*Genet. Res., Camb.* (2003), **81**, *pp.* 133–144. *With 5 figures.* © 2003 Cambridge University Press
DOI: 10.1017/S0016672303006153 Printed in the United Kingdom

# Estimating allelic number and identity in state of QTLs in interconnected families

JEAN-LUC JANNINK[1]* AND XIAO-LIN WU[1,2]
[1] *Department of Agronomy, Iowa State University, Ames, IA 50011-1010, USA*
[2] *Center for Life Science Research, Hunan Agricultural University, Changsha, Hunan, 410218, China*

(*Received 31 July 2002 and in revised form 16 December 2002*)

## Summary

When multiple related families derived from inbred lines are jointly analysed to detect quantitative trait loci (QTLs), the analysis should estimate allelic effects as accurately as possible and estimate the probability that different parents carry alleles that are identical in state. Analyses exist that assume that all parents carry unique alleles or that all parents but one carry the same allele. In practice, many configurations are possible that group different parents according to their identity-in-state condition at a putative QTL allele. Here, we propose a variable model Bayesian analysis that selects among possible identity-in-state configurations and jointly estimates the allelic effects of identical-in-state parents. We contrast this analysis with a fixed model analysis that estimates unique allelic effects for all parents. We analyse two simulated mating designs: an experimental design in which three inbred parents were crossed to generate two families of 150 doubled haploid lines; and a breeding design in which 20 inbred parents were crossed to generate 60 families of 20 doubled haploid lines, with each parent contributing to six families. In all cases where some parents were simulated to carry alleles of identical effect (that is, they were identical in state), the variable analysis estimated allelic effects with lower mean-squared error than the fixed analysis. The variable analysis showed that, unless each family contains many individuals (more than 100), there is insufficient information in DNA-marker and phenotypic data to determine with high probability the QTL allelic number.

## 1. Introduction

Plant geneticists and breeders need to perform quantitative-trait-locus (QTL) analysis in multiple related families derived from inbred lines in two contexts. First, multiple experimental QTL mapping families exist for numerous crops (e.g. barley, oat, soybean, corn), and specific inbred parents are often shared across families (e.g. Brummer *et al.*, 1997; Kianian *et al.*, 1999; Orf *et al.*, 1999). Second, in breeding situations, parents are often mated in diallel designs to generate families from which to select. The number of families produced is then often greater than the number of inbred parents. Several methods have been proposed to map QTLs within multiple families. Xu (1998) and Yi & Xu (2000) propose models that parameterize allele substitution effects or allelic effects themselves as random effects. These models assume a specific parametric distribution of the random effects, usually normal with mean zero and estimated variance. Fixed effect models can relax the assumption of normally distributed random effects and accommodate the desire to estimate fixed QTL allelic values.

Fixed effect models have been presented by Rebaï & Goffinet (1993, 2000) and by Liu & Zeng (2000). A major issue with these models relates to the number of parameters they are required to estimate. The least-squares models of Rebaï & Goffinet (2000) assume that all parents contributing to the families evaluated carry different alleles such that numerous QTL effect parameters must be estimated. The maximum-likelihood methods of Liu & Zeng (2000) examine a series of likelihood-ratio tests in which the value of the allele carried by each parent in turn is estimated. These tests therefore allow for two allelic value

* Corresponding author. Tel: +1 515 294 4153. Fax: +1 515 2946505. e-mail: jjannink@iastate.edu

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

parameters in the specific configuration that one parent carries a value different from all others (Liu & Zeng, 2000). More than two allelic value parameters might be required accurately to model data on progeny derived from several parents even when some parents carry identical QTL alleles. An ideal statistical model would fit the data well while estimating fewer allelic effect parameters than required by the assumption that all parents carry distinct alleles. If the identities of QTL alleles were known, such a model could be constructed by pooling parents carrying the same allele into a single class. In practice, although the identities of alleles are not observable, the phenotypic effect of alleles provides information that can guide the specification of statistical models estimating the correct number of alleles and pooling information appropriately. Selection among such models would allow statistical assessment of identity in state at a QTL among parents. Furthermore, determining identity in state among parents would allow models to fit the data well without estimating separate allelic effects for all parents. For example, if alleles carried by parents A and B are contrasted in one family, and alleles carried by parents A and C in another, estimated substitution effects will indicate the similarity between the alleles carried by parents B and C. Information of this type can form the basis of a model-selection procedure in which different models represent different numbers of alleles segregating among parents and different configurations of parents classified together as identical in state.

The non-nested nature of such models presents difficulties for model selection within a maximum-likelihood framework. Here, we propose a Markov chain Monte Carlo (MCMC) method capable of generating a Bayesian posterior distribution of the number of QTL alleles segregating among parents, and of estimating the posterior probabilities of the different possible identity-in-state configurations. We apply the method to simulated data to determine its ability to estimate the true number of alleles present and their configuration within parents, and the impact of this modelling procedure on the estimates of allelic effect relative to procedures that assume all parents carry unique alleles.

## 2. Methods

### (i) *QTL model*

Consider $f$ families derived from crosses among $P$ inbred founders and containing a total of $n$ progeny individuals. Among the founders, there are $l \leqslant P$ distinct alleles at a putative QTL. The vector of observed phenotypes $\mathbf{y}$ is modelled as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{QCa} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{X}$ is an $n \times f$ design matrix relating progeny to families, $\boldsymbol{\beta}$ is a $f \times 1$ vector of family means, $\mathbf{Q}$ is an $n \times P$ matrix indicating from which parent a progeny received QTL alleles (for any given progeny row, only the two columns of $\mathbf{Q}$ that correspond to the progeny's parents can be non-zero), $\mathbf{C}$ is a $P \times l$ configuration matrix linking the parental origin of each allele with its effect, $\mathbf{a}$ is a $l \times 1$ vector of those QTL allelic effects, and $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I}\sigma^2)$ is an $n \times 1$ vector of residuals.

The QTL allele configuration $\mathbf{C}$ is equivalent to an identity-in-state configuration among parents. The $l$ alleles segregating among $P$ inbred parents define $l$ groups, with all parents within a group sharing a common allele at the QTL. For example, assuming five parents and three alleles, a possible configuration groups parents 1 and 4, and parents 2 and 3, whereas parent 5 carries a unique allele. The elements of $\mathbf{C}$, $c_{ij}$, are set to 1 if parent $i$ carries allele $j$, and to 0 otherwise. Thus, in the case of inbred parents studied here, each row has a single non-zero element, and parents grouped by sharing a common QTL allele have 1 in the same column. For the above example, the configuration matrix is

$$\mathbf{C}_{5\times 3} = \{c_{ij}\} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{2}$$

In specifying $\mathbf{C}$, the identity of the parents matter but the identity of the alleles do not. That is, referring back to the matrix, permuting columns in the matrix results in configurations that are deemed identical, whereas permuting rows changes the configuration because it changes which parents are grouped together. Naturally, also, shifting a 1 from one column to another changes the configuration.

We define $\kappa(P, l)$ as the number of distinct allele configurations given $P$ and $l$. Notice that a configuration conveys information about the grouping of parents as carrying the same allele but not information about the order of alleles. The function $\kappa(P, l)$ can be defined recursively as

$$\kappa(P, l) = \frac{l^P - \sum_{i=1}^{l-1} \kappa(P, i) \dfrac{l!}{(l-i)!}}{l!}. \tag{3}$$

The first term $l^P$ represents a complete ordered enumeration of $l$ alleles among $P$ parents. Within this enumeration are configurations that contain fewer than $l$ alleles and we must eliminate these configurations from the tally. To eliminate the configurations that contain $i$ alleles, we use the configuration number $\kappa(P, i)$. There are $l! \div (l-i)!$ orderings of $i$ distinct alleles using $l$ allelic symbols. We therefore multiply

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

$\kappa(P, i)$ by that ratio. Finally, we divide the whole by the number of orderings of $l$ alleles, $l!$.

The joint posterior density of all unobservables $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{Q}, \mathbf{C}, \mathbf{a}, l, \sigma^2)$ given the observables ($\mathbf{y}$ and $\mathbf{X}$) and prior information can be expressed as

$$p(\boldsymbol{\theta} \,|\, \mathbf{y}, \mathbf{X}) \propto p(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{a}, \mathbf{Q}, \mathbf{C}, l, \sigma^2) \times p(\boldsymbol{\beta}) p(\mathbf{a}) p(\mathbf{Q}) p(\mathbf{C} \,|\, l) p(l) p(\sigma^2), \tag{4}$$

where $p(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{a}, \mathbf{Q}, \mathbf{C}, l, \sigma^2)$ represents the likelihood assuming multivariate normality:

$$p(\mathbf{y} \,|\, \boldsymbol{\beta}, \mathbf{a}, \mathbf{Q}, \mathbf{C}, l, \sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \times \exp\left( - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{QCa})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{QCa})}{2\sigma^2} \right) \tag{5}$$

and $p(*)$ is the prior distribution for parameter $*$, with the prior for configuration, $p(\mathbf{C} \,|\, l)$, being conditioned on allelic number.

### (ii) *MCMC sampling procedures*

To implement a MCMC model, we used a scalar Metropolis–Hastings procedure in which each parameter in $\boldsymbol{\theta}$ is sampled in turn, considering all other parameters fixed (Gilks *et al.*, 1996). Briefly, a candidate $\tilde{\boldsymbol{\theta}}$ for MCMC sample $\boldsymbol{\theta}_{t+1}$ is generated from a proposal density $p(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}_t)$. With probability $\alpha(\boldsymbol{\theta}_t \,|\, \tilde{\boldsymbol{\theta}})$, $\boldsymbol{\theta}_{t+1} = \tilde{\boldsymbol{\theta}}$ and otherwise $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t$, where

$$\alpha(\boldsymbol{\theta}_t, \tilde{\boldsymbol{\theta}}) = \min\left( \frac{p(\boldsymbol{\theta}_t \,|\, \tilde{\boldsymbol{\theta}}) p(\tilde{\boldsymbol{\theta}}) p(\mathbf{y} \,|\, \tilde{\boldsymbol{\theta}})}{p(\tilde{\boldsymbol{\theta}} \,|\, \boldsymbol{\theta}_t) p(\boldsymbol{\theta}_t) p(\mathbf{y} \,|\, \boldsymbol{\theta}_t)}, 1 \right). \tag{6}$$

Allelic values and QTL genotypes are initialized given the information from the linked markers. Iterations then consisted of the steps:

- update QTL inheritance matrix $\mathbf{Q}$;
- update QTL allelic effects $\mathbf{a}$;
- update family means $\boldsymbol{\beta}$ and residual variance $\sigma^2$;
- update QTL allelic configuration conditional on allelic number, $\mathbf{C} \,|\, l$;
- update the number of alleles $l$.

Details of the prior probabilities, candidate proposal densities and calculation of the acceptance probability follow.

1. *QTL inheritance matrix* $\mathbf{Q}$. The prior for QTL genotype derives from the rules of Mendelian segregation and recombination relative to flanking markers. QTL genotypes are updated independently for each progeny and are sampled directly from their full conditional posterior (Jansen, 1994).

2. *Allelic effects* $\mathbf{a}$. Denote $a_j$ as the effect of allele $j$ at the QTL. The prior for $a_j$ was normal with zero mean and standard deviation equal either to the phenotypic standard deviation or to one half the phenotypic standard deviation. The proposal density for $\tilde{a}_j$ was uniform centred on the previous parameter value $\tilde{a}_j \,|\, a_{j(t)} \approx unif(a_j - r_a, a_j + r_a)$, where $r_a$ is the radius of possible change in allelic value. The radius $r_a$ was empirically tuned to obtain acceptance probabilities between 0·2 and 0·5 (Roberts, 1995). The acceptance probability $\alpha(a_j, \tilde{a}_j)$ is

$$\alpha(a_j, \tilde{a}_j) = \min\left( \frac{p(\tilde{a}_j) p(\mathbf{y} \,|\, \tilde{\boldsymbol{\theta}})}{p(a_j) p(\mathbf{y} \,|\, \boldsymbol{\theta})}, 1 \right), \tag{7}$$

where $\tilde{\boldsymbol{\theta}}$ is identical to $\boldsymbol{\theta}$ except that $a_j$ is replaced by $\tilde{a}_j$.

3. *Family means* $\boldsymbol{\beta}$ *and residual variance* $\sigma^2$. The prior distribution for the family mean was uniform positive with a maximum at twice the phenotypic mean. The prior for the residual variance was uniform positive with a maximum at ten times the phenotypic variance. These vague priors were chosen to avoid constraining the posterior distributions of the parameters. We denote the mean for family $i$ as $\beta_i$. The $\beta_i$ values were sampled for each family in turn using proposal densities and acceptance probabilities similar to the allelic values described above. The residual variance was assumed to be either equal across or independent for each family, depending on the family size as described below. Otherwise, it was sampled similarly to the family means and to the allelic values.

4. *QTL configuration matrix conditional on allelic number* $\mathbf{C} \,|\, l$. To obtain adequate acceptance rates for changing $\mathbf{C}$, small changes were proposed at each iteration. Among the $l$ alleles, consider that $l_2$ are carried by more than one parent, of which one (say, allele $u$) is randomly chosen. The proposal consists of shifting one random parent from carrying this allele to carrying a different allele (say, allele $v$). In terms of $\mathbf{C}$, this operation corresponds to identifying columns with more than one '1' entry, picking one of those columns (the column for $u$) and one of its '1' entries, and moving the entry to a different column (the column for $v$). Only columns with more than one '1' entry can be used for allele $u$ so that, after the operation, there are no null columns. The acceptance probability $\alpha(\mathbf{C}, \tilde{\mathbf{C}})$ for this is

$$\alpha(\mathbf{C}, \tilde{\mathbf{C}}) = \min\left( \frac{\eta_u l_2 p(y \,|\, \tilde{\boldsymbol{\theta}})}{\eta_v \tilde{l}_2 p(y \,|\, \boldsymbol{\theta})}, 1 \right), \tag{8}$$

where $\eta_i$ is the number of parents carrying allele $i$, and $\tilde{l}_2$ is the number of alleles carried by more than one parent in the proposal (the number of columns in $\tilde{\mathbf{C}}$ with more than one '1' entry).

5. *QTL allelic number* $l$. With equal probability, a proposal is made either to increase or to decrease the allelic number. An increase is proposed only if the current number of alleles is less than the number of founders ($l < P$). A decrease is proposed only if the current number of alleles is greater than two ($l > 2$).

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

Changing $l$ also changes $\mathbf{C}$. We first describe proposals to increase the allelic number.

To increase allelic number, an allele (say, allele $u$) carried by more than one parent is randomly chosen and the parents are randomly split into two groups. A new allele is created (say, allele $v$) with an allelic value chosen from the allelic value prior distribution, and one group of parents is shifted to carrying this new allele. Given $l_2$ alleles carried by more than one parent and $(2^{\eta_u-1}-1)$ distinct random splits into two groups for allele $u$, the proposal probability for increasing allelic number by one is

$$p(l+1 \mid l)=\frac{p(\tilde{a}_v)}{(2^{\eta_u-1}-1)l_2}. \tag{9}$$

To decrease the allelic number from $l+1$ back to $l$, two different alleles are randomly chosen from the $l+1$ alleles at the QTL, and these are combined into one allele. The probability for this proposal is

$$p(l \mid l+1)=2 \div (l+1)l. \tag{10}$$

Bringing all these terms together, the acceptance probability $\alpha(l,\ l+1)$ for an increase in allelic number is

$$\alpha(l,\ l+1)=\min\left(\frac{2(2^{\eta_u-1}-1)l_2}{(l+1)l}\times\frac{p(\tilde{\mathbf{C}} \mid l+1)p(l+1)}{p(\mathbf{C} \mid l)p(l)} \times\frac{p(\tilde{\boldsymbol{\theta}}^-)}{p(\boldsymbol{\theta}^-)p(\tilde{a}_v)}\times\frac{p(\mathbf{y} \mid \tilde{\boldsymbol{\theta}})}{p(\mathbf{y} \mid \boldsymbol{\theta})}, 1\right), \tag{11}$$

where $\boldsymbol{\theta}^-$ includes all elements of $\boldsymbol{\theta}$ except for $l$ and $\mathbf{C}$. Notice that $\tilde{\boldsymbol{\theta}}^-$ contains the same parameters as $\boldsymbol{\theta}^-$ plus the parameter $\tilde{a}_v$ so that its prior is $p(\tilde{\boldsymbol{\theta}}^-)=p(\boldsymbol{\theta}^-)p(\tilde{a}_v)$, and the third term in $\alpha(l, l+1)$ cancels out. To determine the prior $p(\tilde{\mathbf{C}} \mid l+1)$, consider that there are $\kappa(p, l+1)$ possible configurations of $l+1$ alleles among $p$ parents. In the absence of information leading to a preference for any given configuration, the prior for a specific configuration conditional on $l+1$ alleles is therefore $\kappa(p, l+1)^{-1}$. Similarly, $p(\mathbf{C} \mid l)$ is $\kappa(p, l)^{-1}$. In the analysis, we evaluated both uniform and Poisson priors for $l$. With the uniform prior, $p(l+1)$ and $p(l)$ cancel out. With a Poisson prior, their ratio is $\lambda \div (l+1)$, where $\lambda$ is the prior mean of the Poisson distribution. These considerations simplify the second term in $\alpha(l, l+1)$ to $\kappa(p, l) \div \kappa(p, l+1)$ and $\kappa(p, l)\lambda \div \kappa(p, l+1)(l+1)$ for the uniform and Poisson cases, respectively. In the terminology of reversible-jump MCMC (Green, 1995; Waagepetersen & Sorensen, 2001), the parameters $l$ and $\mathbf{C}$ are categorical variables that are model indicators. With respect to continuous variables, the jump satisfies dimension matching because $\dim(\tilde{\boldsymbol{\theta}}^-)=\dim([\boldsymbol{\theta}^-, \tilde{a}_v])$. That is, $\tilde{\boldsymbol{\theta}}^-$ has one more parameter than $\boldsymbol{\theta}^-$. Further, to conform to reversible-jump MCMC, the acceptance probability needs to be multiplied by the Jacobian of the function $\tilde{\boldsymbol{\theta}}^-=g(\boldsymbol{\theta}^-, \tilde{a}_v)$. However, $g(\boldsymbol{\theta}^-, \tilde{a}_v)$ is the identity function and its Jacobian is 1.

Similar considerations for the acceptance probability for a decrease in allele number $\alpha(l, l-1)$ lead to the following

$$\alpha(l, l-1)=\min\left(\frac{(l-1)l}{2(2^{\eta_u+\eta_v-1}-1)\tilde{l}_2} \times\frac{p(\tilde{\mathbf{C}} \mid l-1)p(l-1)}{p(\mathbf{C} \mid l)p(l)}\times\frac{p(\mathbf{y} \mid \tilde{\boldsymbol{\theta}})}{p(\mathbf{y} \mid \boldsymbol{\theta})}, 1\right), \tag{12}$$

where $\eta_u$ and $\eta_v$ are the number of parents that carry the two alleles randomly chosen to be combined. The second term in $\alpha(l, l-1)$ simplifies to $\kappa(p, l) \div \kappa(p, l-1)$ and $\kappa(p, l)l \div \kappa(p, l-1)\lambda$ for the uniform and Poisson cases, respectively.

### (iii) *Simulations and analyses*

We simulated two mating designs, one typical of experimental QTL-mapping families and one typical of plant-breeding families. For the first design, three parents ($P_1$, $P_2$, and $P_3$) were mated to produce two families, $P_1 \times P_2$ and $P_2 \times P_3$ ($P_2$ was common to both families). Each family consisted of 150 doubled haploid (DH) progeny. For the second design, a total of 20 parents were mated in a circulant diallel to produce 60 families, each of 20 DH progeny.

A 10 cM chromosome segment flanked by markers was simulated with a QTL located at 4 cM. In the first design, three true QTL configurations were simulated. Using '=' to indicate identity in state at the QTL, these configurations were $P_1=P_3\neq P_2$, $P_1=P_2\neq P_3$ and $P_1\neq P_2\neq P_3$. In all cases, for those families in which the QTL segregated, the variance caused by it was either 10% or 20% of the phenotypic variance. In the second design, two true QTL configurations were simulated, one with four alleles each carried by five parents and the other with 20 alleles, each parent carrying its own unique allele. Allelic values were assigned to be evenly spread in terms of normal percentiles and standardized so that the average variance caused by the QTLs over the 60 families was either 6% or 12% of the phenotypic variance. For the simulated chromosomal segment, markers and QTL genotypes in the progeny were randomly assigned following the rules of Mendelian segregation and recombination using Haldane's mapping function. A random normal deviate was added to the genetic value conferred by the QTL to obtain the QTL heritabilities given above.

All Markov chains started with a burn in of 1000 iterations. We ran test analyses using either a single chain of 100 000 iterations initialized with parameters estimated from the data, or with ten chains of 10 000

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

iterations with starting points chosen at random from the parameter prior distributions. A preliminary analysis of the two approaches indicated that they did not differ in estimating parameter means or in predicting identity in state configurations among parents. We therefore ran further analyses using the single long-chain approach. To determine error in parameter estimates caused by MCMC sampling, we used the 'batch means' method (Roberts, 1995) with 100 batches, each 1000 iterations long. For estimating allelic number, the error variance associated with MCMC sampling was ~4% of the variance among analyses of independently simulated data. The proportion of the variance caused by MCMC sampling was higher for allelic number than for other parameters. We deemed this level of MCMC sampling error acceptable and continued running Markov chains with a total of 100 000 iterations. For each parameter combination, 15 replicate analyses were performed, with new data simulated for each analysis. Factors that differentiated the parameter combinations were mating design, true QTL configuration, QTL heritability, prior distribution for allelic values and prior distribution for allelic number (see sampling procedures above). We compared analyses allowing for a variable number of QTL alleles (variable model) with analyses that assumed each parent carried a unique allele (fixed model). Analyses assumed the presence of a QTL (we did not obtain posterior probabilities for the presence of a QTL). The position of the QTL analysed was also fixed at its simulated position (we did not obtain posterior distributions of the QTL position).

Model performance was evaluated according to the estimated allelic number $l$ and according to the mean squared error (MSE) of the estimated allelic values $a$. The MSE for a given allelic value is calculated from $T$ MCMC iterations as

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T}(\hat{a}_t - a)^2, \tag{13}$$

where $\hat{a}_t$ is the estimate for iteration $t$ and $a$ is the true value. The MSE can in turn be decomposed into a term caused by the sampling variance and a term caused by the bias of the estimate

$$\mathrm{MSE} = \mathrm{var}(\hat{a}) + (\mathrm{bias}(\hat{a}))^2, \tag{14}$$

where $\mathrm{bias}(\hat{a}) = \frac{1}{T}\sum_{t=1}^{T}\hat{a}_t - a$. For a given analysis, we calculated the MSE, sampling variance and bias of $a$ for each parent and averaged these responses over all parents.

As an exploratory analysis, the posterior means for $l$ and for the MSE for $a$ were analysed using analysis of variance. Analyses were conducted separately for each mating design and each QTL allele configuration within mating design. Factors in the analysis were allelic number prior (either Poisson or uniform), allelic value prior (with standard deviation of 0·5 or 1 as described above) and QTL heritability (low or high as described above).

## 3. Results

### (i) *Estimation of QTL allelic number*

Analysis of the influence of the prior distributions and the effect of segregating QTL revealed straightforward patterns in the case of the experimental mating design with three inbred parents of two families. Results for the two possible configurations in which only two alleles were segregating ($P_1 = P_3 \neq P_2$ versus $P_1 = P_2 \neq P_3$) were similar (data not shown) and we pooled them. With only two QTL alleles segregating, the analysis was sensitive to the assumed prior distribution of the QTL allelic number. For the Poisson distribution with mean parameter of 2 but truncated to [2..3], the prior mean is 2·4, whereas the uniform distribution of [2..3] has a mean of 2·5. This small difference in prior mean was reflected in the posterior mean, which increased by 0·063, averaged over QTL heritability and allelic value prior factors (Table 1). The analysis was most affected by the prior assumed for the allelic value parameter. The two priors used for the allelic value were narrow (prior standard deviation of half of the phenotypic standard deviation) or wide (prior standard deviation equal to the phenotypic standard deviation). Compared with the narrow prior for allelic number, the wide prior led to a lower posterior mean for the allelic number.

The pattern of the estimate of allele number was simpler when three QTL alleles were segregating. In that case, when the QTL had high heritability, the model virtually always selected a three-allele model, so that the estimate of allelic number was three with low standard deviation (Table 1). When the QTL had lower heritability, the analysis at times attempted to group either parents $P_1$ and $P_2$ or $P_2$ and $P_3$, such that the estimate of allelic number was slightly lower (Table 1). The priors for allelic number and allelic value did not affect the estimates of allelic number in these cases.

The most immediate observation from the posterior means for allelic number for the breeding mating design involving 20 parents and 60 families is that, when the true number of alleles segregating was 20 (that is, each parent carried its own unique allele), the estimate of that number was much lower than the true number. In estimating the number of alleles, the effects of the QTL heritability and of the priors were qualitatively similar to the three-parent mating design but were stronger and interacted with each other (Table 2). Given that there were 20 parents, the truncated Poisson prior had mean 4·3, whereas the mean of the

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

Table 1. *Posterior mean of allelic number as affected by QTL heritability ($H^2Q$), allelic number prior and allelic value prior. Standard deviations for posterior means given in parentheses*

| | Allelic number prior | Poisson distribution‡ | | Uniform distribution [2..*P*] | |
|---|---|---|---|---|---|
| | Allelic number prior† | 0·5 | 1·0 | 0·5 | 1·0 |
| Parents : alleles§ | $H^2Q$ | | | | |
| 3 : 2 | 0·10 | 2·30 (0·09) | 2·22 (0·13) | 2·37 (0·10) | 2·25 (0·10) |
| 3 : 2 | 0·20 | 2·31 (0·14) | 2·19 (0·10) | 2·39 (0·10) | 2·26 (0·15) |
| 3 : 3 | 0·10 | 2·89 (0·20) | 2·82 (0·21) | 2·98 (0·03) | 2·88 (0·16) |
| 3 : 3 | 0·20 | 3·00 (0·00) | 3·00 (0·00) | 3·00 (0·01) | 3·00 (0·00) |
| 20 : 4 | 0·06 | 3·58 (0·50) | 2·62 (0·24) | 4·17 (1·35) | 2·66 (0·40) |
| 20 : 4 | 0·12 | 4·23 (0·66) | 3·29 (0·43) | 5·83 (1·44) | 3·86 (0·54) |
| 20 : 20 | 0·06 | 4·05 (0·62) | 2·91 (0·45) | 4·85 (1·49) | 3·14 (0·61) |
| 20 : 20 | 0·12 | 5·16 (0·66) | 4·31 (0·69) | 8·97 (1·48) | 4·88 (0·91) |

† Prior was normal with a standard deviation equal to the fraction of the phenotypic standard deviation given.

‡ The Poisson prior was truncated to [2..*P*]. The mean parameter for the prior was 2 for the three-parent design and 4 for the 20-parent design.

§ Number of parents and simulated number of alleles carried by parents.

Table 2. *Model factors affecting inference of the number of alleles at a QTL as determined by analysis of variance*

| Number of parents | 3 | | 3 | | 20 | | 20 | |
|---|---|---|---|---|---|---|---|---|
| True allele number | 2 | | 3 | | 4 | | 20 | |
| Model factor | *F* value | | *F* value | | *F* value | | *F* value | |
| Allelic number prior (pNA) | 15·72 | ** | 3·41 | NS | 22·63 | ** | 61·59 | ** |
| QTL heritability ($H^2Q$) | 0·00 | NS | 24·28 | ** | 50·19 | ** | 147·54 | ** |
| Allelic value prior (pAV) | 45·15 | ** | 3·89 | NS | 82·85 | ** | 128·15 | ** |
| pNA × $H^2Q$ | 0·32 | NS | 3·37 | NS | 6·75 | NS | 23·83 | * |
| pNA × pAV | 0·51 | NS | 0·07 | NS | 7·01 | NS | 30·75 | * |
| $H^2Q$ × pAV | 0·32 | NS | 3·99 | NS | 0·59 | NS | 9·23 | NS |

NS, not significant; *, significant with $P<0{\cdot}001$; **, significant with $P<0{\cdot}0001$.

uniform prior was 11. The posterior mean was affected by this difference in prior mean, increasing by 1·03, averaged over QTL heritability, allelic value prior and true allelic number factors. Increasing the QTL heritability strongly increased the estimated number of alleles. For two cases in which four alleles were segregating, this increase was surprising because the resulting estimated allelic number was greater than the true number (Table 1). Thus, increasing QTL heritability and therefore QTL information content did not necessarily bring the estimated allelic number closer to the true number. As in the three-parent design, increasing the standard deviation of the prior for allelic value decreased the estimated number of alleles. The effects of allelic number prior and QTL heritability worked synergistically, in that increasing the prior mean for allelic number and increasing the QTL heritability together had a greater effect on the estimated number of alleles than predicted from either factor in isolation. Finally, increasing the standard deviation of the prior for allelic value tended to suppress the effects of increasing the prior for allelic number or of increasing QTL heritability.

### (ii) *Prediction of identity in state at the QTL*

To evaluate the effectiveness of the analysis at determining identity in state among parents, we checked whether the analysis grouped as identical in state pairs of parents simulated to carry alleles of equal effect and distinguished as not identical-in-state pairs of parents simulated to carry alleles of differing effects. Consider the prior probability that any two parents will be declared identical in state, $P_{\mathrm{IIS}}$. $P_{\mathrm{IIS}}$ is closely related to the allelic number assumed by the analysis; at the limit, if the analysis only considered one allele then $P_{\mathrm{IIS}}=1$; if it considered the same number of alleles as parents then $P_{\mathrm{IIS}}=0$. It is therefore more useful to consider the prior probability ($P_{\mathrm{IIS}}|l$), where $l$ is the allelic number estimated for an analysis. For

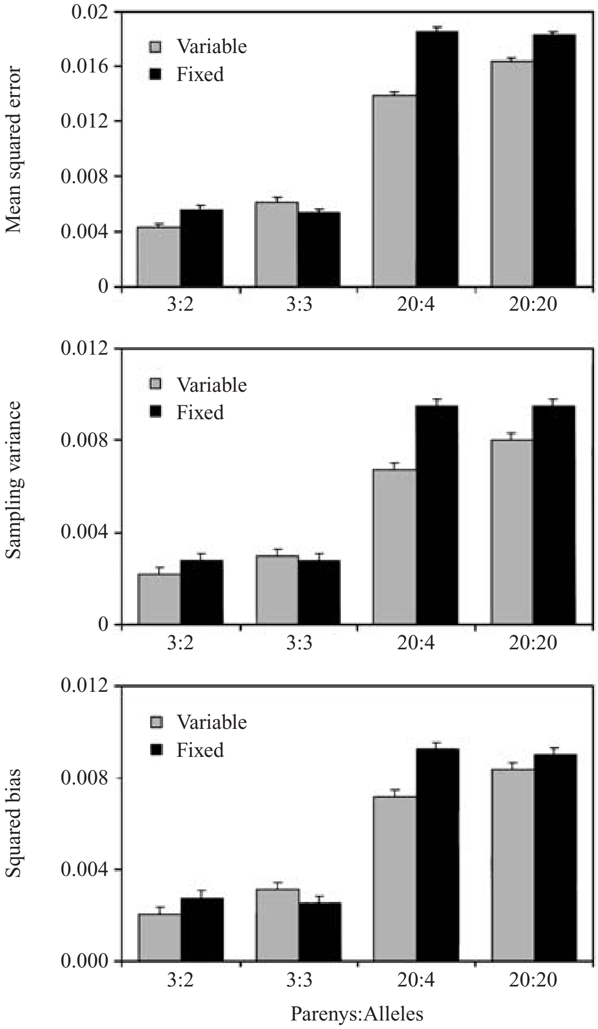

Fig. 1. Mean squared error, sampling variance and squared bias in the estimation of allelic values using either the variable-allele-number model or the fixed-allele-number model.

parents that were in fact identical in state, we also calculated a posterior probability of identity in state ($P_{\mathrm{IIS}} \mid l, \mathbf{y}$). For the three-parent, two-family mating design, ($P_{\mathrm{IIS}} \mid l$) averaged 0·24, whereas ($P_{\mathrm{IIS}} \mid l, \mathbf{y}$) averaged 0·63 and 0·70 for QTLs explaining 10% and 20% of the phenotypic variance, respectively. For the 20-parent, 60-family mating design, ($P_{\mathrm{IIS}} \mid l$) averaged 0·32, whereas ($P_{\mathrm{IIS}} \mid l, \mathbf{y}$) averaged 0·56 and 0·63 for QTLs explaining 6% and 12% of the phenotypic variance, respectively. Similarly, we define the prior probability that two parents are declared not identical in state, conditional on the estimated allelic number, as ($P_{\mathrm{NIS}} \mid l$)$=1-$($P_{\mathrm{IIS}} \mid l$). For pairs of parents that are in fact not identical in state, we also calculated ($P_{\mathrm{NIS}} \mid l, \mathbf{y}$). For the three-parent mating design, ($P_{\mathrm{NIS}} \mid l, \mathbf{y}$) averaged 0·96 and 1·00 for QTLs explaining 10% and 20% of the phenotypic variance, respectively. For the 20-parent mating design, ($P_{\mathrm{NIS}} \mid l, \mathbf{y}$) averaged 0·71 and 0·82 for QTLs explaining 6% and 12% of the phenotypic variance, respectively.

### (iii) *Estimation of QTL allelic value*

Unlike the estimate of allelic number, the posterior MSE of the estimate of allelic value was not sensitive to differences in QTL heritability, allelic number prior or allelic value prior (data not shown). We investigated whether the variable model estimates allelic values with lower MSE than the fixed model. In the three-parent mating design, when compared with the fixed model, the variable model produces a lower MSE when only two alleles were segregating but a higher MSE when three alleles were segregating (Fig. 1). The responses in the MSE were mirrored by qualitatively identical responses in the sampling variance and bias components of the MSE (Fig. 1). In the 20-parent mating design, the variable model estimated allelic values with the lowest MSE, irrespective of whether there were fewer alleles than parents or the same number of alleles as parents (Fig. 1). This result was surprising both because it contrasted with the result from the three-parent design and because the variable model was in fact incorrect when all parents carried a unique allele. Results for the sampling variance and bias components of the MSE were again qualitatively identical (Fig. 1).

## 4. Discussion

### (i) *Influence of the prior for allelic value on the posterior for allelic number*

Compared with the narrow prior for allelic number, the wide prior led to a lower posterior mean for the allelic number. In determining the posterior for allelic number, the analysis must select from models with different numbers of allelic value parameters and different possible identity-in-state configurations. Evaluating the effect of the allelic value prior on this model selection analytically would be a formidable task and we provide only an intuitive discussion. Formally, the posterior probability for a given model M is $P(M \mid y) \propto P(M)P(y \mid M)$, where the second term derives from an integration of the likelihood over the prior for the parameters in M: $P(y \mid M) = \int_{\boldsymbol{\theta}_M} P(y \mid \boldsymbol{\theta}_M)P(\boldsymbol{\theta}_M \mid M)d\boldsymbol{\theta}_M$. To see how the prior for $\boldsymbol{\theta}_M$ affects this likelihood, consider model $M_1$ with two parents grouped to carry the same allele and contrast it with model $M_2$ in which the parents carry different alleles. In $M_1$, there is no QTL substitution effect between the two parents. In $M_2$, the estimated QTL substitution $\hat{\delta}$ depends on the allelic value parameters $\hat{a}_1$ and $\hat{a}_2$ of the two parents: $\hat{\delta} = \hat{a}_1 - \hat{a}_2$. Assume now that $M_2$ is correct, such that there is some allelic substitution effect $\delta = \hat{a}_1 - \hat{a}_2 \neq 0$. If the prior for allelic value is vague, within a large fraction of the prior for

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

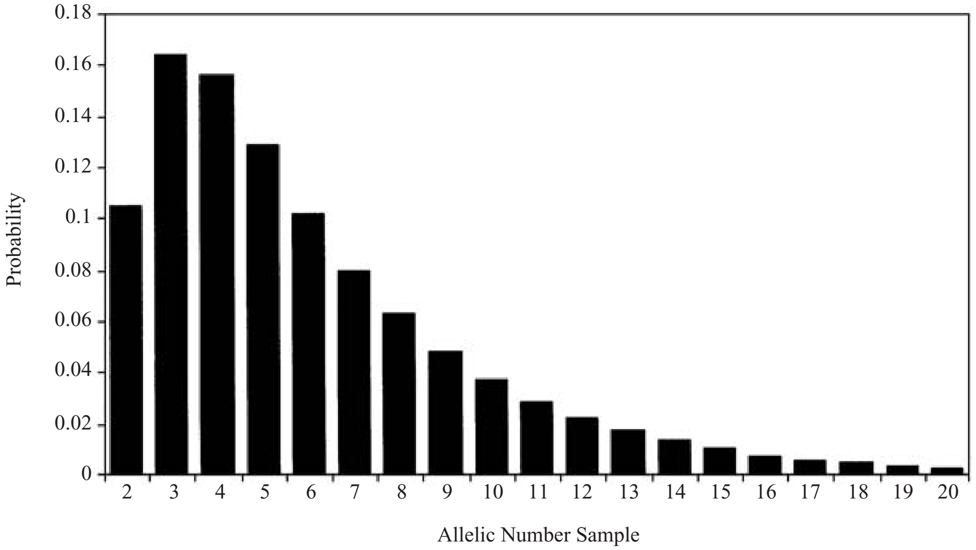

Fig. 2. Typical posterior distribution of allelic number for analyses of the 20-parent mating design with four alleles segregating, QTL heritability of 12%, uniform allelic number prior and standard deviation of allelic value prior of one-half of the phenotypic standard deviation.

$(\hat{a}_1, \hat{a}_2)$, the QTL substitution effect $\hat{\delta}$ will cause a lower likelihood than assuming no QTL substitution effect. Such decreased likelihood will occur when $\mathrm{sign}(\hat{\delta}) \neq \mathrm{sign}(\delta)$ and when $\mathrm{abs}(\hat{\delta}) \gg \mathrm{abs}(\delta)$. Thus, even when $M_2$ is correct, a vague prior for the allelic effects will cause the Bayesian analysis to favour $M_1$. Extending this idea to mating designs with many parents, models with fewer QTL alleles group more parents, and the posterior probability of these models will increase as the prior for allelic value becomes more vague. Sorensen & Gianola (2002) treat a simpler case analytically to demonstrate the influence of parameter priors on Bayesian selection between simple versus complex models. In selection between a normal model with known mean and a normal model with estimated mean, they show that a vague prior on the mean favours selection of the simple over the complex model and, in the extreme case of an improper prior for the mean, the posterior probability of the simple model is 1. Despite the simplicity of their example, it captures the essence of our observation that a wide prior on QTL allele effect leads to a lower inferred number of QTL alleles than a narrow prior.

In the specific case of a 20-parent mating design, when each parent carries a unique allele, the allelic values conferred by some pairs of parents will be quite close, such that grouping those parents will result in an imperceptible decrease in the likelihood. Thus, the data can be adequately modelled assuming the existence of fewer alleles than are in fact segregating. In some sense, what the Bayesian analysis provides is an estimate of the effective number rather than the true number of segregating alleles. Distinct alleles that nevertheless have similar effects contribute little to increasing the number of effective alleles.

### (ii) *Information content in the data to determine allelic number*

In the context of the 20-parent breeding mating design, three lines of evidence indicate that phenotypic and DNA-marker data contain relatively little information that enables an estimate of QTL allele number. First, we have shown that the priors for allelic number and allelic value strongly influence the estimate of allelic number (Table 1). Second, the posterior distribution for allelic number has high variability, indicating that, during MCMC sampling, there is little likelihood penalty to either underestimating or overestimating the allelic number (Fig. 2). Finally, the posterior estimate for allelic number was highly variable across replicate simulations, particularly under the uniform prior distribution for allelic number (Table 2, Fig. 3). Three sources of variation contribute to between-replicate differences: MCMC sampling error; genetic sampling of the evaluated lines; and environmental error on each line. Because the error caused by MCMC sampling was small (see Methods) between-replicate differences indicate that the last two sources strongly influence the allelic-number estimate. These observations beg the question of the family size necessary to obtain reliable estimates of allele number. We ran replicate simulations of the 20-parent mating design under the uniform prior for allelic number with different family sizes (Fig. 4). The non-monotonic behaviour of the mean estimate of allelic number as the family size increases is noteworthy. It suggests that, as family size increases, different information sources dominate the analysis, presumably as follows. With very small families (i.e. two or three individuals per family), the

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

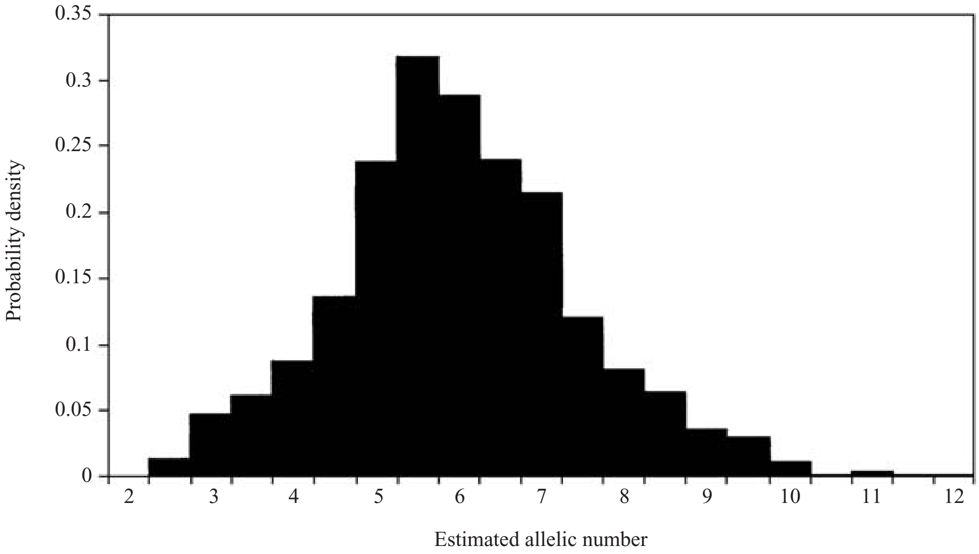

Fig. 3. Distribution of estimated allelic numbers from 2000 replicate simulations of the 20-parent mating design with four alleles segregating, QTL heritability of 12 %, uniform allelic number prior and standard deviation of allelic value prior of one-half of the phenotypic standard deviation.

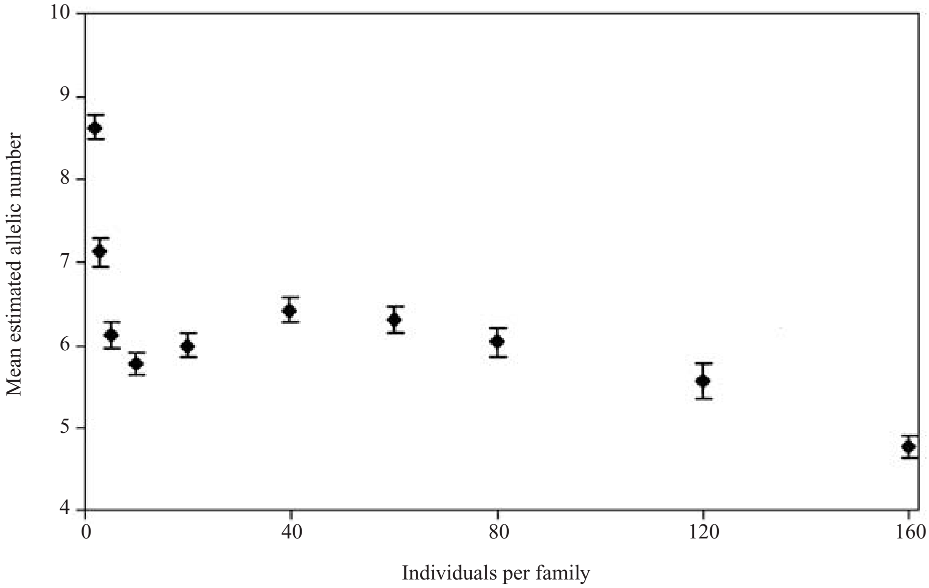

Fig. 4. Mean estimated allelic numbers from replicate simulations of the 20-parent mating design with four alleles segregating, QTL heritability of 12 %, uniform allelic number prior and standard deviation of allelic value prior of one-half of the phenotypic standard deviation with different numbers of individuals per family.

estimate for allelic number is dominated by the prior information on allelic number. With small families (5–20 individuals), the likelihood of models with high allelic numbers decreases because of the effect of the prior information on allelic value, as described in the preceding section. With medium-sized families (40–80 individuals), information from the data causes the likelihood of configurations that falsely group parents carrying different alleles to drop. Consequently, the posterior probability of models with allelic numbers lower than the true allelic number becomes small. Finally, with large families ( > 150 individuals), data information consistently favours not only distinguishing parents that carry different alleles but grouping parents that carry the same allele. The estimated allelic number therefore approaches the true number.

Although this explanation is intuitively appealing, a more rigorous analysis is desirable. In particular, we show here that Bayesian model selection does not necessarily pick most parsimonious models and that the relationship between data information content and selected-model parsimony is complex. Given the growing interest in Bayesian selection across models

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

of different dimensions, particularly in the context of QTL mapping (e.g. Sillanpää & Arjas, 1998; Yi & Xu, 2001), greater understanding of what determines the selection of simple or complex models is needed. In the case discussed here, interactions between prior information and data information are complicated. With small family sizes, the estimated allelic number will reflect these interactions and probably has no simple interpretation. Only if the data at hand include 150 or more individuals per family can the estimated allelic number be interpreted as reflecting the true allelic number.

### (iii) *Prior information on identity in state among parents*

In the analyses presented, we have assumed no prior information on identity in state among parents. In fact, such information could derive from the pedigree of the parents if it was known. Using the pedigree, probabilities of identity by descent can be calculated (Lynch & Walsh, 1998) and further refined using linkage methods if marker data is available on the parents' ancestors (Fernando & Grossman, 1989; Goddard, 1992; van Arendonk *et al.*, 1994). A second source of prior information could derive from linkage disequilibrium among markers surrounding the QTL being analysed. Sufficient linkage disequilibrium creates a relationship between the parent's marker haplotype and the identity of the QTL allele that it carries. This relationship could be modelled to give the probability of QTL identity between two parents as a function of the similarity between their marker haplotypes. Linkage disequilibrium and marker haplotype information have been proposed in random-effect QTL models (Meuwissen & Goddard, 2000) and in fixed effect QTL models (Jansen *et al.*, 2003). Incorporating such prior information should improve estimates of allelic effects.

### (iv) *Estimation of allelic value*

When the variable-model groups parents as carrying the same allele, it also pools the progeny derived from those parents to obtain an estimate of allelic value. Presumably, combining observations on alleles of similar value so that more observations contribute to a single parameter enables the variable model to estimate allelic value with lower sampling variance. Pooling observations to estimate parameters proved to be important in the 20-parent design because of its small family size.

Despite the fact that the variable model estimated allelic value with lower MSE and lower squared bias averaged across all parents, we found that it introduces a systematic bias. In particular, estimates of the allelic values of alleles with extreme effects are smaller than the true effects. This moderation of the estimates arises because, when different alleles are incorrectly grouped, extreme alleles must inevitably be grouped with less-extreme alleles, pulling their resulting allelic value estimate toward the mean. We regressed estimated allelic value on true allelic value and found regression coefficients significantly lower than one for the variable model but not significantly different from one for the fixed model (Table 3, Fig. 5). Regression coefficients were little affected by the actual number of alleles segregating (data not shown). This paradoxical combination of lower mean bias with systematic bias can be explained by a linear model of the estimated on the true allelic value: $\hat{a}_{ij} = b_0 + b_1 a_i + e_{ij}$, where $\hat{a}_{ij}$ is the estimated allelic value for allele $i$ in analysis $j$, $a_i$ is the true value for that allele and $e_{ij}$ is a residual. The intercept for this linear model was consistently very close to zero and we omit it from the analysis below.

Table 3. *Regression of estimated allelic value on the true allelic value as affected by analysis model and QTL heritability. The residual variance about regression is given in parentheses*

| Number of parents | QTL heritability | Variable model | Fixed model |
|---|---|---|---|
| 3 | 0·10 | 0·962 (0·0029) | 0·992 (0·0029) |
| 3 | 0·20 | 0·969 (0·0018) | 0·970 (0·0025) |
| 20 | 0·06 | 0·755 (0·0058) | 0·984 (0·0093) |
| 20 | 0·12 | 0·882 (0·0076) | 0·987 (0·0090) |

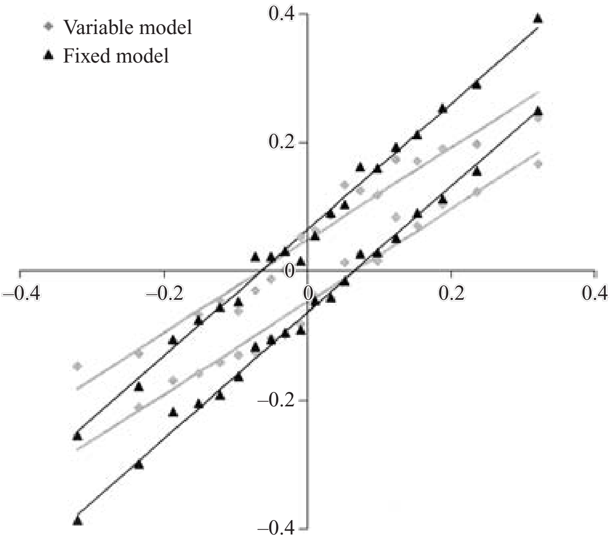

Fig. 5. Regression of 25% and 75% quartiles of estimated allelic value on the true allelic value (triangles and black line for variable model, and diamonds and grey line for fixed model). QTL heritability was 6% with the configuration in which all parents carried unique alleles; analyses combined over allelic value and allelic number priors.

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

We are interested in the expectation of the squared bias:

$$
\begin{aligned}
E\{[\mathrm{bias}(\hat{a})]^2\} &= E[(\hat{a}_{ij}-a_i)^2] \\
&= E\{[(b_1-1)a_i+e_{ij}]^2\} \\
&= (b_1-1)^2E(a^2)+E(e^2).
\end{aligned}
$$

This equality follows from the property of regression that the predictor variable has zero covariance with the residual. Finally

$$
\begin{aligned}
E\{[\mathrm{bias}(\hat{a})]^2\} &= (b_1-1)^2E(a^2)+E(e^2) \\
&= (b_1-1)^2\mathrm{var}(a)+\mathrm{var}(e), \qquad (15)
\end{aligned}
$$

because $E(a)=E(e)=0$. Eqn 15 allows us to compare the sources of bias in the variable and fixed models. The var($a$) term is fixed within a QTL heritability class. For the fixed model, $(b_1-1)^2 \approx 0$, whereas, for the variable model, $(b_1-1)^2>0$ (Table 3). Because, despite this systematic effect, the squared bias for the variable model is lower than that for the fixed model, the residual variance around the regression must also be lower. We did in fact observe lower residual variances in the variable model than in the fixed model (Table 3). These relationships are illustrated in Fig. 2. The variable model shows a shallower slope but also a narrower interquartile range. This narrower interquartile range gives the variable model its advantage over the fixed model.

### (v) *Extension to other mating designs and unknown QTL locations*

We have simplified the development of this analysis by assuming inbred parents and progeny. Extensions to mating designs with non-inbred parents or progeny do not necessitate the introduction of new concepts. If parents are not inbred, the dimensions of the allelic configuration matrix **C** would expand to $2P \times l$ so that the matrix could keep track of the identity of both maternally and paternally derived alleles of each parent. A vector of dominance interactions between all allele pairs would be needed to model the phenotypes of non-inbred progeny. Increasing the allelic number from $l$ to $l+1$ would require $l$ new dominance parameters on top of the one additive parameter. All new allelic-effect parameters could be drawn from their priors as we have done in the simple case described. In that case, calculation of the Metropolis–Hastings acceptance probability would be exactly the same as for the simpler additive model.

We have also simplified the analysis by assuming the QTL position to be known and focussing all MCMC iterations on that single position. An analysis that also estimated the posterior density of QTL position would require a procedure to update QTL position and would estimate allelic number conditional on position. Several procedures to update QTL position based on the Metropolis–Hastings algorithm have been published (Satagopan *et al.*, 1996; Sillanpää & Arjas, 1998; Yi & Xu, 2000) and could be adapted to the present context. Estimating allelic number conditional on position would require saving MCMC realizations in a vector indexed by QTL position, as has been proposed for estimating QTL effect or variance conditional on position (Sillanpää & Arjas, 1998; Yi & Xu, 2001). To obtain adequate estimates of QTL parameters conditional on position, methods that estimate QTL position naturally require more MCMC iterations than methods that assume the position known. The increase in the number of iterations might not be that great, however, given that most iterations will occur at positions of high posterior probability, focusing the iterations in much the same way as restricting the QTL to a known position.

We thank Rohan L. Fernando and Hal S. Stern for valuable discussion. Two anonymous reviewers also provided helpful feedback. This research was funded by USDA-NRI, CSREES Project Award No. 2001-35301-10848, and supported by Hatch Act and State of Iowa.

## References

Brummer, E. C., Graef, G. L., Orf, J. H., Wilcox, J. R. & Shoemaker, R. C. (1997). Mapping QTL for seed protein and oil content in eight soybean populations. *Crop Science* **37**, 370–378.

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

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall.

Goddard, M. E. (1992). A mixed model for analyses of data on multiple genetic markers. *Theoretical and Applied Genetics* **83**, 878–886.

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

Jansen, R. C. (1994). Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Jansen, R. C., Jannink, J.-L. & Beavis, W. D. (2003). Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. *Crop Science* (in press).

Kianian, S. F., Egli, M. A., Phillips, R. L., Rines, H. W., Somers, D. A., Gengenbach, B. G., Webster, F. H., Livingston, S. M., Groh, S., LS, O. D., Sorrells, M. E., Wesenberg, D. M., Stuthman, D. D. & Fulcher, R. G. (1999). Association of a major groat oil content QTL and an acetyl-CoA carboxylase gene in oat. *Theoretical and Applied Genetics* **98**, 884–894.

Liu, Y. & Zeng, Z.-B. (2000). A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. *Genetical Research* **75**, 345–355.

Lynch, M. & Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits.* Sunderland, MA: Sinauer Associates.

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press

Meuwissen, T. H. E. & Goddard, M. E. (2000). Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. *Genetics* **155**, 421–430.

Orf, J. H., Chase, K., Jarvik, T., Mansur, L. M., Cregan, P. B., Adler, F. R. & Lark, K. G. (1999). Genetics of soybean agronomic traits: I. Comparison of three related recombinant inbred populations. *Crop Science* **39**, 1642–1651.

Rebaï, A. & Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theoretical and Applied Genetics* **86**, 1014–1022.

Rebaï, A. & Goffinet, B. (2000). More about quantitative trait locus mapping with diallel designs. *Genetical Research* **75**, 243–247.

Roberts, G. O. (1995). Markov chain concepts related to sampling algorithms. In *Markov Chain Monte Carlo in Practice* (ed. W. R. Gilks, S. Richardson & D. J. Spiegelhalter), pp. 45–58. London: Chapman & Hall.

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

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

Sorensen, D. & Gianola, D. (2002). *Likelihood, Bayesian and MCMC Methods in Genetics*. Berlin: Springer-Verlag.

van Arendonk, J. A. M., Tier, B. & Kinghorn, B. P. (1994). Use of multiple genetic markers in prediction of breeding values. *Genetics* **137**, 319–329.

Waagepetersen, R. & Sorensen, D. (2001). A tutorial on reversible jump MCMC with a view toward QTL-mapping. *International Statistical Review* **69**, 49–61.

Xu, S. Z. (1998). Mapping quantitative trait loci using multiple families of line crosses. *Genetics* **148**, 517–524.

Yi, N. J. & Xu, S. Z. (2000). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics* **156**, 411–422.

Yi, N. J. & Xu, S. Z. (2001). Bayesian mapping of quantitative trait loci under complicated mating designs. *Genetics* **157**, 1759–1771.

https://doi.org/10.1017/S0016672303006153 Published online by Cambridge University Press