University of Wisconsin — Madison
Department of Statistics

Markov Chain Monte Carlo Based Inference for Quantitative Trait Loci in Plant Breeding

Jaya M. Satagopan
Brian S. Yandell
Michael A. Newton
Department of Statistics

Technical Report # 925
August, 1994

Department of Statistics
University of Wisconsin–Madison
1210 West Dayton Street
Madison WI 53706-1685

Phone: 608/262-2598
Fax: 608/262-0032
Internet: yandell@stat.wisc.edu

Markov chain Monte Carlo based inference for quantitative trait loci in plant breeding

Jaya M. Satagopan   Brian S. Yandell
Michael A. Newton
University of Wisconsin–Madison

August 22, 1994

Abstract

We present a Markov chain Monte Carlo (MCMC) method for infering the presence and position of quantitative trait loci (QTLs). The missing quantitative trait genotypes at a putative locus are treated as augmented data (Tanner and Wong 1987). The Monte Carlo methods allow one to study the properties of the marginal distribution of the putative QTL without making asymptotic assumptions. Properties of the estimated QTL position are obtained using time series methods. This provides a viable alternative to interval mapping (Lander and Botstein 1989), which profiles the QTL using the EM method and appeals to a chi-square approximation of the likelihood ratio for inference. The approach is illustrated with a simple genetic system in plant breeding, in which one can readily obtain large sample sizes. A simulated experiment is presented for such a double haploid population, a system with two alleles per locus and every individual is homozygous at every locus.

1 Introduction

We consider the problem of estimating the loci of polygenes associated with a quantitative trait given genetic marker data. We assume that the marker data have been organized into a linkage map composed of disjoint linkage groups. The single QTL model for one linkage group is used as a test bed to demonstrate the use of the Markov chain Monte Carlo method as an alternative to interval mapping or regression methods in common use today.

Statistical methods for estimating single quantitative trait locus (QTL) in a linkage group are fairly well understood when sufficient assumptions are made. For example, the interval mapping LOD score approach (Lander and Botstein 1989) uses the EM method to profile the locus across the linkage map. Recently, this idea has been extended to estimating several quantitative trait loci (QTLs) using likelihood and/or regression methods (Zeng 1994; Jansen and Stam 1994; Churchill and Doerge 1994a). However, these approaches become much more complicated as larger models are considered.

Markov chain Monte Carlo (MCMC) methods promise to extend the domain of such inference by weakening assumptions, such as allowing multiple QTLs and relaxing the normal distribution of traits. MCMC methods, particularly the Metropolis-Hastings algorithm and Gibbs sampling, have recently been applied to several genetic problems including ordering of markers on linkage maps (Stephens and Smith 1993) and linkage analysis in complex pedigrees (Thompson 1991). Markov chain Monte Carlo ideas have been developed in recent articles by Tierney (1991), Besag and Green (1993) and Smith and Roberts (1993). Geyer (1992) discusses practical considerations to reduce bias and autocorrelation and to assess error variation using time series methods.

In this paper we develop a MCMC approach for estimating a single QTL in one linkage group for a simple breeding system by augmenting the marker and trait data with the missing QTL genotypes at a putative QTL. While this problem can be readily handled by profile likelihood computations, it demonstrates the power of these methods. Analagous MCMC methods are proposed for polygene (multiple QTL) inference in more complicated settings.

Section 2 outlines a plant breeding scheme which allows us to only consider haploid type. Section 3 presents a model for a single QTL, detailing the relationship between trait data, marker data on the linkage map, the missing genotypes and model parameters. The interval mapping method is briefly outlined at the end of this section. Section 4 develops the MCMC method for a single QTL. Section 5 proposes extensions of the MCMC method to polygenes and other potential generalizations. Finally, simulations in Section 6 examine the behavior of the MCMC method relative to interval mapping.

2 Double Haploid (DH) Plant Population

We consider a segregating population of plants which have been artificially developed such that every locus in the genome is homozygous. Effectively, these plants can be treated as haploid, simplifying the genetics considerably. We briefly review assumptions about the association between recombination and distance which are used in later sections of the paper.

A haploid-like population was created by first crossing two homozygous parents of species *Brassica napus*, Major (P_1) and Stellar (P_2), having very different characteristics. Pollen from a single F_1 individual from this cross was collected and treated with colchicine to double the chromosomes. The surviving double haploid pollen grains developed into a population of double haploids (DH). Ferreira, Williams, and Osborn (1994) developed 105 DH lines of *B. napus* which were scored by 132 restriction fragment length polymorphisms (RFLPs). In addition, they measured several phenotypic traits, some qualitative and some quantitative. Ferreira, Satagopan, Yandell, Williams, and Osborn (1994) created a linkage map and then estimated gene loci for these phenotypic measurements.

Figure 1 shows schematically how recombination events can occur during meiosis in the development of a DH population. Consider two loci M_1 and M_2 along a chromosome. There is some chance R of recombination between these two loci during the process of meiosis which forms the pollen. Thus there are 4 possible haploid types for the 2 markers M_1M_2 on a single chromosome strand: AB, Ab, aB and ab (Figure 1 shows the recombinants). Since DH progeny are homozygous at every locus, we need only consider haplotypes.

One would expect to see the four possible haplotypes in the sample in the following proportions.

genotype	Major (P_1)	recombinants		Stellar (P_2)
haplotype	AB	Ab	aB	ab
expected frequency	$(1-R)/2$	$R/2$	$R/2$	$(1-R)/2$

Loci with small recombination rates tend to be nearby. If the recombination rate between two loci is 1/2 or greater, they are considered to be unlinked. One estimates the rate as $\hat{R}$ = #recombinants/#progeny. The linkage map distances among loci tends to differ slightly from pair-wise recombination rates, due to missing data and assumptions about dependence, or interference, among recombination events (Ott 1991; but see Evans, McPeek, and Speed 1993). For convenience, we adopt the Haldane map function, which assumes no

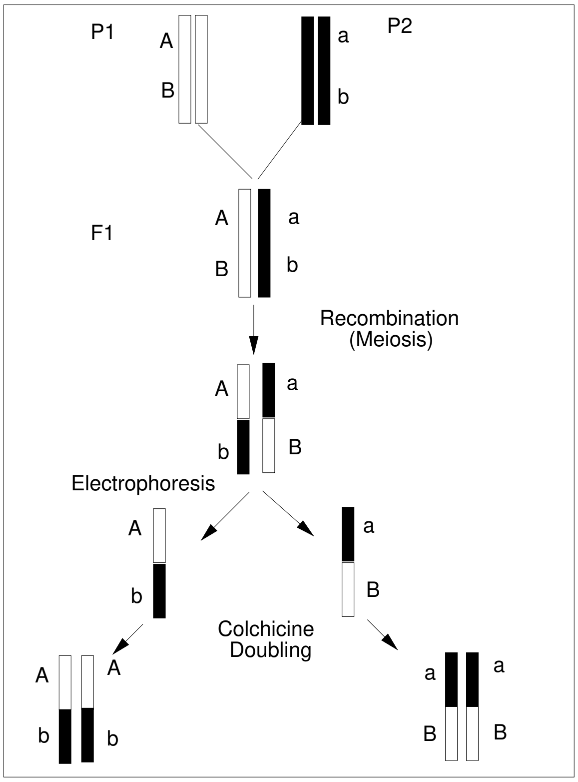

Figure 1: *Recombination event during DH development.*

interference, or complete independence of recombination events. With this assumption, one can express the map distance D between loci, as measured in centi-Morgans (cM), in terms of recombination rate R,

$$\begin{aligned} D &= -50\ln(1-2R)\ , \quad 0 \le R < 0.5\ , \\ R &= 0.5[1-\exp(1-0.2D)]\ , \quad 0 \le D\ . \end{aligned}$$

The loci in a linkage map may be any characteristics which can be unambiguously scored as one of the parent types. In practice, one uses a combination of DNA-derived genetic markers such as RFLPs and directly observable phenotypic traits, including qualitative traits such as the presence/absence of hairs and quantitative traits such as days to budding. Quantitative phenotypic traits require more care, and more assumptions. The next section presents a model and the interval mapping estimation method for a single quantitative trait locus.

3 Detecting A Single QTL

In this section we consider the problem of detecting a single quantitative trait locus (QTL) using interval mapping. Section 3.1 develops the normal mixture model for a trait given genetic marker information. Section 3.2 reviews the interval mapping approach and its relation to profile likelihood. Throughout, we use the notation M_j (or M_{ji}) to signify a marker and its genotype, and similarly for the quantitative trait locus Q (or Q_i). We further associate 1 with capital alleles $(A, B, \cdots)$ and 0 with lower case alleles $(a, b, \cdots)$.

Consider a quantitative trait which is expressed by a single gene but subject to environmental variation. Let $\mu_0 = \mu$ and $\mu_1 = \mu + \delta$ be the phenotypic means when the QTL has genotype $Q = 0$ and $Q = 1$, respectively. The observed phenotype y thus is a random draw from a distribution with mean μ_0 or μ_1 plus some error. For convenience, assume the error is normal with variance σ^2. One could write this as a linear model

$$y = \mu + Q\delta + \epsilon \ , \quad \epsilon \sim N(0, \sigma^2) \ ,$$

or as a conditional density,

$$\pi(y|\mu, \delta, \sigma, Q) = \phi\left(\frac{y - \mu - Q\delta}{\sigma}\right) / \sigma \ , \tag{1}$$

with ϕ being the standard normal density.

3.1 Normal Mixture Model

In practice, we know the phenotype y but not the genotype Q. Suppose one has developed a linkage map for a genome (here reduced to one linkage group for convenience) using a sample of size n from a DH population, and one has quantitative trait measurements $\boldsymbol{y} = \{y_i\}_{i=1}^n$. Suppose there are $m + 1$ markers $\{M_1, \cdots, M_{m+1}\}$ already ordered as a linkage group with corresponding map distances (in cM) $\boldsymbol{D} = \{D_0 = 0, D_1, \cdots, D_m\}$, with D_j being the map distance between M_1 and M_{j+1}. Denote the marker genotypes for individual i by M_{ji}, with $\boldsymbol{M} = \{M_{ji}\}_{j,i=1}^{m+1,n}$, and let the missing genotypes at the QTL be $\boldsymbol{Q} = \{Q_i\}_{i=1}^n$. Let λ be the map distance of the `locus` of Q along the linkage group. That is, the QTL is in some interval j between "flanking markers" M_j and M_{j+1}. Denote the model `parameters` by $\theta = (\mu, \delta, \sigma)$.

Our aim is to determine the conditional distribution of y given the position and marker information, and then use this to estimate λ. Suppose the QTL is between the first two markers M_1, M_2, which are a map distance $D = D_1$ apart. The Haldane assumption of independence implies that the genotype of the QTL given the "flanking markers" M_1 and M_2 is independent of all other marker loci. Further, the phenotypes $\boldsymbol{y}$ are conditionally independent of the marker data $(\boldsymbol{D}, \boldsymbol{M})$ given the genotypes $\boldsymbol{Q}$. This induces the following normal mixture likelihood for the ith individual,

$$\begin{aligned} \pi(y_i|\lambda, \theta, \boldsymbol{D}, \boldsymbol{M}) &= \pi(y_i|\lambda, \theta, D, M_{1i}, M_{2i}) \\ &= \gamma_i\, \pi(y_i|\theta, Q_i = 1) + (1-\gamma_i)\, \pi(y_i|\theta, Q_i = 0) \ , \end{aligned}$$

with mixing parameter

$$\gamma_i = \pi(Q_i = 1|\lambda, D, M_{1i}, M_{2i}) \ . \tag{2}$$

The mixing parameter γ_i under the assumption of independent recombination events comes from the following recombination table.

Table 1. Mixing Parameter under Haldane Model

		M_{2i}	
	γ_i	0	1
M_{1i}	0	$r_1 r_2/(1-R)$	$r_1(1-r_2)/R$
	1	$(1-r_1)r_2/R$	$(1-r_1)(1-r_2)/(1-R)$

Here R, r_1 and r_2 are the recombination rates corresponding to $D = D_2$, λ, $D - \lambda$, respectively, using the Haldane map function. While this formulation assumes independence, one could modify this approach to accomodate other multilocus recombination models which incorporate interference.

The likelihood for n individuals is thus

$$L(\theta|\lambda, \boldsymbol{y}, \boldsymbol{D}, \boldsymbol{M}) = \prod_{i=1}^{n} [\gamma_i\, \pi(y_i|\theta, Q_i = 1) + (1-\gamma_i)\, \pi(y_i|\theta, Q_i = 0)] \ .$$

3.2 Profile Interval Mapping

One can draw inference about the presence of a QTL at a `locus` λ using the likelihood ratio statistic. By allowing λ to range over the entire genome (one linkage group for convenience here), one can test the null hypothesis of no QTL, which is equivalent to $H_o : \mu_0 = \mu_1 = \mu$, or $H_o : \delta = 0$. Human geneticists adopted the logarithm of odds (LOD) score (Ott 1991) to test this hypothesis and to develop estimates of the `locus` λ. The LOD score is given by

$$LOD(\texttt{locus}) = \log_{10} \frac{\max L(\texttt{parameters}|\texttt{locus}, \texttt{data})}{\max L_o(\texttt{parameters}|\texttt{data})} \ , \tag{3}$$

where the maximum is over the `parameters`. The denominator is the likelihood under the null hypothesis of no QTL,

$$L_o(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} \phi\left(\frac{y_i - \mu}{\sigma}\right) / \sigma \ .$$

The numerator of the LOD score is the profile likelihood of the `locus` λ, which maximizes the likelihood L over the `parameters` $\theta = (\mu, \delta, \sigma)$ given the presumed `locus` λ and the `data` $(\boldsymbol{y}, \boldsymbol{D}, \boldsymbol{M})$,

$$PL(\lambda|\boldsymbol{y}, \boldsymbol{D}, \boldsymbol{M}) = \max_{\theta} L(\theta|\lambda, \boldsymbol{y}, \boldsymbol{D}, \boldsymbol{M}) = L(\hat{\theta}|\lambda, \boldsymbol{y}, \boldsymbol{D}, \boldsymbol{M}) \ , \tag{4}$$

where $\hat{\theta}$ is the maximum likelihood estimate of θ.

The most commonly used method to detect a QTL is the interval mapping approach (Lander and Botstein 1989). This method involves a "brute force" scan across the genome, using an EM algorithm to iteratively maximize the profile likelihood. The algorithm was incorporated into the public domain computer program Mapmaker/QTL (cf. Lander et al. 1987). More detailed developments of the EM algorithm for this problem can be found in Luo and Kearsey (1991), vanOoijen (1992) and Churchill and Doerge (1994b).

When the number of individuals n is large, the maximum LOD score, at the estimated QTL `locus` $\hat{\lambda}$, is approximated by $LOD(\hat{\lambda}) \approx \chi_1^2/2log(10)$ under H_o. One concludes there is no QTL if the maximum LOD is too small, usually $LOD(\hat{\lambda}) < 3$ in practice. However, see Doerge and Churchill (1994) and Churchill and Doerge (1994a) for another approach based on permutation tests. An approximate confidence interval is constructed of all loci with LOD "close" to the maximum, $\{\lambda : LOD(\hat{\lambda}) - LOD(\lambda) < 1\}$. Darvasi et al. (1993) present simulations which suggest this works well if one has a sample size of $n = 200$ to 500.

4 MCMC for QTLs

In this section, we present an MCMC method for inferring the location of single QTL by augmenting the missing data of the QTL genotypes (Tanner and Wong 1987; Besag and Green 1993). For convenience, we drop $\boldsymbol{D}$, recognizing that `map distance` is usually derived directly from the `markers` $\boldsymbol{M}$. We implement a Markov chain whose equilibrium distribution is the joint distribution of the QTL `locus` λ, `missing genotypes` $\boldsymbol{Q}$ and model `parameters` θ given the `data` $(\boldsymbol{y}, \boldsymbol{M})$. The `locus` λ is updated by a Metropolis-Hastings algorithm, while the `missing genotypes` $\boldsymbol{Q}$ and `parameters` θ are updated using the Gibbs sampler. The basic idea is to learn about the joint distribution of `locus`, `parameters` and `missing genotypes` given the `data`, which can be expressed, using Bayes theorem, as

$$\begin{aligned} \pi(\lambda, \theta, \boldsymbol{Q} | \boldsymbol{y}, \boldsymbol{M}) &= \pi(\boldsymbol{y} | \lambda, \theta, \boldsymbol{Q}, \boldsymbol{M}) \; \pi(\lambda, \theta, \boldsymbol{Q} | \boldsymbol{M}) \; / \; \pi(\boldsymbol{y} | \boldsymbol{M}) \\ &\propto \pi(\boldsymbol{y} | \lambda, \theta, \boldsymbol{Q}, \boldsymbol{M}) \; \pi(\lambda, \theta, \boldsymbol{Q} | \boldsymbol{M}) \; . \end{aligned} \tag{5}$$

We condition on the markers $\boldsymbol{M}$, as is done in most other QTL studies (but see Stephens and Smith (1993)). The joint density of the `locus`, `parameters` and `missing genotypes` given the `markers` can be decomposed as

$$\pi(\lambda, \theta, \boldsymbol{Q} | \boldsymbol{M}) = \pi(\lambda | \boldsymbol{M}) \; \pi(\theta) \; \pi(\boldsymbol{Q} | \lambda, \boldsymbol{M}) \, . \tag{6}$$

These need to be specified in advance, based on model assumptions. For instance, the distribution of λ may be uniform over the genome, Q_i is determined by recombination rates with flanking markers as in equation (2), and $\theta = (\mu, \delta, \sigma)$ can be decomposed with vague priors on each parameter.

Knowing the joint distribution of (λ, θ) allows us to examine, for instance, the marginal distribution of the `locus` given the `data`,

$$\pi(\lambda | \boldsymbol{y}, \boldsymbol{M}) = \int \int \pi(\lambda, \theta, \boldsymbol{Q} | \boldsymbol{y}, \boldsymbol{M}) \, d\boldsymbol{Q} \, d\theta \; ,$$

as an alternative to the profile likelihood, which maximizes over θ.

A Gibbs sampler would draw successively from the full conditional distributions of λ, $\boldsymbol{Q}$ and θ given the other unknowns and the `data`. However, sampling from these full conditional distributions may be difficult. Instead, we construct a Monte Carlo Markov chain, taking advantage of simplification in conditioning wherever possible, leading to

$$\begin{aligned}
\pi(\lambda|\theta, \boldsymbol{Q}, \boldsymbol{y}, \boldsymbol{M}) &\propto \pi(\boldsymbol{Q}|\lambda, \boldsymbol{M})\ \pi(\lambda|\boldsymbol{M})\ , \\
\pi(\boldsymbol{Q}|\lambda, \theta, \boldsymbol{y}, \boldsymbol{M}) &\propto \pi(\boldsymbol{y}|\theta, \boldsymbol{Q})\ \pi(\boldsymbol{Q}|\lambda, \boldsymbol{M})\ , \\
\pi(\theta|\lambda, \boldsymbol{Q}, \boldsymbol{y}, \boldsymbol{M}) &\propto \pi(\boldsymbol{y}|\theta, \boldsymbol{Q})\ \pi(\theta)\ .
\end{aligned} \tag{7}$$

With suitable priors, the conditional distribution for θ can be written explicitly. That is, letting μ and δ have normal priors and σ^2 have an inverse gamma prior leads to simple formula (cf. Tierney 1991). The conditional distributions for Q_i are exactly Bernoulli,

$$\pi(Q_i = 1|\lambda, \theta, \boldsymbol{y}, \boldsymbol{M}) = \frac{\gamma_i \pi(y_i|\theta, Q_i = 1)}{\gamma_i \pi(y_i|\theta, Q_i = 1) + (1 - \gamma_i)\pi(y_i|\theta, Q_i = 0)}\ .$$

However, the conditional distribution of λ is known only up to a normalizing constant, which led us to implement the Metropolis-Hastings approach. One needs a proposal density q for updating λ in $(0, D_{m+1})$. The two natural choices are "independence," which selects a new locus uniformly along the linkage group, or random walk, which is a normal density centered at the current value with some tunable dispersion hyperparameter. In practice, one uses the independence proposal density to establish if the equilibrium distribution is well behaved, and then switches to the random walk proposal for a more efficient algorithm. For step $k+1$ of the Markov chain, one draws λ^* from the proposal density $q(\lambda^*, \lambda^k)$ and determines the Hastings ratio, using (7), as

$$\alpha = \frac{\pi(\lambda^*|, \theta^k, \boldsymbol{Q}^k, \boldsymbol{y}, \boldsymbol{M})\ q(\lambda^*, \lambda^k)}{\pi(\lambda^k|, \theta^k, \boldsymbol{Q}^k, \boldsymbol{y}, \boldsymbol{M})\ q(\lambda^k, \lambda^*)}\ . \tag{8}$$

One then updates $\lambda^{k+1} = \lambda^*$ with probability $\min(\alpha, 1)$, or otherwise sets $\lambda^{k+1} = \lambda^k$. Note that q cancels out in the numerator and denominator of α due to the symmetry of both the independence and random walk proposal density. However, if one modifies the proposal slightly to simplify computing by choosing a distance Δ in $(1, m+1)$, leading to interval $J = [\Delta]$ and `locus` $\lambda = D_{J-1} + (D_J - D_{J-1})(\Delta - J)$, then

$$\frac{q(\lambda^*, \lambda^k)}{q(\lambda^k, \lambda^*)} = \frac{(D_{J^*} - D_{J^*-1})}{(D_{J^k} - D_{J^k-1})}\ .$$

After updating λ^{k+1}, new draws of $\boldsymbol{Q}^{k+1}$ and θ^{k+1} from their respective full conditional densities (7) complete the step. The resulting Markov chain $\{(\lambda_k, \boldsymbol{Q}^k, \theta^k)\}_{k=1}^N$ has the desired equilibrium distribution (6).

5 Locating Two QTLs

Quantitative traits are often believed to be controlled by two or more genes, or polygenes, (Tanksley 1993). Recent work (cf. Jansen and Stam (1994), Zeng (1994), Churchill and Doerge (1994a) Haley, Knott, and Elsen (1994) and Knapp (1994) as well as references therein) highlight shortcomings of fitting a model with one QTL when in fact there are two. The "animal model" allows for an arbitrary number of genes by including a random effect incorporating the covariance among individuals in a population (cf. Foulley, Gianola, and Im 1990).

Polygene models become progressively more complicated, particularly if one employs a profile likelihood interval mapping approach. In addition, approaches to date employ a "brute force" search for two or more loci, which is very computer intensive. The often recommended "stepwise" approach, in which one fixes the major QTL then searches for the next most important QTL, can sometimes mislead one to detect a single "ghost" QTL rather than two on the same linkage group.

Here we consider modelling two genes on the same linkage group, investigating the joint distribution of their `loci` $\lambda = (\lambda_1, \lambda_2)$. Associated with each locus λ_ℓ are `parameters` δ_ℓ and `missing genotypes` $\boldsymbol{Q}_\ell$. One can further allow interaction (epistasis or interference) between the two genes with an additional parameter β_{12}. We develop a Markov Chain Monte Carlo simulation analagous to the previous section, updating the pairs of `loci` sequentially, and similarly for the `parameters` and `missing genotypes`. Given the `missing genotypes`, the model for $\boldsymbol{y}$ generalizes (1) as

$$\pi(y_i|\theta, Q_{i1}, Q_{i2}) = \phi\left(\frac{y_i - \mu - Q_{i1}\delta_1 - Q_{i2}\delta_2 - Q_{i1}Q_{i2}\beta_{12}}{\sigma}\right)/\sigma\,.$$

The full conditionals for updating components can be generalized from (7) using suitable extensions to full conditional distributions of unknowns given the `markers` (6),

$$\pi(\lambda, \theta, \boldsymbol{Q}|\boldsymbol{M}) = \pi(\lambda_1, \lambda_2|\boldsymbol{M})\ \pi(\theta)\ \pi(\boldsymbol{Q}_1, \boldsymbol{Q}_2|\lambda_1, \lambda_2, \boldsymbol{M})\ .$$

If the `loci` are in the same marker interval, one needs to modify the distribution of $\boldsymbol{Q}$,

$$\pi(\boldsymbol{Q}_1, \boldsymbol{Q}_2 | \lambda_1, \lambda_2, \boldsymbol{M}) = \pi(\boldsymbol{Q}_1 | \lambda_1, \lambda_2, \boldsymbol{Q}_2, \boldsymbol{M}) \; \pi(\boldsymbol{Q}_2 | \lambda_2, \boldsymbol{M}) .$$

The first term involves recombination rates with only one marker and the second QTL under the Haldane independence assumption.

This idea can be generalized in an obvious way to polygenes with more than two loci. One could further generalize by allowing the number of loci L to be unknown, using perhaps a truncated Poisson prior on L and updating

$$\begin{aligned} \pi(L|\lambda, \theta, \boldsymbol{Q}, \boldsymbol{y}, \boldsymbol{M}) &\propto \pi(\boldsymbol{y}|\lambda, \theta, \boldsymbol{Q}, L, \boldsymbol{y}, \boldsymbol{M}) \; \pi(\lambda, \theta, \boldsymbol{Q}, L, \boldsymbol{y}, \boldsymbol{M}) \\ &\propto \pi(\boldsymbol{y}|\theta, \boldsymbol{Q}, L) \pi(\theta|L) \pi(\boldsymbol{Q}|\lambda, L, \boldsymbol{M}) \pi(\lambda|L, \boldsymbol{M}) \pi(L|\boldsymbol{M}) \; . \end{aligned} \tag{9}$$

Extensions in another direction allow for generalized linear models such as Poisson or binomial counting traits, rather than normal quantitative traits, by modifying $\pi(\boldsymbol{y}|\theta, \boldsymbol{Q})$ of (1) and possibly $\pi(\theta)$. Further extensions allow modelling of experimental design details, such as blocking and treatment structures, which are unavailable in present packages which examine quantitative traits. Other breeding schemes could be incorporated in a natural way as well, with some modification of notation. Missing marker data M_{ji} can be incorporated in a natural way by further conditioning, following Gelfand, Hills, Racine-Poon, and Smith (1990).

6 Simulation Results

We present a simulation result for the DH plant breeding system described in Section 2. A genome of length 100 cM containing 6 equidistant markers and a single QTL halfway between markers 4 and 5 (70cM) was simulated for a range of effects δ and sample sizes $n = 100$ and 200. Throughout $\mu = 0$ and $\sigma = 1$. Single realizations had Markov chains of $N =$ 1,000,000, recording every $100th$ point (1% subsampling). Summary tables and plots from 100 independent realizations use Markov chains of length $N =$ 100,000 with 10% subsampling. All plots are of locus estimates presented on a relative scale, with 20cM between markers, for ease of comparison, even though calculations were made using the Haldane map function.

The difference in means between genotypes were of the form $\delta = s \times 2\sigma/\sqrt{n}$, being roughly s standard errors apart at the QTL if $n_0 = n_1 = n/2$ (see Table 2). One would not expect to pick up effects for $s = 1$ or 2 standard errors using a two-sample t-test if one knew the true gene locus. With $s = 5$ standard errors, one would have slightly more than 95% power of detecting differences at the true locus using a t-test, while one would be virtually certain of detecting differences of $s = 10$ standard errors.

Table 2. Values of δ

		standard errors s			
		1	2	5	10
n	100	0.200	0.400	1.000	2.000
	200	0.141	0.283	0.707	1.414

Figure 2 shows the marginal densities of the estimated QTL locus from simulated data sets using the MCMC method over the range of sample size n and gene effects δ considered. The profile LOD scores were back-transformed and rescaled as densities for comparison. Note that one would not expect to pick up a locus for effects of $s = 1$ or 2 standard errors, leading to densities for locus λ spread over most of the linkage group. As the size of effect (s or δ) increases, the marginal and profile densities become more concentrated around the mean. Note, however, that the profile density (and hence the profile LOD) sometimes has peaks which are not found in the marginal density.

The joint distribution of effect δ and locus λ are examined in Figure 3. Note again how λ becomes more concentrated around the actual locus as the effect increases.

One hundred independent realizations allow us to examine the the sampling error of the the Monte Carlo Markov chain methodology. Figure 4 shows scatter plots of the means for effect δ and locus λ. The variation reflects sample-to-sample variation, giving an idea of the range of possilbe realizations from one "true" model. Note again the poor fit of λ when s is small, as expected. However, the estimates of δ are fairly good.

Confidence intervals for parameters were constructed based on percentiles (e.g. 75% or 90%) of the subsampled points from the Markov chains. These had correct coverage rates within a few percent for the 100 realizations (not shown). We suspect that confidence intervals based on the likelihood would behave similarly.

We have begun empirical examination of the two QTL model. Our initial work uses the (wrong) single QTL model when the right model has QTLs near 30cM and 70cM, with effects for the two QTL about $\delta_1 = \delta_2 = 2$ with a sample size of $n = 100$. Figure 5 shows (a) the histogram of λ for one realization, overlaid by the profile LOD density and (b) the joint density of λ and δ. Note that there are "gaps" between the peaks for λ, and a lower peak for the "ghost" position between QTLs for the marginal density. The profile density characteristically gives a larger peak at ghost QTL.

In addition, examination of time series plots of λ for two QTL data with the MCMC method show that the procedure remains in an interval for a long stretch, rarely jumping between these three intervals. This is very different from the behavior of single QTL data, suggesting one can use the time series as diagnostic for the presence of multiple QTL. One should in fact incorporate multiple QTL into the method, as we are currently attempting.

7 Acknowledgements

The authors wish to acknowledge helpful discussions with Thomas C. Osborn, Charles Geyer and Rebecca W. Doerge. This work was supported in part by grants from the USDA/CSRS.

References

Besag, J. and P. J. Green (1993). Spatial statistics and Bayesian computation. *Journal of the Royal Statistical Society 55*, 25–37.

Churchill, G. and R. Doerge (1994a). Empirical threshold values for quantitative trait mapping. *Genetics 000*, 000–000.

Churchill, G. and R. Doerge (1994b, August). Mapping quantitative trait loci in experimental populations. In *International Biometrics Conference Proceedings.*

Darvasi, A., A. Weinreb, V. Minke, J. Weller, and A. Soller (1993). Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. *Genetics 134*, 943–951.

Doerge, R. and G. Churchill (1994). Issues in genetic mapping for quantitative trait loci. In *Second Joint Plant Breeding Symposium.*

Evans, S. N., M. S. McPeek, and T. P. Speed (1993). Crossover models and map functions. *Theoretical Population Biology 43*, 80–90.

Ferreira, M., J. Satagopan, B. Yandell, P. Williams, and T. Osborn (1994). Genetics and mapping of loci controlling vernalization requirement and days to flower in rapeseed (brassica napus l.). *Molecular and General Genetics 00*, 000–000. (submitted).

Ferreira, M., P. Williams, and T. Osborn (1994). Rflp mapping of Brassica napus using F1-derived doubled haploid lines. *Theoretical and Applied Genetics 00*, 000–000. (submitted).

Foulley, J. L., D. Gianola, and S. Im (1990). Genetic evaluation for discrete polygenic traits in animal breeding. In D. Gianola and K. Hammond (Eds.), *Adv. Ser. Agric. Sci. 18: Advances in Statistical Methods for Genetic Improvement of Livestock*, pp. 361–409.

Gelfand, A. E., S. E. Hills, A. Racine-Poon, and A. F. M. Smith (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *J. of the Amer. Stat'l. Assn. 85*, 972–985.

Geyer, C. (1992). Practical Markov chain Monte Carlo. *Statistical Science 7*, 437–511. (with discussion).

Haley, C., S. Knott, and J.-M. Elsen (1994). mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics 136*, 1195–1207.

Jansen, R. C. and P. Stam (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics 136*, 1447–1455.

Knapp, S. (1994). Mapping quantitative trait loci. In R. L. Phillips and I. K. Vasil (Eds.), *DNA-Based Markers in Plants.*

Lander, E., P. Green, J. Abrahamson, A. Barlow, M. Daly, S. Lincoln, and L. Newburg (1987). Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics 1*, 174–181.

Lander, E. S. and D. Botstein (1989). Mapping Mendelian factors underlying quantitative traits using Rflp linkage maps. *Genetics 121*, 185–199.

Luo, Z. W. and M. J. Kearsey (1991). Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. Ii. Application to backcross and doubled haploid populations. *Heredity 66*, 117–124.

Ott, J. (1991). *Analysis of Human Genetic Linkage* (revised ed.). John Hopkins U.

Smith, A. and G. Roberts (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society 55*, 3–23.

Stephens, D. and A. Smith (1993). Bayesian inference in multipoint gene mapping. *Annals of Human Genetics 57*, 65–82.

Tanksley, S. D. (1993). Mapping polygenes. *Ann. Rev. Genet. 27*, 205–233.

Tanner, M. A. and W. H. Wong (1987). The calculation of posterior distributions by data augmentation (c/r: P541-550). *J. of the Amer. Stat'l. Assn. 82*, 528–540.

Thompson, E. (1991). Probabilities on complex pedigrees. The Gibbs sampler approach. *8 Proc. of Computer Sci. and Stat.: 8th Annual Symp. on the Interface 23*, 371–378.

Tierney, L. (1991). Exploring posterior distributions using Markov chains. *8 Proc. of Computer Sci. and Stat.: 8th Annual Symp. on the Interface 23*, 563–570.

vanOoijen, J. W. (1992). Accuracy of mapping quantitative trait loci in autogamous species. *Theoretical and Applied Genetics 84*, 803–831.

Zeng, Z.-B. (1994). Precision mapping of quantitative trait loci. *Genetics 136*, 1457–1468.

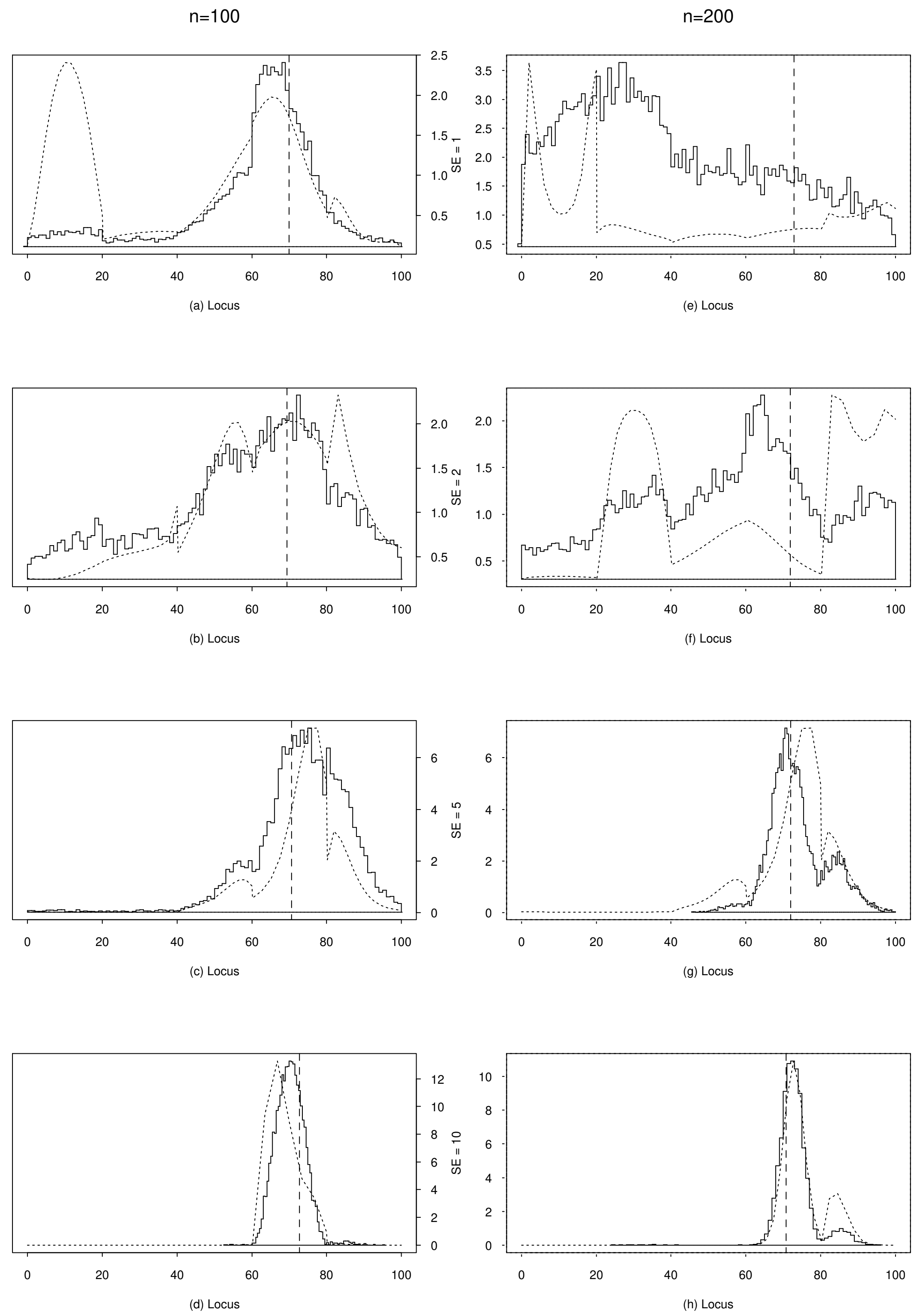

Figure 2: Marginal and Posterior Density of Locus λ. Dashed = posterior (LOD), Solid = marginal (MCMC).

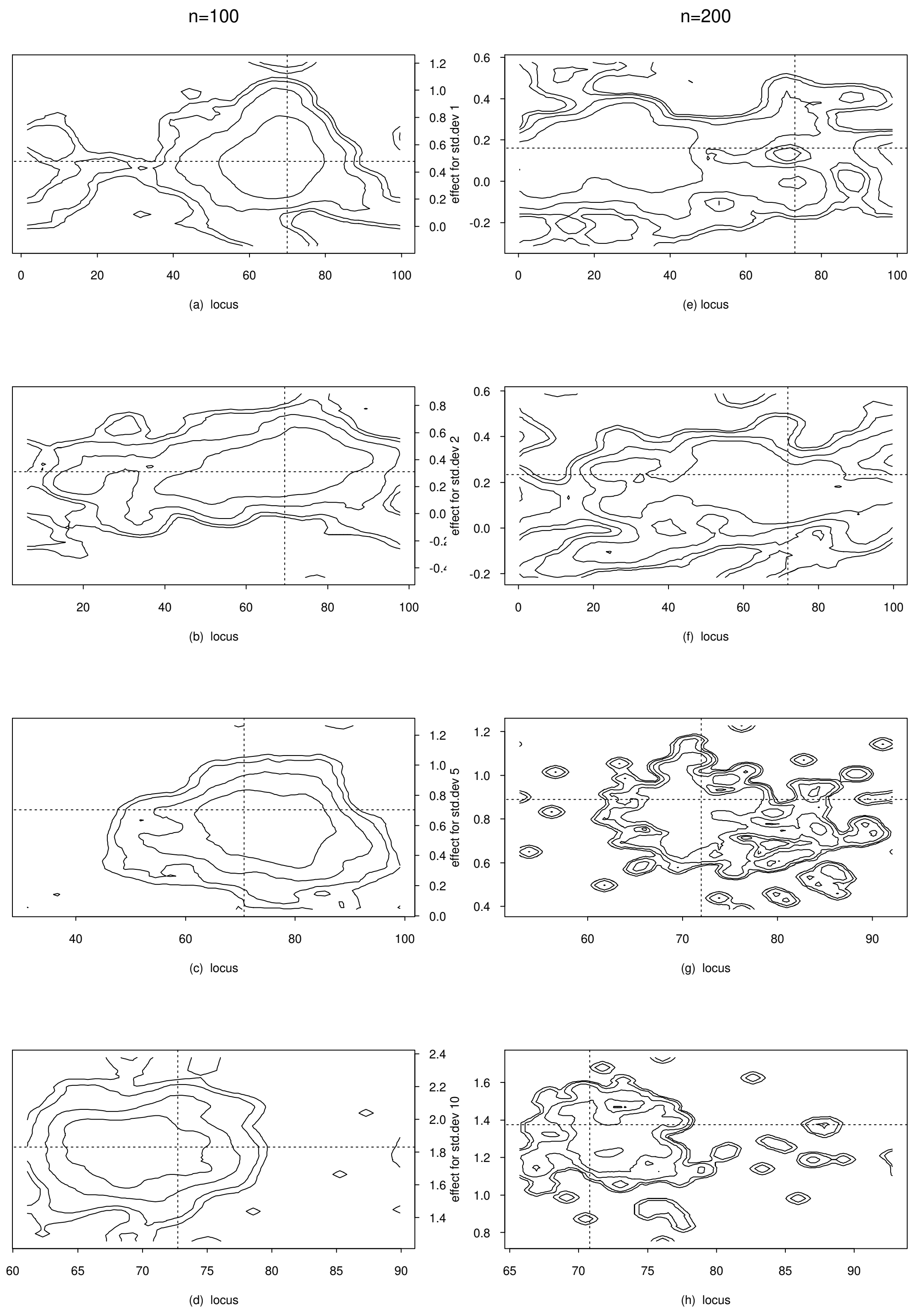

Figure 3: Joint density of locus λ and effect δ. Contours for 50%, 75%, 90% and 95%.

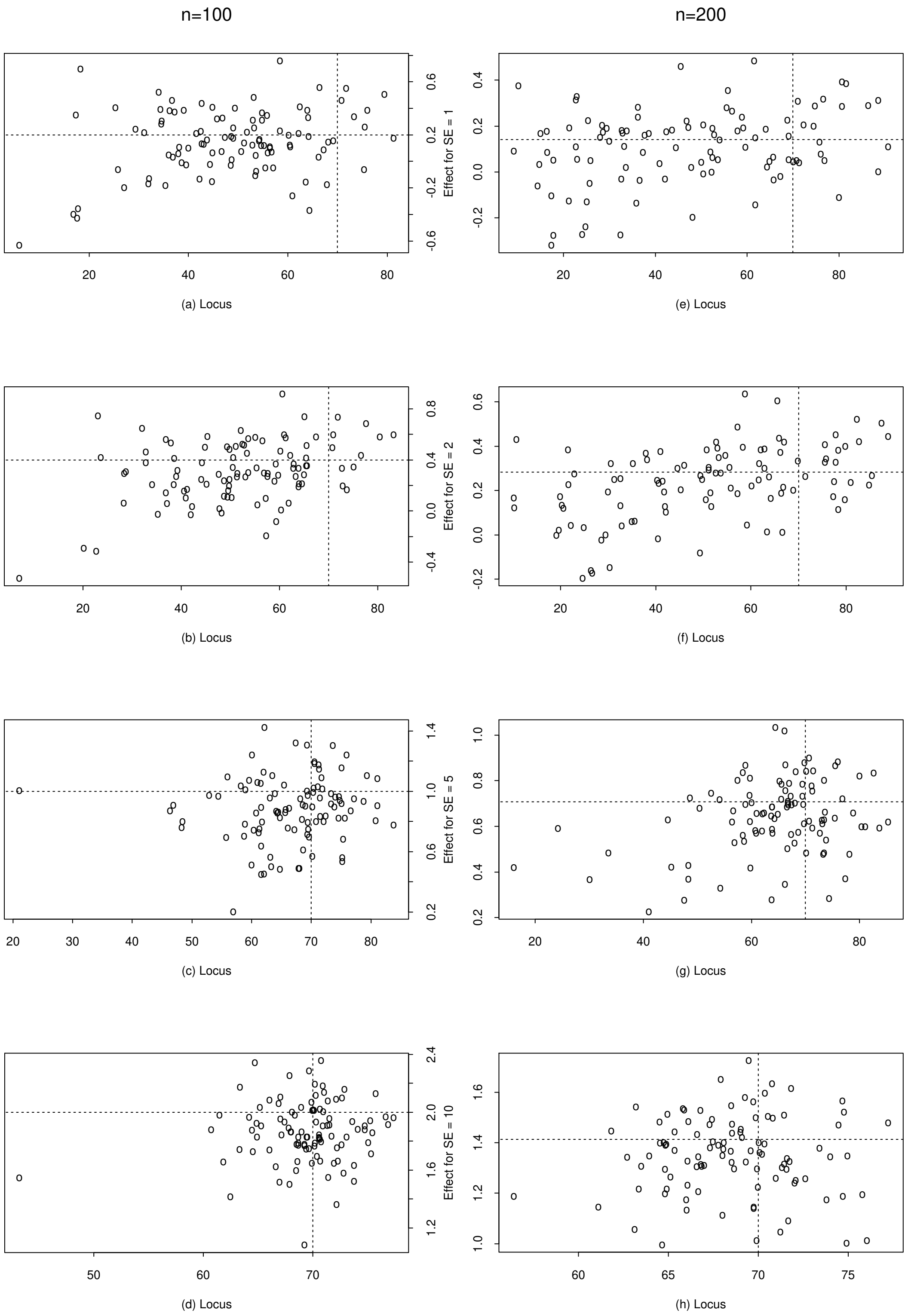

Figure 4: MCMC Means for 100 Independent Realizations.

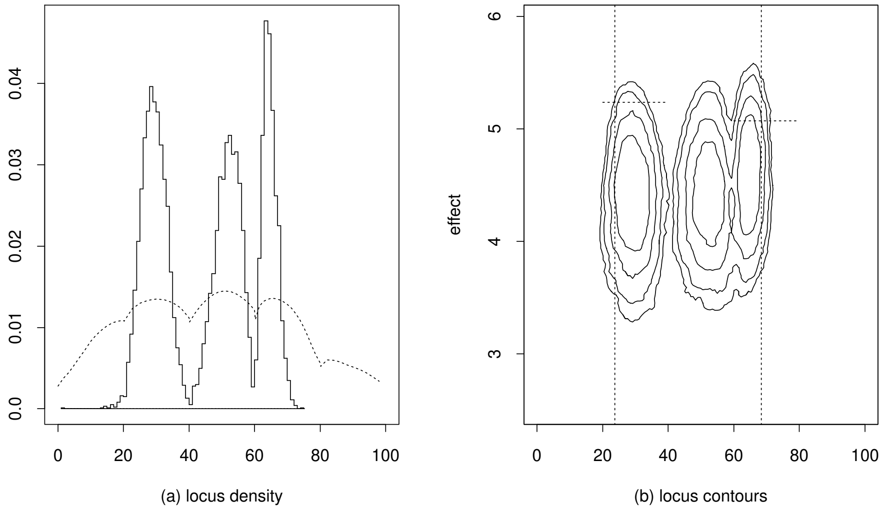

Figure 5: Joint density of locus λ and effect δ when there are two QTLs.