

Genome-wide evaluation for quantitative trait loci under the variance component model

Lide Han · Shizhong Xu

Received: 26 April 2010 / Accepted: 1 September 2010 / Published online: 12 September 2010
© The Author(s) 2010. This article is published with open access at Springerlink.com

Abstract The identity-by-descent (IBD) based variance component analysis is an important method for mapping quantitative trait loci (QTL) in outbred populations. The interval-mapping approach and various modified versions of it may have limited use in evaluating the genetic variances of the entire genome because they require evaluation of multiple models and model selection. In this study, we developed a multiple variance component model for genome-wide evaluation using both the maximum likelihood (ML) method and the MCMC implemented Bayesian method. We placed one QTL in every few cM on the entire genome and estimated the QTL variances and positions simultaneously in a single model. Genomic regions that have no QTL usually showed no evidence of QTL while regions with large QTL always showed strong evidence of QTL. While the Bayesian method produced the optimal result, the ML method is computationally more efficient than the Bayesian method. Simulation experiments were conducted to demonstrate the efficacy of the new methods.

Keywords Bayesian analysis · Genome selection · Markov chain Monte Carlo · Maximum likelihood

Introduction

Identical-by-descent (IBD) based variance component method is often used to map quantitative trait loci (QTL) for outbred populations (Goldgar 1990; Amos 1994). The commonly used method is the interval mapping where two markers are used at a time to infer the IBD matrix for any positions bracketed by the two markers (Fulker and Cardon 1994). The model usually contains one QTL and a polygenic effect so that the variance of the QTL, the polygenic variance and the residual variance are the only variance components to be estimated. If multiple QTL exist, this interval mapping approach will produce biased estimate for the QTL variance. When the entire genome is scanned, the total genetic variance (sum of all variances of detected QTL) is often greater than the total phenotypic variance. This phenomenon always occurs in interval mapping, regardless whether the random model for an outbred population or the fixed model for a line cross is used. The reason for that is that QTL effects or QTL variances of different locations are estimated using different models. To scan the entire genome, multiple analyses are conducted, one for each putative location. None of the single QTL models is correct if multiple QTL exist. Therefore, the optimal method should be a multiple variance component model in which all QTL are included in a single model.

Multiple variance components may be difficult to estimate if the number of QTL included in the model is extremely large. However, the popular MCMC implemented Bayesian method is designed to handle large saturated models and it is the ideal method for multiple variance component estimation (Uimari and Hoeschele 1997). The maximum likelihood method may also be sufficient to handle large saturated models under the random model framework; we just never thought of placing one

Electronic supplementary material The online version of this article (doi:10.1007/s10709-010-9497-1) contains supplementary material, which is available to authorized users.

L. Han · S. Xu (✉)
Department of Botany and Plant Science, University
of California, Riverside, CA 92521, USA
e-mail: xu@genetics.ucr.edu

QTL in every few cM of the genome. Meuwissen et al. (2001) first attempted to evaluate the entire genome using a high dense marker map under the popular Bayesian approach. Their method actually treats the positions of QTL as fixed and only estimates the QTL variances and other parameters. Meuwissen et al. (2001) placed many QTL in the model. As a result QTL positions may not be relevant because the whole genome is already well covered by the proposed QTL. Yi and Xu (2000) used the reversible jump MCMC to infer the number of QTL under the random model framework. Only large QTL were eventually included in the model and the entire genome may not be evaluated thoroughly due to the slow mixing behavior of the reversible jump MCMC.

In this study, we proposed to cover the entire genome by QTL and estimated the QTL variances simultaneously within a single model. As long as the extra QTL placed in regions of the genome that do not contain QTL have estimated QTL variances close to zero, we can put as many QTL as we want to make sure that the entire genome is evaluated fairly. We investigated both the ML method and the Bayesian method and showed the pros and cons of each method.

Methods

Linear model and likelihood

Consider N independent families and the size of the j th family is n_j for $j = 1, \dots, N$. For simplicity, let us assume that the family size is constant across families so that $n_j = n$ for all $j = 1, \dots, N$. Assume that we want to put M quantitative trait loci (QTL) in the linear model described below,

$$y_j = 1\mu + \sum_{k=1}^M \gamma_{jk} + \varepsilon_j \quad (1)$$

where y_j is an $n \times 1$ vector for the phenotypic values of a quantitative trait for family j , μ is the population mean of the trait, 1 is an $n \times 1$ unity vector, γ_{jk} is an $n \times 1$ vector for the additive genetic effects (breeding values) for family j at the k th QTL and ε_j is an $n \times 1$ vector for the environmental effects. The number of QTL proposed in the model is not the actual number of QTL but a larger number determined by the investigator based on the size of the genome, the marker density and the population size. If the marker density is not high, say one marker in every 10 centiMorgan (cM), the number of proposed QTL can take the number of markers. If the marker density is high, we may choose to place one QTL among a few markers. If the marker density is low, we may insert virtual markers

between two consecutive markers. The bottom line is to put one QTL in every d cM. The number of QTL proposed should be sufficiently large to make sure that the entire genome is well evaluated without any large gaps. If a proposed QTL is nearby a true QTL, the effect of the true QTL will be absorbed by the proposed QTL. If a proposed QTL is further away from a true QTL, its estimated effect will be close to zero.

The expectation and variance–covariance matrix of y_j are

$$E(y_j) = 1\mu \quad (2)$$

and

$$\text{var}(y_j) = V_j = \sum_{k=1}^M \Pi_{jk} \sigma_k^2 + I\sigma^2 \quad (3)$$

respectively, where Π_{jk} is an $n \times n$ IBD matrix for the j th family at the k th QTL, σ_k^2 is the genetic variance for the k th QTL and σ^2 is the environmental error variance. Assume that $\gamma_{jk} \sim N(0, \Pi_{jk} \sigma_k^2)$ and $\varepsilon_j \sim N(0, I\sigma^2)$, the log likelihood function for the j th family is

$$L_j(\theta) = -\frac{1}{2} \ln |V_j| - \frac{1}{2} (y_j - 1\mu)^T V_j^{-1} (y_j - 1\mu) \quad (4)$$

where $\theta = \{\mu, \sigma_1^2, \dots, \sigma_M^2, \sigma^2\}$ is the parameter vector. The overall log likelihood function for the entire population is

$$L(\theta) = -\frac{1}{2} \sum_{j=1}^N \ln |V_j| - \frac{1}{2} \sum_{j=1}^N (y_j - 1\mu)^T V_j^{-1} (y_j - 1\mu) \quad (5)$$

Maximum likelihood estimation

The challenge for the genome-wide evaluation is that, for a large genome, the number of proposed QTL can be very large and majority of the proposed QTL should have estimated variance components close to zero. This will cause problems in the parameter estimation. The EM algorithm is the first candidate method for the variance component model (Thompson and Shaw 1990). However, it is sensitive to the initial values of the parameters. We cannot choose zero as the initial value for σ_k^2 , although most σ_k^2 are in fact zero. Other initial values are hard to choose. Therefore, we decide to directly maximize the log likelihood function using a sequential approach by updating one variance component at a time, conditional on the values of all other variance components. When a single variance component is considered, maximizing the log likelihood function is a one-dimension problem; the bisection or any other simple algorithm can be used when one parameter is updated. When all parameters are updated, we go back to the first parameter and update the value again. The sequential algorithm

requires iterations nested within other iterations until a certain criterion of convergence is satisfied. The iterations within an iteration are called the inner iterations while the iterations outside are called the outer iterations. This algorithm requires more iterations than an algorithm that updates all parameters simultaneously, but choosing the initial value for the parameter of interest becomes trivial, i.e., $\sigma_k^2 = 0$ can be used as initial for all $k = 1, \dots, M$. The sequential approach of Xu (2007) was adopted here, where V_j^{-1} and $|V_j|$ are calculated only once for each outer iteration. For large family sizes, much of the computing burden comes from calculating V_j^{-1} and $|V_j|$. Therefore, the sequential algorithm can save computing time substantially, in addition to ease the choice of initial values.

Estimation of QTL positions

The fixed position approach described previously requires a full coverage of the genome by the proposed QTL. We now introduce a method that can update the positions of the proposed QTL. When the QTL positions are estimated, we can place a smaller number of QTL but still maintain a high probability that regions containing true QTL are visited frequently by the proposed QTL. Let λ_k be the position of the k th proposed QTL for $k = 1, \dots, M$. The parameter vector is now defined as

$$\theta = \{\mu, \sigma_1^2, \dots, \sigma_M^2, \lambda_1, \dots, \lambda_M, \sigma^2\} \tag{6}$$

The QTL positions can move along the genome, but the order of the QTL remains unchanged, as denoted by $\lambda_1 < \lambda_2 < \dots < \lambda_M$. The connection between the log likelihood function and the QTL positions is through the IBD matrices. We first calculate the IBD matrix for each putative position of the genome (Gessler and Xu 2000). If a QTL moves to a new position, the IBD matrix for the new position is used to evaluate the log likelihood function. The search for QTL positions is also sequential, i.e., we update one position at a time, given positions of all other QTL. For the k th QTL, we use a grid search between λ_{k-1} and λ_{k+1} with 2 cM increment. When the iterations converge, all parameters, including the QTL positions, will remain unchanged. We now have the MLE of all parameters, including the MLE for the QTL positions.

Bayesian estimation of parameters

The maximum likelihood method provides a point estimate for each parameter. The main purpose of the genome evaluation is to examine the entire genome for possible association with a quantitative trait. The Bayesian method is adopted here because it gives a chance to evaluate every putative location of the genome. To compare the Bayesian

method with the maximum likelihood method, we choose uniform prior for each parameter, including the population mean, the variance components, QTL positions and the environmental variance. The prior distribution for λ_k is also uniform but within the range defined by $\lambda_{k-1} < \lambda_k < \lambda_{k+1}$. The posterior distribution for the population mean is normal with mean

$$E(\mu|\dots) = \left[\sum_{j=1}^N 1^T V_j^{-1} 1 \right]^{-1} \left[\sum_{j=1}^N 1^T V_j^{-1} y_j \right] \tag{7}$$

and variance

$$\text{var}(\mu|\dots) = \left[\sum_{j=1}^N 1^T V_j^{-1} 1 \right]^{-1} \tag{8}$$

from which a realization of μ is sampled. Other parameters do not have explicit forms of a distribution, and thus they are sampled based on the Metropolis-Hastings rule (Metropolis et al. 1953; Hastings 1970). For each of the parameters sampled using the M-H rule, the proposal distribution is a uniform distribution centered in the parameter value of the previous cycle. For example, the proposed value of λ_k in cycle $t + 1$ is

$$\lambda_k^* = \lambda_k^{(t)} + u \tag{9}$$

where

$$p(u) = U(u|\lambda_k^{(t)} - \delta, \lambda_k^{(t)} + \delta) \tag{10}$$

and δ is a small positive number, say $\delta = 2$ cM. If λ_k^* is accepted, $\lambda_k^{(t+1)} = \lambda_k^*$, otherwise, $\lambda_k^{(t+1)} = \lambda_k^{(t)}$. In most situations, the Metropolis algorithm is sufficient, but when the value of a parameter is near the boundary, the Hastings adjustment is required to ensure that the parameter is not trapped to a fixed point. For the sampling of QTL position, the Hastings adjustment can be found in Wang et al. (2005). The variance component for each QTL is bounded between zero and the phenotypic variance present in the data, where zero is a legal value of the variance component. The residual variance is also bounded between zero and the phenotypic variance present in the data, but zero is excluded. The posterior sample consists of the observations after burn—in deletion and chain thinning.

In addition to the uniform prior for each variance component, we also considered the following hierarchical prior distribution for the QTL variances. The following exponential distribution was assigned to σ_k^2 ,

$$p(\sigma_k^2|\tau^2) = \frac{\tau^2}{2} \exp\left(-\frac{\tau^2}{2}\sigma_k^2\right) \tag{11}$$

where the parameter τ^2 was also assigned a Gamma prior,

$$p(\tau^2|a, b) = \frac{b^a}{\Gamma(a)} (\tau^2)^{a-1} \exp(-b\tau^2) \tag{12}$$

The values of the hyper parameters $a > 0$ and $b > 0$ can be chosen arbitrarily, e.g. $(a,b) = (0.5, 0.1)$.

Results

Setup of simulation experiments

We designed the following simulation experiment to evaluate the performance of the proposed ML and Bayesian methods. We simulated a single large chromosome of 1,000 cM in length. The genome was covered by 101 evenly spaced markers with 10 cM per marker interval. The population size was $N \times n = 500 \times 3 = 1,500$. The parental alleles of markers were randomly sampled from five different alleles with an equal frequency. Eight QTL were placed in the genome with positions and QTL variances shown in Table 1. The parental alleles of the QTL were sampled from an infinite number of alleles, i.e., each parental allele was different from any other parental alleles. The genetic effect of the k th QTL was the sum of the two allelic effects while each allelic effect was sample from $N(0, \frac{1}{2}\sigma_k^2)$. The positions of these simulated QTL varied in terms of distances from the nearest markers, some overlapping with a marker and some residing in the middle of an interval bracketed by two markers. The heritability of an individual QTL (proportion of the phenotypic variance contributed by the QTL) ranged from 0.30 to 40.0%. The overall mean and the residual variance were set at $\mu = 10$ and $\sigma^2 = 1.0$, respectively. The overall proportion of the

phenotypic variance contributed by all the eight QTL was 79.78%.

The simulation experiment with this setup is called the standard setup. Some parameters were eventually altered relative to the standard setup in the extended simulation experiments. For example, the marker density was later decreased from 10 cM per interval to 20 and 40 cM per interval. The sampling strategy was also extended to $750 \times 2 = 1,500$ and $375 \times 4 = 1,500$. When one experimental parameter was altered, the remaining parameters were fixed at the values in the standard setup.

Results of data analysis

Standard setup

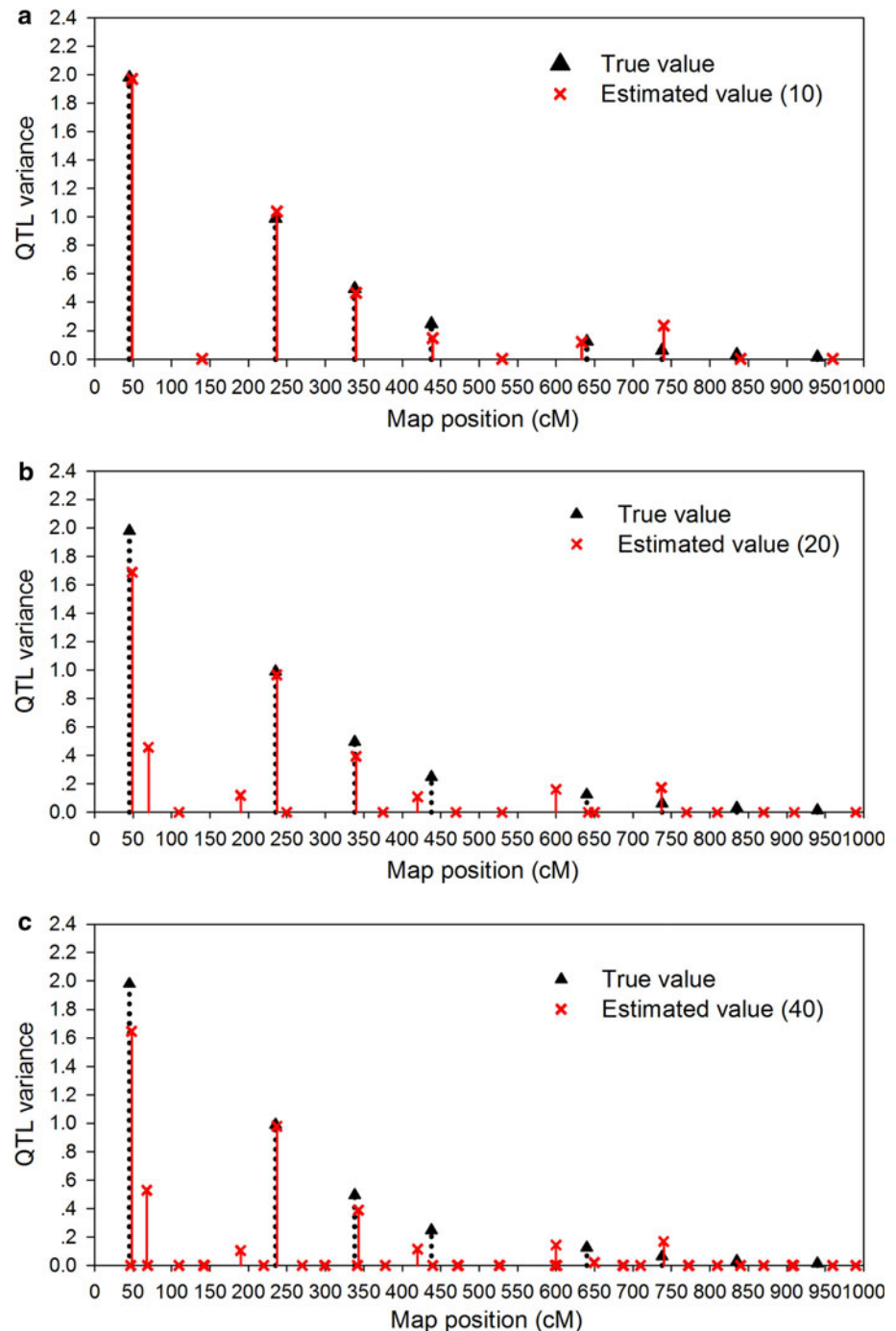
Under the standard setup (10 cM per marker interval and 3 siblings per family), we simulated one dataset, which was analyzed under the following three models: (a) 10 proposed QTL, (b) 20 proposed QTL and (c) 40 proposed QTL. The real number of QTL is eight, which is smaller than the proposed number of QTL in all situations. The maximum likelihood method was used to analyze the data. The QTL positions were also treated as parameters and were searched along with other parameters. The estimated QTL variance and the positions are shown in Fig. 1. All three models have correctly identified the three largest QTL, i.e., QTL whose contributions to the phenotypic variance are greater than or equal to 10%. The Akaike's information criterion (AIC) is a simple and very useful criterion for

Table 1 Comparison of the multiple variance component model with the interval mapping approach under the standard setup

| True parameter | | | Estimated parameter | | | | | | | |
|----------------------|----------|------------------|------------------------------|---------------|-----------|------------------|------------------|---------------|-----------|------------------|
| Position | Variance | Heritability (%) | Multiple variance components | | | | Interval mapping | | | |
| | | | Position | Variance | Power (%) | Heritability (%) | Position | Variance | Power (%) | Heritability (%) |
| 45 | 1.980 | 40.00 | 47.0 (3.2) | 1.485 (0.296) | 100 | 29.5 (5.8) | 47.9 (2.6) | 1.856 (0.306) | 100 | 38.2 (5.5) |
| 235 | 0.990 | 20.00 | 234.6 (7.0) | 0.707 (0.266) | 100 | 14.1 (5.3) | 238.4 (4.3) | 1.184 (0.277) | 97 | 24.8 (5.4) |
| 340 | 0.495 | 10.00 | 338.9 (10.0) | 0.431 (0.206) | 85 | 8.5 (4.0) | 342.1 (9.2) | 0.969 (0.222) | 84 | 20.7 (4.6) |
| 440 | 0.248 | 5.00 | 439.4 (12.1) | 0.323 (0.180) | 71 | 6.4 (3.6) | 440.7 (11.4) | 0.847 (0.174) | 54 | 18.2 (3.4) |
| 640 | 0.124 | 2.50 | 637.7 (11.6) | 0.272 (0.189) | 63 | 5.4 (3.7) | 638.7 (13.4) | 0.868 (0.168) | 46 | 18.5 (3.2) |
| 740 | 0.062 | 1.25 | 737.0 (13.7) | 0.251 (0.196) | 45 | 5.0 (4.0) | 741.2 (14.3) | 0.831 (0.184) | 41 | 17.8 (3.8) |
| 835 | 0.031 | 0.62 | 832.5 (10.2) | 0.181 (0.123) | 52 | 3.6 (2.4) | 838.2 (12.7) | 0.825 (0.127) | 39 | 17.6 (2.7) |
| 940 | 0.015 | 0.30 | 937.5 (13.1) | 0.281 (0.161) | 41 | 5.5 (3.1) | 942.1 (15.3) | 0.792 (0.133) | 39 | 17.0 (2.9) |
| Residual variance | 1.000 | | 0.448 (0.162) | | | | - | | | |
| Phenotype variance | 4.945 | | 5.032 (0.229) | | | | - | | | |
| Number of iterations | | | 98.6 (24.6) | | | | - | | | |

The means and standard deviations (in parentheses) of the estimates are calculated from 100 replicated simulation experiments

Fig. 1 The estimated QTL variances and positions along the genome under three different models: **a** 10 proposed QTL, **b** 20 proposed QTL and **c** 40 proposed QTL. The true locations and variance of eight simulated QTL are also shown in the plots



selecting the best model among alternative models (Wada and Kashiwagi 1990). The AIC values for the three models (10 QTL, 20 QTL and 40 QTL) are 3,851.81, 3,879.45 and 3,922.46, respectively. The minimum AIC occurs for the model with 10 proposed QTL, and thus this model is the best. However, in real data analysis, the actual number is unknown and the proposed number of QTL is often larger than the true number of QTL. The extra QTL placed in the model should be closed to zero for the estimated QTL variances and this has been demonstrated by Fig. 1, where

all the superfluous QTL have very small estimated variances unless they are close to a true large QTL. Therefore, it is safe to place more QTL in the model than the actual number of QTL and let the program shrink the superfluous QTL to zero. When the proposed number of QTL is too big, a real large QTL may be split by two or more proposed QTL in the neighborhood of the true QTL, as demonstrated in Fig. 1b, c. This presents no problem because we can choose more appropriate model among a few different models with the AIC criterion.

We now further extend the analysis by including 100 proposed QTL in the model. Under this analysis, we investigated two situations: (a) the positions of the 100 proposed QTL were estimated along with the QTL variances, labeled “100 moving”; (b) the positions of the 100 proposed QTL were fixed and evenly placed along the genome, labeled “100 fixed”. The estimated QTL variances and their locations are demonstrated in Fig. 2. Both methods work very well regarding the ability to identify large QTL (contribution greater than 10%). Again, a large QTL is often split by a few proposed QTL in the neighborhood of the true QTL. This analysis shows that if a large number of QTL are placed in the model, the positions of the proposed QTL do not have to be estimated. The moving position and fixed position approaches generate almost identical results.

To compare the result with the interval mapping of Xu and Atchley (1995), we also analyzed the same dataset with the interval mapping approach. The result is shown in Fig. 3. The interval mapping only detected the two largest QTL. The third largest QTL (10% contribution to the phenotypic variance) was not detectable while this QTL has been detected by the multiple variance component model.

Under the standard setup (10 cM per interval and 500 family each with three siblings), we replicated the

experiment 100 times and each replicated dataset was analyzed with two methods. One method is the multiple variance component model proposed in this study, where the proposed number of QTL included in the model was 20 and the positions of the 20 proposed QTL were also estimated using the maximum likelihood method. The other method is the interval mapping of Xu and Atchley (1995) in which a single QTL and a polygenic effect were included in the model. Since the multiple variance component model has no test for a chromosome location, we simply examined the estimated QTL variance in the neighborhood of a true QTL. When the estimated QTL variance in the neighborhood (within 20 cM) of a true QTL is sufficiently large (larger than any peak appearing in a non-QTL region), the QTL was claimed to be detected. For each simulated true QTL, the mean estimate and the standard deviation across the 100 replicated simulations were calculated. The empirical statistical power for each simulated QTL was also calculated as the proportion of the replicated experiments that the QTL was detected. It appears to be subjective, but the multiple variance component model usually provides very small estimated QTL variances for regions that are not placed for any QTL. Therefore, any region that has a noticeable estimated QTL variance

Fig. 2 The estimated QTL variances and positions along the genome under two models: **a** 100 proposed QTL with positions also estimated along with the QTL variances and **b** 100 proposed QTL with positions fixed evenly along the genome

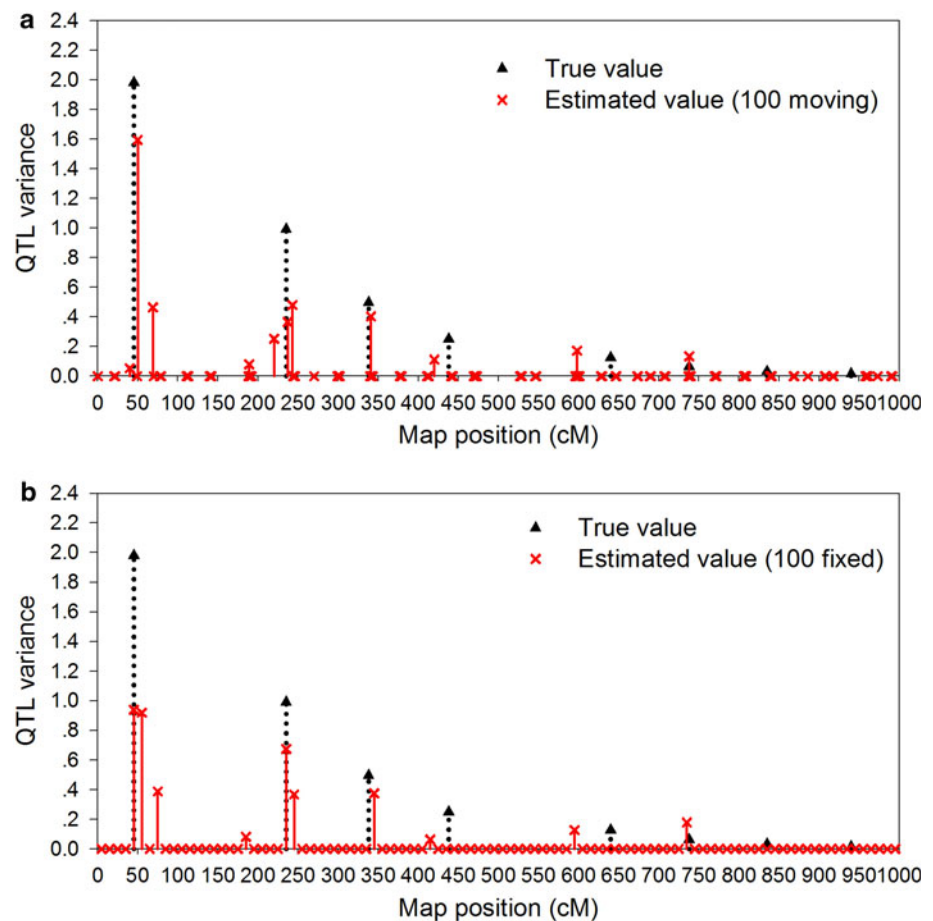


Fig. 3 Likelihood ratio test statistics and the estimated variance components for the interval mapping approach. The threshold value is 5.99

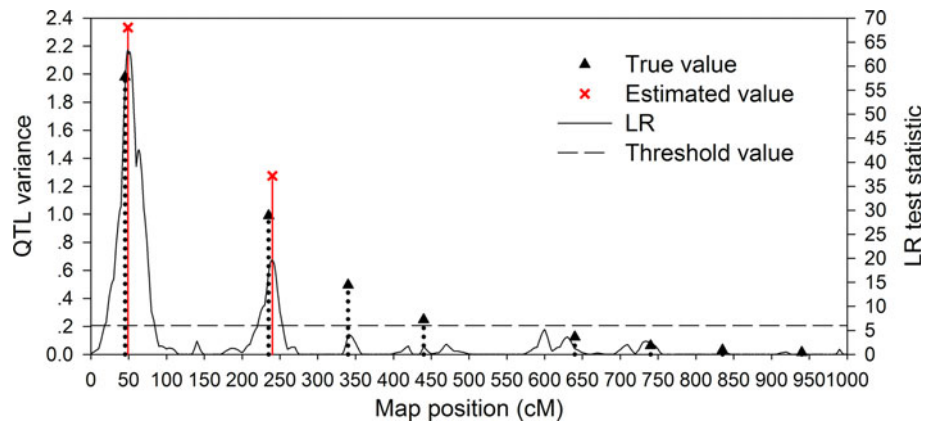


Table 2 Estimated QTL parameters by the multiple variance component model under two extended family structures

| True parameter | | | Estimated parameter | | | | | | | |
|----------------------|----------|------------------|-----------------------------|---------------|-----------|------------------|-----------------------------|---------------|-----------|------------------|
| Position | Variance | Heritability (%) | $N \times n = 750 \times 2$ | | | | $N \times n = 350 \times 4$ | | | |
| | | | Position | Variance | Power (%) | Heritability (%) | Position | Variance | Power (%) | Heritability (%) |
| 45 | 1.980 | 40.00 | 46.5 (4.0) | 1.414 (0.326) | 100 | 28.1 (6.4) | 46.4 (2.7) | 1.506 (0.221) | 100 | 29.6 (4.3) |
| 235 | 0.990 | 20.00 | 234.9 (7.3) | 0.670 (0.318) | 89 | 13.3 (6.3) | 234.3 (5.4) | 0.756 (0.227) | 99 | 14.9 (4.4) |
| 340 | 0.495 | 10.00 | 337.2 (10.2) | 0.519 (0.277) | 80 | 10.3 (5.4) | 340.0 (8.2) | 0.445 (0.171) | 95 | 8.7 (3.3) |
| 440 | 0.248 | 5.00 | 438.3 (11.9) | 0.424 (0.215) | 62 | 8.4 (4.3) | 439.3 (9.3) | 0.284 (0.149) | 80 | 5.6 (2.9) |
| 640 | 0.124 | 2.50 | 635.6 (12.3) | 0.288 (0.217) | 46 | 5.7 (4.3) | 640.0 (10.6) | 0.209 (0.120) | 66 | 4.1 (2.3) |
| 740 | 0.062 | 1.25 | 737.4 (12.1) | 0.311 (0.208) | 38 | 6.2 (4.1) | 735.8 (13.3) | 0.226 (0.164) | 53 | 4.4 (3.2) |
| 835 | 0.031 | 0.62 | 831.0 (10.1) | 0.299 (0.220) | 35 | 5.9 (4.4) | 834.1 (9.6) | 0.164 (0.096) | 47 | 3.2 (1.9) |
| 940 | 0.015 | 0.30 | 937.0 (12.1) | 0.314 (0.189) | 30 | 6.3 (3.8) | 936.7 (12.9) | 0.154 (0.116) | 53 | 3.0 (2.3) |
| Residual variance | 1.000 | | 0.395 (0.165) | | | | 0.467 (0.146) | | | |
| Phenotype variance | 4.945 | | 5.026 (0.216) | | | | 5.082 (0.220) | | | |
| Number of iterations | | | 124.4 (33.2) | | | | 101.3 (19.2) | | | |

The means and standard deviations (in parentheses) of the estimates are calculated from 100 replicated simulation experiments

indicates that a true QTL is nearby. For the interval mapping of Xu and Atchley (1995), the likelihood ratio test statistic was used to claim the significance of a QTL. If an estimated QTL variance nearby a true QTL (within 20 cM) is significant, this QTL was claimed to be detected. The estimated QTL variances and QTL positions for the interval mapping are compared with those obtained from the multiple variance component model (see Table 1 for the comparison). Overall the multiple variance component model performs better than the interval mapping. The interval mapping provided biased (upward) estimates for all the QTL variances, especially when the true QTL variance was small. Because of the large biases for the estimated QTL variances, they do not add up, i.e., the sum of all the estimated QTL variances is greater than the total phenotypic variance. Therefore, the multiple variance component model outperforms the interval mapping approach.

We now extended the simulation to examine the effect of different family structure on the result of the multiple variance component model. The two additional family structures were 750×2 and 375×4 . Other parameter settings were the same as the previous experiment, i.e., 10 cM per marker interval and 20 proposed QTL included in the model. The experiment was replicated 100 times. The results are shown in Table 2. Result of structure 375×4 appears to be better than structure 750×2 in terms of smaller estimation errors and higher statistical power. Therefore, the multiple variance component model performs better with small number of large families.

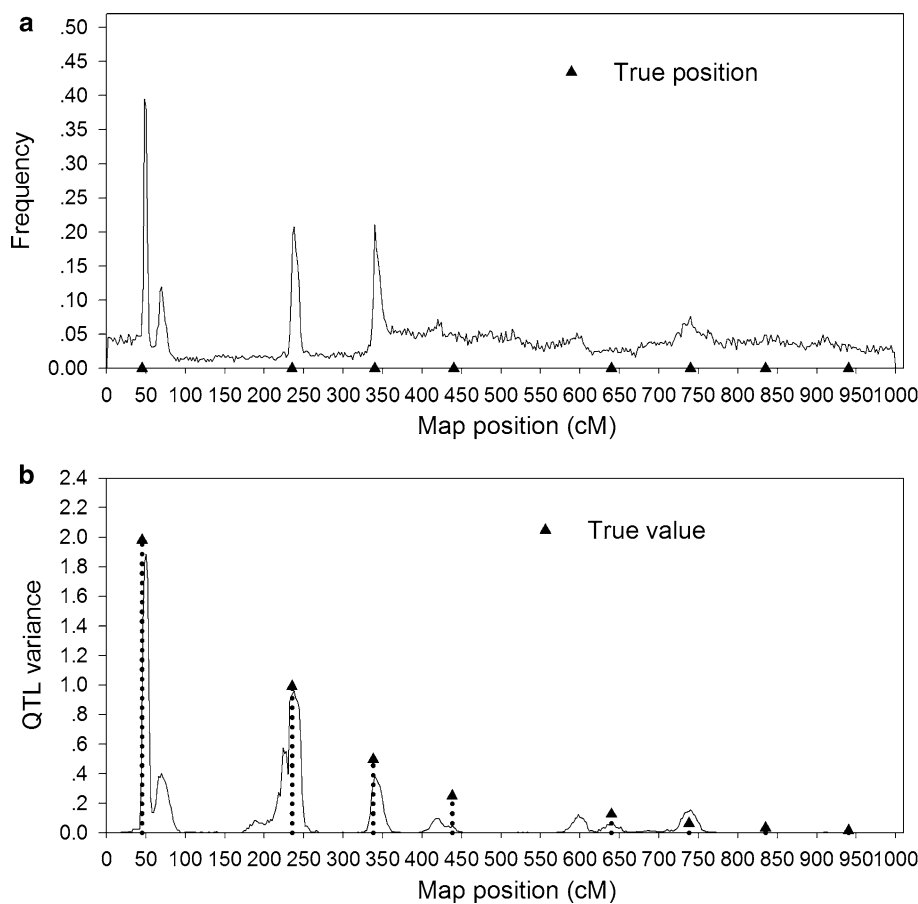
Further extension was made in terms of varying the marker density given that other experimental parameters were the same as those of the standard setup. We examined two more different marker densities, one is 20 cM per marker interval and the other is 40 cM per marker interval. The family

Table 3 Estimated QTL parameters by the multiple variance component model under two extended marker densities

| True parameter | | | Estimated parameter | | | | | | | |
|----------------------|----------|------------------|---------------------|---------------|-----------|------------------|---------------|---------------|-----------|------------------|
| | | | 20 cM | | | | 40 cM | | | |
| Position | Variance | Heritability (%) | Position | Variance | Power (%) | Heritability (%) | Position | Variance | Power (%) | Heritability (%) |
| 45 | 1.980 | 40.00 | 43.8 (4.3) | 1.423 (0.386) | 100 | 28.6 (7.5) | 41.7 (7.1) | 1.393 (0.459) | 100 | 28.3 (9.1) |
| 235 | 0.990 | 20.00 | 236.7 (5.8) | 0.794 (0.282) | 99 | 16.0 (5.6) | 236.6 (7.3) | 0.891 (0.366) | 95 | 18.1 (7.2) |
| 340 | 0.495 | 10.00 | 340.7 (8.0) | 0.444 (0.232) | 91 | 8.9 (4.6) | 339.3 (16.6) | 0.458 (0.249) | 73 | 9.3 (5.0) |
| 440 | 0.248 | 5.00 | 440.2 (12.6) | 0.380 (0.246) | 73 | 7.6 (4.8) | 442.2 (16.1) | 0.385 (0.212) | 62 | 7.8 (4.3) |
| 640 | 0.124 | 2.50 | 642.3 (12.3) | 0.309 (0.167) | 65 | 6.3 (3.4) | 640.5 (16.7) | 0.308 (0.236) | 44 | 6.3 (4.8) |
| 740 | 0.062 | 1.25 | 742.3 (14.0) | 0.240 (0.159) | 52 | 4.8 (3.1) | 737.3 (17.9) | 0.247 (0.172) | 44 | 5.1 (3.6) |
| 835 | 0.031 | 0.62 | 834.0 (10.1) | 0.218 (0.111) | 30 | 4.4 (2.3) | 840.8 (17.0) | 0.314 (0.188) | 25 | 6.4 (3.8) |
| 940 | 0.015 | 0.30 | 943.0 (15.2) | 0.233 (0.131) | 42 | 4.7 (2.6) | 939.9 (18.6) | 0.297 (0.168) | 48 | 6.1 (3.5) |
| Error variance | 1.000 | | 0.279 (0.168) | | | | 0.139 (0.156) | | | |
| Phenotype variance | 4.945 | | 4.976 (0.208) | | | | 4.912 (0.210) | | | |
| Number of iterations | | | 126.9 (22.4) | | | | 226.5 (123.2) | | | |

The means and standard deviations (in parentheses) of the estimates are calculated from 100 replicated simulation experiments

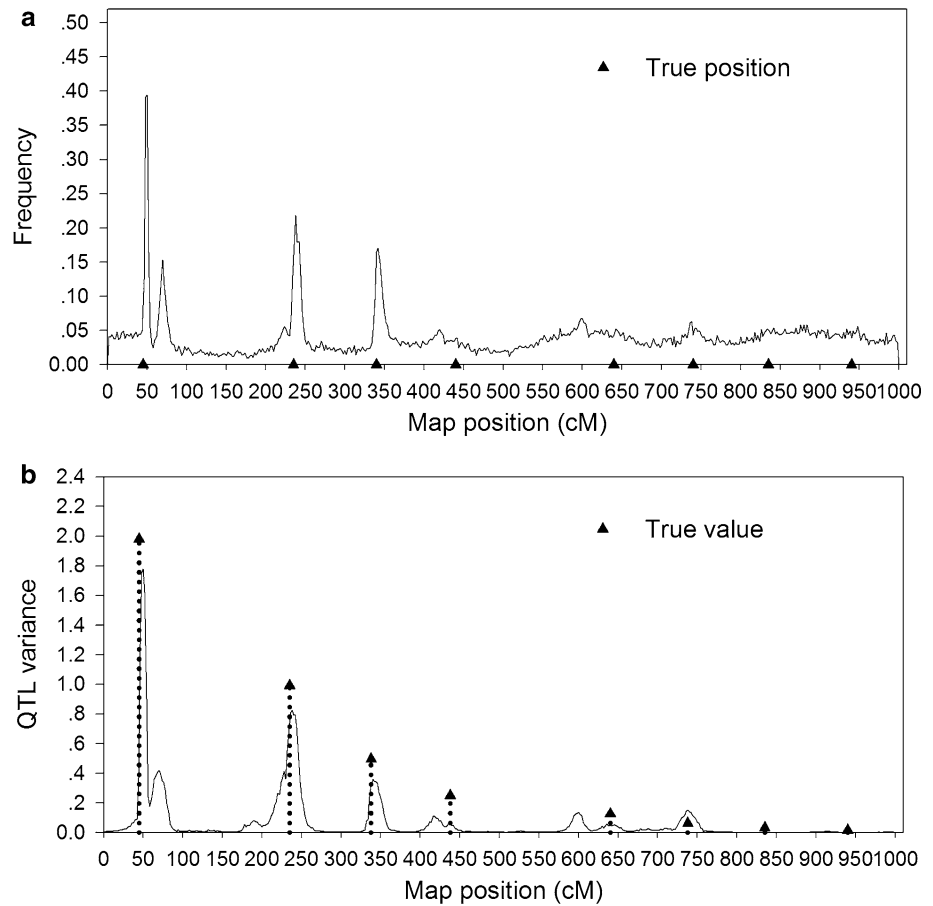
Fig. 4 MCMC implemented Bayesian analysis with 20 proposed QTL in the model. The uniform prior distribution is assigned to each parameter. **a** The *top panel* represents the QTL frequency profile, **b** the *bottom panel* represents the estimated QTL variance profile



structure was 500×3 and the proposed number of QTL was 20. The results are given in Table 3, showing that higher marker density has improved the performance of the method.

Finally, we examined the MCMC implemented Bayesian analysis for the simulated data in the standard setting. Again, we placed 20 QTL on the genome. First, we used

Fig. 5 MCMC implemented Bayesian analysis with 20 proposed QTL in the model. Exponential prior is assigned to each QTL variance and Gamma (0.5,0.1) is used for the parameter of the exponential prior. **a** The *top panel* represents the QTL frequency profile, **b** the *bottom panel* represents the estimated QTL variance profile



the uniform prior for each parameter, including the variances and positions of proposed QTL. The first 2000 iterations were treated as burn-in and thereafter one observation was saved in every 20 iterations to reduce the serial correlation. The posterior sample contained 2,000 observations for post MCMC analysis. The result of the MCMC implemented Bayesian analysis is shown in Fig. 4. The frequency profile is shown in the top panel while the QTL variance profile is shown in the bottom panel. The estimated QTL variances are very close to the true values. From Fig. 4b, we can see that all QTL but the two smallest ones are detectable. This demonstrates the advantage of the Bayesian method over the maximum likelihood method.

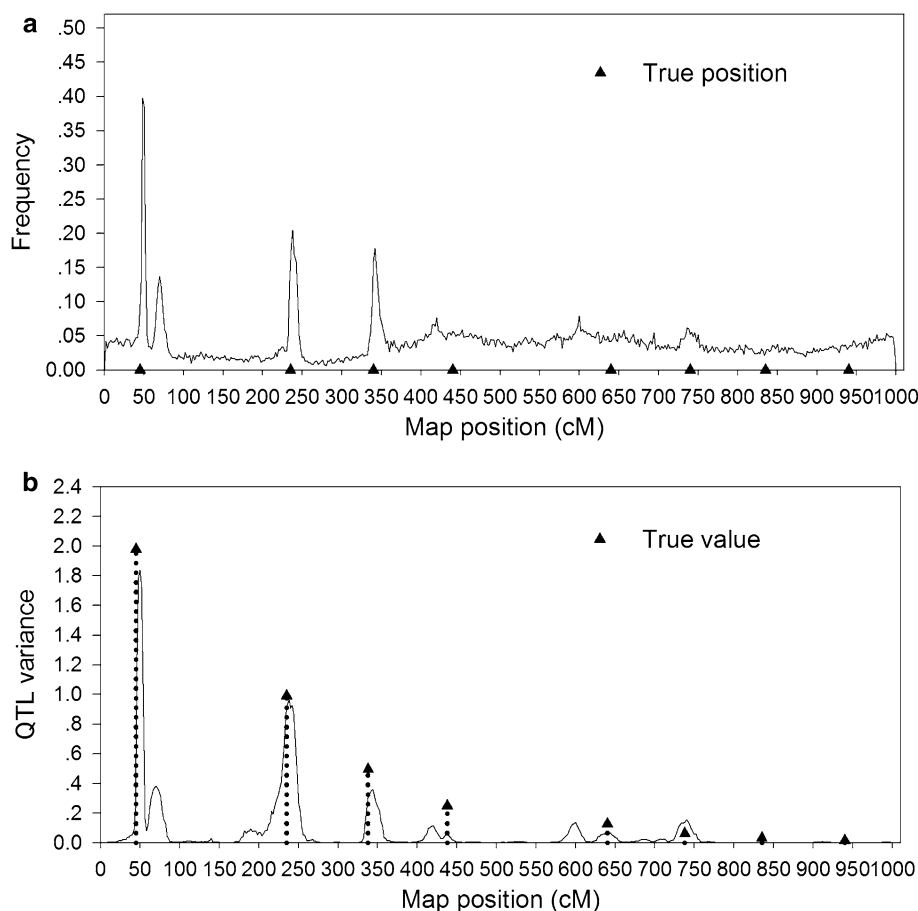
We further examined the MCMC implemented Bayesian method under the hierarchical modeling with exponential prior for each QTL variance and the parameter of the exponential prior was further assigned a Gamma prior with parameter a and b , $\text{Gamma}(a,b)$. The result of $\text{Gamma}(0.5, 0.1)$ is shown in Fig. 5. We also choose $\text{Gamma}(0.5, 0.01)$ and the result is given in Fig. 6. The hierarchical models with two different sets of hyper parameters are similar to each other, meaning that the choice of the hyper parameters (a and b) does not have too much influence on the result.

These results (hierarchical modeling) are also similar to the result of the uniform prior. Overall, the MCMC implemented Bayesian analysis performs better than the ML analysis. However, the ML method is computationally more efficient than the Bayesian method.

Discussion

We examined two different methods for genome-wide evaluation of QTL in outbred populations. The ML method is an extension of the interval mapping of Xu and Atchley (1995) to handle multiple QTL. The MCMC implemented Bayesian method is an extension of the Bayesian shrinkage analysis of Wang et al. (2005) for line crosses to outbred populations. Similar random model methodology has been proposed by Yi and Xu (2000) who used the reversible jump MCMC algorithm for model selection. In Yi and Xu (2000), the QTL number was treated as a parameter and sampled along with other parameters. In this study, we emphasize genome evaluation rather than QTL mapping. The difference between genome evaluation and QTL mapping is that the former tries to evaluate the entire

Fig. 6 MCMC implemented Bayesian analysis with 20 proposed QTL in the model. Exponential prior is assigned to each QTL variance and $\text{Gamma}(0.5, 0.01)$ is used for the parameter of the exponential prior. **a** The *top panel* represents the QTL frequency profile, **b** the *bottom panel* represents the estimated QTL variance profile



genome, including regions that have no QTL, while the latter emphasizes detecting regions of the genome that have QTL. We purposely placed more QTL than necessary to give the method a better chance to evaluate the entire genome. For regions of the genome that contain no QTL, the proposed QTL in those regions often have very small estimated variances. Another advantage of the genome evaluation is that it has avoided model selection, which is still a hot topic for discussion in the literature (Kadane and Lazar 2004).

We used multiple full-sib families as an example to demonstrate the method. Extension to multiple complicated pedigrees is straightforward, at least, theoretically because the method requires only the IBD matrices for each putative location of the genome. Methods to calculate the IBD matrix using marker information are available for arbitrarily complicated pedigrees (Amos et al. 1990; Almasy and Blangero 1998). The programs Lokie (Heath 1997) and SimWalk2 (Sobel et al. 2001) are the most well known software packages for IBD matrix calculation.

Surprisingly, the multiple variance component model has very low false positive rate (also called the Type I error). Although we did not actually calculate the Type I

error in our simulation experiments, just by visual inspection on the QTL variance profiles, we can see that regions of the genome that contain no QTL rarely show any noticeable peaks while regions with large QTL always have strong signals. This observation implies that the multiple variance component model has great power and small Type I error. Of course, statistical power and Type I error are concepts of frequentists, not of Bayesians. Another surprising discovery is that the Bayesian method is very robust to the prior choice for the QTL variance components. We examined the uniform prior and hierarchical prior (exponential and Gamma), they all generated similar results.

Finally, genome evaluation has two purposes: identifying the regions of the genome for association with the variance of a trait (similar to QTL mapping) and evaluating the genetic effect for each individual animal or plant (marker assisted selection). This study emphasizes the first purpose. To estimate the genetic effects (breeding values) for all individuals in a particular family, the best linear unbiased prediction (BLUP) technology can be applied. For example, to calculate the BLUP estimate for the k th QTL for all individuals in the j th family, the following BLUP equation can be used,

$$\hat{\gamma}_{jk} = \hat{\sigma}_k^2 \Pi_{jk} V_j^{-1} (y_j - 1\hat{\mu}) \quad (13)$$

The variance–covariance matrix of this BLUP estimate is

$$\text{var}(\hat{\gamma}_{jk}) = \hat{\sigma}_k^2 (I - \Pi_{jk} V_j^{-1} \Pi_{jk} \hat{\sigma}_k^2) \quad (14)$$

The overall breeding values for all members of the j th family is

$$\hat{\gamma}_j = \sum_{k=1}^M \hat{\gamma}_{jk} \quad (15)$$

Individuals can be ranked based on the estimated breeding values and selected as candidates for breeding. This is referred to as marker assisted selection.

Supplementary materials

The SAS/IML programs for the maximum likelihood method and Bayesian method are posted on the journal website along with sample data.

Acknowledgments We thank two anonymous reviewers for their comments on an early version of the manuscript and suggestions on the improvement of the manuscript. This project was supported by the Agriculture and Food Research Initiative (AFRI) of the USDA National Institute of Food and Agriculture under the Plant Genome, Genetics and Breeding Program 2007-35300-18285.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62:1198–1211

- Amos CI (1994) Robust variance-components approach for assessing genetic linkage in pedigrees. *Am J Hum Genet* 54:535–543
- Amos CI, Dawson DV, Elston RC (1990) The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. *Am J Hum Genet* 47:842–853
- Fulker DW, Cardon LR (1994) A sib-pair approach to interval mapping of quantitative trait loci. *Am J Hum Genet* 54:1092–1103
- Gessler DDG, Xu S (2000) Multipoint genetic mapping of quantitative trait loci with dominant markers in outbred populations. *Genetica* 105:281–291
- Goldgar DE (1990) Multipoint analysis of human quantitative genetic variation. *Am J Hum Genet* 47:957–967
- 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 of oligogenic models. *Am J Hum Genet* 61:748–760
- Kadane JB, Lazar NA (2004) Methods and criteria for model selection. *J Am Statist Assoc* 99:279–290
- Metropolis N, Rosenbluth AW, Rosenbluth MN et al (1953) Equations of state calculations by fast computing machines. *J Chem Phys* 21:1087–1091
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* 157:1819–1829
- Sobel E, Sengul H, Weeks DE (2001) Multipoint estimation of identity-by-descent probabilities at arbitrary positions among marker loci on general pedigrees. *Hum Hered* 52:121–131
- Thompson EA, Shaw RG (1990) Pedigree analysis for quantitative traits: variance components without matrix inversion. *Biometrics* 46:399–413
- Uimari P, Hoeschele I (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* 146:735–743
- Wada Y, Kashiwagi N (1990) Selecting statistical models with information statistics. *J Dairy Sci* 73:3573–3582
- Wang H, Zhang YM, Li X et al (2005) Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* 170:465–480
- Xu S (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* 63:513–521
- Xu S, Atchley WR (1995) A random model approach to interval mapping of quantitative trait loci. *Genetics* 141:1189–1197
- Yi N, Xu S (2000) Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics* 156:411–422