Mapping QTLs for Grain Yield in Maize: A Bayesian Approach Using Reversible Jump MCMC

Elisabeth Regina de Toledo *
UFV; LEA/UFMS, Corumbá, Brazil – erdtoledo@gmail.com
Antônio Policarpo Souza Carneiro
UFV, Viçosa, Brazil – policarpo@ufv.br
Cláudio Lopes de Souza Junior
ESALQ/USP, Piracicaba, Brazil – clsouza@esalq.usp.br
Anete Pereira de Souza
UNICAMP, Campinas, Brazil – anete@unicamp.br

Abstract

Many traits that are important to agriculture, human health and evolutionary biology, in its majority, can be classified as quantitative, whose phenotypic expression presents continuous variation, attributed to the simultaneous segregation of many genes, in definite regions as QTLs (Quantitative Trait Loci). Efficient and robust identification and mapping onto genomic positions of those genes is a very important goal in quantitative genetics. Mapping QTL is identify its position in the genoma and to estimate its effect. The availability of genome-wide molecular markers provides the means for us to locate and map those QTL in a systematic way. Since the publication of Lander and Botstein (1989), that first proposed interval mapping method for a genome-wide scan of QTL, many statistical methods have been proposed and developed to map QTL , great part of them present Classical boarding. This work will present Bayesian approach. Adopting a Bayesian approach a model is fit to quantitative trait and molecular marker data, instead of fitting to locus. The output of a Bayesian analysis is a posterior distribution on the parameters, fully incorporating prior beliefs and parameter uncertainty. Reversible Jump MCMC (RJMCMC) is used in this work. Bayes Factor is used to estimate the number of QTL. Inference summaries for the locations of the QTL and their effects are derived from the corresponding marginal posterior densities. Through Bayesian interval, significant associations between markers and QTLs were obtained in four chromosomes and five QTLs has been mapped, which explained 11.23% of genetic variation. Most alleles that contributed positively in trait came from parental strain L02-02D. The latter had the highest yield rate. We examine grain yield in maize to illustrate the applied method.

Keywords: Quantitative Trait Loci; Bayesian Approach; Reversible Jump MCMC; Bayes Factor

1. Introduction

Grain yield and its components in maize are controlled by many loci and present high interaction with environments. Until recently inheritance studies of these traits used statistical-genetic models based on the net effects of the segregating loci in the populations. With the advent of molecular markers and new statistical-genetic models, well-satured genetic maps could be developed allowing the mapping of the loci (QTLs) that control these traits Bento (2006). Thus, the number of loci, their genomic position, and the genetic effects of individual QTLs could be estimated.
Mapping QTL is identify its position in the genoma and esteem its effect. Some methods of mapping exist, great part of them present Classical boarding. The objective of this work is to map QTLs for grain yield in a tropical maize population via Bayesian Interval Mapping (BIM).

2. Material and Methods

The data set used in the work of grain yield data, obtained from the evaluation of 256 maize progenies $F_{2:3}$. Both lines were developed in the Department of Genetics, ESALQ/USP. Genotypic data of molecular markers for all progenies used in the experiment were obtained from the Laboratory of Molecular Biology and Genetic Analysis, located at the Molecular Biology and Genetic

Engineering Center (CBMEG), the State University of Campinas/UNICAMP. (LIMA *et al.*, 2006; BENTO, 2006). The experimental design was lattice 16 *x* 16 with two replications each.

The QTLs were mapped based on the adjusted means across the environment by the Bayesian Interval Mapping Method (BIM) proposed by Satagopan *et al.* (1996), using the QTLCartographer 2.5 version software for Windows. The experimental module Bmapqtl was used for implements Bayesian analysis in this work.

Bayesian Interval Mapping for controlled experiment provides a nice complement to the classical analysis for mapping QTLs. Considers the same statistical model as multiple interval mapping, which is an extension of composite interval mapping Zeng (1993); Zeng (1994). The model:

$$y_i = \mu + \sum_{j=1}^{P} a_j x_{ij}^* + \sum_{j=1}^{P} d_j z_{ij}^* + e_i$$

where y_i is observed phenotype for the *i-th* individual in a sample of size *n* (i = 1, …, *n*), μ is the grand mean of the trait, $e_i \sim N(0,\sigma^2)$ is a error random, a_j e d_j , denote the additive and dominance effects of the p loci, respectively. j; (j = 1, …, p). The *j-th* QTL contributes for individual i, (Q_{ij}) are defined by the Dummy variable x_{ij}^* e z_{ij}^* . The indicator variable x_{ij}^* is associate to additive effect, assume 1 with one homozygous parent type, -1 with the other homozygous parent type for *qq* e *QQ*, respectively; z_{ij}^* is associate to dominance effect, assume and 0 with the heterozygote *Qq*; $z_{ij} = 1 - |x_{ij}^*|$ is equal to 1 if $x_{ij}^*{}_{ij} = 0$ and $x^*{}_{ij} = \pm 1$, if $z^*{}_{ij} = 0$. $z^*{}_{ij} = 0$. Equivalent as:

$$\mathbf{y} = \mathbf{X}^{(p)}\Theta^{(p)} + \boldsymbol{\varepsilon} \qquad (1)$$

$\mathbf{y}$ is a vector of trait values for the *n* individuals; $\theta^{(p)} = [\mu, \sigma^2, a_1, d_1, a_2, d_2, ..., a_p, d_p]^T$ is a vector of unknow parameters; denote the effects; a_j e d_j are the additive and dominance effects for each QTL, with $2p+1$ colluns (if the number of QTLs is random, the length of vector $\theta^{(p)}$ is random too)**;** $\boldsymbol{\varepsilon}$ is the error vector, with normal distribution, $\boldsymbol{\varepsilon} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}\sigma^2)$**;** $\mathbf{X}^{(p)} = [1 \quad x_1 z_1 \quad x_2 z_2 \quad \dots \quad x_p z_p]$, is the design matrix which includes the QTL genotype information (*n* x 2*p*+1), where $x_j = [x_{1j}, x_{2j}, \ldots, x_{nj}]^T$ and $z_j = [z_{1j}, z_{2j}, \ldots, z_{nj}]^T$ are indicator variable vectors for the genotypes of the QTL putative constructed using flanking markers. Gaffney (2001); Satagopan *et al.*, (1996).

The number of QTLs, and hence the length of the vector of additive and dominance effects, may be either fixed or random. If the number of QTL is unknown, it is sampled are slightly more complicated, since changing the number of QTL involves switching between models-adding or dropping other model parameters for additive and/or dominance effects and QTL genotypes. We employ the Reversible Jump MCMC Green (1995); Richardson & Green (1997). In this work was mapping QTL for grain yield with number (*P*) of QTL random. Bayes Factor is used to compare the models and obtain as estimate for the number of QTL affecting the trait. We employ the "Reversible Jump MCMC", RJMCMC Green (1995); Sorensen; Gianola (2002) which extends the MCMC approach to such situations. In standard MCMC, the dimension of the parameter space is fixed. RJMCMC allows the dimension of the parameter space to vary and the chain can move between different models.

Bmapqtl currently uses a control file to set additional parameters that are used by the RJMCMC sampling. Collects 400,000 RJMCMC samples, saving every 100th value (thin), after a burnin of 2,000 samples, with a pre-burnin to overfit the model of 1,000 samples, was obtained a final

random sample with length 15,315. As part of our Bayesian analysis, there is a probability distribution over the genetic parameters. RJMCMC is used to study the resulting joint posterior of the parameters given the data, and estimated marginal posteriors are the basis for QTL estimates. Initially, there are 0 QTL. The prior for the grand mean is normal distribution, N (0; $0.5s^2$), while the prior for the variance is inverse gamma distribution, IG (3; s^2). Priors for any additive or dominance effects are both set to Beta (2; 10). The prior for the number of QTL is Poisson (3).

The Bayes Factor (BF) is used to estimate the number of QTL, defined as

$$BF(M_i, M_j) = \frac{\left(\frac{P(M_i \mid y)}{P(M_j \mid y)}\right)}{\left(\frac{P(M_i)}{P(M_j)}\right)},$$

$P(M_i \mid y)$ e $P(M_j \mid y)$ are conditionals posteriors probabilities conditionals to observations for models, $P(M_i)$ e $P(M_j)$ are probabilities priors to models with *j* and *j+1* QTLs, i.e., M_j e $M_{(j-1)}$ M_i e M_j . Table 1 present interpretation of Bayes Factor and decisions about evidence for M_j with respect $M_{(j-1)}$ (RAFTERY, 1995):

TABLE 1 - Interpretation of Bayes Factor and decisions about evidence for M_j with respect $M_{(j-1)}$, as given Jeffreys (1935)

Values of $B(M_j, M_{(j-1)})$	**Evidence in favor of model M_j**
$1 \le B(M_j, M_{(j+1)}) \le 3$	Not worth more than a bare mention
$3 < B(M_j, M_{(j+1)}) \le 10$	Substantial
$10 < B(M_j, M_{(j+1)}) \le 100$	Strong
$B(M_j, M_{(j+1)}) > 100$	Decisive

Raftery (1995).

3. Results and Discussion

Calculated the Bayes Factor, to estimate BF favored a five QTL model was fitting to data. Hence we infer that the five QTL model (M_5) is appropriate. Were calculated its position (λ) and genetics effects and were presents 95% credibility intervals for them. Next, were calculated its position (λ) on chromosome chosen and estimated the effects of each QTL: additive effect (*a*) and dominance (*d*) and allelic level of dominance (LD). Table 2 shows the summary of the results. The names of QTLs are comprised of code indicating the trait followed by a number indicating the chromosome is QTL putative; letter and number identifying QTLs located on the same chromosome. For example, the qprod5 is QTL for grain yield trait located on chromosome five. The letters "a", "b" identify the various QTLs located on the same chromosome.

The sign of the estimates was used to identify the favorable alleles contributed by each parent. For QTL main effects, positive and negative signs of the estimates indicate that L20-01F and L20-01F, respectively, contributed towards higher value alleles for the trait. Average levels of dominance of each QTL and across all QTL were calculated as the $|d|/|a|$ ratio, with average level of dominance was partial dominance estimated for the F_2 population. Most of the alleles that contribute favorably to the character comes from the L02-03D parental strain, with the highest yield. The contribution through favorable alleles for this strain was effective in four of the five mapped QTLs. The QTLs detected by BIM located in the corn genome can be seen from Figure 1. The $F_{2:3}$ evaluated showed average production of 43.84 g plant^{-1}. The coefficient of heritability ($\hat{h}^2$) was 15%, indicating low heritability for the trait. For grain yeld, 5 QTLs were detected on chromosomes 1 (one QTL), 3 (one QTL), 5 (two QTLs) and 9 (one QTL) (Fig. 1, 2 and Table 2). These 5 QTLs together explain 1.66 % of the phenotypic variation and explain 11.23% of the genetic variation. The proportional phenotypic

variation ($\hat{R}^2_{F*}$) explained for QTLs accounted for 0.08%–0.6 %. The proportional genetic variation ($\hat{R}^2_{G*}$) explained for QTLs accounted for 0.53%–4.04 %. Table 3 presents 95% credibility intervals for position and genetics effects additive (*a*) and dominance (*d*).

TABLE 2 - Chromosomic location. effects and types of individual gene action of QTLs for grain yield.

	QTL	Position		Genetic Efects						
QTL	Chro[a]	cM	Marker Interval	*a*	*d*	$\lvert\hat{d}\rvert/\lvert\hat{a}\rvert$	LD[b]	Direction	$\hat{R}^2_{F*}$(%) [c]	$\hat{R}^2_{G*}$(%) [d]
qprod1a	1	106.02	U1917-U1558	0.51	0.19	0.38	PD	L02-03D	0.19	1.31
qprod3a	3	107.34	P073-B197	0.81	0.09	0.11	A	L02-03D	0.47	3.15
qprod5a	5	75.52	M0282-U1524	0.66	0.25	0.37	PD	L02-03D	0.33	2.20
qprod5b	5	98.05	U1524-U2013	0.92	-0.16	0.18	A	L02-03D	0.60	4.04
qprod9a	9	64.84	P065-B1714	-0.23	0.34	1.45	OD	L20-01F	0.08	0.53
					$\hat{ALD}$[e]	0.28	DP	R^2Total	1.66	11.23

[a]Chromosomic location of the QTL putative; [b]LD (level of dominance): A= Additive ($\lvert\hat{d}\rvert/\lvert\hat{a}\rvert<0.2$). PD= partial dominance ($0.2<\lvert\hat{d}\rvert/\lvert\hat{a}\rvert<0.8$). D= dominance ($0.8<\lvert\hat{d}\rvert/\lvert\hat{a}\rvert<1.2$). OD= overdominance ($\lvert\hat{d}\rvert/\lvert\hat{a}\rvert>1.2$); [c] $\hat{R}^2_{F*}$(%): proportion of phenotypic variance explained for QTL; [d] $\hat{R}^2_{G*}$(%): Total proportion of phenotypic variance explained for QTL; [e] $\hat{ALD}$ (average level of dominance) $\hat{ALD}=\frac{\sum LD_i\hat{R}^2_{G_i}}{\sum\hat{R}^2_{G_i}}$.

TABLE 3 - Credibility intervals for position (λ) and genetics effects. (*a*) e (*d*) of QTLs for grain yield

		Effects	
QTL	λ	*a*	*d*
qprod1a	(12.56; 276.47)	(-1.95; 3.10)	(-1.61; 2.17)
qprod3a	(11.028; 275.27)	(-2.10; 3.70)	(-1.93; 2.36)
qprod5a	(4.61; 153.48)	(-2.36; 3.84)	(-1.93; 2.56)
qprod5b	(3.64; 240.11)	(-2.21; 3.76)	(-2.64; 1.96)
qprod9a	(2.95; 141.29)	(-2.64; 3.37)	(-2.03; 2.73)

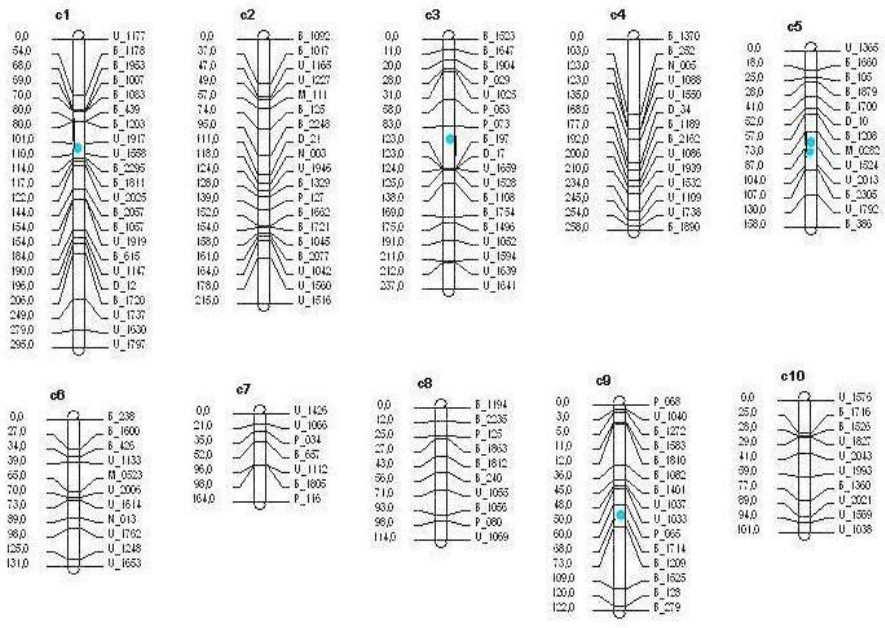

FIGURE 1 - Genetic map of 10 chromosomes in a population of tropical maize with 139 markers showing the location of the five QTLs associated with grain yield obtained for Bayesian Interval Mapping analysis. indicated with circles Toledo *et al.* (2008).

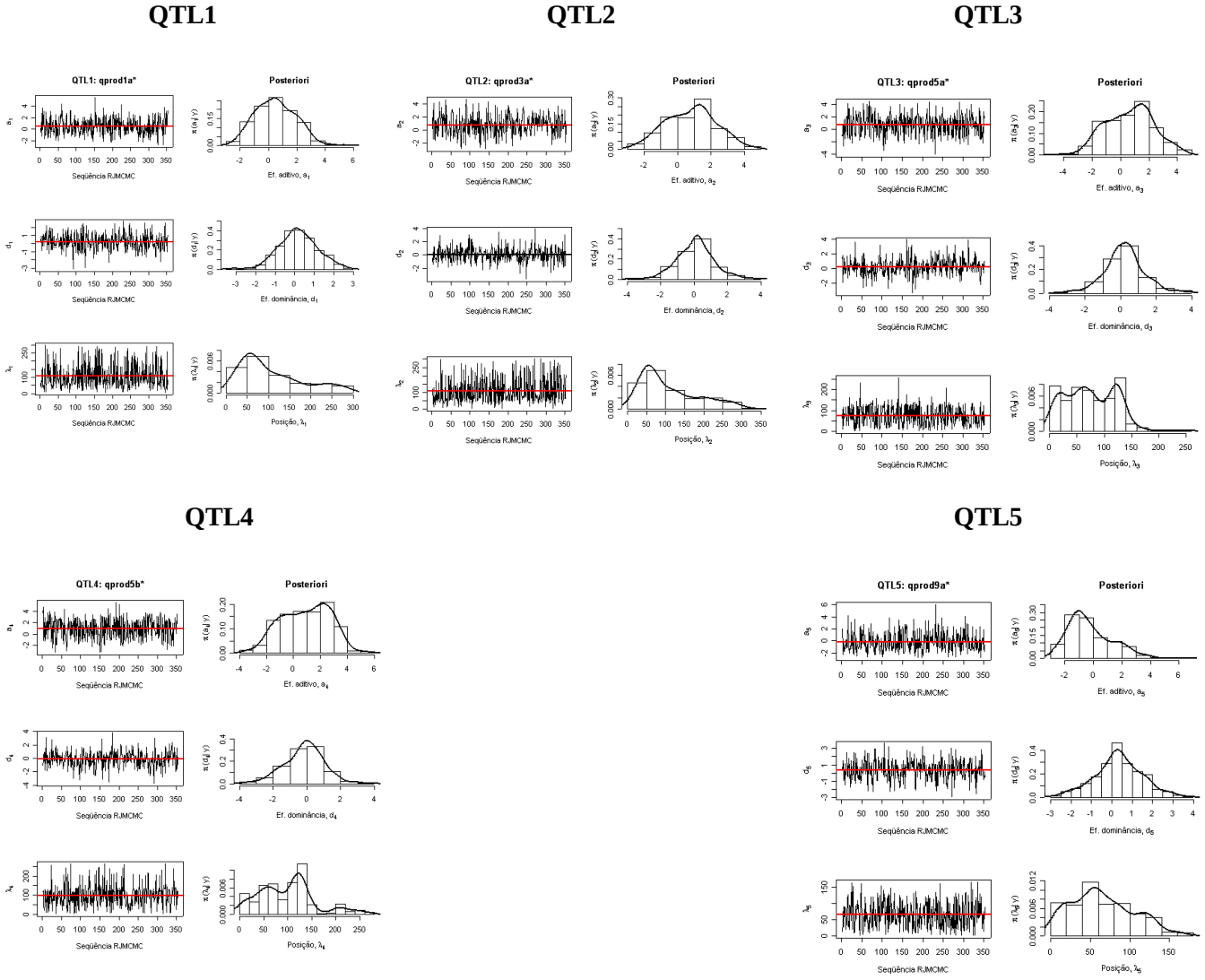

FIGURE 2 - RJMCMC Chains and Posteriors Densities for additive and dominance effects of the locus and location putative QTL.

5. Conclusions

A total of five QTLs were identified for grain yield: 1 QTL on chromosomes one. three and nine; 2 QTLs on chromosome five.

A weakness observed in the Bayesian methodology interval mapping (BIM) was regarding the specification of priors. The output of a Bayesian analysis is a posterior distribution on the parameters, fully incorporating prior beliefs and parameter uncertainty via RJMCMC. When making the mapping of QTLs. through sensitivity analysis where a priori densities have changed the parameters or in distributions. it was found that the parameter estimates were significantly affected. We have seen above that the resulting modification in the prior can be quite substantial in terms of the prior mean. Changes in the prior can affect the posterior distribution for the number of QTL. Were observed that the variance effect can affect the Bayes Factors too. The prior for the effect (additive or dominance) may depend on the number of QTL. It seems intuitive that some QTL may account for a higher percentage of the variance than will others Gaffney (2001). Should therefore be cautious when specifying a priori densities.

Acknowledgements

We would like to thanks Departamento de Genética. Escola Superior de Agricultura ‘‘Luiz de Queiroz’’. Universidade de São Paulo (ESALQ/USP) for providing the data analyzed in this work. Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG). Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) for financial support .

References

Bento. D. *Mapeamento de QTLs para produção de grãos e seus componentes em uma população de milho tropical*. Piracicaba. Dissertation (Doctor of Genética e Melhoramento de Plantas). Universidade de São Paulo. Escola Superior de Agricultura “Luiz de Queiroz”.

Gaffney. P.J. *An efficient Reversible Jump Markov Chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses. Madison*. Dissertation (Doctor of Philosophy in Statistics). University Of Wisconsin. 2001.

Green. P. J Reversible Jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika* 82. 711-732. 1995.

Lima. M.L.A; Souza Junior. C.L; Bento. D.A.V; Souza. A.P; Carlini-Garcia. L.A.. Mapping QTL for Grain Yeld and Plant Traits in a Tropical Maize Population. *Molecular Breeding* 17. 227-239. 2006.

Raftery. A. E. Bayesian model selection in social research. Sociological Methodology. Washington. v.25. p.111-163. 1995.

Satagopan. J.M..Yandell. B.S.;Newton. M. A.; Osborn. T.C. A Bayesian Approach To Detece Quantitative Trait Loci Using Markov Chain Monte Carlo. *Genetics* 805-816. 1996.

Sillanpää. M. J.; Arjas. E. Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data. *Genetics* 148. 1373-1388.

Sorensen. D.; Gianola. D. *Likelihood. Bayesian. and MCMC Methods in Quantitative Genetics*. ed. Springer-Verlag. USA. 2002.

Toledo. E.R.; Leandro. R.A.; Souza Junior. C.L.; Souza. A.P.; Mapeamento de QTLs: Uma Abordagem Bayesiana. *Revista Brasileira de Biometria* 26. n. 2. 107-114. 2008.