Theor Appl Genet (2002) 104:751–762
DOI 10.1007/s00122-001-0796-x

M.C.A.M. Bink · P. Uimari · M.J. Sillanpää
L.L.G. Janss · R. C. Jansen

Multiple QTL mapping in related plant populations via a pedigree-analysis approach

Received: 1 February 2001 / Accepted: 20 June 2001 / Published online: 7 March 2002

Abstract QTL mapping experiments in plant breeding may involve multiple populations or pedigrees that are related through their ancestors. These known relationships have often been ignored for the sake of statistical analysis, despite their potential increase in power of mapping. We describe here a Bayesian method for QTL mapping in complex plant populations and reported the results from its application to a (previously analysed) potato data set. This Bayesian method was originally developed for human genetics data, and we have proved that it is useful for complex plant populations as well, based on a sensitivity analysis that was performed here. The method accommodates robustness to complex structures in pedigree data, full flexibility in the estimation of the number of QTL across multiple chromosomes, thereby accounting for uncertainties in the transmission of QTL and marker alleles due to incomplete marker information, and the simultaneous inclusion of non-genetic factors affecting the quantitative trait.

Keywords Bayesian approach · Markov chain Monte Carlo analysis · QTL mapping

Communicated by J.W. Snape

M.C.A.M. Bink (✉) · R.C. Jansen
Business unit Biometry,
Plant Research International B.V., P.O. Box 16,
6700 AA Wageningen, The Netherlands
e-mail: m.c.a.m.bink@plant.wag-ur.nl
Fax: +31-317-418094

M.C.A.M. Bink · L.L.G. Janss
Department of Genetics and Reproduction, ID-Lelystad,
P.O. Box 65, 8200 AB Lelystad, The Netherlands

P. Uimari
CSC-Scientific Computing, P.O. Box 405, 02101 Espoo, Finland

M.J. Sillanpää
Rolf Nevenlinna Institute, P.O. Box 4,
00014 University of Finland, Finland

Introduction

Breeders and geneticists have developed statistical methods to identify quantitative trait loci (QTL) by utilising molecular markers. These methods have sought to answer basic questions concerning QTL (e.g. number, mode of action and size of action) and to map QTL on the genome to facilitate their manipulation for breeding purposes. In plants, populations derived from single crosses of inbred lines have predominantly been used in QTL mapping experiments (Jansen 2001). Major incentives do exist to study more complex populations those derived from multiple founders or collected from ongoing breeding programs. These incentives are:

1) Highly improved exploration of QTL variation since multiple alleles are present at a high probability when a population arises from many founders.
2) Applied context of identified QTL alleles since experimental line crosses often do not represent the (commercial) breeding populations (Tanksley and Nelson 1996).
3) Improved cost effectiveness of QTL mapping by using available phenotypes from selection experiments, since the cost of obtaining marker data likely continues to decline and evaluating phenotypes becomes relatively more expensive. Breeding programs routinely evaluate the phenotypes of many progeny with replication at several locations.

These incentives should convince plant geneticists and breeders that they should better exploit the data from complex populations. However, the analysis of this type of data has been hampered by the absence of flexible and robust statistical tools and methods. Important criteria for QTL mapping in complex data may be summarised as:

1) Robustness and flexibility to possible structures in the data, especially in pedigree; i.e. individuals may cover multiple generations, the population may cover multiple families, with large differences in size and relationships in between.

2) The number of QTL, across a single chromosome and across all chromosomes, is in fact unknown and should be treated as such in the analysis. Also, the mode of action of QTL is unknown and/or may interact with the (genetic or environmental) background in which it is expressed.
3) Partial marker information; this holds on multiple levels, i.e. DNA on an individual may be absent, markers are partially scored on an individual, markers may be partially informative (e.g. dominant scoring) or markers may not be informative on an individual (e.g. homozygous for parents).
4) Environmental factors may contribute to the observed phenotypic variation in the quantitative trait. Pre-correction for these factors may eliminate uncertainty in these factors and can introduce bias in parameter estimates. Simultaneous analysis seems to be more appropriate.

Here we attempt to accommodate all these criteria by the application of a Bayesian approach. That is, a Bayesian framework with Markov chain Monte Carlo (MCMC) algorithms provides a powerful tool for estimating the chromosomal location and contribution of genes affecting complex traits and, potentially, gene-by-gene and gene-by-environment interactions as well. The introduction of the reversible-jump MCMC algorithm by Green (1995) solved the problem of estimating the number of QTL in a Bayesian framework (see also Waagepetersen and Sorensen 2001). This has successfully been applied in simple line crosses or a single outbred family in plants (Satagopan and Yandell 1996; Sillanpää and Arjas 1998, 1999; Stephens and Fisch 1998), humans (Heath 1997; Thomas et al. 1997; Lee and Thomas 2000; Uimari and Sillanpää 2001) and animals (Uimari and Hoeschele 1997; George et al. 2000). The aim of this paper is to show plant geneticists and breeders that complex data from plant breeding schemes can be utilised for QTL mapping by making use of Bayesian and MCMC methodology. Full utilisation of this type of data with non-Bayesian methodology is currently not possible, and extensive comparison of our method to others was not possible. This type of comparison, frequentist versus Bayesian, has been performed on previously published experimental data from an outbred progeny of a single cross between two apple cultivars (Maliepaard et al. 2001). Next to the description of the genetic model and sampling procedures, the application of the method to real data is described to illustrate the potential of the method. Important aspects and extensions of the Bayesian methodology accommodating plant population characteristics are discussed.

Methods and material

Terminology

Consider a diploid mapping population of N individuals containing N_f founder individuals and N_{nf}, non-founder individuals and expressing some arbitrary but known pedigree structure. A non-founder individual has both parents known and present in the pedigree, while a founder individual does not. The haplotype is defined as the allele configuration (at different loci) received from one parent. Under Hardy-Weinberg equilibrium, the population frequencies of the maternal and paternal alleles at a locus are considered to be independent (unrelated). Let a phenotype be an observable or measurable trait of an individual, then the likelihood (or "penetrance") function, $f(y|g)$, is the conditional probability of observing phenotype y given QTL genotypes g, which is assumed to be a Normal distribution. Here, we consider bi-allelic QTL and elaborate on the multi-allelic QTL in the Discussion section.

The genotypes for QTLs are determined together by the alleles of founder individuals (**G**) and by segregation indicators of non-founder individuals (**S**) (Lange and Matthysse 1989; Thompson 1994; Sobel and Lange 1996). The segregation indicators uniquely describe the flow of genes through a pedigree. An example hereof is given in Fig. 1(a, b), were segregation indicator '0' indicates the paternally inherited allelic state and '1' indicates the maternally inherited one. For example, individual 4 with genotype Aa has segregation indicators [0, 0] since it received the paternal allele from its father and also the paternal allele from its mother. Note that the paternal segregation indicator of individual 4 is not unique in the sense that it cannot be inferred uniquely from the observed marker data (homozygous parent). The segregation indicators are arranged into a matrix where each non-founder individual has two columns, i.e. the first column pertaining to the grandparental origin of the paternal allele and the second column pertaining to the grandparental origin of the maternal allele. The rows of this matrix correspond with all loci (markers and QTL). Observed marker data (**M**) may not be decisive in the segregation of alleles from parents to offspring, especially the so-called grandparental origin may be blurred. Here, we will use marker haplotypes (**H**) that uniquely determine the grandparental origin or linkage phase in parents, where the indicator $P(\mathbf{M}|\mathbf{H})$ equals one if the haplotype is consistent with observed marker data and equals zero otherwise.

Model

The model and underlying assumptions follow very closely the genetic model presented by Uimari and Sillanpää (2001). Here, QTL – QTL and QTL-environment interactions are not considered. A quantitative trait is modeled as being genetically controlled by N_{QTL} different QTLs and possibly multiple environmental factors. For QTL i, genotypes QQ, Qq and qq have effects a_i, d_i and $-a_i$, respectively. The additive (a_i) and dominance (d_i) effects for QTL i are collected together in the vector $\alpha_i \equiv (a_i d_i)^T$. Let $\mathbf{Q}_i$ ($n \times 2$) denote the (unknown) incidence matrix for the i^{th} QTL for a pedigree with n individuals. The elements of $\mathbf{Q}_i$ are derived directly from the genotypes for the i^{th} QTL, and these genotypes are typically unknown and are inferred from (observed) genotypes at flanking markers. A particular row of $\mathbf{Q}_i$ takes values of {–1,0}, {0,1} or {1,0} in its first and second column. The model for the quantitative trait values **y** ($n \times 1$ vector) is

$$\mathbf{y} = \sum_{i=1}^{k} \mathbf{Q}_i \alpha_i + \mathbf{X}\beta + \mathbf{e}$$

where, β is an ($m \times 1$) vector of environmental effects (including an intercept μ) and **e** is an ($n \times 1$) vector of normally distributed residual effects; **X** ($n \times m$), is a known incidence matrix.

In a Bayesian context, the model may best be described by a directed acyclic graph (DAG) with the explanation of parameters given in Fig. 2. Here, the dependencies among observed (or known) variables (in boxes) and unobserved variables (in ovals) become apparent. The joint posterior density of the parameters and unobserved data, given observed phenotypic and marker data and environmental factors is

$$P(\theta, N_{QTL}, \mathbf{G}, \mathbf{S}, \mathbf{H} | \mathbf{y}, \mathbf{X}, \mathbf{M}) \propto P(\theta | N_{QTL}) P(N_{QTL}) P(\mathbf{G}|p) P(\mathbf{H}|q) P(\mathbf{S}|l, \mathbf{H}) P(\mathbf{M}|\mathbf{H}) P(\mathbf{y}|\theta, N_{QTL}, \mathbf{G}, \mathbf{S}, \mathbf{X})$$

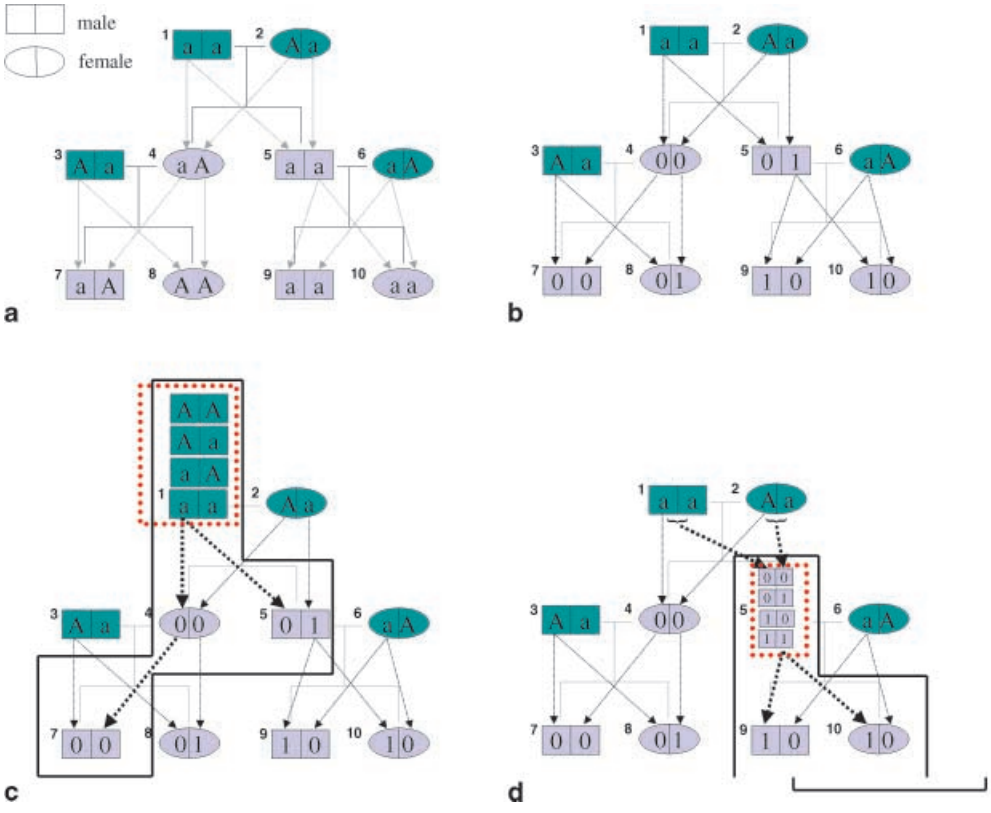

Fig. 1 Graphical representation of the flow of genes (alleles A and a) in a pedigree of ten individuals. **a** Pedigree and genotype configuration at a locus. **b** Distinction between founder individuals (not having parents identified in pedigree) with alleles representing their genotype and non-founder individuals with segregation indicators representing their genotype. The segregation indicators determine the (grand-parental) origin of the alleles of non-founder individuals; *value 0 (1)* indicates the paternal (maternal) allele of a parent where, within an individual, the first (second) allele comes from the father (mother). **c** Updating genotype of founder individual *1*; *dashed box* indicating the four possible genotypes, *bold box* indicating all individuals contributing to sampling probabilities. **d** Updating 'genotype' of a non-founder individual *5*; *dashed box* indicating the four possible segregation indicator patterns, *bold box* containing all individuals contributing to sampling probabilities

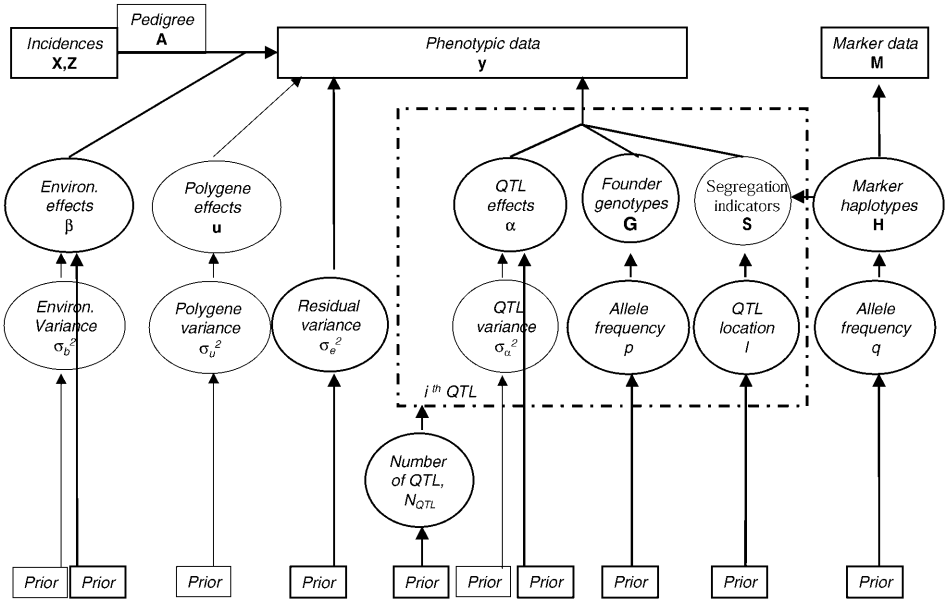

Fig. 2 Hierarchical structure (directed acyclic graph, DAG) of the model. *Boxes* refer to known variables and *ellipses* to random variables. The variables in the *top layer* have been observed: the variables in the *bottom layer* are pre-defined (by the user); the variables in the *intermediate layers* are to be inferred by sampling. The *solid-lined objects* in the DAG are used in the analysis of this study, whereas the *gray-lined objects* may be included in other analyses as outlined in the Discussion. These are polygene effects for individuals to account for background genes that are not linked to marker loci and the additional distributions for environmental and allelic effects when these distributions are treated unknown *a priori*

where θ represent the unknown parameters of interest, that is, $\theta = \{\alpha, l, p, \sigma_e^2, \beta, \sigma_b^2\}$. Note that if the number of founders is small, interest may be in **G** rather than in allele frequency p (which is the case in our data, see later). All other parameters were treated as nuisance variables.

Prior assumptions

In Bayesian analysis, definition of prior knowledge on the model variables is necessary (e.g.Fig. 2). Here, we assign uniform proper priors to all unknowns (except for N_{QTL}), reflecting our ignorance of prior knowledge of these variables. The prior distribution of N_{QTL} is assumed to be a truncated Poisson with mean λ with a predefined maximum (Sillanpää and Arjas 1998). We use two sets of mean and maximum values (Table 2) for this prior to check interference with posterior knowledge (e.g. Satagopan and Yandell 1996; Stephens and Fisch 1998). We refer to Uimari and Sillanpää (2001) for a full description and explanation of the priors on marker and QTL haplotypes.

Map positions of markers were assumed to be known, and Haldane's mapping function was used to convert genetic distances

Table 1 Pedigree and number of informative markers (paternal and maternal) of experiment

Individual	Name	Pedigree		Markers	
		Father	Mother	Paternal	Maternal
1	*mcd*167	–	–		
2	*mcd*178	–	–		
3	*sh*111	–	–		
4	*sh*2988	–	–		
5	*sh*223	–	–		
6–71[a]	*mcd*167×*sh*111	1	3	62[b]	61[b]
72–117	*mcd*167×*sh*2988	1	4	64	60
118–168	*mcd*167×*sh*223	1	5	64	60
169–248	*mcd*178×*sh*111	2	3	51	58
249–315	*mcd*178×*sh*2988	2	4	41	51
316–373	*mcd*178×*sh*223	2	5	30	39

[a] Individuals 6 through 71 shared the same parents; i.e., father *mcd*167 and mother *sh*111

[b] In this set of individuals, 62 (61) markers were informative for paternally (maternally) inherited alleles; i.e. the father (mother) was heterozygous for these markers

into recombination fractions. We assume Hardy-Weinberg equilibrium and linkage equilibrium for all loci among the founder individuals of the pedigrees.

Data

At Plant Research International 24 crosses between cultivated and wild genotypes were tested on resistance to *late blight* (*Phytophthora infestans*) (Colon et al. 1995). Six of these crosses were subsequently genotyped for molecular markers (Sandbrink et al. 2000). The genotypes *mcd*167 and *mcd*178 of the wild South American potato species *Solanum microdontum* were selected as 'resistant' parents of a set of segregating populations on the basis of 3-year averages of field infection by *P. infestans*. These two genotypes were crossed with three 'susceptible' diploid *S. tuberosum* clones, *sh*111, *sh*2988 and *sh*223. In the analysis, these five founders were assumed to be unrelated (Table 1). The progeny size per cross varied between 46 and 80 individuals (Table 1): individuals 72–117, and individuals 169–248, respectively.

One integrated linkage map was available, comprising 12 chromosomes with in total 174 markers covering 983 cM (Bink et al. 2001). The majority of the markers were dominantly-scored amplified fragment length polymorphism (AFLP) markers; others were restriction fragment length polymorphisms (RFLPs) which implied relatively low information per marker. The maximum number of markers being informative for paternally and maternally inherited alleles was 64 and 61, respectively, while the minimum number was 30 and 39, respectively (Table 1). Some chromosomes were not covered with informative markers for certain offspring. For example, chromosome 6 did not contain any informative markers for all three *sh* parents; similarly, chromosomes 8 and 11 did not contain any informative markers for *mcd*178 (see also Fig. 1; Bink et al. 2001). The quantitative trait was field resistance to *Phytophthora infestans*, after artificial inoculation. Each offspring had a single observation, giving 368 phenotypes with values ranging from 0.02 to 51.94. The phenotypic mean and variance were 23.80 and 133.16, respectively. More details on the production of the populations and on disease testing can be found in Colon et al. (1995).

Markov chain Monte Carlo (MCMC) simulation

In each of the MCMC analyses, a single chain was run for l0^6 iterations. No values were deleted because of burn-in. The chain was thinned (saving values in every 200th iteration) to reduce serial correlation in the stored samples (and to reduce storage); the number of stored samples was 5×10^3 Four models were studied (P5A, P5D, P8A and P8D), differing in mode of action of the QTL (A or D) and in the prior for N_{QTL} (''5" or ''8"). When fitting a purely additive acting QTL, the dominance effect was kept fixed at zero, otherwise both additive and dominance was fitted (Table 2). All chains were initiated with $N_{QTL} = 0$, the initial values for elements in β were also zero, except for the overall mean having its initial value computed as the average of all trait phenotypes $\hat{\mu}$, and the residual variance was set equal to the variance of all trait phenotypes.

Table 2 Specification of four MCMC chains

	λ	max(N_{QTL})	QTL effects	
			Additive	Dominance
P5A	5	10	Fitted	Fixed at zero
P8A	8	12	Fitted	Fixed at zero
P5D	5	10	Fitted	Fitted
P8D	8	12	Fitted	Fitted

Identification of QTL location and effects

Following Sillanpää and Arjas (1998), we used the posterior QTL intensity as a probabilistic measure for the localisation of QTL. During the MCMC sampling, we did not restrict the (chromosomal) order of the QTL in order to label them since this way of implementation was easier.

The design of our data affected the estimation of the QTL effects when allowing dominance in the model: the small number of founders and the availability of only two generations of individuals sometimes resulted in the absence of a class of genotypes in the second generation (for which phenotypes are available): e.g. absence of qq, hence prohibiting the estimation of some contrasts. For example, when parent MCD167 was heterozygous for a QTL and all other four founders were homozygous for the same allele, then it was not possible to distinguish additive effect and dominance effect of this QTL. Therefore, the MCMC output was scrutinised on the presence of all three genotypes in the offspring when estimating additive and dominance effects of a QTL. If Q and q denoted the two possible alleles at a QTL (where Qq represents also qQ) and *mcd*, *sh* refer to parents, then the following cases were useful: (1) $mcd_{Qq} \times sh_{Qq}$; (2) $mcd_{Qq} \times (sh_{QQ} + sh_{qq})$; (3) $sh_{Qq} \times (mcd_{QQ} + mcd_{qq})$. At a given MCMC-iteration, zero, one or two of these cases may occur, of which the zero-cases were excluded when estimating additive and dominance effects.

Results

Genome-wide N_{QTL}

Fitting the additive models P5A and P8A (Table 2) resulted in a posterior estimate of N_{QTL} 3.56 and 3.64, respectively. For the dominance models P5D and P8D, these numbers were 4.59 and 4.45, respectively. Obviously, differences in these estimates were larger due to the type of actions fitted for the QTL than due to the prior assumptions. Also, the shapes of the posterior distributions of N_{QTL} were very similar within the 'type-of-action' models, where as a clear shift was observed between the 'type-of-action' models (additive versus domi-

Table 3 Posterior probability estimates for the number of QTL per chromosome, P(N_{QTL} |**y**,**M**). Four models were applied, varying in *a priori* expected number of QTL (P5A+P5D or P8A+P8D) and varying in exclusion (P5A+P8A) or inclusion (P5D+P8D) of dominance as mode of action for the QTL

	Chromosome											
	1	2	3	4	5	6	7	8	9	10	11	12
P(N_{QTL} |**y**,**M**)												
P5A												
0	0.99	1.00	0.93	0.00	0.00	0.62	0.98	1.00	0.98	0.00	1.00	0.96
1	0.01	0.00	0.07	0.99	0.99	0.36	0.02	0.00	0.02	0.99	0.00	0.04
2	0.00	0.00	0.00	0.01	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.00
P5D												
0	0.99	0.98	0.53	0.00	0.00	0.32	0.98	0.99	0.99	0.00	0.97	0.96
1	0.01	0.02	0.46	0.97	0.91	0.67	0.02	0.01	0.01	0.91	0.03	0.04
2	0.00	0.00	0.01	0.03	0.09	0.01	0.00	0.00	0.00	0.07	0.00	0.00
P8A												
0	0.98	0.99	0.93	0.00	0.00	0.60	0.99	0.97	0.99	0.00	0.98	0.98
1	0.02	0.01	0.07	0.99	0.97	0.39	0.01	0.03	0.01	0.98	0.02	0.02
2	0.00	0.00	0.00	0.01	0.03	0.01	0.00	0.00	0.00	0.02	0.00	0.00
P8D												
0	0.99	1.00	0.55	0.00	0.00	0.23	0.98	1.00	0.99	0.00	0.94	0.97
1	0.01	0.00	0.44	1.00	0.95	0.76	0.02	0.00	0.01	0.99	0.06	0.03
2	0.00	0.00	0.01	0.00	0.05	0.01	0.00	0.00	0.00	0.00	0.00	0.00

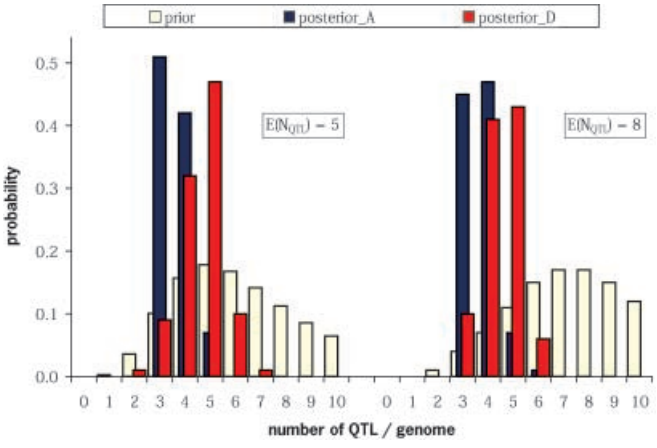

Fig. 3 Prior and posterior probability density on N_{QTL} across the genome (12 chromosomes) with respect to the four different models (*Posterior_A* = additive QTL, *Posterior_D* = additive and dominance QTL, as defined in Table 2)

nance) (Fig. 3). In all models, the posterior distribution was clearly more peaked than the prior distribution with three or four values for N_{QTL} covering more than a 0.98 probability in the additive and dominance models, respectively (Fig. 3).

Chromosome-wise N_{QTL}

The estimated posterior distributions of N_{QTL} for all 12 chromosomes are given in Table 3 and these clearly indicate that chromosomes 1, 2, 7, 8, 9, 11 and 12 did not contain QTL that were segregating in the five parents. The presence of at least one QTL on chromosomes 4, 5 and 10 was approved, irrespective of which model was fitted, where the probability of two QTL on the same chromosome was relatively small; this occured only in the dominance models. The probability of a segregating QTL on chromosomes 3 and 6 was clearly higher for the dominance models than for the additive models (Table 3). For chromosome 6, QTL were only approved if segregation occurred in the *mcd* parents, since the *sh* parents did not have informative markers on this chromosome.

QTL locations

The posterior distributions (QTL intensity according to Sillanpää and Arjas 1998) of map locations of QTL are given in Fig. 4 for models P5A (left panel) and P8D (right panel). The posterior distributions of P8A were very similar to those of P5A, and those of P5D were very similar to those of P8D (results not shown). Posterior knowledge of the map location of QTL varied from dense to vague for the QTL on chromosomes 4 and 3, respectively (Fig. 4). When evidence for presence of a QTL was strong, the posterior distributions of QTL locations were very similar for the additive and dominance models. For the P8D model, the posterior modes were 13, 39, 47, 9 and 14 cM for QTL at chromosomes 3, 4, 5, 6 and 10, respectively. Again, these modes were very similar for all models, although the magnitude of the peaks varied between the additive and dominance models (Fig. 4).

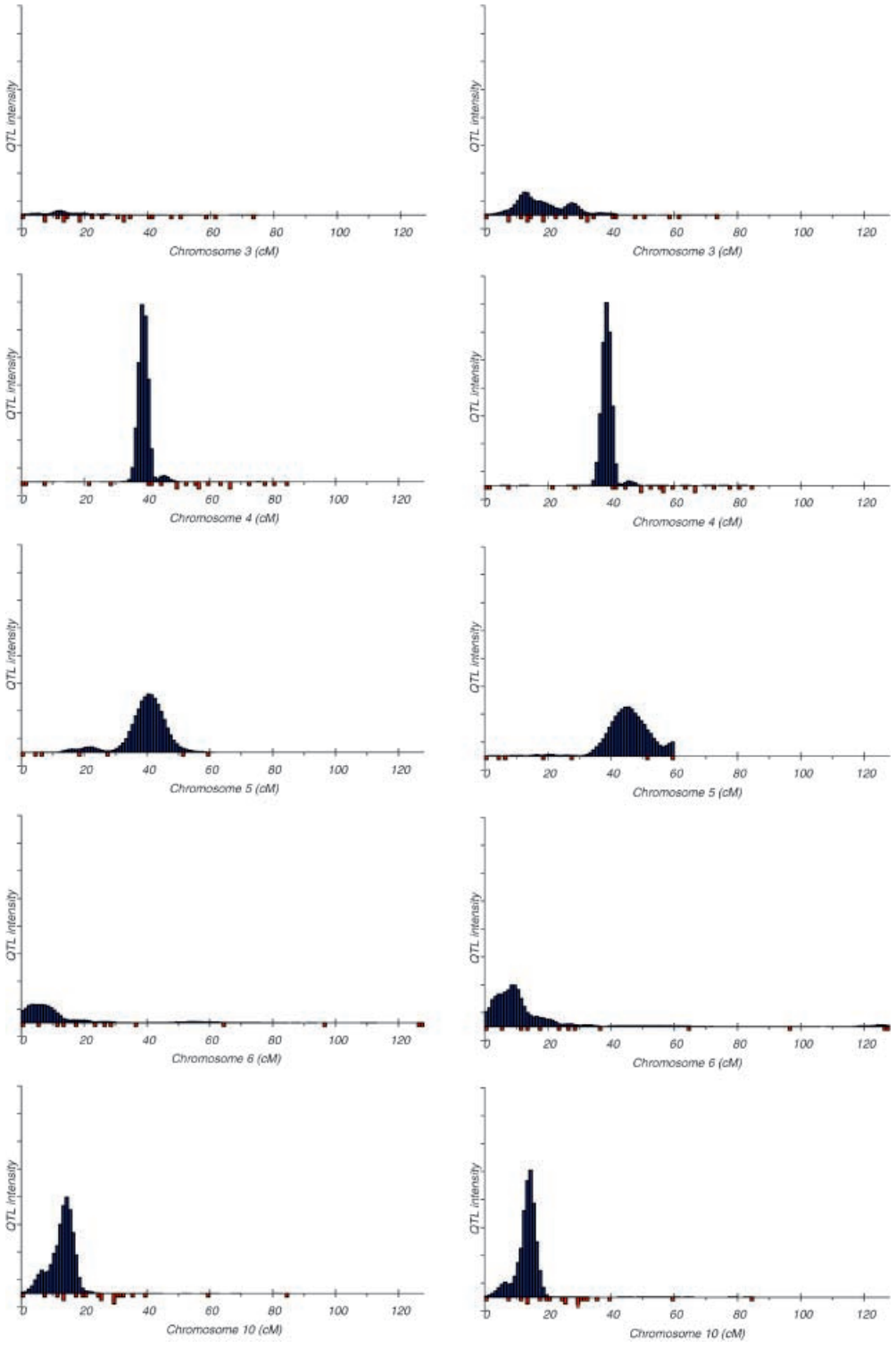

Fig. 4 Posterior QTL intensity (scale equal but not shown) for those chromosomes with $P(N_{QTL} \geq 1 | \mathbf{y}, \mathbf{M}) > 0.10$ for model P5A (*left panel*) and P8D (*right panel*)

QTL heterozygosity of founders

For the QTL on chromosome 4, only one of the founders (*mcd*167) was segregating; all others were homozygous for the same allele (Table 4). Apparently, individual *mcd*167 was segregating for a very large QTL that did not or could not be segregating in other parents. Inference on heterozygosity for the QTL on chromosome 10 was also clear; i.e. both *mcd*167 and *mcd*178 segregated for this QTL while the other founders did not, although the posterior probability of *sh*223 was relatively high for the P8D model. The inferences on segregation of the QTL on chromosome 5 were consistent within the models fitted, but differences between the additive and dominance models were clearly present. The *mcd* parents were very likely segregating for the QTL, although the

Table 4 Posterior probability estimates for parents being heterozygous for a QTL

Chromosome	Model[a]	P(QTL|**y**, **M**)[b]	P(heterozygous parent | QTL, **y**, **M**)[c]				
			*mcd*167	*mcd*178	*sh*111	*sh*2988	*sh*223
3	P5A	0.07	0.06	0.09	0.05	0.97	0.13
3	P8A	0.07	0.15	0.07	0.07	0.95	0.15
3	P5D	0.58	0.10	0.01	0.02	0.97	0.03
3	P8D	0.46	0.10	0.02	0.03	0.98	0.02
4	P5A	1.00	1.00	0.00	0.00	0.00	0.00
4	P8A	1.00	1.00	0.00	0.00	0.00	0.00
4	P5D	1.00	0.98	0.00	0.00	0.01	0.01
4	P8D	1.00	1.00	0.00	0.00	0.00	0.00
5	P5A	1.00	0.92	0.99	0.92	0.94	0.71
5	P8A	1.00	0.89	0.98	0.89	0.90	0.68
5	P5D	1.00	0.84	0.92	0.21	0.59	0.65
5	P8D	1.00	0.90	0.97	0.20	0.62	0.58
6	P5A	0.39	0.93	0.27	0.42	0.63	0.17
6	P8A	0.41	0.90	0.18	0.41	0.67	0.14
6	P5D	0.67	0.97	0.76	0.25	0.33	0.11
6	P8D	0.78	0.97	0.58	0.30	0.29	0.17
10	P5A	1.00	0.96	1.00	0.00	0.01	0.16
10	P8A	1.00	0.95	0.98	0.00	0.01	0.15
10	P5D	1.00	0.89	0.93	0.01	0.06	0.12
10	P8D	1.00	0.99	1.00	0.00	0.01	0.22

[a] Models: see Table 2
[b] P(QTL|**y**, **M**) is the posterior probability of at least one QTL being present on a chromosome
[c] P(heterozygous parent | QTL, **y**, **M**) is the posterior probability of parent being heterozygous for a QTL, conditional on the presence of that QTL

probability tended to decrease when dominance was allowed in the model. In the dominance model, the probability of segregation of QTL in the *sh* parents was considerably lower, especially for parent *sh*111 (0.2 instead of 0.9 in the additive model). The QTL on chromosome was mostly likely segregating in parent *sh*2988, irrespective of the model, while all other parents most likely did not segregate for this QTL. Parent *mcd*167 was most probably segregating for the QTL on chromosome 6, while parents *mcd*178, *sh*111 and *sh*2988 had a moderate probability of segregation. For these latter three parents, the probabilities were different for the additive and dominance models. Here it should be noted (again) that parents *sh*111 and *sh*2988 did not have informative markers on chromosome 6 and, consequently, instead of multipoint QTL mapping a (major gene) segregation analysis was actually performed for these individuals.

QTL effects

Additive models

The largest substitution effect was found for the QTL on chromosome 4, which was approximately three times as large as the effects estimated for on chromosome 5 and 10 (not shown). These latter substitution effects were also estimated accurately, while posterior distributions of the substitution effects of QTL on chromosomes 6 and 3 were less accurate (results not shown).

Dominance models

As explained previously, it was not possible to estimate dominance and additive effects for all QTL. For the QTL found on chromosomes 3, 5 and 10, we computed the highest posterior density (HPD) regions along the chromosome using 1 cM – bins (Fig. 5). Note that we did not account for the probability of occurrence of the QTL in a particular bins: to do this one can weight the estimates with the estimated QTL intensity from Fig. 4 (or one could impose a threshold on QTL intensity). The estimated HPD regions for substitution effects were very similar to those in the additive model (latter results not shown). Dominance was substantial for the QTL on chromosomes 3 and, in particular 5, whereas the QTL on chromosome 10 acted predominantly in an additive mode. The HPD region results for the P5D and P8D models were highly similar (Fig. 5), although the range along chromosomes 3 and 5 was somewhat extended in the P5D model (but the QTL intensity in these additional ranges was very low).

Interpretation of Bayesian results

Our marker data comprised a linkage map of 12 chromosomes with 174 markers. While these figures suggest that a rather dense marker map was available, the information content of the markers was highly variable. For example, we had a few RFLP markers that were informative on all five founder parents and that were scored for all offspring. However, we also had many AFLP markers that were only informative for one out of five parents and perhaps only for offspring resulting from the mating of this parent with only one of the other parents. Furthermore, a large number of (AFLP) marker scores were missing on offspring (for details, Bink et al. 2001). An important observation in our analysis was that for chromosomes with a low marker information content (e.g. chromosomes 2, 8 and 11). QTL probabilities were

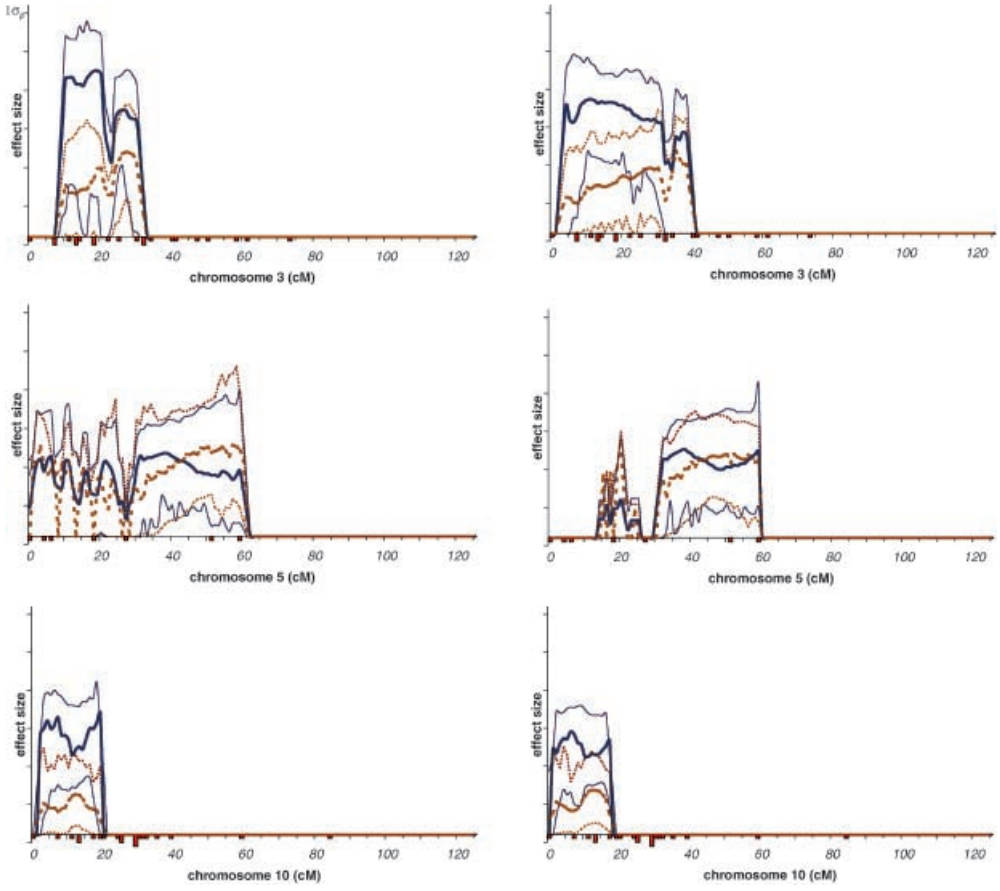

Fig. 5 Highest Posterior Density regions for dominance (*dashed lines*) and additive (*solid lines*) effects for QTL on chromosomes 3, 5, and 10 for the models P5D (*left panel*) and P8D (*right panel*). The *bold line* represent the posterior median (50% quantile) the *thinner lines* represent the lower bound (5% quantile) and upper bound (95% quantile) of the 90% HPD region

low. The Bayesian multiple-QTL methods apparently avoid false-positive (or ghost) QTL (e.g. Wright and Kong 1997). Also, in general, Bayesian methods seem to be well suited to detect multiple QTLs on a linkage group since these are modelled explicitly. This is supported by simulation studies (Sillanpää and Arjas 1998, 1999).

The simultaneous screening of all chromosomes and the recognition of parental relationships in the Bayesian analysis resulted in somewhat different results than those previously obtained from analyses of the same data (Sandbrink et al. 1999; Bink et al. 2001). In these previous analyses, data on a single group of offspring of two parents (Sandbrink et al. 1999) or on a single chromosome (Bink et al. 2001) were analysed with single-QTL models. Bink et al. (2001) in particular reported many QTL that were not consistent across different groups of offspring from the same parent. Fitting the same parental alleles for these groups of offspring automatically eliminated these inconsistencies. One may argue that a QTL may interact with either background genes (epistasis) or with environmental factors (G×E interaction), a factor that we currently ignore in our model. These types of interactions may be incorporated in the Bayesian model (see later); however, we think that the potato marker data was insufficient to estimate parameters of these extended models accurately.

Accurate estimation of dominance and additive effects for some QTL was already hampered in the current analysis because of the small number of founders (and small number of generations). For the QTL on chromosomes 3, 5 and 10 we plotted HPD regions for the size of the allelic effects across the chromosome (Fig. 5). These HPD regions can be very useful to plant breeders: they do not only provide an point estimate but also clearly express the remaining uncertainty in parameter estimates. For example, in the case of risk avoidance, the choice for a QTL for implementation in a breeding scheme may be based on its relatively high HPD lower bound rather than on its relatively lower posterior mean estimate. The HPD regions were rather constant over a substantial length of the chromosomes, suggesting a low discrimination power in the location of the QTL. However, the QTL intensity plots (Fig. 4) should be the most determining factor in assigning the most likely map location of the QTL (Sillanpää and Arjas 1998; Xu and Yi 2000). One could impose a minimum threshold of QTL intensity for plotting the HPD regions for allelic effects of the QTL.

The high similarity of results indicated that posterior inferences were not sensitive to prior assumptions on N_{QTL} i.e. the *a priori* expected number of QTL. In other words, the data contained sufficient information to overwhelm the prior expectations. Note that inconsistency in posterior estimates P5A versus P8A, and P5D versus P8D) could have been caused by both convergence failure or by sensitivity to prior assumptions. In addition, we ran additional MCMC simulations with an extreme value

for λ– 25 – and the posterior mean estimates for the number of QTL were 3.91 and 4.61 for P25A and P25D, respectively. The posterior distributions for N_{QTL} in these models were very similar to those presented in Fig. 3.

Discussion

We describe a Bayesian method for QTL mapping in complex plant populations and report results from its application to a potato data set. This Bayesian method was originally developed for the analysis of data in human genetics (Uimari and Sillanpää 2001). Uimari and Sillanpää (2001) evaluated their method on simulated (and real) data in human genetics, and their results on effects and locations of QTLs were consistent with the values used in simulation. Here, the same method proved to be adequate for complex plant populations as well. In the following we will discuss issues on: (1) Bayesian hierarchical modeling; (2) extension to typical plant populations and (3) revenues from a Bayesian QTL analysis.

Bayesian hierarchical modeling

The directed acyclic graph that was presented in Fig. 2 presents all dependencies among known quantities (data or prior assumption) and unknown quantities (model parameters). This graph directly shows how the modal can be extended or modified to allow additional variables in the model; e.g. a polygenic component for all individuals accounting for the joint additive effect (**u**) of genes unlinked to any of the markers (background genes). These additive effects, one for each individual, follow a Normal prior, i.e., $N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, where **A** is the numerator relationship matrix (Henderson 1976) derived from the known pedigree, and σ_u^2 is the variance due to these unlinked genes (also treated as an unknown). One may as well introduce an extra layer in the model; for example, the allelic effects of the QTL are no longer distributed following a pre-specified prior distribution (bounded Uniform) but following a Normal prior, $N(\mathbf{0}, \mathbf{I}\sigma_\alpha^2)$, A similar extension can be made for environmental factors (see Fig. 2).

In this study, we introduced QTL with two alleles, which has been commonly used in previously presented Bayesian analyses as well (Uimari and Hoeschele 1997, Heath 1997). When the true number of alleles at a QTL is larger than two, this may be accommodated in the biallelic QTL model by allowing multiple QTL at the same chromosomal segment. In our study, we did not prohibit QTL to be very close to each other; however, we did not observe this phenomenon in our analysis, suggesting that the biallelic QTL model seemed sufficient for these data. The extension from biallelic to a multi-allelic QTL model in the Bayesian analysis is also straightforward since this only affects the allele frequency (**p**) and the founder genotypes (**G**). Instead of a scalar **p** representing the allele frequency of one of the two QTL alleles, we now have a vector **p** representing the frequencies of all possible QTL alleles. There is, however, a potential danger when using highly multi-allelic QTL modelsnot all QTL genotypes may be represented among individuals with phenotypes. Then, the posterior estimates for allelic effects that are related to these 'phenotype-empty' genotypes are simply a representation of the prior knowledge, which may be very vague. An appealing alternative seems to be the treatment of the number of alleles per QTL as being similar to the number of QTL – as unknown and to be sampled by using a reversible jump mechanism. Another alternative model for the QTL (not shown in Fig. 2) is the QTL variance component analysis (Fernando and Grossman 1989). In this approach each individual has two unique allelic effects for a QTL, where allelic effects of relatives exhibit a covariance structure that is based on marker haplotypes and the map location of the QTL relative to the known marker positions. This latter Bayesian approach has been implemented for livestock populations (Bink and VanArendonk 1999; Yi and Xu 2000).

Extension to typical plant populations

While we concentrated on outbred plant species in this study, the Bayesian approach can be easily extended to handle fully inbred individuals. Suppose one or more founders are an inbred line of individuals, then the allele frequency for QTL is discrete; i.e. 1 for the allele that is present and 0 for (all) other allele(s). In that case, multiple "families" of founder individuals may be assigned, with unique allele frequencies for each family.

Care must be taken when including an additional polygenic component in the model when inbred founders are present. This may be solved by fitting a finite polygenic model (Thompson and Skolnick 1977), instead of an infinitesimal polygenic model (Fisher 1918). The occurrence of selfing and individuals being the mother of one progeny and being the father of another progeny (diallel designs; Reba and Goffinet 2000) is typical for plant populations and requires some modification of the pedigree approach.

Revenues from a Bayesian QTL analysis

Genome-wide multiple QTL mapping

Simultaneously screening of all chromosomes avoids the use of cofactors. The use of cofactors is not truly Bayesian if cofactor selection is based on the same data; the use of cofactors or unlinked QTLs may also be problematic if markers are not informative in all parents. Cofactors cannot easily be used in multiple family situation because linked alleles (in linkage disequilibrium with the QTL) can be different in different families. The effects of cofactors may be nested within families, which substantially increases the number of parameters. However,

unlinked QTLs or polygenic components may be used instead.

Multiple generations of individuals

Bayesian QTL analysis has the ability to analyse data on multiple generations of genotyped individuals and unbalanced family sizes and mating designs. For example, our Bayesian method was used to analyze data on pig selection lines with almost 4,500 individuals whose pedigree covered more than seven generations (Bink et al. 2000).

Uncertainty in linkage phase of parents

Our sampling of the grand-parental origin of alleles in all individuals naturally reflected the uncertainty in linkage phase of parents. This approves its application to pedigrees with small numbers of offspring per parent as well as for cases with (very) incomplete marker data, which was the case in our potato data.

Missing marker genotypes

Our multi-point QTL mapping procedure utilises all markers neighboring a putative QTL. The informative markers closest to the QTL are used when calculating the segregation probabilities of QTL alleles (avoiding data augmentation).

Non-QTL influences

Any environmental factors or residual polygenes can be included directly into the QTL mapping model, such that phenotypic pre-adjustments are not needed and, moreover, uncertainties in their estimation can be fully accounted.

Acknowledgements The manuscript has benefited from reviewers' comments. M. Bink acknowledges financial support of the Acadamy of Finland (research grant no. 38352) when visiting the Rolf Nevanlinna Institute.

Appendix

Markov chain Monte Carlo simulation

Inferences about estimated parameters were based on marginal posterior distributions that were achieved (approximated) through MCMC simulations. One iteration of the simulation scheme was

A) Marker variables

1) Update blocks of ordered marker loci haplotypes **H** individual by individual. Size of a block was four and the initial block started randomly across all loci;
2) Update marker allele frequencies *q*.

B) Genetic variables

1) Update number of QTLs N_{QTL}, i.e., propose deletion or addition of a QTL with equal probability;
2) Update gene effects α_i for each QTL *i*;
3) Update map position l_i for each QTL *i*;
4) Update ordered founder genotypes **G** for each QTL *i*;
5) Update segregation indicators **S** for each QTL *i*,
6) Joint update **S** with marker haplotypes **H** (at every 10^{th} iteration, see also Figure 2).
7) Update QTL allele frequency *p* for each QTL *i*.

C) Non-genetic variables

1) Update covariate effects β;
2) Update variances σ^2_β and σ^2_e.

The parameters (**H**, **G**, **S**, *p*, q, β, σ^2_β, σ^2_e) were updated by use of Gibbs steps since sampling from their full conditional densities proved to be simple and efficient. Founder genotypes and segregation indicators were updated a single individual at a time using a Gibbs step, as illustrated in Fig. 3(C and D). The full posterior distribution of genotype of individual 1 (Fig. 3C) is determined by the allele frequency (*p*) and possible phenotypes of individuals 1, 4, 5 and 7, since the latter three individuals inherited the first (paternal) allele of individual 1. Note that by use of segregation indicators, the maternal allele of individual 2 being A or a is entirely determined by the allele frequency. Sampling the segregation indicators for individual 5 (Fig. 3D), four possible combinations, involves contributions from markers flanking the trait locus of individuals 1, 2 and 5, and possible phenotypes of individuals 5, 9 and 10. Note that the segregation indicators of these latter individuals do not change when updating individual 5, however, their genotypes may change. The parameters (α_i, l_i) were updated by use of Metropolis-Hastings steps. The number of affecting loci (N_{QTL}) was updated through reversible jump sampling (Green 1995). Here, we adopted the implementation proposed by Sillanpää and Arjas (1998). As a basic strategy only single-step moves were allowed, i.e., only one locus may be added or deleted during an updating cycle. In the locus addition proposal (=birth step), new values for *I*, *p* and α were generated from their priors, allowing multiple QTLs within the same marker interval. To increase the probability of acceptance for the addition step, we used a scaling factor *w* on the prior distribution of α. New founder genotypes were proposed from the prior conditional on the new allelic frequencies (*p*). Similarly, new segregation indicators for non-founder individuals were created conditional on the new location of the QTL; that is, incorporating information from its flanking markers (or neighboring QTL if these were closer to the new location). Given a truncated Poisson prior distribution on the number of QTL with parameter λ, the acceptance ratio of the birth step reduces to (see also Sillanpää and Arjas 1998)

$$A = \frac{P(\mathbf{y}|S', G', \alpha', others)}{P(\mathbf{y}|S, G, \alpha, others)} \times \frac{\lambda}{(N'_{\mathrm{QTL}})^2} \times w$$

where ‘indicates the old values plus proposed values for the new locus (and “others” refer to all other random variables, which are constants in the numerator and denominator). If a deletion (death step) was proposed the locus to be deleted was chosen randomly. The acceptance ratio for a deletion step is $1/A$.

MCMC simulation: convergence and mixing

The order of MCMC updating may be crucial in obtaining proper and efficient mixing. There are two strong dependency relations in the offspring data: the vertical dependency between parents and their offspring, and the horizontal dependency between adjacent loci in each individual. Single-site-updating, i.e. individual by individual and locus by locus, may cause the sampler to explore only a fraction of the sample space because of these dependencies (Sheehan and Thomas 1993; Janss et al. 1995; Jensen and Sheehan 1998). To improve its mixing behavior the proposed Bayesian method was implemented by: (1) replacing QTL genotypes for all individuals by genotypes for founders and segregation indicators for non-founders; (2) joint sampling of map position and segregation indicators of the QTL (3) Omitting data augmentation for untyped or uninformative markers; (4) updating several markers jointly within a single block, where the block was randomly chosen along a chromosome (Uimari and Sillanpää 2001). The contribution of each of these steps or their overall contribution to proper mixing compared to single-site updating schemes, was not evaluated in this study and remains an interesting area of research.

In this study we performed MCMC simulations of 10^6 iterations per analysis to obtain reliable posterior inferences; i.e., to explore to entire parameter space adequately. Whether this number of iterations was really sufficient is hard to say since assessment of convergence of a MCMC chain is a rather difficult task. This holds especially here because the dimension of the model changed with high frequency and consequently the identity of the QTL also changed. We calculated the effective number of samples (Geyer 1992; Sorensen et al. 1995; Lee and Thomas 2000) for those parameters that were always in the model (e.g. β, N_{QTL}, and σ_e^2); however, their validity may not hold in a variable dimension problem. The minimum of these calculated effective numbers among the four MCMC simulations was always for N_{QTL}, with values of 94 and higher. Another possibility would have been to apply some diagnostic tools, such as CODA (Best et al. 1995). but here also problems due to the variable dimension are severe. An ensuring measure that convergence was likely reached in our simulations were the very similar results obtained for the two additive models (P5A and P8A) and those for the dominance models (P5D and P8D).

References

Best NG, Cowles MK, Vines SK (1995) CODA manual version 0.30. MRC Biostatistics Unit, Cambridge, UK

Bink MCAM, VanArendonk JAM (1999) Detection of quantitative trait loci in outbred populations with incomplete marker data. Genetics 151:409–420

Bink MCAM, Uimari P, Verburg FJ, Jansen RC, Janss LLG (2000) Dissection of genetic variance via Bayesian analysis – application to pig selection lines. In: Book of abstracts of the 51st meeting of EAAP, The Hague, The Netherlands, p 95

Bink MCAM, Sandbrink JM, Jansen RC (2001) Linkage and QTL mapping in related full sib families – a case study in potato. J Agric Genomics (www.ncgr.org/research/JAG).

Colon LT, Jansen RC, Budding DJ (1995) Partial resistance to late blight (*Phytophthora infestans*) in hybrid progenies of four South American *Solanum* species crossed with diploid *S. tuberosum*. Theor Appl Genet 90:691–698

Fernando RL, Grossman M (1989) Marker-assisted selection using best linear unbiased prediction. Gen Sel Evol 21:467–477

Fisher RA (1918) The correlation between relatives on the supposition of mendelian inheritance. Rans R Soc Edinburgh 52: 399–433

George AW, Mengersen KL, Davis GP (2000) Localization of a quantitative trait locus via a Bayesian Approach. Biometrics 56:40–51

Geyer CJ (1992) Practical Markov chain Monte Carlo (with discussion). Stat Sci 7:467–511

Green PJ (1995) Reversible Jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732

Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am J Hum Genet 61:748–760

Henderson CR (1976) A simple method for the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83

Jansen RC (2001) Quantitative trait loci in inbred lines. In: Balding D (ed) Handbook of statistical genetics. Wiley, New York, pp 567–597

Janss LLG, Thompson R, VanArendonk JAM (1995) Application of Gibbs sampling for inference in a mixed model major gene-polygenic inheritance model in animal populations. Theor Appl Genet 91:1137–1147

Jensen CS, Sheehan N (1998) Problems with determination of noncommunicating classes for Monte Carlo Markov chain applications in pedigree analysis. Biometrics 54:416–425

Lange K, Mathysse S (1989) Simulation of pedigree genotypes by random walks. Am J Hum Genet 45:959–970

Lee JK, Thomas DC (2000) Performance of Markov chain-Monte Carlo approaches for mapping genes in oligogenic models with an unknown number of loci. Am J Hum Genet 67: 1232–1250

Maliepaard C, Sillanpää MJ, Van Ooijen J, Jansen RC, Arjas E (2001) Bayesian versus frequentist analysis of multiple quantitative trait loci with an application to an outbred apple cross. Theor Applied Genet (in press).

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

Sandbrink JM, Colon LT, Wolters PJCC, Stiekema WJ (2000) Two related genotypes of *Solanum microdontum* carry different segregating alleles for field resistance to *Phytophthora infestans*. Mol Breed 6:215–225

Satagopan JM, Yandell BS (1996) Estimating the number of quantitative trait loci via Bayesian model determination. In: Special contributed Paper Sess Genet Anal Quant Traits Complex Dis, Biomet Sess, Joint Stat Meet.

Sheehan N, Thomas A (1993) On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics 49:163–175

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

Sillanpää MJ, Arjas E (1999) Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics 151:1605–1619

Sobel E, Lange K (1996) Descent graphs in pedigree analysis: application to haplotyping, location scores, and marker sharing statistics. Am J Hum Genet 58:1323–1337

Sorensen DA, Andersen S, Gianola D, Korsgaard I (1995) Bayesian inference in threshold models using Gibbs sampling. Gen Sel Evol 27:229–249

Stephens DA, Fisch RD (1998) Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics 54:1334–1347

Tanksley SD, Nelson JC (1996) Advanced backcross QTL analysis: a method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite lines. Theor Appl Genet 92:191–203

Thomas DC, Richardson S, Gauderman J, Pitkäniemi J (1997) A Bayesian approach to multipoint mapping in nuclear families. Genet Epidemol 14:903–908

Thompson EA (1994) Monte Carlo likelihood in genetic mapping. Stat Sci 9:355–366

Thompson EA, Skolnick MH (1977) Likelihoods on complex pedigrees for quantitative traits. In: Pollack E, Kempthorne O, Bailey TB. Jr (eds) Prac Int Conf Quant Genet. Iowa State University Press, Rmes, pp 815–818

Uimari P, Hoeschele I (1997) Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics 146:735–743

Uimari P, Sillanpää MJ (2001) Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genet Epidemiol (in press)

Waagepetersen R, Sorensen D (2001) A tutorial on Reversible Jump MCMC with a view toward applications in QTL mapping. Int Stat Rev 69:49–62

Wright FA, Kong A (1997) Linkage mapping in experimental crosses: the robustness of single-gene models. Genetics 146:417–425

Xu S, Yi N (2000) Mixed model analysis of quantitative trait loci. Proc Natl Acad Sci USA 97:14542–14547

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