 0.3$ and $\sigma_2^2 = 0.4$, and the dominant variances were assumed to be absent. The QTL additive effects of parents were randomly sampled from $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$. The alleles of the offspring at markers and QTL were inherited from their parents following Haldane's recombination rule. The polygenic variance was set to $\sigma_A^2 = 1.5$. The residual error was sampled from a normal distribution with mean 0 and variance $\sigma_e^2 = 1.5$. The overall population mean was set as 0. The phenotypic values of each offspring were the sum of the population mean, QTL effect, polygenic

effect and residual error. Accordingly, the heritabilities explained by the two QTL were 8.1 and 10.8%. In this experiment, we only investigated the performance of the continuous trait, because the heritability of the simulated QTL is so low that neither the proposed method nor the LS method can give a precise estimate for a binary trait.

In the analysis, the expected QTL number $l_0 = 2$, leading to the maximum QTL number $L = 5$, and thus $p(\gamma_j = 1) = l_0/L = 0.4$. The tuning parameter for QTL position was 2 cM. The thinning interval for MCMC was 10, the burn-in period was 1000 and the length of the complete chain was 21 000.

The BF from our method and the F -value from the LS method are plotted in Figure 4a and the posterior probabilities are in Figure 4b. With the new method, both simulated QTL were successfully detected. The parameter estimates are listed in Table 2, but only one QTL was found by the LS method at a 0.01 significance level. The results suggest that the new method is more powerful than the LS method. The parameter estimates are given in Table 2, and they are mostly close to their true values. The acceptance rate for the joint update of covariance, polygenic variance and residual variance was 23.3%, which also reflected the efficiency of our developed joint-updating strategy.

To test the sensitivity for the maximum QTL number, we set the expected QTL number $l_0 = 1$, leading to the maximum QTL number $L = 3$ and thus $p(\gamma_j = 1) = l_0/L = 0.33$. The results are plotted in Figures 4a and b and show that the profiles from $L = 5$ and $L = 3$ have very little difference.

To further compare the performance of the proposed method and the LS method, a receiver-operating characteristic was used to summarise the true positive rate (tp rate) and false positive rate (fp rate; Fawcett, 2006). Because of the high computational intensity and the difficulties in summarising the results of our method, we only replicated 30 experiments for the proposed method and the LS method. The threshold was divided on average into 20 points from the lowest threshold, which results in the maximum tp rate and minimum fp rate, to the highest threshold, which results in the minimum tp rate and maximum fp rate. In our study, the 20 successive thresholds ranged from 0.7 to 36 for the proposed method and from 1.7 to 7.3 for the LS method. At each threshold, the tp rate and fp rate were summarised for the two simulated QTLs together and are plotted in Figure 5. In our study, the maximum tp rate is not 1 because sometimes both methods failed to generate a clear signal of a QTL at the simulated position; thus, the QTL was treated as missing. As pointed out by

Figure 2 Profiles of the posterior probability from real data analysis using our method.

Table 1 The parameter estimates for the milk production traits using the proposed method

Trait	Position (cM)	HPD region	Posterior probability (%)	σ_j^2	h^2 (%)	σ_A^2	σ_e^2
Milk yield	—	—	—	—	—	0.463 (0.018)	0.018 (0.008)
Fat yield	28 (1.28)	(26,36)	98.66	0.030 (0.018)	4.57	0.598 (0.030)	0.028 (0.014)
Protein yield	—	—	—	—	—	0.444 (0.019)	0.018 (0.007)
Fat percentage	14 (2.22)	(9,22)	99.02	0.037 (0.018)	4.52	0.736 (0.040)	0.045 (0.026)
Protein percentage	32 (3.45)	(27,38)	98.76	0.046 (0.029)	4.45	0.708 (0.072)	0.233 (0.062)
	49 (2.56)	(42,52)	97.64	0.047 (0.028)	4.55		

Abbreviation: HPD, highest posterior density.
s.d. are given in parentheses.

Figure 3 Profiles of the posterior probabilities of population mean (a), polygenic variance (b) and residual variance (c) from real data analysis for the trait MY.

Figure 4 Profiles of test statistics obtained from the simulated single-chromosome data. (a) Profiles of the posterior probability for $L = 5$ and $L = 3$ from the new method. (b) Profiles of BF score from the proposed method for $L = 5$ and $L = 3$, and profile of the F -statistic from the LS method. The true positions of the simulated QTL are represented by upward arrows. The grey dotted line and black dotted line present the threshold of the LS method at a 0.01 significance level and the BF score 3.2 of the new method, respectively.

Table 2 The parameter estimates in the first simulated experiment

QTL no.	True parameter		Bayes estimates							LS estimates	
	Position (cM)	σ_j^2	Position (cM)	HPD region	Posterior probability (%)	σ_j^2	h^2 (%)	σ_A^2	σ_e^2	Position (cM)	Confidence interval
1	15	0.3	14 (2.98)	(3,20)	98.8	0.45 (0.19)	9.25	1.67 (0.57)	1.02 (0.41)	12	(0.40)
2	75	0.4	74 (4.81)	(63,87)	99.1	0.32 (0.16)	13.0			—	—

Abbreviations: HPD, highest posterior density; LS, least square; QTL, quantitative trait loci.

Figure 5 Receiver-operating characteristic (ROC) graph obtained from 30 replicated simulation experiments for the new method (grey line) and the LS method (black line).

Fawcett (2005), a point in receiver-operating characteristic space is better than another if it is to the northwest (tp rate is higher, fp rate is lower or both). Accordingly, our method always outperforms the LS method, as seen in Figure 5. Specifically, we summarised the results at the BF threshold of 3.0 for our method and the F threshold of 0.01 for the LS method. The statistical powers were 0.833 and 0.717 for our method and the LS method, respectively, and the fp rates were 0.033 and 0.067, showing that our method had a higher tp rate and fp rate than the LS method. The average estimates of QTL position over 30 replications were 16 (6.09) and 77.13 (6.95) for the first and second QTL using our method, and 22.25 (5.56) and 78.39 (7.21) with the LS method, where the values in parentheses are the s.d. estimates. The position estimation for the first QTL with our method is closer to the true value of 15 than that obtained using the LS method. An explanation of the results is that a multiple model is used in our method, so it tends to give less-biased estimates for QTL position. The estimates for the second QTL show no clear difference between the two methods. Moreover, the s.d. estimates with the two methods also show no clear differences.

In the second simulation study, we focused on genome-wide mapping and simulated a large genome with length 2000 cM, with 201 evenly spaced markers

covering it with a marker interval of 10 cM. Five QTLs were placed in the genome, and the corresponding positions and effects are listed in Table 3. The residual variance was set as 1.5 and the polygenic variance was absent. The total heritability was 69.4%, and the heritability explained by these QTL ranged from 8.1 to 22.4%. The family structure and other parameters were the same as in the first simulation. To simulate the phenotypic value for a binary trait, the continuous phenotypic values generated above were treated as liabilities and the threshold was set as 0, resulting in 50% incidence. Specifically, if the liability is below 0, the binary phenotypic value is set to 0; otherwise it is set to 1.

In the analysis, we set the expected QTL number $l_0 = 3$, leading to the maximum QTL number $L = 8$ and thus the prior probability of the model indicator $p(\gamma_j = 1) = l_0/L = 0.38$. The tuning parameter for QTL position was set as 20 cM. The thinning interval of MCMC was 10, the burn-in period was 1000 and the length of the complete chain was 21 000. When binary traits were analysed, the residual variance was re-parameterised (Albert and Chib, 1993) and equal to 1.5.

For continuous trait analysis, the posterior probabilities of QTL position from the new method are plotted in Figure 6, and the BF from the new method and F -statistic from the LS method are plotted in Figure 6b. The results for binary-trait analysis are shown in Figure 7. Using our method, all simulated QTLs were identified both for continuous traits and for binary traits. However, by the LS method, one QTL was missed for continuous trait analysis and two were missed for binary trait analysis. The simultaneous modelling of multiple QTL in our method likely contributes to a gain in statistical power. The parameter estimates are given in Table 3. Generally, the position estimates are largely close to their true values, both for continuous trait analysis and for binary trait analysis, but the estimation for the second QTL in binary trait analysis deviates from its true value by ~ 20 cM; the relative lower heritability of this QTL (12.2%) and the information loss in binary trait analysis could explain the biased position estimate. Nonetheless, the true positions of all QTLs are all within their highest posterior density region. The estimates of QTL variance are also given in Table 3, and they also deviate from their true values to some extent. In fact, for mapping QTLs in half-sib design, only paternal genotypes are available and the maternal genotypes are completely unknown; thus, the variance components are difficult to estimate accurately. Finally, we summarised the acceptance rate of M-H run for the joint update of covariance, polygenic variance and residual variance, which equalled 36.8 and 48.8% for continuous and binary traits, respectively, and were near their optimal regions.

Table 3 The parameter estimates in the second simulated experiment

Trait	QTL no.	True parameter		Bayes estimates						LS estimates	
		Position (cM)	σ_j^2	Position (cM)	HPD region	Posterior probability (%)	σ_j^2	h^2 (%)	σ_e^2	Position	Confidence interval
Continuous	1	215	0.40	211 (3.15)	(203, 22)	99.0	0.41 (0.16)	10.8	1.81 (0.37)	217	(203,238)
	2	422	0.60	418 (3.02)	(413,426)	99.7	0.75 (0.38)	19.8		428	(400,460)
	3	853	1.10	856 (3.49)	(845,865)	99.6	0.68 (0.33)	18.0		859	(797,872)
	4	1450	0.50	1455 (3.33)	(1445,1458)	97.7	0.52 (0.28)	13.8		—	—
	5	1843	0.80	1846 (1.44)	(1842,1848)	99.2	1.42 (0.42)	37.6		1846	(1836,1854)
Binary	1	215	0.40	209 (3.64)	(204,227)	97.4	0.65 (0.50)	10.8	—	217	(203,245)
	2	422	0.60	443 (3.29)	(416,465)	99.2	0.58 (0.35)	10.0		—	—
	3	853	1.10	856 (3.24)	(842,867)	98.9	0.68 (0.57)	11.3		—	—
	4	1450	0.50	1454 (4.11)	(1443,1459)	99.6	0.92 (0.62)	15.3		1457	(1403,1504)
	5	1843	0.80	1846 (1.56)	(1842,1849)	99.5	2.19 (0.92)	36.4		1848	(1834,1855)

Abbreviations: HPD, highest posterior density; LS, least square; QTL, quantitative trait loci.

Figure 6 Profiles of statistics from the simulated genome-wide data for continuous trait. (a) Profiles of the posterior probability by the proposed method. (b) Profiles of the BF score by the proposed method and the F -statistic by the LS method. The grey horizontal line represents the thresholds for the LS method at a 0.01 significance level. The standard threshold 3.2 of BF score is used for declaring the QTL by the new method, which is not shown in the figure for clear presentation. The true positions of the simulated QTL are represented by upward arrows.

We also varied the expected QTL number $l_0 = 4$ and 5, leading to the maximum QTL number $L = 11$ and 14 and $p(\gamma_j = 1) = l_0/L = 0.363$ and 0.357. The profiles of posterior probability and BF are similar to that from $L = 8$ (data was not shown). These results further illustrate that our method was not very sensitive to the expected QTL number or the maximum QTL number.

Discussion

We employed the approach reported by Yi *et al.* (2005) to ascertain the maximum-QTL number and found that the results were not very sensitive to the maximum QTL number. Theoretically, the maximum-QTL number may be set as any value as long as it is greater than the actual

QTL number. The simplest method assumes that each marker interval contains one QTL, but the computational burden will be increased because of the large QTL number, and at the same time, the efficiency will be challenged. Xu (2007) developed a method to handle such a large model using a hierarchical prior for γ_j , $p(\gamma_j) = \rho^{\gamma_j}(1-\rho)^{1-\gamma_j}$ and $p(\rho) = \text{Bata}(1,1)$, which is useful when the number of model effect is large, especially for the investigation of QTL epistatic effects and QTL-environment interaction effects.

In our real data analysis, we found that marker BMS143 affects FP traits. This marker has also been reported by others to affect FP (Schnabel *et al.*, 2005), MY (Gomez-Raya *et al.* 1998; Velmala *et al.*, 1999) and PP (Spelman *et al.*, 1996; Zhang *et al.*, 1998; Velmala *et al.*, 1999; Schnabel *et al.*, 2005), but we only confirmed that it affects FP. Moreover, we found two extra QTL

Figure 7 Profiles of statistics obtained from the simulated genome-wide data for a binary trait. (a) Profiles of the posterior probability by the proposed method; (b) profiles of the BF score by the proposed method and the F statistic by the LS method. The grey horizontal line represents the thresholds for the LS method at a 0.01 significance level. The standard thresholds 3.2 of BF score is used for declaring the QTL by the new method, which is not shown in the figure for clear presentation. The true positions of the simulated QTLs are represented by upward arrows.

that affected PP traits. For the first extra QTL, the position estimates were close to the marker ILSTS97, which was also reported by Moiso *et al.* (2000) to affect the PP trait. For the second, the estimated position was close to CSN3 and CSN1, as reported by many researchers (Velmal *et al.*, 1995; Maki-Tanila *et al.*, 1998; Leone *et al.*, 1998; Ikonen *et al.*, 1999), and also close to AFR227, as reported by Velmal *et al.* (1999), and to BP7, as reported by Ashwell *et al.* (1998) and Schnabel *et al.* (2005), all of which were found to affect the PP trait.

The variance component model can be written as $\text{Var}(\mathbf{y}) = \mathbf{V} = \sum_{j=1}^L \gamma_j \Theta_j \sigma_j^2 + \mathbf{A} \sigma_A^2 + \mathbf{R} \sigma_e^2$, and $\mathbf{R}$ is a diagonal matrix with the diagonal element being the inverse of reliability for each individual $1/r_i$. If the difference of r_i between each individual is small, the reliability may have negligible effects on the results; however, if the difference is large, it is necessary to take such information into account. Although we did not consider the reliability in real data analysis, the computational package provided by us (described below) can include such information.

In our real data analysis, we assumed that the markers of parents are in linkage equilibrium, but this assumption may not be reasonable if the marker interval is relatively small. A method that combines the information of linkage disequilibrium (LD) and linkage analyses called LDL can significantly improve the estimate precision for QTL position if the markers of parents are in linkage disequilibrium (Meuwissen and Goddard, 2000, 2001; Meuwissen and Goddard, 2004; Lee and Van der Werf, 2006). LDL also estimates variance components, and therefore the proposed multi-QTL mapping method is completely applicable if the appropriate IBD matrix is constructed.

We only consider QTL main effects in this study and ignore the interacting effects between QTL for simplicity. In our variance component model, the interactions can be included by introducing Hadamard multiplication of matrices. However, when the loci are linked, the IBD covariance components will be difficult to construct and the performance of our method will be affected if such information is ignored. Moreover, we can also extend our method to categorical trait analysis as described in Appendix B.

We investigated the efficiency of the new method in half-sib families. In practice, if the appropriate IBD matrix is constructed, the new method also can be used for a more complex population, such as the mixture of full-sib and half-sib families or more complicated mating designs. However, the method is not useful for large populations because it is infeasible to calculate the inverse of a large (co)variance matrix. One solution is to incorporate an empirical Bayesian procedure to estimate variance components (Lee and Van der Werf, 2006), such as using REML (restricted maximum likelihood) estimates, which need not calculate the inverse of (co)variance matrix.

In the new method, calculating the IBD matrix in MCMC may cost a great deal of computational time. To solve the problem, the half-stored technology was employed. We first calculated and saved the IBD matrix on a hard drive for all positions and then read the IBD for particular positions in the current iteration. As a result, computational time was dramatically reduced. All computational analyses in our research were implemented via our computational package 'BayesMapQTL.exe', which was built in FORTRAN language and is available upon request.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

We thank the three reviewers for their critical comments on the manuscript. This work was financially supported by the Key Development of New Transgenic Breeds Program (2009ZX08009-156B), the Research Fund for the Doctoral Program of Higher Education of China (20070019044), the National Natural Science Foundation of China (31072016, 30972092 and 31001001) Heilongjiang Education Ministry (11541254), and the Beijing Natural Science Foundation of China (6102016).

References

- Albert JH, Chib S (1993). Bayesian analysis of binary and polychotomous response data. *J Am Stat Assoc* **88**: 669–679.
- Ashwell M, Da Y, Vanraden P (1998). Detection of putative loci affecting milk production and composition, health and type traits in a US Holstein population using 44 microsatellite markers. *Anim Genet* **29**(Suppl 1): 61–62.
- Banerjee S, Yandell BS, Yi N (2008). Bayesian QTL mapping for multiple traits. *Genetics* **179**: 2275–2289.
- Browne WJ (1998). Applying MCMC Methods to Multi-level Models. PH.D. Thesis, Department of Mathematical Sciences, University of Bath.
- Chan JSK, Kuk AC (1997). Maximum likelihood estimation for probit-linear mixed models with correlated random effects. *Biometrics* **88**: 86–97.
- Chen HY, Zhang Q, Wang CK, Shu J, Mei G, Yin CC et al. (2005). Mapping QTLs on BTA6 affecting milk production traits in a Chinese Holstein population. *Chinese Sci Bull* **50**: 1737–1742.
- De Koning DJ, Schulman NF, Elo K, Moisio S, Kinso R et al. (2001). Mapping of multiple quantitative trait loci by simple regression in half-sib designs. *J Anim Sci* **9**: 616–622.
- Fang M, Liu SC, Jiang D (2009). Bayesian composite model space approach for mapping quantitative trait loci in variance component model. *Behav Genet* **39**: 337–346.
- Fawcett T (2006). An introduction to ROC analysis. *Pattern Recognit Lett* **27**: 861–874.
- Freyer G, Kühn C, Weikard R, Zhang Q, Mayer M (2002). Multiple QTL on chromosome six in dairy cattle affecting yield and content traits. *J Anim Breed Genet* **119**: 69–82.
- Gaffney PJ (2001). An efficient reversible jump Markov chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. PH.D. Thesis, Department of Statistics, University of Wisconsin.
- Gelman A, Carlin J, Stern H, Rubin D (2004). *Bayesian Data Analysis*. Chapman & Hall: London.
- Gelman A, Roberts GO, Gilks WR (1995). *Bayesian Statistics 5*. Oxford University Press: Oxford.
- Georges M, Nielsen D, Mackinnon M, Mishra A, Okimoto R (1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics* **139**: 907–920.
- George EI, McCulloch RE (1993). Variable selection via Gibbs sampling. *J Am Stat Assoc* **88**: 881–889.
- Godsill SJ (2001). On the relationship between MCMC model uncertainty methods. *J Comput Graph Stat* **10**: 230–248.
- Gomez-Raya L, Klungeland H, Vage DL, Olsaker V, Fimland E, Klemetsdal G et al. (1998). Mapping QTL for milk production traits in Norwegian cattle. *Proc 6th World Congr. Genet Appl Livest Prod* **26**: 429–432.
- Green PJ (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**: 711–732.
- Grignola FE, Hoeschele I, Tier B (1996a). Mapping quantitative trait loci in outcross populations via residual maximum likelihood. I. Methodology. *Genet Sel Evol* **28**: 479–490.
- Grignola FE, Hoeschele I, Zhang Q, Thaller G (1996b). Mapping quantitative trait loci in outcross populations via residual maximum likelihood. II. A simulation study. *Genet Sel Evol* **28**: 491–504.
- Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97–109.
- Heath SC (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* **61**: 748–760.
- Heyen DW, Weller JL, Ron M, Band M, Beever JE (1999). A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiol Genomics* **1**: 165–175.
- Ikonen T, Jala MO, Ruottinen O (1999). Associations between milk protein polymorphism and first lactation milk production traits in Finnish Ayrshire cows. *J Dairy Sci* **82**: 1026–1033.
- Jansen RC (2003). Studying complex biological systems using multi-factorial perturbation. *Nat Rev Genet* **4**: 145–151.
- Kao CH, Zeng ZB, Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**: 1203–1216.
- Kass RE, Raftery AE (1995). Bayes factors. *J Am Stat Assoc* **90**: 773–795.
- Knott SA, Elsen JM, Haley CS (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theor Appl Genet* **93**: 71–80.
- Lee SH, van der Werf JHJ (2006). Simultaneous fine mapping of multiple closely linked quantitative trait loci using combined linkage disequilibrium and linkage with a general pedigree. *Genetics* **173**: 2329–2337.
- Leone P, Scaltriti V, Caroli A, Sangalli S, Samore A, Pagnacco G et al. (1998). Effects of the CASK E variant on milk yield indexes in Italian Holstein Friesian Bulls. *Anim Genet* **29**: 63.
- Liu J, Liu Y, Liu X, Deng HW (2007). Bayesian mapping of quantitative trait loci for multiple complex traits with the use of variance components. *Am J Hum Genet* **81**: 304–320.
- Liu Y, Jansen GB, Lin CY (2004). Quantitative trait loci mapping for dairy cattle production traits using a maximum likelihood method. *J Dairy Sci* **87**: 491–500.
- Maki-Tanila A, deKoning DJ, Elo K, Moisio S, Velmala R (1998). Mapping of multiple quantitative trait loci by regression in half sib designs. *Proceedings of the 6th World Congress on Genetics Applied to Livestock Production* Vol. 25, pp 269–272. Armidale: Australia.
- Metropolis NA, Rosenbluth W, Rosenbluth MN, Teller AH, Teller E (1953). Equation of state calculations by fast computing machines. *J Chem Phys* **21**: 1087–1091.
- Meuwissen TH, Goddard ME (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol* **36**: 261–279.
- Meuwissen THE, Goddard ME (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**: 421–430.
- Meuwissen THE, Goddard ME (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet Sel Evol* **33**: 605–634.
- Moisio SM, Schulman NF, de Koning DJ, Elo K, Velmala R, Virta A et al. (2000). A genome scan for milk production QTL for Finnish Ayrshire cattle. *Proc 27th Conf. Anim Genet*, Minneapolis, MN, p 24.
- Nadesalingam J, Plante Y, Gibson JP (2001). Detection of QTL for milk production on chromosomes 1 and 6 of Holstein cattle. *Mamm Genome* **12**: 27–31.
- Plante Y, Gibson JP, Nadesalingam J, Mehrabani-Yeganeh H, Lefebvre S, Vandervoort G et al. (2001). Detection of quantitative trait loci affecting milk production traits on 10 chromosomes in Holstein cattle. *J Dairy Sci* **84**: 1516–1524.
- Plummer M, Best N, Cowles K, Vines K (2008). Coda: output analysis and diagnostics for MCMC, R package version 0.13-2.

- Raftery AE, Madigan D, Hoeting JA (1997). Bayesian model averaging for linear regression models. *J Am Stat Assoc* **92**: 179–191.
- Roberts GO, Rosenthal JS (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Stat Sci* **16**: 351–367.
- Ron M, Klinger D, Feldmesser E, Seroussi E, Ezra E (2001). Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design. *Genetics* **159**: 727–735.
- Satagopan JM, Yandell BS (1996). Estimating the number of quantitative trait loci via Bayesian model determination. *Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases. Biometric Section, Joint Statistical Meetings* Chicago.
- Schnabel RD, Kim JJ, Ashwell MS, Sonstegard TS, Van Tassell CP (2005). Fine-mapping milk production quantitative trait loci on BTA6: analysis of the bovine osteopontin gene. *Proc Natl Acad Sci USA* **102**: 6896–6901.
- Seaton S, Haley CS, Knott SA, Kearsley M, Visscher PM (2002). QTL express: mapping quantitative trait loci in simple and complex pedigrees. *Bioinformatics* **18**: 339–340.
- Sillanpää MJ, Arjas E (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**: 1373–1388.
- Sorensen DA, Andersen S, Gianola D, Korsgaard I (1995). Bayesian inference in threshold models using Gibbs sampling. *Genet Sel Evol* **27**: 229–249.
- Spelman RJ, Coppieters W, Karim L, van Arendonk JA, 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.
- Stephens DA, Fisch RD (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov Chain Monte Carlo. *Biometrics* **54**: 1334–1347.
- Uimari P, Hoeschele I (1997). Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**: 735–743.
- Velmalä RJ, Vilkkilä HJ, Elo KT, de Koning DJ, Mäki-Tanila AV (1999). A search for quantitative trait loci for milk production traits on chromosome 6 in Finnish Ayrshire cattle. *Anim Genet* **30**: 136–143.
- Velmalä R, Vilkkilä J, Elo K, Mäki-Tanila A (1995). Casein haplotypes and their association with milk production traits in the Finnish Ayrshire cattle. *Anim Genet* **26**: 419–425.
- Wang H, Zhang YM, Li X, Masinde GL, Mohan S et al. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**: 465–480.
- Xu S, Gessler DD (1998). Multipoint genetic mapping of quantitative trait loci using a variable number of sibs per family. *Genet Res* **71**: 73–83.
- Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.
- Xu S (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**: 513–521.
- Yi N (2004). A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. *Genetics* **167**: 967–975.
- Yi N, George V, Allison DB (2003). Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics* **164**: 1129–1138.
- Yi N, Xu S (2000). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics* **156**: 411–422.
- Yi N, Yandell BS, Churchill GA, Allison DS, Eisen EJ (2005). Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. *Genetics* **170**: 1333–1344.
- Zeng ZB (1994). Precision mapping of quantitative trait loci. *Genetics* **136**: 1457–1468.
- Zhang Q, Boichard D, Hoeschele I, Ernst C, Eggen A (1998). Mapping quantitative trait loci for milk production and health of dairy cattle in a large outbred pedigree. *Genetics* **149**: 1959–1973.

Appendix A

Updating QTL variance and QTL position

The new proposal variance $\sigma_j^{2(*)}$ is generated using the same method as that of the polygenic variance and residual variance and accepted with probability $\min(1, r)$ and if $\gamma_j = 1$,

$$r = \frac{L_1 \cdot p(\sigma_j^{2(*)} | s_j^2, \omega_j)}{L_2 \cdot p(\sigma_j^{2(0)} | s_j^2, \omega_j)} \cdot hr_j,$$

where L_1 and L_2 denote the likelihood conditional on the new proposal variance $\sigma_j^{2(*)}$ and the old one $\sigma_j^{2(0)}$, respectively; otherwise

$$r = \frac{f(\sigma_j^{2(*)} | s_j^2, \omega_j)}{f(\sigma_j^2 | s_j^2, \omega_j)} \cdot hr_j,$$

and the proposal ratio hr_j sees Equation (3). The prior distributions of QTL variance, polygenic variance and residual variance are truncated at a certain value, usually the phenotypic variance (Yi and Xu, 2000). The QTL position λ_j is also updated by the M–H algorithm. The new position is proposed uniformly around the old one and accepted with the new calculated IBD (identity-by-descent) matrix with probability $r = L_1/L_2$ for $\gamma_j = 1$; and $r = 1$ for $\gamma_j = 0$, and L_1 and L_2 denote the likelihood conditional on the new IBD and on the old IBD.

Appendix B

Extension to multiple ordered categorical traits

Suppose that the discrete phenotypic value w_i takes one of c ordered categories, $1, \dots, c$. In a traditional threshold model, w_i is affected by continuous liability y_i and the fixed thresholds, $t_1, t_2, \dots, t_{c-1}$, truncate y_i to determine the observed categories. If $w_i = k$ is observed, $t_{k-1} < y_i \leq t_k$ ($k = 1, \dots, c$). Let $t_0 = -\infty$, $t_c = +\infty$, and $t_1 = 0$ (Albert and Chib, 1993), then the estimates $\mathbf{t} = (t_2, t_3, \dots, t_{c-1})$. The joint posterior probability of unknowns can be expressed as

$$p(\mathbf{y}, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\lambda} | \mathbf{w}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}) \propto f(\mathbf{w}, \mathbf{y}, \mathbf{t}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}) p(\boldsymbol{\theta}, \boldsymbol{\lambda}) p(\mathbf{t}) p(\mathbf{y}) \quad (\text{B1})$$

where the likelihood

$$f(\mathbf{w} | \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{A}, \mathbf{X}, \mathbf{Z}, \mathbf{M}) = \prod_{i=1}^n \left[\sum_{j=1}^c I(t_{j-1} < y_i < t_j) I(w_i = j) \right], \quad (\text{B2})$$

and $I(x)$ is an indicator function taking the value 1 if x is true and 0 otherwise.

The prior probability of $\mathbf{t}$ is uniformly distributed, $p(\mathbf{t}) \propto \text{constant}$, and the prior of the liability follows normal distribution, $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ (Albert and Chib, 1993;

Sorensen *et al.*, 1995). Then the conditional posterior of the thresholds also follows uniform distribution,

$$f(t_j | \mathbf{y}, \mathbf{w}, t_{j+1}, t_{j-1}) = \frac{1}{\min[\min(\mathbf{y} | \mathbf{w} = j + 1), t_{j+1}] - \max[\max(\mathbf{y} | \mathbf{w} = j), t_{j-1}]} \quad (\text{B3})$$

where $\min(\mathbf{y} | \mathbf{w} = j + 1)$ indicates the minimum value of the liabilities within observations in category $j + 1$; similarly, $\max(\mathbf{y} | \mathbf{w} = j)$ denotes the maximum value of liabilities for observations in category j (Albert

and Chib, 1993; Sorensen *et al.*, 1995). The liability y_i is sampled from a doubly truncated normal distribution with mean

$$E(y_i) + \text{Cov}(y_i, \mathbf{y}_{-i})[\text{var}(\mathbf{y}_{-i})]^{-1}[\mathbf{y}_{-i} - E(\mathbf{y}_{-i})],$$

and variance

$$\text{var}(y_i) - \text{Cov}(y_i, \mathbf{y}_{-i})[\text{var}(\mathbf{y}_{-i})]^{-1}\text{Cov}(\mathbf{y}_{-i}, y_i),$$

where, $\mathbf{y}_{-i}$ denotes all elements except the i th (Chan and Kuk, 1997).

Supplementary Information accompanies the paper on Heredity website (<http://www.nature.com/hdy>)