INTRODUCTION TO BAYESIAN QUANTITATIVE TRAIT LOCUS ANALYSIS FOR POLYPLOIDS

Dachuang Cao

Bruce A. Craig

R. W. Doerge

Recommended Citation

Cao, Dachuang; Craig, Bruce A.; and Doerge, R. W. (2004). "INTRODUCTION TO BAYESIAN QUANTITATIVE TRAIT LOCUS ANALYSIS FOR POLYPLOIDS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1158

INTRODUCTION TO BAYESIAN QUANTITATIVE TRAIT LOCUS ANALYSIS FOR POLYPLOIDS

Dachuang Cao[1], Bruce A. Craig[1], & R.W. Doerge[1,2]

[1]Department of Statistics, Purdue University, West Lafayette, IN 47907
[2]Department of Agronomy, Purdue University, West Lafayette, IN 47907

Abstract

Quantitative Trait Locus (QTL) mapping in polyploids is complicated by the unobservable parental QTL configuration, especially the number of copies (dosage) of the QTL. Existing techniques estimate the parental QTL configuration using a profile likelihood approach and do not address the uncertainty in the estimates. In this paper, a Bayesian method is proposed to jointly model the parameters including the parental QTL configuration, QTL location, and QTL effects. Inference for parameters is obtained by integrating the posterior distribution of the parameters via a Markov chain Monte Carlo (MCMC) sampler, which is a hybrid of the Metropolis-Hastings, Gibbs, and reversible jump samplers. Here, because the size of the parameter space varies for different parental QTL dosages, the reversible jump is utilized in order to allow the sampler to move between parameter spaces with different dimensionalities. Additional advantage of this Bayesian technique resides in its flexibility to incorporate prior information and treat missing data augmented. As an example, our method is applied to alfalfa experimental data to identify QTL related to winter hardiness.

Key words: Polyploid, Bayesian QTL mapping

1 Introduction

Quantitative Trait Locus (QTL) mapping detects genomic regions which are associated with the variation of an inheritable quantitative trait of interest. Many advanced QTL mapping methods have been developed for diploid species, which contain two complete chromosome sets in one cell. A complete review of these statistical methods can be found in Doerge (2002). A polyploid has more than two complete chromosome sets, and the number of chromosome sets is called ploidy level (K). QTL mapping for polyploids is much less advanced due to the complications in statistical modeling induced by the complex genetic structure of a polyploid. More specifically, a polyploid may undergo either bivalent pairing (two homologues pair) or multivalent pairing (more than two homologues pair) (Rieseberg and Doyle 1989). Furthermore, with polyploids, the number of alleles for each locus, how many copies for each allele, and the linkage phase between loci in the parents can not be completely observed. Due to this incomplete information, there are multiple possible parental configurations from which the same set of observed data can be generated, and thus follow multiple possible statistical models.

Existing QTL mapping methods for polyploids identify the underlying parental configuration and locate putative QTL by selecting the model which optimizes or maximizes certain criterion, such as the one providing the maximal test statistic among all the possible models (Doerge and Craig 2000; Hackett et al. 2001; Cao et al. 2003). These studies have demonstrated the performance of the QTL mapping method depends on how informative the underlying parental configuration is and how distinct it is from other possible parental configurations. Also, different parental configurations may result in different parameter estimates, such as the QTL location and effects. If the underlying parental configuration is not very informative, the underlying parental configuration may not be the one that provides the maximal test statistic among all the possible models, although its maximal test statistic may be close to the global maximum. Therefore, besides identifying the most likely parental configuration, one would also like to know how confident one is in claiming the identified configuration is the correct configuration; also whether there are other configurations with high probability besides the most likely one. One way to answer this question is to estimate the probability of the identified model (and the other possible models) given the observed data using a Bayesian approach (Bayes 1783).

In this paper, it is assumed that the polyploid of interest performs random bivalent pairing (equally likely to pair with each homologue), and only dominant markers are available. In practice, a co-dominant marker system often provides more information on the parental configuration than a dominant marker system, but it still cannot remove the problem of incomplete information in parental configuration. However, co-dominant markers available for autopolyploids are relatively rare, and dominant markers have the advantages of being both rich in polyploids and easily scoreable. Also the method presented in this paper can be readily intended to co-dominant markers. Therefore, this paper will be presented based on dominant markers under a pseudo-doubled backcross experiment (Doerge and Craig 2000) or pseudo-testcross experiment (Grattapaglia and Sederoff 1994; Brouwer and Osborn 1999). Thus, each locus will be assumed to be biallelic.

2 Method

In a pseudo-doubled backcross experiment, the experimental data are collected on F_1 progeny, the first generation. The F_1 progeny are obtained by crossing an informative parent P_1 and a non-informative parent P_2 under the assumption that the informative parent has at least one, and at most, half the ploidy dose of the dominant allele at each locus, and that the non-informative parent only contains recessive allele at each locus. The non-informative parental can be produced by doubling the half non-informative chromosomes in the informative parent, thus the name of pseudo-doubled backcross experiment. An example of a pseudo-doubled backcross experiment for a tetraploid with two markers is shown in Figure 1. In what follows, for each locus the upper case is used to denote both the locus name and its dominant allele, and the lower case stands for the recessive allele (e.g., A and a). The dosage of a locus means the dosage of the dominant allele at that locus. When a marker is present in an individual, *at least* one dose of the dominant allele for that marker is observed.

Under a pseudo-doubled backcross experiment, the parental configuration will stand for the configuration of the informative parent , since the non-informative parent only contains recessive alleles. Let C denote the (informative) parental configuration, $C = (M, Q)$, where M is the parental marker configuration, and Q is the parental QTL configuration. Take a tetraploid with two marker loci A and B, and one QTL locus Q as an example. To simplify the notation, the parental QTL configuration and QTL locus are both represented by Q, and the meaning of Q will be explained if confusion could occur. Given that one only knows that both markers are present in the informative parent, there are five possible parental marker configurations; and similarly, given that the QTL is present, there are three possible parental QTL configurations (Figure 2).

Let n denote the number of progeny, n_m denote the number of markers, and K the ploidy level. The observable data are the progeny marker presence/absence data $\boldsymbol{I} = (\boldsymbol{I}_1, \boldsymbol{I}_2, \cdots, \boldsymbol{I}_n)$, and the progeny quantitative trait data $\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$, where $\boldsymbol{I}_i = (I_{i1}, I_{i2}, \cdots, I_{i,n_m})$ with $I_{ih} = 1$ meaning the h^{th} marker is present for the i^{th} individual, and 0 otherwise; and y_i is the observed quantitative trait value for the i^{th} individual. Assume a genetic map for a polyploid is given, and provides estimated marker order $\boldsymbol{\lambda}$, marker recombination fractions $\boldsymbol{r}$, and parental marker configuration M. In what follows, $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, ..., \zeta_{n_m-1})$ will be used to represent marker locations in the genetic map, where ζ_h stands for the genetic distance between the the $h + 1^{th}$ marker and the first marker. The proposed Bayesian QTL mapping method for polyploids is to jointly model the parental QTL configuration Q, QTL location η, and QTL effects given the observable data. It is assumed that the QTL effect is a function of its dosage. Let D denote the parental QTL dosage corresponding to the parental QTL configuration Q. The quantitative trait Y has a normal distribution with mean μ_j, and variance σ_j^2 for the progeny having a QTL dosage of j, ($i.e.$, j copies of the QTL), $j = 0, 1, \cdots, D$. The vector $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma})$ will be used to denote the set of QTL effect parameters.

In the proposed Bayesian method, the statistical inference of the parameters is based on their joint posterior distribution, which is proportional to the product of the likelihood function and the prior distribution of the parameters. The posterior distribution reflects the updated beliefs about the parameters combining prior beliefs about the parameters with the information on the parameters contained in the likelihood function. In the setting of QTL mapping, usually there is no prior information on the parental QTL configuration Q, one can then assume all the possible QTL configurations are equally likely. Let all the possible parental QTL configurations be indexed from 1 to Q_T, where Q_T is the total number of parental QTL configurations. Then the parental QTL configuration Q admits a discrete uniform distribution on $\{1, 2, \cdots, Q_T\}$. Similarly, without prior knowledge regarding the location of the QTL, a uniform distribution for η in the interval $(0, \zeta_{n_m-1})$ is a natural choice for the prior distribution of η, where ζ_{n_m-1} denotes the genetic distance between the ${n_M}^{th}$ marker and the first marker in the linkage group. Furthermore, it is assumed that $\mu_j \sim N(\xi_j, \tau_j^2)$ and $\sigma_j^{-2} \sim Gamma(\alpha_j, \beta_j)$, where $\boldsymbol{\xi} = (\xi_0, \cdots, \xi_{n_d})$, $\boldsymbol{\tau} = (\tau_0, \cdots, \tau_{n_d})$, $\boldsymbol{\alpha} = (\alpha_0, \cdots, \alpha_{n_d})$, and $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{n_d})$ are prespecified hyper parameters.

The prior distributions for the Bayesian model are summarized below,

$$\begin{aligned}
Q &\sim \text{ Discrete Uniform on } \{1, 2, \cdots, Q_T\}, && (1)\\
\eta &\sim U(0, \zeta_{n_m-1}), && (2)\\
\mu_j | \xi_j, \tau_j^2 &\sim N(\xi_j, \tau_j^2), && (3)\\
\sigma_j^{-2} | \alpha_j, \beta_j &\sim Gamma(\alpha_j, \beta_j), && (4)
\end{aligned}$$

$i.e.$,

$$\begin{aligned}
p(Q = q) &= \frac{1}{Q_T},\\
p(\eta) &= I(0 < \eta < \zeta_{n_m-1}) / \zeta_{n_m-1},\\
p(\mu_j | \xi_j, \tau_j^2) &= \frac{1}{\sqrt{2\pi}\tau_j} \exp\left\{ -\frac{1}{2} \left(\frac{\mu_j - \xi_j}{\tau_j} \right)^2 \right\},\\
p(\sigma_j^{-2} | \alpha_j, \beta_j) &= \frac{\beta_j^{\alpha_j} x^{\alpha_j - 1} \exp\{-\beta_j \sigma_j^{-2}\}}{\Gamma(\alpha_j)} I(\sigma_j^{-2} > 0).
\end{aligned}$$

The likelihood of the parameters $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}), Q$, and η from the i^{th} individual is given by

$$\begin{aligned} p(y_i, \boldsymbol{I}_i, M, \boldsymbol{\zeta}|\boldsymbol{\theta}, Q, \eta) &= \prod_{j=0}^{K/2} p(y_i, \boldsymbol{Z}_i = j|\boldsymbol{I}_i, M, Q, \boldsymbol{\zeta}) \\ &= \sum_{j=0}^{K/2} p(y_i|\boldsymbol{Z}_i = j, \mu_j, \sigma_j) p(Z_i = j|\boldsymbol{I}_i, M, Q, \eta, \boldsymbol{\zeta}), \end{aligned}$$

where $\boldsymbol{Z}_i = (Z_{i0}, Z_{i1}, \cdots, Z_{i,K/2})$ is the indicator vector of QTL dosage for the i^{th} individual ($i.e.$, if $Z_{ij} = 1$, then the i^{th} individual has QTL dosage j, which is written as $\boldsymbol{Z}_j = j$); and $y_i|\boldsymbol{Z}_i, \mu_{\boldsymbol{Z}_i}, \sigma_{\boldsymbol{Z}_i} \sim N(\mu_{\boldsymbol{Z}_i}, \sigma^2_{\boldsymbol{Z}_i})$. Assuming the quantitative trait data $\boldsymbol{y}$ are independent observations, the likelihood from all the n individuals becomes

$$p(\boldsymbol{y}, \boldsymbol{I}, M, \boldsymbol{\zeta}|\boldsymbol{\theta}, Q, \eta) = \prod_{i=1}^{n} \sum_{j=0}^{K/2} p(y_i|\boldsymbol{Z}_i = j, \mu_j, \sigma_j^2) p(\boldsymbol{Z}_i = j|\boldsymbol{I}_i, M, Q, \eta, \boldsymbol{\zeta}). \tag{5}$$

With data augmentation of the unobservable progeny QTL dosages $\boldsymbol{Z}$, the likelihood of $\boldsymbol{\theta}$, η, Q, and $\boldsymbol{Z}$ is then

$$\begin{aligned} & p(\boldsymbol{y}, \boldsymbol{I}, M, \boldsymbol{\zeta}|\boldsymbol{\theta}, Q, \eta, \boldsymbol{Z}) \\ = \; & p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\boldsymbol{Z}|\boldsymbol{I}, M, Q, \eta, \boldsymbol{\zeta}) \\ = \; & \prod_{i=1}^{n} \prod_{j=0}^{K/2} \left[p(y_i|\boldsymbol{Z}_i = j, \mu_j, \sigma_j) p(\boldsymbol{Z}_i = j|\boldsymbol{I}_i, M, Q, \eta, \boldsymbol{\zeta})\right]^{Z_{ij}}. \end{aligned} \tag{6}$$

Assume the prior distributions are independent. With the notation for conditioning on the parental marker configuration M and $\boldsymbol{\zeta}$ suppressed, the joint posterior distribution of $(Q, \eta, \boldsymbol{\mu}, \boldsymbol{\sigma})$, and augmented data $\boldsymbol{Z}$ (progeny QTL dosages) is given by

$$\begin{aligned} & p(Q, \eta, \boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{I}) \\ \propto \; & p(\boldsymbol{y}|\boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\boldsymbol{Z}|\boldsymbol{I}, Q, \eta) p(Q, \eta, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ = \; & \left[\prod_{i=1}^{n} p(y_i|\boldsymbol{Z}_i, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\boldsymbol{Z}_i|\boldsymbol{I}_i, Q, \eta)\right] p(Q, \eta, \boldsymbol{\mu}, \boldsymbol{\sigma}) \\ = \; & \left[\prod_{i=1}^{n} p(y_i|\boldsymbol{Z}_i, \boldsymbol{\mu}, \boldsymbol{\sigma}) p(\boldsymbol{Z}_i|\boldsymbol{I}_i, Q, \eta)\right] p(Q) p(\eta) p(\boldsymbol{\mu}) p(\boldsymbol{\sigma}) \\ \propto \; & \frac{I(0 < \eta < \zeta_{n_m - 1})}{Q_T \, \zeta_{n_m - 1}} \prod_{j=0}^{K/2} \frac{\exp\left\{-\frac{1}{2}\left(\frac{\mu_j - \xi_j}{\tau_j}\right)^2\right\}}{\sqrt{2\pi}\tau_j} \prod_{j=0}^{K/2} \frac{\beta_j^{\alpha_j} \sigma_j^{-2(1+\alpha_j)} \exp\{-\frac{\beta_j}{\sigma_j^2}\}}{\Gamma(\alpha_j)} \\ & \prod_{i=1}^{n} \prod_{j=0}^{K/2} \left[P(Z_{ij} = 1|\eta, \boldsymbol{I}) \frac{1}{\sqrt{2\pi}\sigma_j} \exp\left\{-\frac{1}{2}\left(\frac{y_i - \mu_j}{\sigma_j}\right)^2\right\}\right]^{Z_{ij}}. \end{aligned} \tag{7}$$

Since the posterior distribution lives in a high-dimensional product space, a MCMC sampler (Gilks et al. 1996) is designed to sample from this space for Bayesian statistical inference. First the chain is started at an arbitrary point $(Q^0, \eta^0, \boldsymbol{\mu}^0, \boldsymbol{\sigma}^0, \boldsymbol{Z}^0)$. Then the five unknowns $Q, \eta, \boldsymbol{\mu}, \boldsymbol{\sigma}$, and $\boldsymbol{Z}$ are updated to yield $(Q^1, \eta^1, \boldsymbol{\mu}^1, \boldsymbol{\sigma}^1, \boldsymbol{Z}^1)$. Repeat this N times. A MCMC sample of size N, $\{(Q^t, \eta^t, \boldsymbol{\mu}^t, \boldsymbol{\sigma}^t, \boldsymbol{Z}^t)\}_{t=0}^N$, can then be obtained by repeating the iteration N times.

At each step, the five unknowns can be updated sequentially or randomly using a hybrid sampler of the Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970) and Gibbs samplers (Geman and Geman 1984). More specifically, with conjugate priors specified, Gibbs sampler can be used to update $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and the progeny QTL dosage $\boldsymbol{Z}$ using their marginal posterior distributions (or full conditionals). The other parameters (*i.e.*, the parental QTL configuration Q and the QTL position η) can be updated using the Metropolis method. However simulation shows the chain converges quickly if the parental QTL configuration is fixed, but when the chain is allowed to visit different parental QTL configurations, it mixes slowly in the parameter space. The slow mixing is caused by a varying dimensionality of the sampling space as the parental QTL dosage changes. For example, suppose the current parental QTL dosage is d, $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are $d+1$ dimensional. Since all the progeny could have at most d doses of QTL, the current $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{Z}$ do not provide information concerning $\boldsymbol{\mu}_{d+1}$ and $\boldsymbol{\sigma}_{d+1}$. Therefore, it is very unlikely for the chain to accept a proposed move to a higher QTL dosage $d+1$. A reversible jump (Green 1995; Richardson and Green 1997) then needs to be incorporated into the MCMC sample in order to improve the mixing efficiency of the chain when it moves between QTL configurations with different QTL dosages.

To differentiate parameter spaces with different dimensions, parental configurations are classified into groups according to the parental QTL dosage level D. For example, for a tetraploid, two configurations with one dose of the QTL $\{Qq, qQ\}$ form one group, and one configuration with two doses of the QTL forms another group $\{QQ\}$. The corresponding QTL configuration indices for the two groups are $\{1, 2\}$, and $\{3\}$, respectively. Without any prior information, the prior distribution for parental QTL dosage D can be taken to be proportional to the group size. In the case of a tetraploid under a pseudo-doubled backcross experiment,

$$p(D = d) = \begin{cases} 2/3 & \text{if } d = 1, \\ 1/3 & \text{if } d = 2. \end{cases}$$

Given a parental QTL dosage $D = d$, the parental QTL configuration Q is randomly selected from the set of parental QTL configurations, which have d doses of the QTL. For a tetraploid, the sampling distribution of Q is

$$p(Q = q | D = d) \stackrel{\triangle}{=} p(q|d) = \begin{cases} 0.5 & \text{if } d = 1 \text{ and } q = 1 \text{ or } 2, \\ 1 & \text{if } d = 2 \text{ and } q = 3. \end{cases}$$

The unconditional prior for QTL configuration is still discrete uniform.

At any stage, the model comprises the following parameters: parental QTL dosage

$D = d$ and configuration $Q = q$, progeny QTL dosages $\boldsymbol{Z}$, QTL position η, mean $\boldsymbol{\mu} = (\mu_0, \mu_1, \cdots, \mu_d)$, and variance $\boldsymbol{\sigma} = (\sigma_0, \sigma_1, \cdots, \sigma_d)$. Assume the current state is $\boldsymbol{s} = (d, q, \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\sigma}), \boldsymbol{u}, \boldsymbol{Z})$, and the proposed state is $\boldsymbol{s}^* = (d^*, q*, \boldsymbol{\theta}^* = (\boldsymbol{\mu}^*, \boldsymbol{\sigma}^*), \boldsymbol{u}^*, \boldsymbol{Z}^*)$, where $\boldsymbol{u}$ and $\boldsymbol{u}^*$ are used to match the dimensionality of the two parameter spaces associated with $\boldsymbol{s}$ and $\boldsymbol{s}^*$. To achieve a higher acceptance rate, the parental QTL dosage is allowed to change only by one. If d is equal to 1, then d^* could remain 1 or increase by 1; or if d is equal to $K/2$, then d^* remains $K/2$ or decrease by 1; otherwise, d^* could be any one of the following $d-1$, d, or $d+1$. The proposal distribution of the parental QTL dosage is denoted by $p(d \to d^*)$,

$$p(d \to d^*) = \begin{cases} 0.5 & \text{if } d^* = d \text{ and } d = 1 \text{ or } K/2, \\ 0.5 & \text{if } d^* = d+1 \text{ and } d = 1, \\ 0.5 & \text{if } d^* = d-1 \text{ and } d = K/2, \\ 1/3 & \text{if } d^* \in \{d, d-1, d+1\} \text{ and } 1 < d < K/2. \end{cases}$$

To implement this MCMC sampler, four possible move types are proposed, where the split and merge steps are typical move types for the reversible jump MCMC (Richardson and Green 1997). Due to the limitation of space, the technical details of the steps will not be presented here, but are referred to Cao (2004).

1. Split – The parental QTL dosage is proposed to increase by one (*i.e.*, propose $d^* = d+1$). A parental QTL configuration q^* is randomly selected from the corresponding parental QTL configuration group. And component means $\boldsymbol{\mu}^*$, variances $\boldsymbol{\sigma^{2}}^*$ of the mixture distribution, as well as the progeny QTL dosages $\boldsymbol{Z}^*$ are proposed.

2. Merge – The parental QTL dosage is proposed to decrease by one (*i.e.*, propose $d^* = d-1$). A parental QTL configuration q^* is randomly selected from the corresponding parental QTL configuration group. And component means $\boldsymbol{\mu}^*$, variances $\boldsymbol{\sigma^{2}}^*$ of the mixture distribution, as well as the progeny QTL dosages $\boldsymbol{Z}^*$ are proposed.

3. Shift – The parental QTL dosage remains the same (*i.e.*, propose $d^* = d$). A parental QTL configuration q^* is randomly selected from same QTL configuration group. And component means $\boldsymbol{\mu}^*$, variances $\boldsymbol{\sigma^{2}}^*$ of the mixture distribution, as well as the progeny QTL dosages $\boldsymbol{Z}^*$ are proposed.

4. Update the QTL location η.

3 Data Analysis

The proposed Bayesian polyploid QTL mapping method is employed to identify the putative QTL associated with winter hardiness for tetraploid alfalfa (Brouwer and Osborn 1999; Brouwer et al. 2000). Alfalfa is a tetraploid ($K = 4$) that undergoes bivalent random pairing

(Bingham and McCoy 1988; Cao et al. 2004). Two genotypes, B17 and P13, representing the extremes for each trait are crossed, and a single F_1 plant is crossed to each parent to create two populations of 101 individuals each. Each population was scored for 82 single-dose restriction fragment (SDRF) loci. The composite map and the original analysis of the trait data and marker data can be found in Brouwer *et al.* (1999, 2000). Based on the genetic map, seven linkage groups are identified and corresponded to seven out of the eight chromosomes. Because only B17 markers are considered for the P13 x F_1 backcross population, and similarly only P13 markers for the B17 x F_1 backcross, each backcross population is equivalent to a pseudo-doubled backcross population.

In the experiment, multiple quantitative traits were measured as related traits of the complex trait winter hardiness. The quantitative trait fall growth, measured by height of vertical regrowth in centimeters early October 1995, of the B17 backcross population is chosen for analysis to demonstrate the proposed Bayesian method, and this trait will be referred to as FG95. The proposed Bayesian method is applied to one interval at a time. In each interval, a MCMC sampler of $100,000$ cycles is run and is sampled every 10 cycles, with $10,000$ initial burn-in. This working set of $10,000$ states is then used to estimate the posterior quantities.

In each interval, let $\{\boldsymbol{s}^{(t)} = (Q^{(t)}, \eta^{(t)}, \boldsymbol{\mu}^{(t)}, \boldsymbol{\sigma}^{(t)}, \boldsymbol{Z}^{(t)})\}_{t=1}^{N}, N = 10,000$ denote the $10,000$ MCMC sample. The marginal posterior distribution of the QTL location, η, can be estimated from $\{\eta^{(t)}\}_{t=1}^{N}$. This way, the $\{\eta^{(t)}\}_{t=1}^{N}$ are treated as if they are sampled from the same distribution. However, considering the fact that these sample points are actually from different parental QTL configurations (*i.e.*, Qq, qQ, and QQ), and the distribution of η may be different for different underlying parental configurations, the marginal distribution of the QTL location η is estimated conditioning on the underlying parental QTL configuration. Therefore, in each interval, three density curves are estimated for the three possible parental QTL configurations Qq, qQ, QQ (Figure 3-(a)). The total area under the three curves is one. The estimated interval-wise posterior probability for each parental QTL configuration is shown in Figure 3-(b), where the height of the bar under the index of a parental QTL configuration ($i.e.$, $1 = Qq, 2 = qQ, 3 = QQ$) represents the magnitude of the corresponding posterior probability .

If the QTL is located in an interval, the vicinity of the QTL location will be visited more frequently than the other area. On the other hand, if there is no QTL in this interval, all the locations within the interval are equally likely to be visited. As a result, the posterior distribution of the QTL location in this interval should be uniform. Therefore, the putative QTL can be identified by locating the peaks in the posterior distribution of the QTL location within each interval. Furthermore, the Kolmorogov goodness-of-fit test can be used to measure the deviation between the estimated posterior distribution with a uniform distribution. Let the significance level be $\alpha = 0.05$, then the Bonferroni correction is $\alpha_B = 0.05/47 = 0.0011$ with forty-seven intervals in the genome. For the quantitative trait fall growth measured in 1995 (FG95), six peaks under the parental QTL configuration QQ

are selected based on the Bonferroni correction, which are in the first interval on chromosome 1, the sixth interval on chromosome 2, the third, fifth, and seventh interval on chromosome 5, and the first interval on chromosome 7. Similarly, for the quantitative trait fall growth measured in 1996 (FG96), three peaks under parental QTL configuration QQ are selected, which are in the first interval on chromosome 1, and the first and the fourth interval on chromosome 7.

For each peak, one can examine the posterior probability of the associated parental QTL configuration and estimate the QTL location and effects using the corresponding MCMC sample. Let us illustrate this using the peak under the parental QTL configuration QQ in the first interval on chromosome 1 for the quantitative trait fall growth measured in 1995 (FG95) (Figure 3). Among the $10,000$ recorded MCMC sample points in this interval, the parental configuration QQ is visited 9995 times, which yields an estimated posterior probability of QQ being 0.9995. Based on the $9,995$ sample points of the relative QTL location, the the posterior mode is 0.86, then given the length of this interval 0.348M, the estimate QTL location is $\hat{\eta} = 0.859 \times 0.358 = 0.308$M. With the corresponding $9,995$ MCMC sample points of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, the posterior modes of these two vectors are $\hat{\boldsymbol{\mu}} = (12.9, 21.5, 26.4)$ and $\hat{\boldsymbol{\sigma}} = (2.9, 2.2, 1.1)$, which are taken as the Bayesian estimates. The estimated putative QTL positions and effects corresponding to the peaks selected for the quantitative trait fall growth measured in 1995 (FG95) are listed in Table 1.

4 Summary

Due to the complex nature of the posterior distribution for QTL mapping for polyploids, a hybrid MCMC sampler consisting of the reversible jump MCMC sampler, the MH sampler, and the Gibbs sampler has been employed to sample from the posterior distribution for Bayesian statistical inference. The approach taken to analyze the experimental alfalfa data (Brouwer et al. 2000) is to apply the MCMC sampler locally to each interval on the genetic map. Posterior distributions of the parental QTL configuration, QTL location, and QTL effects are then obtained based on the MCMC sample. This approach has shown to be effective in analyzing the experimental alfalfa data.

Just as the parental QTL configuration plays its role in affecting the estimated QTL location and effects, parental marker configuration and intermarker recombination fractions are also important factors that need to be considered when the accuracy of the final QTL mapping results are assessed. Taking a Bayesian approach allows one to model the uncertainty in the estimated genetic map, and to incorporate this uncertainty into the framework of QTL mapping. Here, the Bayesian QTL mapping approach for polyploids is presented based on the assumption that a specific estimated genetic map is given, which provides an estimated marker order, intermarker genetic distances, and the associated estimated parental marker configuration. Building a Bayesian model for genetic map estimation will be future work with the final goal being a combination of these two pieces in assessing the accuracy of the estimated QTL location as well QTL effects.

Acknowledgment We thank Dr. Tom Osborn for providing the data set.

References

Bayes, T. (1783). An essay towards solving a problem in the doctring of chances. *Philosophical Transactions of the Royal Society of London 53*, 370–418.

Bingham, E. T. and T. J. McCoy (1988). Cytology and cytogenetics of alfalfa. In A. A. Hanson, D. K. Barnes, and R. R. Hill, Jr. (Eds.), *Alfalfa and alfalfa improvement*, pp. 737–776. American Society of Agronomy, Madison, Wisconsin.

Brouwer, D. and T. Osborn (1999). A molecular marker linkage map of tetraploid alfalfa. *Theoretical and Applied Genetics 99*, 1194–1200.

Brouwer, D. J., S. H. Duke, and T. C. Osborn (2000). Mapping genetic factors associated with winter hardiness, fall growth, and freezing injury in autotetraploid alfalfa. *Crop Science 40*, 1387–1396.

Cao, D. (2004). *Quantitative Trait Locus Analysis in Polyploids*. Ph. D. thesis, Purdue University.

Cao, D., B. A. Craig, and R. W. Doerge (2003). Interval mapping for autopolyploids. Proceedings of the 15th Annual Kansas State University Conference on Applied Statistics in Agriculture. Kansas State University.

Cao, D., T. C. Osborn, and R. W. Doerge (2004). Correct estimation of preferential chromosome pairing in autotetraploids. *Genome Research 14*, 459–462.

Doerge, R. W. (2002). Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics 3*, 43–52.

Doerge, R. W. and B. A. Craig (2000). Model selection for quantitative trait locus analysis in polyploids. *Proceedings of the National Acadmy of Sciences 97*(14), 7951–7956.

Geman, S. and D. Geman (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence 6*, 721–741.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (1996). *Markov chain Monte Carlo in Practice* (First ed.). Chapman and Hall/CRC.

Grattapaglia, D. and R. Sederoff (1994). Genetic linkage maps of *eucalyptus grandis* and *eucalyptus urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. *Genetics 137*, 1121–1137.

Green, P. J. (1995, December). Reversible jump Markov chain Monte Carlo computation and bayesian model determination. *Biometrika 82*(4), 711–732.

Hackett, C. A., J. E. Bradshaw, and J. W. McNicol (2001). Interval mapping of quantitative trait loci in autotetraploid species. *Genetics 159*, 1819–1932.

Hastings, W. K. (1970, Apr.). Monte Carlo sampling methods using markov chain and their applications. *Biometrika 57*(1), 97–109.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys. 21*, 1087–1092.

Richardson, S. and P. J. Green (1997). On bayesian analysis of mixtures with an unknown number of components. *Journal of Royal Statistical Society B 59*(4), 731–792.

Rieseberg, L. H. and M. F. Doyle (1989). Tetrasomic segregation in the naturally occurring autotetraploid *allium nevii* (aliaceae). *Hereditas 111*, 31–36.

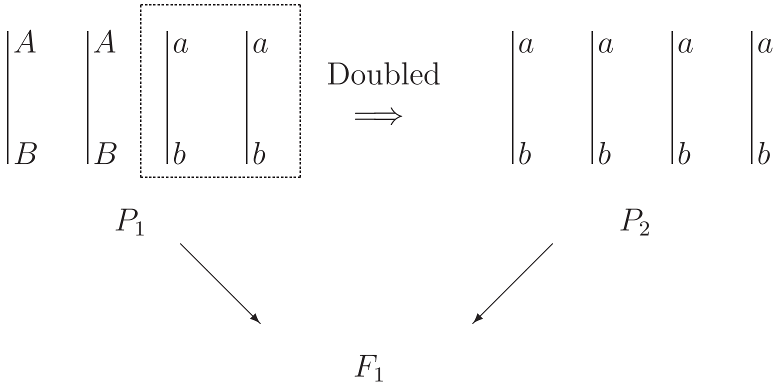

Figure 1: A pseudo-doubled backcross experiment for a tetraploid with two loci. Each locus is assumed to be biallelic: A/a and B/b, where A and a denote two alleles of the first locus and A is the dominant allele, and similarly for B and b. P_1 is the informative parent with two dose of each dominant allele. P_2 is the non-informative doubled-haploid produced by doubling a haploid of P_1, which only contains recessive alleles.

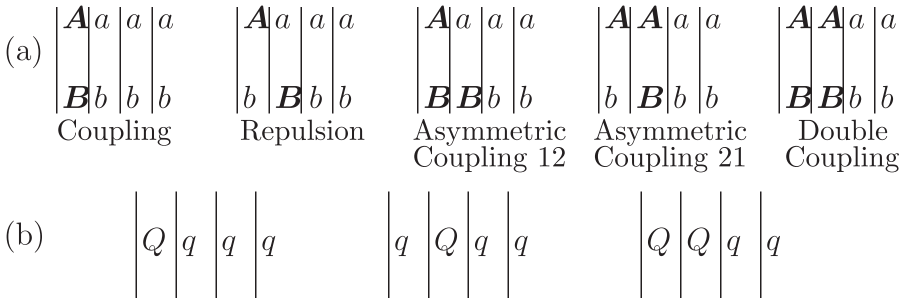

Figure 2: (a) Parental marker configuration for a tetraploid under a pseudo-doubled backcross experiment with two marker loci A/a and B/b. (b) Parental QTL configuration for a tetraploid under a pseudo-doubled backcross experiment with one QTL locus Q/q.

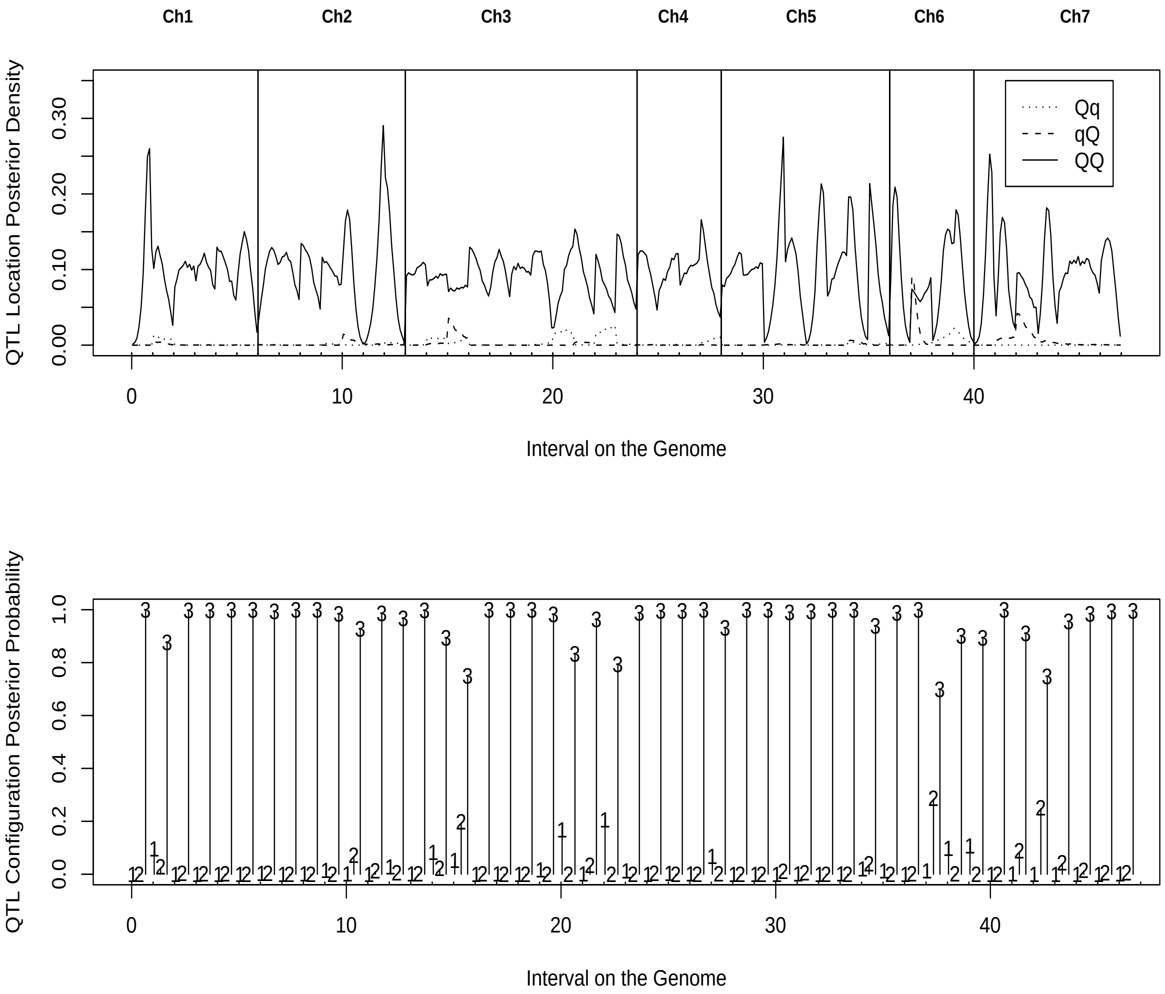

Figure 3: (a). Interval-wise posterior density of the relative location of the QTL affecting alfalfa fall growth measured in 1995 (FG95). The vertical bars separate the linkage groups. (b). Estimated marginal posterior distribution of the parental QTL configuration for alfalfa fall growth (FG) measured in 1995. The numbers denote the binary score of the parental QTL configurations: $1 = Qq, 2 = qQ, 3 = QQ$. In both plots, each interval on the genome is expressed by an $[0, 1]$ interval, and the actual position of the QTL is the product of the relative position and the total length of the corresponding interval.

Table 1: Estimated locations and effects of QTL associated with fall growth measured in 1995 (FG95) for tetraploid alfalfa using the Bayesian method.

Year	[a]Ch	[b]Pos	[c]Qconfig([d]Prob)	[e]$\hat{\mu}$	[f]$\hat{\sigma}$	[g]p-value
95	1	30.8	QQ (0.99)	12.9,21.5,26.4	2.9,2.2,1.1	0.0007
	2	62.0	QQ (0.98)	12.3,22.4,21.4	4.5,2.5,2.3	0.0006
	5	35.7	QQ (0.99)	11.3,21.9,21.6	2.0,3.0,1.4	0.0002
	5	48.0	QQ (0.99)	12.6,21.5,25.7	2.7,2.1,1.1	0.0007
	5	60.2	QQ (0.94)	12.8,21.3,25.9	2.7,2.2,1.1	0.0013
	7	6.60	QQ (1.00)	11.0,21.4,26.1	1.9,2.3,1.3	0.0003

[a] Ch denotes the chromosome number.
[b] Pos denotes the putative QTL positon (cM) in the composite map.
[c] Qconfig denotes the estimated parental QTL configuration.
[d] Prob denotes the posterior probability of the parental QTL configuration.
[e] $\hat{\mu} = (\mu_0, \mu_1, \cdots, \mu_d)$ is the estimated mean vector.
[f] $\hat{\sigma} = (\sigma_0, \sigma_1, \cdots, \sigma_d)$ is the estimated standard deviation vector.
[g] The p-value for the Kolmogorov goodness-of-fit test.