Theor Appl Genet (2004) 108:1434–1442
DOI 10.1007/s00122-003-1569-5

ORIGINAL PAPER

Xiao-Lin Wu · Jean-Luc Jannink

Optimal sampling of a population to determine QTL location, variance, and allelic number

Received: 28 May 2003 / Accepted: 4 December 2003 / Published online: 23 January 2004

Abstract In a population intended for breeding and selection, questions of interest relative to a specific segregating QTL are the variance it generates in the population, and the number and effects of its alleles. One approach to address these questions is to extract several inbreds from the population and use them to generate multiple mapping families. Given random sampling of parents, sampling strategy may be an important factor determining the power of the analysis and its accuracy in estimating QTL variance and allelic number. We describe appropriate multiple-family QTL mapping methodology and apply it to simulated data sets to determine optimal sampling strategies in terms of family number versus family size. Genomes were simulated with seven chromosomes, on which 107 markers and six QTL were distributed. The total heritability was 0.60. Two to ten alleles were segregating at each QTL. Sampling strategies ranged from sampling two inbreds and generating a single family of 600 progeny to sampling 40 inbreds and generating 40 families of 15 progeny each. Strategies involving only one to five families were subject to variation due to the sampling of inbred parents. For QTL where more than two alleles were segregating, these strategies did not sample QTL alleles representative of the original population. Conversely, strategies involving 30 or more parents were subject to variation due to sampling of QTL genotypes within the small families obtained. Given these constraints, greatest QTL detection power was obtained for strategies involving five to ten mapping families. The most accurate estimation of the variance generated by the QTL, however, was obtained with strategies involving 20 or more families. Finally, strategies with an intermediate number of families best estimated the number of QTL alleles. We conclude that no overall optimal sampling strategy exists but that the strategy adopted must depend on the objective.

Communicated by P. Langridge

X.-L. Wu · J.-L. Jannink (✉)
Department of Agronomy, Iowa State University,
Ames, IA 50011-1010, USA
e-mail: jjannink@iastate.edu
Tel.: +1-515-2944153
Fax: +1-515-2946505

X.-L. Wu
Center for Life Science Research, Hunan Agricultural University,
410128 Hunan, China

Introduction

In a population intended for breeding and selection, questions of interest relative to a specific segregating QTL are the variance it generates in the population, and the number and effects of its alleles. Standard QTL mapping methods that employ progeny developed from a cross between two inbred parents deal with a maximum of two alleles per QTL so that they cannot, in general, address the questions of allelic number and effects at the population level. One possible approach to address these questions would be to extract several inbreds from the population and use them to generate multiple mapping families. In addition to making the above questions answerable (Jannink and Wu 2003), joint analysis of these families could lead to gains in detection power and mapping resolution (Muranty 1996; Xu 1998; Rebaï and Goffinet 2000; Jannink and Jansen 2001). Appropriate statistical techniques to identify QTL in multiple related families and evaluation of optimal sampling strategies for the development of such families are therefore needed.

Statistical methods that combine multiple families into one analysis based on fixed QTL allele effect models have been proposed (Rebaï and Goffinet 1993, 2000; Rebaï et al. 1994; Liu and Zeng 2000). A major concern with these methods is that the number of parameters to be estimated increases with the number of families analyzed. In contrast, random allele effect models (Xie et al. 1998; Xu 1998) estimate only a single QTL variance so that the number of model parameters per QTL is independent of the number of families. In addition, random allele effect models deal with multiple family QTL mapping more naturally, shifting the focus from the effects of specific

alleles to the variance generated by the QTL over the original population. Regardless of the fixed or random nature of the QTL effect, current methods all assume that each inbred parent carries a distinct allele. In other words, these methods assume that the number of alleles at a QTL is equal to the number of the parents used to generate mapping families. Unless an infinite allele model holds, there is a non-null probability that two inbred parents extracted from a population will carry the same QTL allele, that is, they will be identical in state (IIS) at the QTL. Knowledge of allelic IIS relationships among parents, while of interest in and of itself, also provides a means to estimate the number of alleles segregating at the QTL (Jannink and Wu 2003). To obtain and use this information, we proposed a single-QTL model in which the number and IIS configuration of QTL alleles were treated as unknown and subjected to Bayesian analysis (Jannink and Wu 2003). Here, we extend this method to multiple QTL that have not been previously mapped, and to the estimation of the variance expected to be generated by the QTL.

The analysis presented here assumes that random individuals have been extracted from the population and inbred by doubling one of their gametes to create a doubled-haploid (DH) parent. These parents are then crossed to generate DH-QTL mapping families. Given random sampling of parents, sampling strategy may be an important factor determining the power of the analysis and its accuracy in estimating QTL variance and allelic number. Uncertainty in sampling QTL alleles will take place at two levels: when sampling founders to produce the F_1 parents, and during Mendelian segregation of the F_1 parents to produce DH progeny families. Previous reports have examined sampling strategy effects on QTL detection using maximum likelihood approaches. Xu (1998) showed that different family number versus family size sampling strategies affected QTL mapping results using a random- or fixed-allele effect model. The power of QTL detection was higher with an intermediate number (10–20) of families than with a small number (1–6) of large families or a large number (50–100) of small families. QTL mapping using a single family caused the most severe loss in power and large bias and errors of QTL estimation. Using more than a single family for QTL mapping clearly reduced type II error in QTL mapping (Xu 1998). The cause of the loss of QTL detection power when analyzing a large number of small families was unclear. Using an identity by descent (IBD) variance component method, Xie et al. (1998) found that the likelihood ratio at the simulated QTL with 25 sibs per family was 500% higher than that with only two sibs per family. Their results again indicated that large family size is desirable to increase QTL detection power. IBD-based QTL analysis relies on modeling the covariance of sibs given the marker alleles they received, such that using two progeny per family did not give enough among-sib information for analysis. Rao and Li (2000) investigated the effect of sampling strategy (family number vs family size) for mapping categorical traits. They found that QTL detection power was greatest using a small number of large families (2–5 families each with 250–100 progeny), especially when QTL effect was small. All of these analyses made specific assumptions concerning QTL allelic number and allele frequencies. If allele frequencies are intermediate, then allelic diversity is likely to be captured within a small number of parents. If allele frequencies are more extreme, however, more parents will be needed to obtain a representative sample. While the effect of sampling strategy on QTL detection power has been examined, sampling strategy implications for the accuracy of estimation of QTL allelic number and variance have not. In what follows we describe extensions of the methodology presented in Jannink and Wu (2003). These extensions enable the mapping of multiple QTL and the estimation of their variances. We apply the complete methodology to simulated data sets to determine optimal strategies for estimating these parameters.

Materials and methods

We describe the linear models and algorithms used to perform the analyses below. The software package we developed and used is available at http://www.public.iastate.edu/~jjannink/Research/Software.htm.

QTL model

The QTL model is presented based on DH mapping families where only additive effects of QTL are involved. Consider f F_1 individuals derived from P founder parents that are inbred and unrelated to each other. The F_1 parents were used to produce a total of n DH progeny individuals. Assuming that the trait is affected by n_{qtl} QTL, the vector of observed phenotypes y can be modeled as

$$y = X\beta + \sum_{j=1}^{n_{qtl}} Q_j C_j a_j + \varepsilon \quad (1)$$

where X is an $n\times f$ design matrix relating progeny to F_1 parents (families), β is a $f\times 1$ vector of family means, Q_j is a $n\times P$ inheritance matrix indicating from which founder parent a DH progeny received alleles at QTL j, C_j is a $P\times l_j$ allelic configuration matrix linking a founder parent to a specific allele at QTL j (l_j is the number of alleles at QTL j), $a_j \sim N\left(0, I\sigma_j^2\right)$ is a $l_j\times 1$ vector of allelic effects at QTL j, and $\varepsilon \sim N\left(0, I\sigma_e^2\right)$ is a $n\times 1$ vector of the residuals.

Observable model variables include trait values y and marker genotypes M. Denoting $Q=\{Q_j\}$, $C=\{C_j\}$, $a=\{a_j\}$, $l=\{l_j\}$, and $\sigma_a^2 = \left\{\sigma_j^2\right\}$, the unobservable parameters in the model are $\theta = \left(\beta, Q, C, \lambda, a, l, \sigma_a^2, \sigma_e^2\right)$, where λ is a $n_{qtl}\times 1$ vector of QTL positions. The joint posterior density of all unobservables given the observables and prior information is

$$p(\theta|y) \propto p(y|\theta)p(\beta)p(\lambda)p(Q|\lambda, M)p(l)p(C|l)p\left(a|\sigma_a^2\right)p\left(\sigma_a^2\right)p\left(\sigma_e^2\right) \quad (2)$$

where $p(y|\theta)$ is the likelihood assuming $\varepsilon \sim N\left(0, I\sigma_e^2\right)$, $p(*)$ is the prior distribution for parameter *, $p(C|l)$ is the prior distribution for QTL configuration conditional on the number of QTL alleles, $p(\theta|\lambda,M)$ is the prior distribution for QTL genotypes conditional on a QTL location and marker genotypes, and is derived from the rules of Mendelian segregation and recombination, and $p\left(a|\sigma_a^2\right)$ is the prior distribution for allelic values conditional on QTL variance and is multivariate normal. The prior distribution for QTL allelic number is assumed to be a truncated Poisson distribution between 2

and the number of founder parents. In MCMC analyses described below we used 4 as the mean of this Poisson distribution. The prior for QTL position is uniform over the genome. The prior distributions for $\{\beta_f\}$, $\left\{\sigma_j^2\right\}$, and σ_e^2 are assumed to be uniform on predefined intervals, where the prior for the family mean is positive with a maximum of twice the phenotypic mean, and the priors for QTL variance and the residual variance are positive with a maximum of twice the phenotypic variance.

Markov chain Monte Carlo sampling procedures

Bayesian inference was used to obtain the marginal posterior probability for each parameter of interest. We used the Markov chain Monte Carlo (MCMC) algorithm to generate samples from the joint posterior density, from which a marginal distribution was inferred. A scalar Metropolis Hastings procedure was used in all the following steps except step 5, with each parameter in θ sampled in turn considering all other parameters fixed (Gilks et al. 1996). In step 5, QTL location and genotypes were updated simultaneously. After initializing unobserved variables by sampling them from their priors, the MCMC iterations consisted of the following steps:

1. Update the number of QTL alleles for each QTL j
2. Update allelic configuration C_j at each QTL conditional on the number of alleles
3. Update QTL variance σ_j^2 at each QTL
4. Update QTL allelic effects a_j at each QTL
5. Update QTL position λ_j and QTL genotypes Q_j jointly for each QTL
6. Update family means β_f
7. Update residual variance σ_e^2

Step 1 involves changing the number of alleles via a reversible jump algorithm and leads to necessary changes in the dimensions of matrix C_j. Unlike Sillanpää and Arjas (1998), we do not estimate the number of QTL in this analysis but keep it fixed at n_{qtl}. In the analyses described below, Markov chains were initialized by placing one QTL at the center of each of seven chromosomes, thus fixing n_{qtl} at seven.

Updating QTL allelic number and QTL configuration

QTL allelic number and configuration are updated locus by locus, as discussed by Jannink and Wu (2003). Briefly, in each simulation cycle, we give equal probability to adding a new allele into the model or deleting an existing allele from the model. We propose an increase of one allele only if the current number of alleles is less than the number of founders ($l_j < P$), and a decrease of one allele only if the current number of alleles is greater than two ($l_j > 2$). To increase allelic number, an allele carried by more than one parent is randomly chosen and the parents carrying this allele are randomly divided into two groups. A new allele is created with a new value generated from its prior distribution. One group of parents is then shifted to carrying this new allele. To decrease allelic number, two alleles are chosen at random. The parents that carry them are grouped together and assumed to carry a single allelic effect.

To update QTL configuration conditional on the number of alleles ($C_j \mid l_j$), we randomly select one allele among those that are carried by more than one parent at QTL j. The proposal consists of shifting one random parent from carrying this allele to carrying an arbitrary different allele. This shift does not change the number of alleles at the locus.

Updating QTL variances and QTL effect

To update QTL variances, a proposal $\tilde{\sigma}_j^2$ is sampled at QTL j from a symmetric uniform density around the previous value σ_j^2

$$\tilde{\sigma}_j^2 \left| \sigma_j^2 \right. \sim \text{unif}\left[\max\left(0, \sigma_j^2 - d\right), \sigma_j^2 + d\right] \tag{3}$$

where d is the radius of change in QTL variance. As a reverse move, $\sigma_j^2 \left| \tilde{\sigma}_j^2 \right.$ is also a symmetrically uniform around $\tilde{\sigma}_j^2$. The proposal is accepted with the following probability

$$\alpha\left(\sigma_j^2, \tilde{\sigma}_j^2\right) = \min\left(1, \frac{\prod_{k=1}^{l_j}\prod_{r=1}^{P_k} p\left(a_{k.j}^{(r)} \left| \tilde{\sigma}_j^2 \right.\right)}{\prod_{k=1}^{l_j}\prod_{r=1}^{P_k} p\left(a_{k.j}^{(r)} \left| \sigma_j^2 \right.\right)} \times \frac{p\left(\sigma_j^2 \left| \tilde{\sigma}_j^2 \right.\right)}{p\left(\tilde{\sigma}_j^2 \left| \sigma_j^2 \right.\right)}\right) \tag{4}$$

where P_k is the number of founder parents carrying allele k, $a_{k.j}^{(r)}$ is the value of allele k at QTL j that is carried by parent r, and $p\left(a_{k.j}^{(r)} \left| \sigma_j^2 \right.\right)$ and $p\left(a_{k.j}^{(r)} \left| \tilde{\sigma}_j^2 \right.\right)$ are normal proposal density of allelic values given old and new QTL variance, respectively.

Allelic value is updated for each allele at each QTL. The prior for $a_{k.j}$ is normal with zero mean and variance σ_j^2. The proposal density for allelic values is uniform centered on the previous parameter value using a strategy similar to (3). The acceptance probability is

$$\alpha\left(a_{k.j}, \tilde{a}_{k.j}\right) = \min\left(1, \frac{p\left(\mathbf{y} \left| \tilde{\theta} \right.\right) p\left(\tilde{a}_{k.j} \left| \sigma_j^2 \right.\right)}{p\left(\mathbf{y} \left| \theta \right.\right) p\left(a_{k.j} \left| \sigma_j^2 \right.\right)}\right) \tag{5}$$

where $\tilde{\theta}$ is identical to θ except that $a_{k.j}$ is replaced by $\tilde{a}_{k.j}$.

Updating QTL position and QTL inheritance matrix jointly

To update genotypes and position jointly at a QTL j involves the proposal $q\left(\tilde{Q}_j, \tilde{\lambda}_j \left| Q_j, \lambda_j \right.\right)$ where λ_j and $\tilde{\lambda}_j$ are the current and candidate position for QTL j, and Q_j and $\tilde{Q}_j$ are the current and candidate QTL genotypes of all progeny at this QTL. First, $\tilde{\lambda}_j$ is generated from a uniform distribution $[\max(\lambda_{j-1}, \lambda_j - d), \min(\lambda_{j+1}, \lambda_j + d)]$, where d is a tuning parameter that affects the proposal probability and the acceptance rate. Next, QTL genotypes are sampled independently for all DH progeny conditional on flanking loci but independently of phenotypic data. A flanking locus can be a marker or a QTL, whichever is closer to QTL j. The proposal is accepted with the following probability:

$$\alpha\left(Q_j, \lambda_j, \tilde{Q}_j, \tilde{\lambda}_j\right) = \min\left(1, \frac{q\left(Q_j, \lambda_j \left| \tilde{Q}_j, \tilde{\lambda}_j \right.\right)}{q\left(\tilde{Q}_j, \tilde{\lambda}_j \left| Q_j, \lambda_j \right.\right)} \times \frac{p\left(\tilde{Q}_j, \tilde{\lambda}_j \left| y, M, \theta^- \right.\right)}{p\left(Q_j, \lambda_j \left| y, M, \theta^- \right.\right)}\right) \tag{6}$$

where θ^- corresponds to all elements in θ without the position and all genotypes at QTL j. This ratio can be simplified as follows:

$$q\left(Q_j, \lambda_j \left| \tilde{Q}_j, \tilde{\lambda}_j \right.\right) = p\left(\lambda_j \left| \tilde{\lambda}_j \right.\right) p\left(Q_j \left| \lambda_j, M \right.\right) \tag{7}$$

since λ_j is independent of $\tilde{Q}_j$, and Q_j is independent of $\tilde{\lambda}_j$ and $\tilde{Q}_j$. Similarly,

$$q\left(\tilde{Q}_j, \tilde{\lambda}_j \left| Q_j, \lambda_j \right.\right) = p\left(\tilde{\lambda}_j \left| \lambda_j \right.\right) p\left(\tilde{Q}_j \left| \tilde{\lambda}_j, M \right.\right) \tag{8}$$

For the posterior terms, we have

$$\begin{aligned} p\left(\tilde{Q}_j, \tilde{\lambda}_j \left| y, M, \theta^- \right.\right) &\propto p\left(\tilde{Q}_j, \tilde{\lambda}_j, y, M, \theta^-\right) \\ &= p\left(y \left| \tilde{Q}_j, \theta^- \right.\right) p\left(\tilde{Q}_j \left| \tilde{\lambda}_j, M \right.\right) p\left(\tilde{\lambda}\right) p(M) p\left(\theta^-\right) \end{aligned} \tag{9}$$

and

$$\begin{aligned} p\left(Q_j, \lambda_j \left| y, M, \theta^- \right.\right) &\propto p\left(Q_j, \lambda_j, y, M, \theta^-\right) \\ &= p\left(y \left| Q_j, \theta^- \right.\right) p\left(Q_j \left| \lambda_j, M \right.\right) p(\lambda) p(M) p\left(\theta^-\right) \end{aligned} \tag{10}$$

Putting these terms in Eq. 6 gives

$$\alpha\left(Q_j, \lambda_j, \tilde{Q}_j, \tilde{\lambda}_j\right) = \min\left(1, \frac{p\left(\lambda_j \left| \tilde{\lambda}_j \right.\right) p\left(y \left| \tilde{Q}_j, \theta^- \right.\right)}{p\left(\tilde{\lambda}_j \left| \lambda_j \right.\right) p\left(y \left| Q_j, \theta^- \right.\right)}\right) \tag{11}$$

Updating family means and residual variance

Denote β_f as the mean for family f. A family mean is sampled with a proposal density similar to (3). The proposal is accepted with the following acceptance probability

$$\alpha(\beta_f, \tilde{\beta}_f) = \min\left(1, \frac{p(\mathbf{y}|\tilde{\theta})}{p(\mathbf{y}|\theta)} \times \frac{p(\beta_f|\tilde{\beta}_f)}{p(\tilde{\beta}_f|\beta_f)}\right) \quad (12)$$

where $\tilde{\theta}$ is identical to θ except that β_f is replaced by $\tilde{\beta}_f$.

We update the residual variance similarly to updating the family mean, except that residual variance is assumed to be equal across all families. The proposal is sampled from a symmetric uniform density around the previous value. The residual variance is updated with a probability defined similarly to (12).

Calculating marginal posterior distributions

Following Sillanpää and Arjas (1998), the location-wise posterior QTL density supplies evidence for the presence of a QTL. We divide each chromosome into 2 cM bins (denoted Δ_1, Δ_2, ..., Δ_t, ...). Let

$$\hat{I}_t = \left[\frac{1}{n_s}\sum_{s=1}^{n_s}\sum_{j=1}^{n_{\text{qtl}}} 1_{\left\{\lambda_j^{(s)} \in \Delta_t\right\}}\right] / 0.02 \quad (13)$$

be the estimated posterior QTL intensity on interval Δ_t obtained from the Monte Carlo simulation, where n_s is the number of saved MCMC iterations, $\sum_{j=1}^{n_{\text{qtl}}} 1_{\left\{\lambda_j^{(s)} \in \Delta_t\right\}}$ is the number of QTL in bin Δ_t in iteration s. The intensity $\hat{I}_t$ integrates to n_{qtl} over the genome.

For assessing QTL variance and allelic number, location-wise posterior densities are defined. Let $f(\Delta_t)$ be the probability density associated with either posterior QTL variance or allelic number in interval Δ_t, the estimate $\hat{f}(\Delta_t)$ is given as

$$\hat{f}(\Delta_t) = \frac{\sum_{s=1}^{n_s}\sum_{j=1}^{n_{\text{qtl}}} \hat{\delta}_j^{(s)} \times 1_{\left\{\lambda_j^{(s)} \in \Delta_t\right\}}}{\sum_{s=1}^{n_s}\sum_{j=1}^{n_{\text{qtl}}} 1_{\left\{\lambda_j^{(s)} \in \Delta_t\right\}}} \quad (14)$$

where $\hat{\delta}_j^{(s)}$ is a posterior estimate of either QTL variance or allelic number at QTL j mapped to this interval at iteration s. For example, we replace parameter $\delta_j^{(s)}$ in equation 14 with $\sigma_j^{2(s)}$ when determining the location-wise posterior density for QTL variance, and replace parameter $\delta_j^{(s)}$ with $l_j^{(s)}$ when determining the location-wise posterior density for allelic number.

Table 1 Map length, marker number and spacing and QTL number in simulated genome

Chromosome	1	2	3	4	5	6	7
Map length, cM	150	150	130	130	150	140	190
Marker number	16	16	14	11	31	8	11
Marker spacing, cM	10	10	10	10	5	20	20
Number of QTL	1	2	0	0	3	0	0

Simulations

Mapping families

The mapping families were simulated using a circulant diallel mating design. Inbred founders were randomly ordered and each founder was mated with its two immediate neighbors to produce F_1 parents, and the latter were used to produce a family of DH progeny. In this mating design, the number of mapping families equaled the number of founder parents from which these mapping families were derived, except that when two parents were used, only a single DH family was obtained.

Genome, markers and QTL

The simulated genome consisted of seven chromosomes, resembling a barley genome (Qi et al. 1996). The whole genome was 1,040 cM in total linkage length and covered by 107 co-dominant markers and six QTL (Table 1). The number of markers varied from chromosome to chromosome, with marker spacing ranging from 5 cM (chromosome 5) to 20 cM (chromosomes 6 and 7). The six QTL were simulated on the genome. Chromosome 1 carried one (QTL 1), chromosome 2 carried two (QTL 2 and 3), and chromosome 5 carried three QTL (QTL 4, 5 and 6). The genetic effects of these QTL differed in terms of the additive variance contributed to the trait. Expressed as a percent of phenotypic variance, the total variance contribution of all the six QTL was 60%, with the largest additive variance (18%) contributed by QTL 6 and the smallest (6%) by QTL 4 and QTL 5. All QTL were assumed to be in linkage equilibrium with each other: any correlations between allelic effects at different QTL were due strictly to sampling effects. Note that QTL 4 and QTL 5 were simulated on the same chromosome and separated by only 10 cM (Table 2). These two QTL were used to test the power of the analysis to distinguish closely linked QTL.

Sampling strategies

Fully informative markers were simulated in each mapping family. In other words, all F_1 individuals were heterozygous at all marker loci (though not necessarily at all QTL). The focus of simulations was on the possible trade-off between QTL mapping family number and family size. The number of families ranged from f=1 to 40 and the family size ranged reciprocally from n_f=600 to 15 so that the total number of progeny was fixed at 600. We denoted the sampling strategies 1:600; 5:120; 10:60; 20:30; 30:20, and 40:15 where the

Table 2 Information about QTL simulated on the genome. The map position was based on Haldane's map function. The allelic number and QTL variance refer to parameter values in the original population. In each simulation, however, the actual allelic number and QTL variance of extracted inbred parents were subject to sampling variation. The QTL variance is expressed as a percent of the total phenotypic variance

QTL	Chromosome	Map position, cM	Allelic number	QTL variance
1	1	75	10	12
2	2	33	4	9
3	2	99	2	9
4	5	45	3	6
5	5	55	6	6
6	5	121	5	18

first number represented mapping family number and the second number represented family size.

Results

Sampling effects

Due to random extraction of inbred parents from the original population, the total number of sampled alleles in parents and subsequent mapping families did not necessarily correspond to the number of alleles in the original population (Table 3). When the number of parental lines used in a sampling strategy was less than the number of QTL alleles, the number of alleles sampled was biased downward in an obvious way. Taking QTL 1 as an example, the number of sampled alleles averaged 1.93 and 3.76 in strategies 1:600 and 5:120, respectively, even though 10 alleles segregated at this locus. As the number of founder parents increased from 10 to 40 the number of sampled alleles approached the true allelic number, and the sampling standard error also decreased. Sampling of parents also affected QTL variance, at times dramatically. Observed total QTL variance in strategy 1:600 averaged 0.366 over 30 replicate runs, much lower than the true variance of 0.6. The standard error of the variance was also large with this strategy (0.130). Sampling more parents rapidly reduced the downward bias in observed QTL variance. In strategies 5:120 and 40:15, observed total QTL variance averaged 0.544 and 0.586, with respective standard errors of 0.094 and 0.033.

Posterior QTL intensity

Posterior QTL intensity peaks were consistently observed at the locations where the QTL were simulated, regardless of sampling strategy (Fig. 1). Maximal posterior QTL intensities on chromosomes where no QTL were simulated were 0.66, 0.60, and 0.51 for the 1:600, 10:60, and 40:15 sampling strategies, respectively. The intensity peaks of QTL were higher at a QTL with a large effect (e.g., QTL 6) than at a QTL with a small effect (e.g., QTL 4 and QTL5). An influence of sampling strategy on QTL mapping was observable in terms of posterior QTL intensity. In general, posterior QTL intensity peaks were higher with large than with small family size (Fig. 1). While posterior QTL intensity peaks were sometimes very high for strategy 1:600, variation in peak height was also largest for this strategy. Averaged over the six QTL, strategy 10:60 had significantly less variability in peak height than either strategies 1:600 or 40:15. The standard deviations in peak height were 4.2, 2.7, and 3.1 for strategies 1:600, 10:60, and 40:15, respectively. Peak intensity and therefore success of QTL mapping was largely subject to sampling of parental alleles. Peak

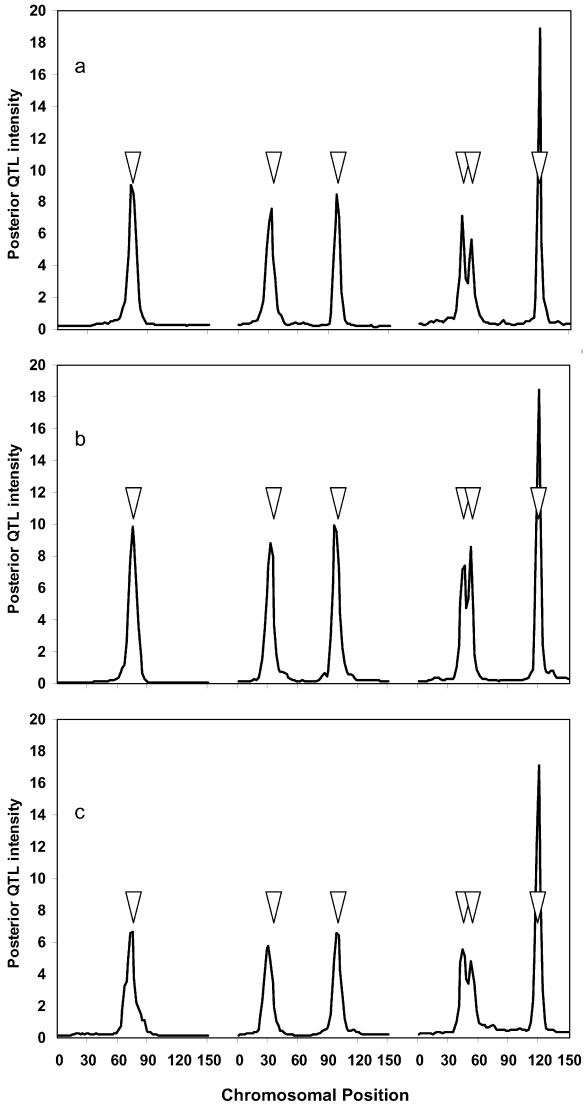

Fig. 1a–c Location-wise posterior QTL intensity under various sampling strategies given as family number:family size. **a** 1:600, **b** 10:60 and **c** 40:15. Chromosomes 1, 2, and 5 are shown. The X-axis indicates map position within each chromosome. *Inverted triangles* show the positions of simulated QTL. Values shown are averages over 30 replicate simulations

Table 3 Sampled number of QTL alleles in mapping families as a function of sampling strategy. The mean ± standard deviation over 30 repeated simulations are shown

Strategy	QTL 1	QTL 2	QTL 3	QTL 4	QTL 5	QTL 6
1:600	1.93±0.24	1.70±0.45	1.57±0.49	1.50±0.49	1.77±0.42	1.80±0.39
5:120	3.76±0.61	3.06±0.57	1.93±0.24	2.63±0.48	3.60±0.84	3.50±0.61
10:60	6.37±0.87	3.70±0.45	2.00±0.00	2.93±0.24	4.97±0.79	4.33±0.59
20:30	8.57±0.71	4.00±0.00	2.00±0.00	3.00±0.00	5.83±0.37	4.93±0.24
30:20	9.43±0.61	4.00±0.00	2.00±0.00	3.00±0.00	5.97±0.17	4.97±0.17
40:15	9.83±0.37	4.00±0.00	2.00±0.00	3.00±0.00	6.00±0.00	5.00±0.00

intensity at a QTL was high when the two parents carried alleles of divergent effect whereas it was low when the two parents carried alleles of similar effect. This type of sampling variation did not occur in strategies involving many parents.

QTL detection power

With repeated simulations, we determined the statistical power for detecting a QTL by counting the number of runs where QTL intensity integrated over a 10 cM interval centered on the simulated QTL was greater than a threshold. Let π_0 be the prior QTL intensity. In the present analysis, π_0=7[QTL]/10.40[Morgans]=0.67[QTL/Morgan]. The arbitrary threshold we chose to declare a QTL present was $3\pi_0 \times 0.10$ where 0.10 is the length in Morgans of the interval over which intensity was integrated. We used a 10 cM interval for all QTL except QTL 4 and QTL 5 for which we used a 20 cM interval covering them both. For these QTL then, the power of QTL detection reflected the power that either or both of the two QTL were detected whereas the power of QTL detection calculated at other intervals applied to only one single QTL.

Various sampling strategies dramatically affected QTL detection power (Fig. 2). Overall, sampling strategies 5:120 and 10:60 were found to be similar to each other and more effective that the other strategies, with strategy 20:30 performing only marginally less well. This intermediate optimum reflects the tradeoff between the need to adequately sample the population's alleles (requiring sampling many parents) and the need to obtain sufficient within-family segregation information (requiring large family size).

We also evaluated the effect of sampling strategy on the power to separate closely linked QTL 4 and QTL 5. Three typical intensity profiles occurred for these QTL (Fig. 3). In the first, two peaks of posterior QTL intensity were clearly observed, indicating that QTL 4 and QTL 5 were detected separately. In the second situation, the two peaks were confounded, mimicking a single QTL located between the true positions of QTL 4 and QTL 5. The third situation, where a QTL intensity peak was found at only one of the two simulated QTL positions, was most frequently observed. We considered QTL 4 and QTL 5 to be separately detected only if the following conditions were both met: the integrated posterior QTL intensity for a 10 cM interval centered at the simulated QTL position was greater than $3\pi_0 \times 0.10$ for both QTL 4 and QTL 5 and the low point between the peaks for QTL 4 and QTL 5 was lower than half the average height of the two peaks and lower than $3\pi_0$. Based on these standards, the optimal sampling strategies were again 5:120 and 10:60 (Fig. 4). Though the power to distinguish QTL 4 and 5 was much lower than that to detect them jointly, the trend in Fig. 4 was very similar to that in Fig. 2.

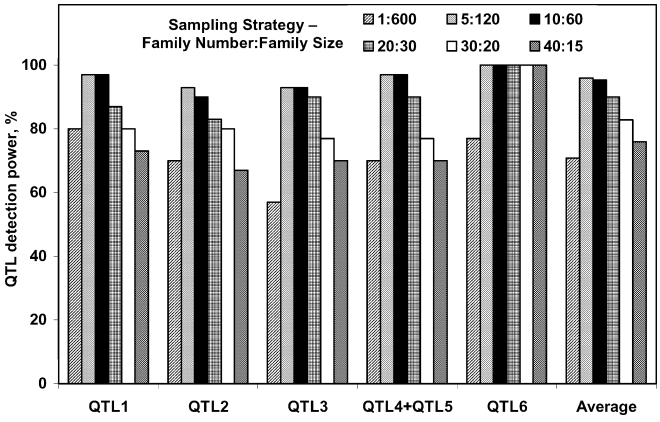

Fig. 2 Power of QTL detection under various sampling strategies. Bar set *QTL4+QTL5* indicates the power to detect at least one of the 2 QTL. Bar set *Average* indicates the average over all six QTL

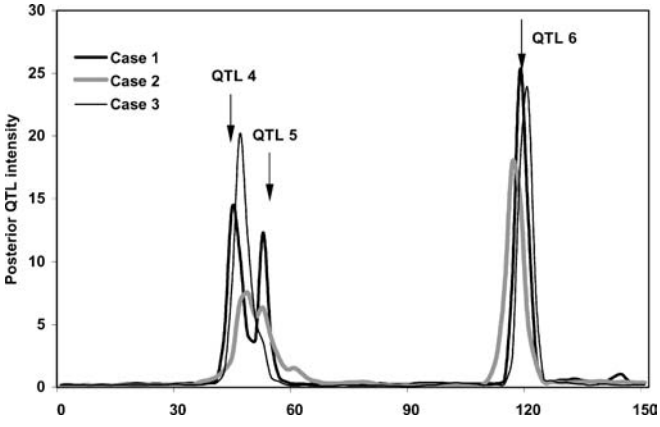

Fig. 3 QTL intensity profiles for closely linked QTL. *Arrows* show the positions of simulated QTL. *Case 1*, two distinct posterior peaks of QTL intensity were observed, and QTL 4 and QTL 5 were detected separately. *Case 2*, two QTL intensity peaks were confounded as a single peak. *Case 3*, a QTL intensity peak was observed at only one of the simulated QTL

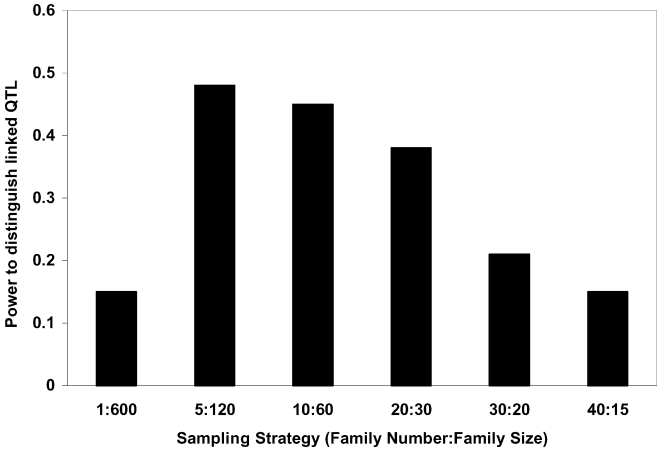

Fig. 4 Power to distinguish QTL 4 from QTL 5 under various sampling strategies

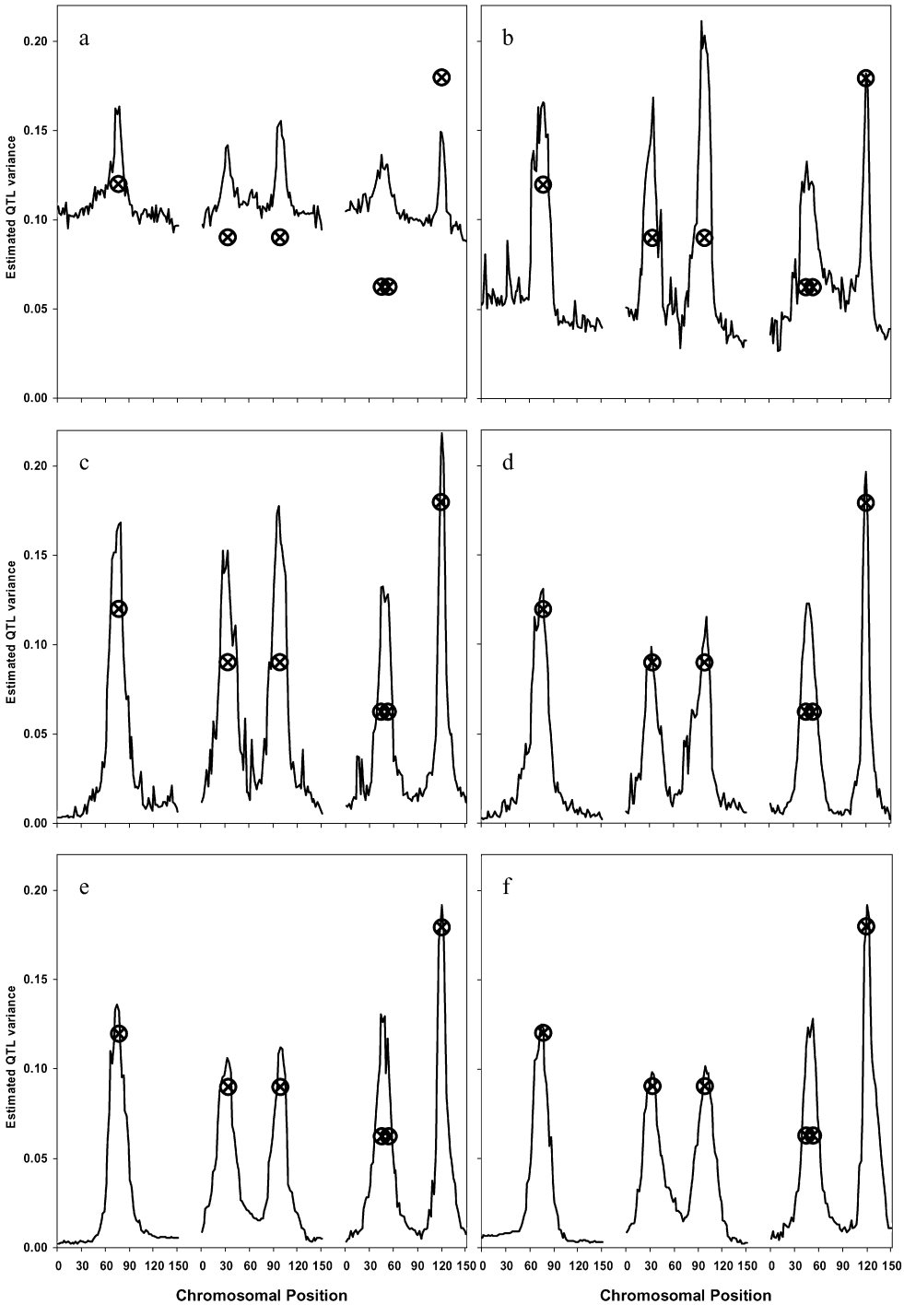

Fig. 5a–f Location-wise posterior QTL variance under various sampling strategies given as family number:family size. **a** 1:600, **b** 5:120, **c** 10:60, **d** 20:30, **e** 30:20 and **f** 40:15. Chromosomes 1, 2, and 5 are shown. The *X-axis* indicates map position within each chromosome. *Circled Xs* indicate simulated QTL position and variance. Values shown are averages over 30 replicate simulations

Posterior QTL variance

Sampling strategies also strongly affected posterior QTL variance (Fig. 5). Location-wise posterior QTL variance for strategies 1:600 and 5:120 was inaccurate and background locations where no QTL were simulated (e.g., between 0 and 50 cM or between 100 and 150 cM on chromosome 1) showed high levels of variance (Fig. 5a, b). This "background variance level" decreased substantially for strategies sampling 10 or more parents (Fig. 5c–f). The cause of high background variance levels in strategies 1:600 and 5:120 can be understood by examining Eq. 4. When few parents are involved, the density of the vector of allelic effects, a_j, is relatively insensitive to the sampled QTL variance $\tilde{\sigma}_j^2$. Consequently, the prior for the variance largely influences the estimated variance. The mean for that prior was 1, much higher than the true variances, and this high prior mean biased all variance estimates upwards, including at

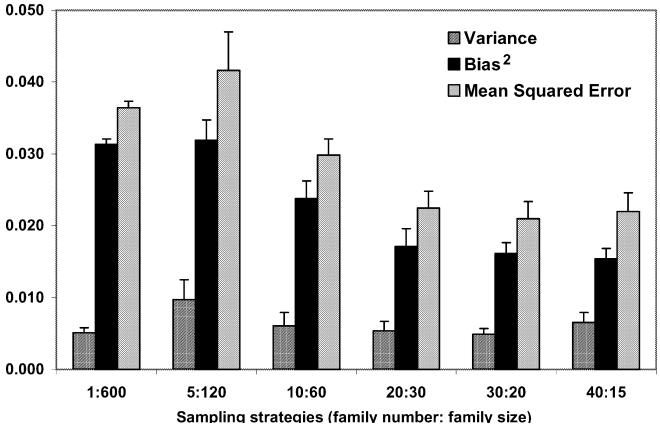

Fig. 6 Measurement variance, squared bias and mean square error of posterior QTL variance under various sampling strategies. Each *bar* represents an average of each statistic over all six QTL and all 30 replicate simulations with the *vertical line* giving the standard error averaged across the six QTL

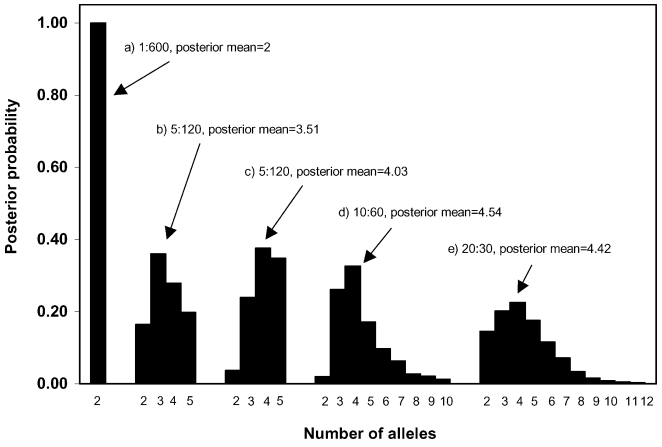

Fig. 7 Posterior distribution of QTL allelic number for QTL 6, with five alleles in the original population, under various sampling strategies given as family number:family size. Four versus five alleles were sampled for bar sets *b* and *c*, respectively

positions devoid of QTL. Consequently, the average posterior QTL variances at the simulated QTL locations were close to their true values only when 20 or more parents were sampled (Fig. 5d–f). Note that the fourth peak of posterior QTL variance corresponded to the sum of variances for QTL 4 and QTL 5.

We quantified the accuracy of QTL variance estimation using its mean squared error, $\mathrm{mse}\left(\sigma_j^2\right) = \mathrm{bias}\left(\sigma_j^2\right)^2 + \mathrm{var}\left(\sigma_j^2\right)$, the sum of the squared measurement bias and the measurement variance. Bias is calculated as $\mathrm{bias}\left(\sigma_j^2\right) = \frac{1}{R}\sum_{r=1}^{R}\left(\hat{\sigma}_j^2\right)_r - \sigma_j^2$ where R is the number of replicate simulations, $\left(\hat{\sigma}_j^2\right)_r$ is the estimate for analysis r, and σ_j^2 is the true value. Measurement variance is the usual variance among replicate estimates. Strategies sampling many parents had much lower mean squared error for QTL variance than strategies sampling few parents (Fig. 6). Counterintuitively, and against this trend, strategy 1:600 had among the lowest measurement variances. Our hypothesis concerning this result refers back to Fig. 5a: the high background observed for this sampling strategy meant that the measurement could not drop below a certain level so that variation in the measurement was restricted.

Posterior QTL allelic number

Sampling strategy affected estimation of the number of alleles in two ways. First, if a small number of parents were used, the estimate of allelic number was biased downward for the sampling reason discussed above (Fig. 7, bar sets a, b, and c). Second, if a large number of parents were sampled, the variance of the allelic number estimate increased (Fig. 7, bar sets d versus e). The prior distribution for allelic number was a truncated Poisson with a lower bound of two alleles and an upper bound equal to the number of sampled parents. Consequently, the variance of the prior distribution also increased for strategies in which a large number of parents were sampled.

Within a given sampling strategy, the posterior estimate of allelic number was somewhat sensitive to the number of alleles actually sampled among the mapping parents (Fig. 7, bar set b versus c). Evaluated across all QTL, the posterior allelic number was positively correlated with the actual number of alleles sampled among mapping parents when the latter ranged from one to five alleles. When the number of alleles sampled ranged above five (as could happen for QTL 1 or QTL 5), the estimated allelic number was no longer affected by the number of alleles (data not shown). When a large number of alleles segregate at a QTL, some of the alleles have similar effects and the analysis would lump them into a single group (Jannink and Wu 2003). When few alleles segregate, their effects diverge sufficiently to be distinguished by the analysis. This reasoning suggests that for analyses evaluating data on 600 progeny, the maximum number of alleles that the analysis can effectively distinguish is five.

Discussion

The QTL mapping design using a single family derived from two inbred parents provides a simple situation for mapping and estimating gene substitution effects using a fixed-model approach (Lander and Botstein 1989). Using a single family, however, reduces the inference space of the estimated parameters to that single cross so that this design is not an optimal strategy to sample a population (Xu 1998): If the QTL is not segregating in the family, it will go undetected. Multiple-family QTL mapping, first introduced by Muranty (1996), can avoid this problem.

Using multiple line crosses broadens the parameter inference space to the original population. From an applied point of view, this analysis will provide relevant estimates of genetic parameters that can be used to predict individual genotypic values in practical breeding programs and will improve QTL detection at the population level.

In this context, sampling strategy (family number vs family size) becomes an important factor in the design of QTL mapping experiments. The results presented above do not point to an unequivocal overall optimal sampling strategy but instead suggest that the strategy adopted must be tailored to the objectives at hand. In particular, we found highest QTL detection power for strategies that employed relatively few parents (5–10) but most accurate estimation of QTL variance for strategies that employed many parents (at least 20, but preferably 30 or more). These results may reconcile to some extent previous disagreements in the literature relative to optimal QTL mapping designs. In particular, Rao and Li (2000) recommended the use of a small number of large families whereas Xu (1998) and Xie et al. (1998) both suggested a larger number of intermediate-sized families. Rao and Li (2000) may have based their recommendation on a fixed-allele model assumption. Indeed, a reanalysis of QTL variances based on allelic values reported in Rao and Li (2000) shows that additive QTL variances using two to five families ranged from 0.14 to 0.27 with standard errors approximately equal to the estimates themselves. As more families were used, however, the standards errors of additive QTL variance decreased, consistent with results reported here.

A mechanism that counteracts the advantage of sampling many families is that in the statistical analysis used here, all QTL information derives from within-family genetic segregation. The analysis discards potential information deriving from QTL effects that contribute to differences among families so that the higher the number of families, the more among-family information is lost. Methods to capitalize on among-family information have recently been presented (Jansen et al. 2003) and could be usefully incorporated into the present analysis. A second refinement that would improve the current analysis when it is applied to many families would be to model family means as random rather than fixed effects. The mean for family i would then be $\beta_j=\mu+\eta_j$, and the single parameter σ^2_β would be of interest rather than the large array of family means. Alternatively, given the relatedness among families, a gametic model could be applied to family means. In that case, the contribution U_i, of inbred parent i, i=1...P to each family mean would be modeled directly as a random effect, and the parameter of interest would be the variance among contributions σ^2_U.

A final approach to improving QTL variance estimation could come from the mating design rather than the statistical analysis. In the simulations presented here, we only investigated a circulant diallel mating design in which each parent is mated to two other parents. For the specific purpose of distinguishing among QTL alleles carried by different parents, the circulant diallel may not be the most effective crossing scheme. Alleles carried by two parents will be contrasted with the greatest power if those two parents are crossed directly. This reasoning suggests that all pair-wise crosses be performed among parents sampled from the original population, leading to a half-diallel design. The drawback of such a design is that it also maximizes the number of mapping families produced relative to the number of parents sampled. For reasons discussed above, analyzing a high number of families may not be best. An alternative would be to cross all sampled parents to the same reference parent. That approach does not allow the direct contrasts among QTL alleles that the half-diallel allows, but instead an indirect contrast relative to a very well characterized reference allele. The effectiveness of different mating designs for QTL analysis merits further research.

Acknowledgements This research was funded by USDA-NRI, CSREES Project Award No. 2001-35301-10848 and supported by the Hatch Act and the State of Iowa.

References

Gilks WR, Richardson S, Spiegelhalter DJ (1996) Introducing Markov chain Monte Carlo. In: Gilks WR, Richardson S, Spiegelhalter, DJ (eds) Markov chain Monte Carlo in practice. Chapman and Hall, London, pp 1–19

Jannink JL, Jansen RC (2001) Mapping epistatic QTL with one-dimensional genome searches. Genetics 157:445–454

Jannink JL, Wu XL (2003) Estimating allelic number and identity in state of QTL in interconnected families. Genet Res 81:133–144

Jansen RC, Jannink JL, Beavis WD (2003) Mapping quantitative trait loci in plant breeding populations: use of parental haplotype sharing. Crop Sci 43:829–834

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP maps. Genetics 121:185–199

Liu Y, Zeng ZB (2000) A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genet Res 75:345–355

Muranty H (1996) Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity 76:156–165

Qi X, Stam P, Lindhout P (1996) Comparison and integration of four barley genetic maps. Genome 39:379–394

Rao S, Li X (2000) Strategies for genetic mapping of categorical traits. Genetica 109:183–97

Rebaï A, Goffinet B (1993) Power of tests of QTL detection using replicated progenies derived from a diallel cross. Theor Appl Genet 86:1014–1022

Rebaï A, Goffinet B (2000) More about quantitative trait locus mapping with diallel designs. Genet Res 75:243–247

Rebaï8 A, Goffinet B, Mangin B, Perret D (1994) Detecting QTLs with diallel schemes. In: van Ooijen JW, Jansen J (eds) Biometrics in plant breeding: applications of molecular markers, 9th meeting of the EUCARPIA, Wageningen, The Netherlands, 1994. CPRO-DLO, pp 170–177

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

Xie C, Gessler DD, Xu S (1998) Combining different line crosses for mapping quantitative trait loci using the identical by descent-based variance component method. Genetics 149:1139–1146

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