Genetics and Molecular Research

A new method of QTL identification for undersaturated maps

A.K.A. Pamplona[1], M. Balestre[1], L.A.C. Lara[2], J.B. Santos[3] and J.S.S. Bueno Filho[1]

[1]Departamento de Ciências Exatas, Universidade Federal de Lavras, Lavras, MG, Brasil
[2]Departmento de Genética, Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, SP, Brasil
[3]Departamento de Biologia, Universidade Federal de Lavras, Lavras, MG, Brasil

Corresponding author: J.S.S. Bueno Filho
E-mail: jssbueno@dex.ufla.br

Genet. Mol. Res. 14 (3): 11462-11479 (2015)
Received February 6, 2015
Accepted May 22, 2015
Published September 25, 2015
DOI http://dx.doi.org/10.4238/2015.September.25.13

ABSTRACT. In many species, low levels of polymorphism prevent the assembly of linkage maps that are used to identify genetic markers related to the expression of quantitative trait loci (QTLs). This study compared two methods of locating QTLs in association studies that do not require a previous estimation of linkage maps. Method I (MI) was a Bayesian multiple marker regression and Method II (MII) combined multiple QTL mapping and "moving away from markers". In this method, markers are not directly regressed to the phenotype, but are used as pivots to search for QTLs along the genome. To compare methods, we simulated 300 individuals from an F_2 progeny with two levels of marker loss (20 and 80%). A total of 165 markers and seven QTLs were spread along 11 chromosomes (roughly emulating the genetic structure of the common bean, *Phaseolus vulgaris*). A real data example with 186 progenies of a $F_{2:4}$ generation of the species was analyzed using 59 markers (17 simple sequence repeats, 31 amplified fragment length polymorphisms, and 11 sequence-related amplified polymorphisms). MII was more precise than MI for both levels of marker loss. For real data, MII detected 17 candidate positions for QTLs,

whereas MI did not detect any. MII is a powerful method that requires further studies with actual data and other designs such as crossover, and genome-wide studies.

Key words: Bayesian regression; QTL analysis; Multiple marker; Genome-wide

INTRODUCTION

The detailed genetic study of the phenotypic expression of quantitative traits is of major interest to geneticists and breeders. Genetic variation in these traits is thought to be controlled by the simultaneous segregation of many genes distributed along the genome, in regions known as quantitative trait loci (QTLs), which are responsible for phenotypic expression.

By using dense genetic maps, it is possible to determine the number of QTLs and quantify their effects and their distribution in the genome. Several methods of QTL mapping exist, which differ in computational requirements, statistical efficiency, type of information extracted, flexibility to deal with different data structures, and ability to map multiple QTLs. The best-known methods are single-marker mapping (Edwards et al., 1987), single-interval mapping (Lander and Botstein, 1989), composite-interval mapping (Jansen, 1993; Zeng, 1993, 1994), multiple-interval mapping (Kao et al., 1999), multiple-marker mapping (Xu, 2003), and multiple-QTL mapping (Wang et al., 2005).

Some of these methods require linkage maps, which are more precise the more saturated they are. However, due to the cost of laboratory techniques, greenhouse space, field plots, marker scoring, and data entry, the question of genome coverage arises (Doerge et al., 1997). Owing to low levels of polymorphism for some characteristics, particularly for mating designs that are widely used in several species, such as beans, the genome is poorly represented. This results in linkage map construction to be inaccurate and sometimes impractical. It is possible to use consensus maps, but this approach is useful only if markers are spread along all of the chromosomes. Therefore, it is necessary to find an approach to search for QTLs along the genome that does not require mapping. One option would be to use multiple-marker regression in association analysis. If the genome is highly saturated, the regression is asymptotically efficient (Xu, 2003), but may be biased.

A mapping method that does not use linkage maps was proposed by Xu (2003), called multiple-marker mapping, which applies a Bayesian shrinkage regression method to simultaneously evaluate the effects of QTLs associated with markers along the genome. It is able to handle situations where the number of estimated parameters is larger than the number of observations. In this approach, each marker is treated as a putative QTL; consequently, the incidence matrix is fully determined and conditioned on genotypes. QTLs with small effects and low variances have their effects shrunk to zero, and QTLs with large effects and high variances are penalized less (Balestre et al., 2012).

According to Wang et al. (2005), the uncertainty of QTL genotypes further complicates QTL mapping, and thus, the incidence matrix is not observed. Furthermore, it is of interest to examine QTL positions. Based on this, the authors suggested extending the Bayesian shrinkage estimation by Xu (2003) to map QTLs where the QTL positions and effects are the parameters to infer. This method, called multiple-QTL mapping, assumes that each interval defined by adjacent markers has a QTL.

Doerge et al. (1997) reviewed the statistics applied to the use of molecular markers and quantitative genetics in QTL searches. One of the analyses they describe is a variation of single-marker regression, where the marker is not the putative QTL. This technique was later called

"moving away from the marker" (MAFM) (Wu et al., 2007). This method consists of sequential tests of the hypothesis that the marker is not associated with a QTL (the frequency of recombination between them is 0.5). In the present study, this technique is extended by adopting Bayesian analysis and adapting multiple-QTL mapping (Wang et al., 2005) and multiple-marker mapping (Xu, 2003).

This study was conducted to find a method of Bayesian analysis for the MAFM technique that does not require a linkage map, to compare the results with those obtained by multiple-marker mapping (Xu, 2003), to evaluate undersaturation levels of the genome, and to compare both techniques in a real data example.

MATERIAL AND METHODS

Simulated data

An F_2 population of 300 individuals with a heritability 0.5 was simulated using the QGene program (Joehanes and Nelson, 2008). The genome generated consisted of 11 chromosomes that were each 120 cM long with 165 single nucleotide polymorphism markers at 10-cM intervals. Seven QTLs were spread along the genome with positions and effects that are listed in Table 1.

Table 1. Positions and additive and dominance effects of simulated quantitative trait loci (QTLs) in chromosomes.

	Chromosome	Position (cM)	Additive effect	Dominance effect
QTL 1	1	76.6	-10.0	25.0
QTL 2	1	102.1	5.0	20.0
QTL 3	5	22.3	-3.0	9.0
QTL 4	7	18.7	-7.0	5.0
QTL 5	7	96.5	15.0	-3.0
QTL 6	8	50.6	10.0	8.0
QTL 7	9	52.8	20.0	5.0

Real data

Real data from QTL searches for resistance to white mold disease (*Sclerotinia sclerotiorum*) in the common bean were obtained from Lara et al. (2014). In all, 186 progenies of a $F_{2:4}$ common bean population were obtained by crossing the CNFC 9506 and RP-2 lines, and were genotyped with 59 markers: 17 simple sequence repeats, 31 amplified fragment length polymorphisms, and 11 sequence-related amplified polymorphisms.

Phenotypic evaluation was conducted using a triple-square lattice design (14 x 14 m) with 10 plants per 1 m^2 plot at Universidade Federal de Lavras. Plots were graded for disease symptoms using a diagrammatic key (1, plant without symptoms, to 9, death of the plant), based on their reaction to the "straw test". Phenotypic analysis was based on the plot means, and assumed the responses followed a normal distribution.

Bayesian version of MAFM

The adaptation of the technique proposed by Doerge et al. (1997) allows QTLs to assume positions varying within an interval (distance) defined by the recombination fraction between the marker and the QTL, rather than fixed positions between two markers as in multiple-QTL mapping.

In this study, we used recombination fractions from 0 to 0.2 to search for QTLs referenced

by the markers. It was crucial that the marker was linked to a QTL. The marker was a pivot. As the QTL moved away from the marker in the interval, different functions of distance were reflected in the posterior probability of the QTL detection and different distributions for its effects.

The linear model can be described as follows. Let y_i, *i = 1, ... n*, be the phenotypic value of the i[th] progeny in a mapping population with three segregating genotypes (e.g., an F_2 population). The linear model was:

(Equation 1) $$y_i = b_0 + \sum_{j=1}^{p} x_{ij} b_j + \sum_{j=1}^{p} w_{ij} d_j + e_i$$

where b_0 is the population mean, *p* is the number of QTLs included in the model (number of markers), x_{ij} is the additive-effect indicator variable of the QTL defined as 1, 0, and -1 for the dominance homozygote, heterozygote, and recessive homozygote, respectively, b_j is the j[th] QTL additive effect, w_{ij} is the dominance-effect indicator variable of the QLT defined as -½, ½, and -½ for the dominance homozygote, heterozygote, and recessive homozygote, respectively, d_j is the j[th] QTL dominance effect, and e_i is the residual with a $N\left(0, \sigma_0^2\right)$ distribution.

For the Bayesian machinery, the observables were $\mathrm{y} = \left\{y_i\right\}$, *i = 1, ... , n*, where *n* is the number of observations, and marker information denoted by $\mathrm{m} = \left\{m_{ij}\right\}$, *i = 1, ... , n* and *j = 1, ..., p*, where *p* is the number of markers. The unobservables included the regression coefficients represented by $\mathrm{c} = \left\{b_0, b_j, d_j\right\}$, *j = 1, ... , p*, the variances represented by $\mathrm{v} = \left\{\sigma_0^2, \sigma_{b_j}^2, \sigma_{d_j}^2\right\}$, *j = 1, ..., p*, the QTL positions $\boldsymbol{\lambda} = \left\{\lambda_j\right\}$, *j = 0, ... , p*, and the QTL genotype indicator variables $\mathrm{x} = \left\{x_{ij}\right\}$ and $\mathrm{w} = \left\{w_{ij}\right\}$, *i = 1, ... , n* and *j = 1, ... , p*.

The prior distributions were:

(Equation 2)
$$p\left(b_0\right) \propto 1,\ p\left(\sigma_0^2\right) \propto \frac{1}{\sigma_0^2},\ p\left(b_j\right) = N\left(0, \sigma_{b_j}^2\right),\ p\left(d_j\right) = N\left(0, \sigma_{d_j}^2\right),$$
$$p\left(\sigma_{b_j}^2\right) \propto \frac{1}{\sigma_{b_j}^2},\ p\left(\sigma_{d_j}^2\right) \propto \frac{1}{\sigma_{d_j}^2},\ j = 1, ..., p$$

The genotype indicator variables x and w were not observed, but they could be inferred from marker information and the positions $\left(\lambda_j\text{'}s\right)$ of the QTLs related to the j[th] marker.

(Equation 3) $$p\left(\boldsymbol{\lambda}, \boldsymbol{m} \mid \boldsymbol{x}, \boldsymbol{w}\right) = p(\boldsymbol{x}, \boldsymbol{w} \mid \boldsymbol{\lambda}, \boldsymbol{m}) p\left(\boldsymbol{\lambda}, \boldsymbol{m}\right)$$

where

(Equation 4) $$p\left(\boldsymbol{\lambda}, \mathbf{m}\right) = \begin{cases} 1/4, & \text{for } AA; \\ 1/2, & \text{for } Aa; \\ 1/4, & \text{for } aa. \end{cases}$$

$$p\left(\boldsymbol{x},\boldsymbol{w}|\boldsymbol{\lambda},\boldsymbol{m}\right)=\begin{bmatrix}\left(1-r^2\right) & 2r\left(2-r\right) & r^2\\ r\left(2-r\right) & \left(1-r^2\right)+\left(1-r^2\right) & r\left(2-r\right)\\ r^2 & 2r\left(2-r\right) & \left(1-r^2\right)\end{bmatrix} \quad \text{(Equation 5)}$$

Fixing up the marker, QTL could vary its position within the specified interval [0, 0.2]. We used flat priors for λ_j (uniformly distributed in this interval).

$$p\left(\lambda_j\right)=1/0.2 \quad \text{(Equation 6)}$$

The joint prior for the unobservable variables was then:

$$p\left(\mathbf{c},\mathbf{v},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\lambda}\right)=p\left(b_0\right)p\left(\sigma_0^2\right)p(\boldsymbol{\lambda},\boldsymbol{m}\mid\boldsymbol{x},\boldsymbol{w})\prod_{j=1}^{p}p\left(b_j\right)p\left(\sigma_{b_j}^2\right)p\left(d_j\right)p\left(\sigma_{d_j}^2\right)p\left(\lambda_j\right) \quad \text{(Equation 7)}$$

The likelihood was described by:

$$\begin{aligned}p\left(\boldsymbol{y},\boldsymbol{m}|\mathbf{c},\mathbf{v},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\lambda}\right)&=\prod_{i=1}^{n}p\left(y_i,\boldsymbol{m}|\mathbf{c},\boldsymbol{x},\boldsymbol{w},\sigma_0^2\right)\\ &\propto\left(\sigma_0^2\right)^{-n/2}exp\left\{-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}\left(y_i-b_0-\sum_{j=1}^{p}x_{ij}b_j-\sum_{j=1}^{p}w_{ij}d_j\right)^2\right\}\end{aligned} \quad \text{(Equation 8)}$$

The posterior distribution was then:

$$p\left(\mathbf{c},\mathbf{v},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\lambda}|\boldsymbol{y},\boldsymbol{m}\right)\propto p\left(\boldsymbol{y},\boldsymbol{m}|\mathbf{c},\mathbf{v},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\lambda}\right)p\left(\mathbf{c},\mathbf{v},\boldsymbol{x},\boldsymbol{w},\boldsymbol{\lambda}\right) \quad \text{(Equation 9)}$$

We used numerical integration to sample values for the parameters from their joint posterior distribution using a Markov Chain Monte Carlo (MCMC) algorithm, based on the Gibbs sampler. The MCMC steps are described in the following sections.

Initialization

The parameters b_0 and σ_0^2 were initialized with the mean and the variance of the phenotypic values of the trait. The genetic effects of all QTLs, b_j and d_j, were initialized with zero. The parameters $\sigma_{b_j}^2$ and $\sigma_{d_j}^2$ were initialized with 0.5. The initial value of λ_j took a random value between 0 and 0.2. The initial values of genotype indicator x_{ij} and w_{ij} were sampled from the probabilities of x_{ij} and w_{ij} conditional on the parameter λ_j and the j[th] marker. The values of all of the unobservable variables were marked with a (k) superscript, which indicated the current iteration, starting from zero.

$$\boldsymbol{I}^{(k)} = \left[b_0^{(k)}, \ldots, b_p^{(k)}, \sigma_0^{2^{(k)}}, \ldots, \sigma_p^{2^{(k)}}, x_{ij}^{(k)}, w_{ij}^{(k)}, \lambda_j^{(k)} \right] \quad \text{(Equation 10)}$$

Updating b_0

The conditional posterior distribution of b_0 was Gaussian with mean $\overline{b_0}$ and variance s_0^2. The sampled b_0 was denoted by $b_0^{(k+1)}$ and replaced $b_0^{(k)}$ in all subsequent steps of the sampling.

$$\overline{b_0} = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}^{(k)} b_j^{(k)} - \sum_{j=1}^{p} w_{ij}^{(k)} d_j^{(k)} \right) \quad \text{(Equation 11)}$$

$$s_0^2 = \frac{1}{n} \sigma_0^{2^{(k)}} \quad \text{(Equation 12)}$$

Updating b_j

The conditional posterior distribution of b_j was Gaussian with mean $\overline{b_j}$ and variance $s_{b_j}^2$, as shown below. The sampled b_j was denoted by $b_j^{(k+1)}$ and replaced $b_j^{(k)}$ in all subsequent steps of the sampling.

$$\overline{b_j} = \left(\sum_{i=1}^{n} x_{ij}^{2^{(k)}} + \sigma_0^{2^{(k)}} / \sigma_{b_j}^{2^{(k)}} \right)^{-1} \sum_{i=1}^{n} x_{ij}^{(k)} \left(y_i - b_0^{(k)} - \sum_{t \neq j}^{p} x_{it}^{(k)} b_t^{(k)} - \sum_{j=1}^{p} w_{ij}^{(k)} d_j^{(k)} \right) \quad \text{(Equation 13)}$$

$$s_{b_j}^2 = \left(\sum_{i=1}^{n} x_{ij}^{2^{(k)}} + \sigma_0^{2^{(k)}} / \sigma_{b_j}^{2^{(k)}} \right)^{-1} \sigma_0^{2^{(k)}} \quad \text{(Equation 14)}$$

Updating d_j

The conditional posterior distribution of d_j was Gaussian with mean $\overline{d_j}$ and variance $s^2_{d_j}$, as shown below. The sampled d_j was denoted by $d_j^{(k+1)}$ and replaced $d_j^{(k)}$ in all subsequent steps of the sampling.

$$\overline{d_j} = \left(\sum_{i=1}^{n} w_{ij}^{2^{(k)}} + \sigma_0^{2^{(k)}} / \sigma_{d_j}^{2\ (k)} \right)^{-1} \sum_{i=1}^{n} w_{ij}^{(k)} \left(y_i - b_0^{(k)} - \sum_{j=1}^{p} x_{ij}^{(k)} b_j^{(k)} - \sum_{t \neq j}^{p} w_{it}^{(k)} d_t^{(k)} \right) \quad \text{(Equation 15)}$$

$$s^2_{d_j} = \left(\sum_{i=1}^{n} w_{ij}^{2^{(k)}} + \sigma_0^{2^{(k)}} / \sigma_{d_j}^{2\ (k)} \right)^{-1} \sigma_0^{2^{(k)}} \quad \text{(Equation 16)}$$

Updating σ_0^2

The residual variance was sampled from a scaled inverted chi-square distribution. The sampled variance $\sigma_0^{2^{(k+1)}}$ replaced $\sigma_0^{2^{(k)}}$.

$$p(\sigma_0^2 | \ldots) \sim \chi^{-2}_{esc}\left(n, FQ\right) \quad \rightarrow \quad \sigma_0^2 = \frac{FQ}{\chi^2_n} \quad \text{(Equation 17)}$$

were $$FQ = \sum_{i=1}^{n} \left(y_i - b_0^{(k)} - \sum_{j=1}^{p} x_{ij}^{(k)} b_j^{(k)} - \sum_{j=1}^{p} w_{ij}^{(k)} d_j^{(k)} \right)^2 \quad \text{(Equation 18)}$$

Updating $\sigma^2_{b_j}$

The $\sigma^2_{b_j}$ was sampled from a scaled inverted chi-square distribution. The sampled variance $\sigma_{b_j}^{2\ (k+1)}$ replaced $\sigma_{b_j}^{2\ (k)}$.

$$p(\sigma^2_{b_j} | \ldots) \sim \chi^{-2}_{esc}\left(1, b_j^{2^{(k)}}\right) \quad \rightarrow \quad \sigma^2_{b_j} = \frac{b_j^{2^{(k)}}}{\chi^2_1} \quad \text{(Equation 19)}$$

Updating $\sigma^2_{d_j}$

The $\sigma^2_{d_j}$ was sampled from a scaled inverted chi-square distribution. The sampled variance $\sigma_{d_j}^{2\ (k+1)}$ replaced $\sigma_{d_j}^{2\ (k)}$.

$$p(\sigma^2_{d_j}|\ldots) \sim \chi^{-2}_{esc}\left(1, d_j^{2^{(k)}}\right) \quad \rightarrow \quad \sigma^2_{d_j} = \frac{d_j^{2^{(k)}}}{\chi^2_1} \qquad \text{(Equation 20)}$$

Updating x_{ij} *and* w_{ij}

Each QTL genotype was sampled from Bernoulli distributions using the j[th] marker information, with the probability shown below.

$$p\left(x_{ij}, w_{ij} \mid \lambda_j^{(k)}, m_{ij}, y_i, \mathbf{c}^{(k)}, \sigma_0^{2(k)}\right) = \qquad \text{(Equation 21)}$$

$$= \frac{p\left(x_{ij}, w_{ij} \mid \lambda_j^{(k)}, m_{ij}\right) p(y_i \mid \mathbf{c}^{(k)}, x_{ij}, w_{ij}, \sigma_0^{2(k)})}{\sum_{l,h} p\left(x_{ij} = l, w_{ij} = h \mid \lambda_j, m_j\right) p(y_i \mid \boldsymbol{c}, x_{ij} = l, w_{ij} = h, \sigma_0^2)}$$

where $l = \{1, 0, -1\}$ and $h = \left\{-1/2, 1/2, -1/2\right\}$.

Updating λ_j

As the λ_j parameter was difficult to sample directly from its conditional posterior distribution, we used a step of the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) within the Gibbs sampler. In the method presented, a uniform distribution was used as an auxiliary function where a new position was sampled $\left(\lambda_{\mathrm{j}}^{(\mathrm{k+1})}\right)$ using the Haldane function (Haldane, 1919) in a limited interval by $max\left(0, \lambda_j^{(k)} - \delta\right)$ and $min\left(0.2; \lambda_j^{(k)} + \delta\right)$, where δ is a constant that defines limits for the random walk within the *j* interval, usually assuming a value of 1 or 2 cM. With this function denoted by $u\left(\lambda_{\mathrm{j}}^{(\mathrm{k+1})}, \lambda_{\mathrm{j}}^{(\mathrm{k})}\right)$, the new position would be accepted in the k[th] iteration with a probability given by $min\left(1, \alpha\right)$.

$$\alpha = \frac{p\left(\lambda_j^{(k+1)} | \mathbf{y}, \mathbf{c}^{(k)}, \sigma_0^{2^{(k)}}, m_j, \mathbf{x}^{(k)}, \mathbf{w}^{(k)}\right) u\left(\lambda_j^{(k+1)}, \lambda_j^{(k)}\right)}{p\left(\lambda_j^{(k)} | \mathbf{y}, \mathbf{c}^{(k)}, \sigma_0^{2^{(k)}}, m_j, \mathbf{x}^{(k)}, \mathbf{w}^{(k)}\right) u\left(\lambda_j^{(k)}, \lambda_j^{(k+1)}\right)} \qquad \text{(Equation 22)}$$

If accepted, a new position was established and a new genotype was suggested for x_{ij} and w_{ij}. The sampling sequence was repeated until reaching a stationary chain. In this final chain, we conducted marginal inference by descriptive statistics and some ad-hoc methods, as described below.

Post-MCMC analysis

In conventional Bayesian mapping analysis, summaries of the marginal posterior

distribution of QTL positions plot the number of hits by QTLs in a short region (bin) against the location where that short region occurs in the genome. The curve produced is called the QTL intensity profile, denoted by $f(\lambda)$, and is a position function (Yang and Xu, 2007; Xu et al., 2009).

In the approach of Wang et al. (2005), each marker interval is associated with a QTL and consequently all intervals are hit by a QTL the same number of times, independently of its effect. However, for an actual QTL to occur within a given interval, the QTL intensity profile will exhibit a peak. However, if the effect is null the intensity profile will be uniform within the interval (Yang and Xu, 2007; Xu et al., 2009; Balestre et al., 2012).

However, the QTL intensity profile cannot be sufficiently informative to infer the QTL location in Bayesian shrinkage analysis. Based on this, Yang and Xu (2007) proposed to weigh the intensity profile by the quadratic terms of the QTL effects.

$$g\left(\boldsymbol{\lambda}\right) = W\left(\boldsymbol{\lambda}\right) f\left(\boldsymbol{\lambda}\right) \quad \text{(Equation 23)}$$

$$W\left(\boldsymbol{\lambda}\right) = \boldsymbol{b}^T \, \mathrm{V}_{\mathbf{b}}^{-1} \, \boldsymbol{b} + \, \boldsymbol{d}^T \mathrm{V}_{\mathbf{d}}^{-1} \boldsymbol{d} \quad \text{(Equation 24)}$$

where *b* and *d* are additive and dominance effects vectors of QTLs, respectively, V_b^{-1} is the inverse of the additive effect variance given by

$$\left(\sum_{i=1}^{n} x_{ij}^2 + \sigma_0^2 \Big/ \sigma_{b_j}^2 \right)^{-1} \sigma_0^2 \quad \text{(Equation 25)}$$

which corresponds to the inverse of the additive effect information matrix, and V_d^{-1} is the inverse of the dominance effect variance given by

$$\left(\sum_{i=1}^{n} w_{ij}^2 + \sigma_0^2 \Big/ \sigma_{d_j}^2 \right)^{-1} \sigma_0^2 \quad \text{(Equation 26)}$$

which corresponds to the inverse of the dominance effect information matrix.

This is equivalent to performing a Wald test on the marginal distribution of the parameters, and it approximately follows a chi-square distribution with two degrees of freedom (Yang and Xu, 2007). This was used in this study to identify meaningful markers and to select them when the value of $W(\lambda)$ (was greater than $\chi^2_{(0,95;2)} = 5,99$.

Analysis

We used two methods to analyze both simulated and real data: Method I was a Bayesian analysis of multiple markers that was proposed by Xu (2003), and Method II was the technique developed in this study.

For the simulated data, an unbalanced coverage of the genome was generated by eliminating 35 and 156 markers throughout the genome (20 and 80%, respectively). For each level of saturation of the genome, the process was repeated 100 times and yielded 100 different unbalancing patterns. To each unbalanced pattern, after selecting significant markers we compared the effectiveness of the methods by the following criteria: for Method I, we compared the distance

between the selected marker and the simulated QTL, the difference between the additive effect of the selected marker and the additive effect of the simulated QTL, and the difference between the dominance effect of the selected marker and the dominance effect of the simulated QTL; for Method II, we compared the distance between the estimated QTL and the simulated QTL, the difference between the additive effect of the estimated QTL and the additive effect of the simulated QTL, and the difference between the dominance effect of the estimated QTL and the dominance effect of the simulated QTL.

Method II was not sensitive to the direction of the QTL search. Therefore, the distance between the estimated QTL and the marker could not measure its relative distance to the simulated QTL. We calculated the direct distance between the posterior mean of the estimated QTL genotype and the simulated QTL using the Kosambi (1944) function, according to the recombination fraction between them.

Figures 1 and 4 presented the QTL detection power for both levels of loss (for each marker in each chromosome that contained a simulated QTL). The power was calculated by:

$$\tau = \frac{F}{d_m} \qquad \text{(Equation 27)}$$

where *F* is the frequency that the QTL was significant among the selections and d_m is the average distance of the simulated QTL.

Additional Figures present features of the estimated QTL by both methods (Figures 2, 3, 5 and 6), such as the relative frequency of detection, which is the number of QTL detection times divided by the number of times selected in 100 unbalanced patterns. This relative frequency is presented in the Figures with an exactly 95% confidence interval for proportion using the ‘‘binom. confint’’ function in the binom R package (Clopper and Pearson, 1934; R Core Team, 2014). The average of the differences in additive effects is represented in the Figures by a black line.

All of the simulations and analyses were conducted using the R software (R Core Team, 2014).

RESULTS

Simulated data: 20% of marker loss

Figure 1 shows the power to detect QTLs by both Methods I (red) and II (blue). The arrows represent the simulated QTLs between the markers, but not their intensities. The higher the number of times that the QTL is found by a marker and the smaller the distance to the simulated QTL the higher are the τ statistics, and consequently the higher the peak. Therefore, if the QTL was extremely close to the real QTL, the distance between them tended to zero and the peak tended to infinity.

For chromosome 1, both methods identified the two simulated QTLs with a very small distance from the real QTLs, but Method II was more powerful. For example, marker 10 was considered a putative QTL by Method I (74 times significant in 82 selections), with an average distance of 11.7 cM from the real QTL; using Method II, this marker found a QTL with an average distance of 1.91 cM from the real QTL and was 71 times significant in 82 selections. For chromosome 5, Method II had greater power to detect QTLs that Method I. Marker 2 was notable, since it was significant 40 times in 74 selections by Method II and 37 times in 74 selections by Method I.

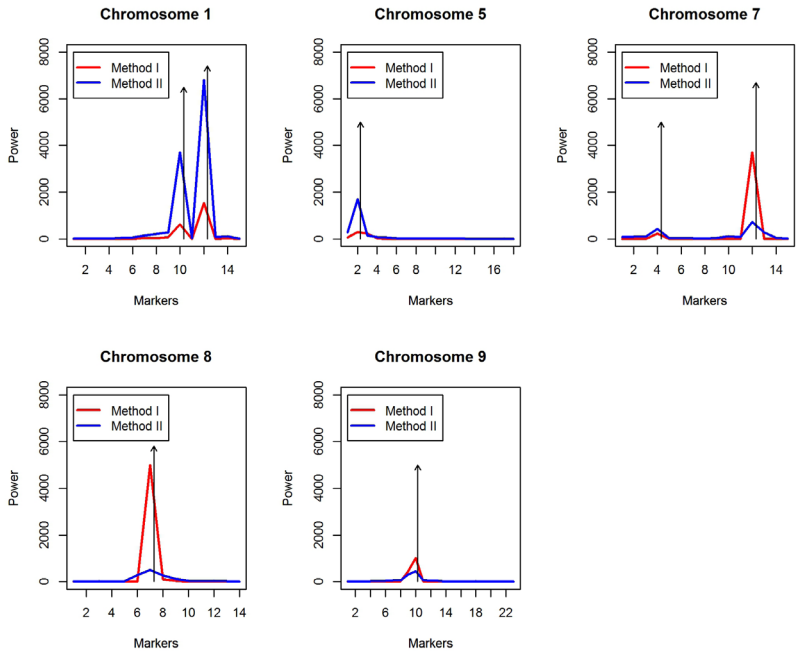

Figure 1. Power of both the methods to detect quantitative trait loci (QTLs) for chromosomes with simulated QTLs, with 20% marker loss.

For the detection of QTLs on chromosome 7, Method I had better results than Method II for marker 12. This may have been because the distance between the marker and the simulated QTL was very short (1.7 cM), and influenced the peak. The simulated QTL was very close to the marker, and was detected using Method I. Using Method II, distances between estimated and simulated QTLs varied between 6.8 x 10^{-11} and 34 cM in 100 unbalanced simulations. The final average distance was 8.2 cM. Therefore, the peak for marker 12 was low, although the frequency of detection was high and close to Method I. It is noteworthy that 28 of the 59 times that marker 12 was significant (for detection) it estimated a QTL with a distance of 6.8 x 10^{-11} cM to the real QTL, demonstrating that it found the position of the simulated QTL.

For marker 4 (chromosome 7) the peak shown in Figure 2 (from Method II) could be higher, since the frequency of detection was higher than in Method I (39 in 85 selections by Method II and 7 in 85 selections by Method I). However, its height was influenced by the average distance of 9 cM, which was a much greater distance than the 3 cM in Method I. Therefore, Figure 2 could be misleading. In 17 of 39 selections, marker 4 was “significant”, and it estimated the QTL with a distance of 4.13 x 10^{-11} cM to the real QTL (perfect identification of the position). Furthermore, the distances found in the selections varied from 4.13 x 10^{-11} to 29 cM.

On chromosome 8, Method I exhibited greater detection power. However, the same misinterpretation that occurred on the seventh chromosome may have occurred here. The simulated QTL was extremely close to marker 7 (distance of 0.1 cM), which greatly influenced the peak using Method I. However, this marker was “significant” only 5 times of 80 selections with this short distance. Using Method II, this marker identified QTLs with an average distance of 7.7 cM to the real QTLs, much higher than that of Method I. However, this marker was significant 39 times in

80 selections (much more frequent than Method I). This demonstrates that Method II can identify real QTLs, although a little far from their actual positions. The Method II distance varied from 1.97 x 10^{-11} to 33 cM. Note that in 20 of the 39 selections the marker was significant using Method II and the exact position was stated.

For chromosome 9, Method I had greater power to detect than Method II. Marker 10 had a detection frequency of 51 of 74 selections, and a distance of 5 cM to the real QTL by Method I. Using Method II, this marker was significant 44 times in 74 selections, with a distance of 9.6 cM. These results are summarized in Figures 2 and 3. High average frequencies with small confidence intervals and small effect differences indicate good detection.

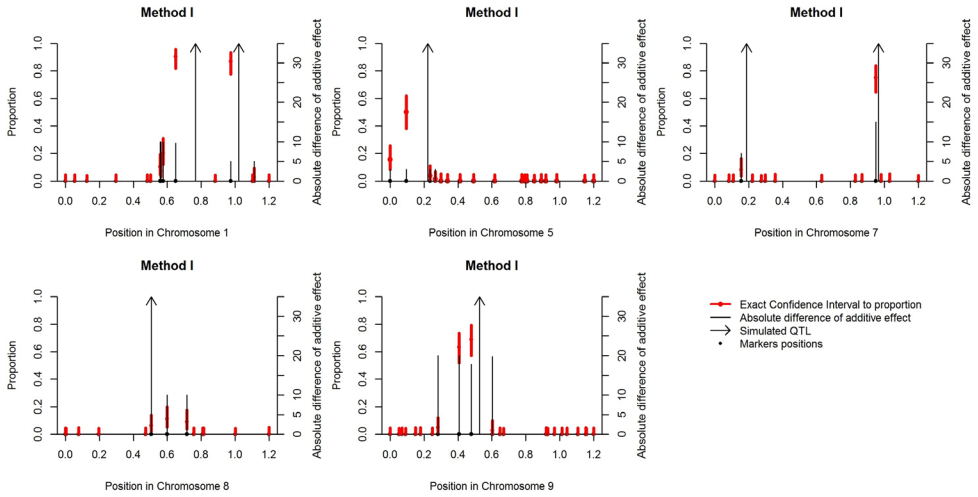

Figure 2. Estimated quantitative trait loci (QTLs) using Method I with respective positions (in Morgan units), relative frequencies with exact confidence intervals to proportion with 5% significance, and absolute difference in additive effects between estimated QTLs and simulated QTLs, with 20% marker loss.

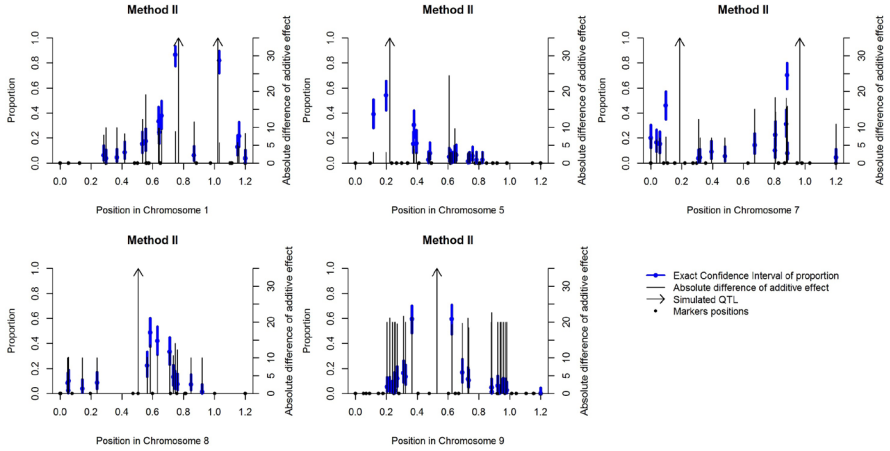

Figure 3. Estimated quantitative trait loci (QTLs) using Method II with respective positions (in Morgan units), relative frequencies with exact confidence intervals to proportion with 5% significance, and absolute difference in additive effects between estimated QTLs and simulated QTLs, with 20% marker loss.

On chromosome 1, markers 10 and 12 found QTLs close to real QTLs using both methods, with high frequencies. The estimated QTLs were closest to the simulated QTLs (arrows) using Method II. The differences in additive effects between these markers were low for both methods. Note that in Method I, markers that were far from the simulated QTL region had detection frequencies with zero values, i.e., they did not identify any QTLs. Using Method II, all of the markers identified QTLs, but the most distant had very low frequencies, indicating that they were distant and unlinked to QTLs.

On chromosome 5, marker 2 identified real QTLs using both methods. However, Method II performed better and had greater detection power (Figure 1). Differences in additive effects were low in both methods.

On chromosome 7, both markers were very close to the simulated QTLs. Using Method I, marker 12 had a high detection frequency, but marker 4 had a low, nonsignificant detection frequency, even within short distances of the simulated QTLs. By Method II, although the distances between the simulated and estimated QTLs were higher than in Method I, the detection frequencies were high for these two markers, demonstrating a better identification of QTLs.

On chromosome 8, marker 7 was very close to the simulated QTL, but had a very low detection frequency by Method I, and the marker was not considered important. Method II estimated QTLs far from their true positions, but identified the right marker. On chromosome 9, markers 9 and 10 identified QTLs more frequently, and marker 10 identified QTLs closest to the real QTLs. The differences in additive effects were high in both methods.

Simulated data: 80% of marker loss

Figure 4 shows the power to detect QTLs by both methods when 81 markers were kept of 165.

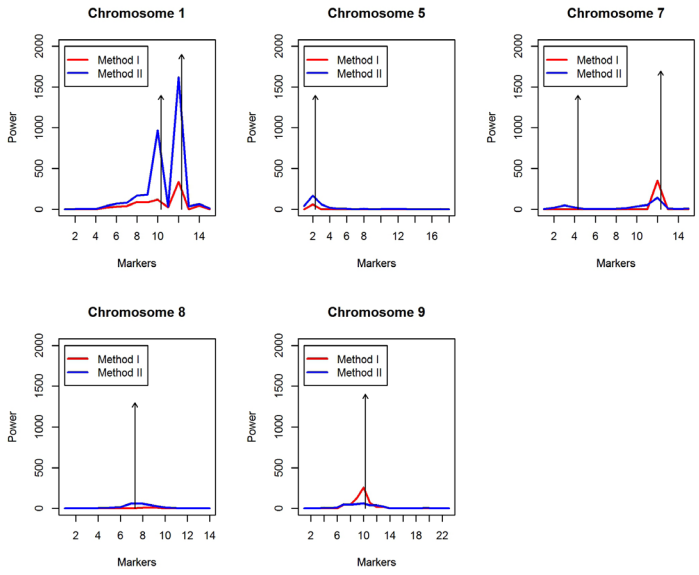

Figure 4. Power of both methods to detect quantitative trait loci (QTLs) in chromosomes with simulated QTLs, with 80% marker loss.

For chromosome 1, Method II found the two simulated QTLs with the greater power, while Method I found the two QTLs but with low power, particularly for the first, which was almost not found. Marker 10 was significant 14 times in 16 selections by Method I, but with a distance of 11.7 cM to the real QTL, which penalized the peak in the graph, making it low. In Method II, this marker found a QTL 16 times in 16 selections, with a distance of 1.7 cM to the real QTL, almost identifying it. It is notable that in 12 of the 16 times in which marker 10 was significant, it identified the QTL with a distance of 3.02 x 10^{-11} cM to the real QTL. For Method I, marker 12 found a QTL 16 times in 17 selections. It was 4.7 cM from the real QTL. In Method II, this marker found QTLs with an average distance of 0.86 cM to real QTLs, being significant 14 times in 17 selections. In 8 of the 14 times marker 12 was significant; it identified the actual QTL position.

For chromosome 5, Method II was more powerful than Method I. The detection frequency for marker 2 by Method I was 8 in 15 selections, with 12.8 cM to the real QTL. Method II had 11 in 15 selections, with an average distance of 6.5 cM to the real QTL. It is notable that in 4 of the 11 times it was significant, and detected a QTL with a distance of 2.71 x 10^{-11} cM to the real QTL.

For chromosome 7, Method II was also better than Method I. Note that in 6 of the 13 times in which marker 12 was significant, it exactly identified the QTL position.

The same results were obtained for chromosome 8; Method II was more powerful than Method I. Using Method I, marker 9 identified the QTL, but with a low frequency (3 of 12 selections), and a distance of 20.9 cM to the real QTL. Using Method II, marker 7 identified QTLs with a frequency of 8 in 19 selections, and an average distance of 12.3 cM to the real QTLs. In 3 of 8 times in which this marker was significant, it identified the QTL position exactly. For chromosome 9, Method I was more powerful, and by Method II the markers 7, 8, 9, 10, 11, and 12 identified the simulated QTLs with the same power intensity, since they were very significant within the selections, although at longer distances.

The characteristics related to the estimated QTLs by both methods are presented in Figures 5 and 6.

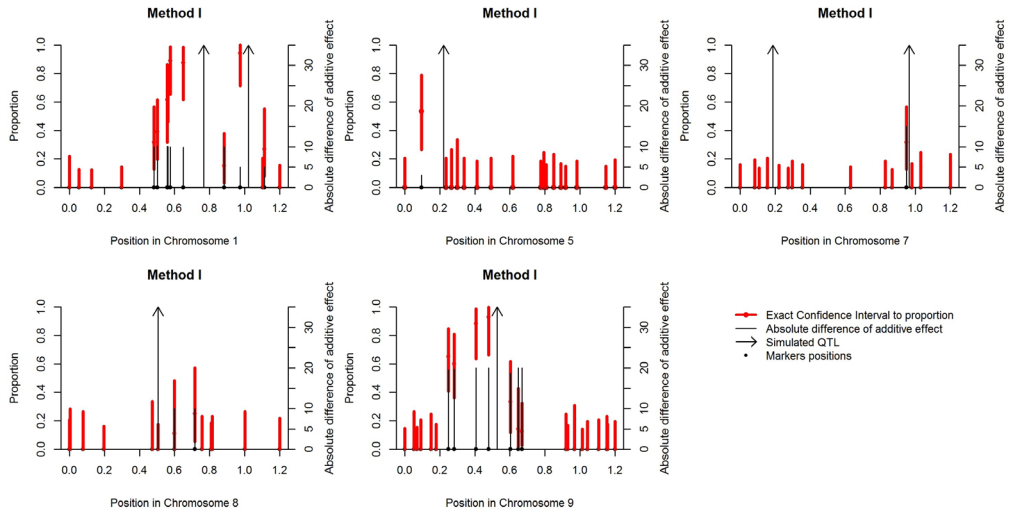

Figure 5. Estimated quantitative trait loci (QTLs) using Method I with respective positions (in Morgan units), relative frequencies with exact confidence intervals to proportion with 5% significance, and absolute difference in additive effects between estimated QTLs and simulated QTLs, with 80% marker loss.

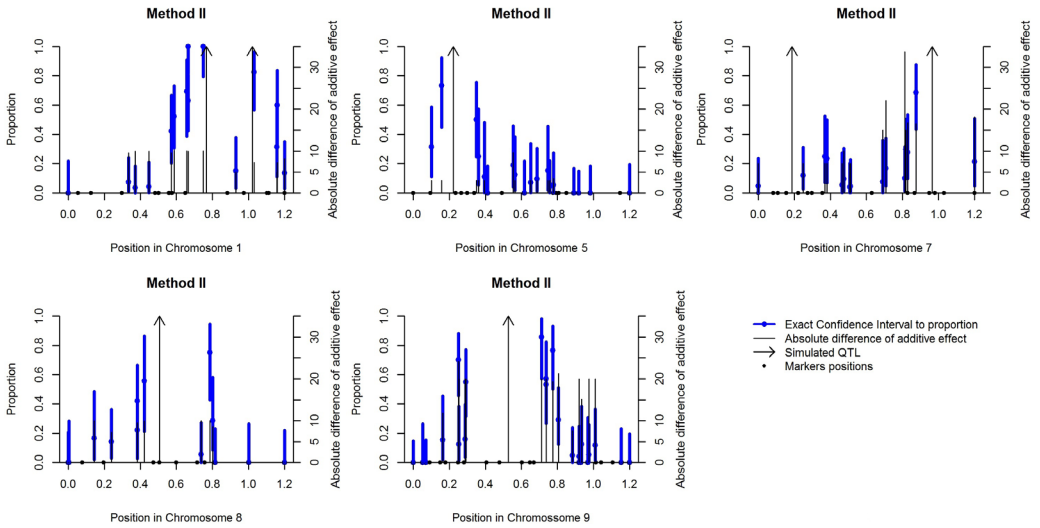

Figure 6. Estimated quantitative trait loci (QTLs) using Method II with respective positions (in Morgan units), relative frequencies with exact confidence intervals to proportion with 5% significance, and absolute difference in additive effects between estimated QTLs and simulated QTLs, with 80% marker loss.

Note that in chromosome 1, the markers 10 and 12 estimated QTLs using both methods. The best estimates were from Method II, since the distances with respect to the simulated QTLs were shorter. The confidence intervals were larger due to the number of low selections, i.e., smaller sample sizes in the simulation. Only one marker did not identify QTLs using Method II, because it was far from the simulated QTL. On chromosome 5, marker 2 estimated closer to the real QTL by Method II than by Method I. The differences in additive effects were low for both methods.

On chromosome 7, marker 12 was very close to the simulated QTL. However, it was rarely detected by Method I. Using Method II, the estimated QTL was more distant, but was frequently detected. The differences in additive effects were high for both methods. Regarding the first simulated QTL, Method I lacked markers to detect it, but using Method II with marker 3 the QTL was identified at a low frequency.

On chromosome 8, the estimated QTL by marker 8 was close to the real QTL using Method II. This resulted in 5 detections out of 9 selections, with an average distance of 8.4 cM. However, marker 9 identified QTLs at greater distances and more often (9 times in 12 selections, with a distance of 27.9 cM). Therefore, using Method II suggested the existence of two QTLs in the chromosome. Although marker 7 was extremely close to the simulated QTL, it did not identify the QTL by Method I, and as the detection frequency of markers 8 and 9 were low, this method seemed to suggest that there were no QTLs on the chromosome.

On chromosome 9, marker 10 frequently identified the real QTL using Method I. The markers 7, 8, and 9 also had high frequencies, but were further from the QTL. By Method II, markers 7 to 12 had high frequencies, but were far from the real QTL, suggesting that they could identify the QTL but could not locate it precisely. The differences in additive effects were high in both methods.

Real data

Using Method I, no marker was important (Figure 7); Wald statistics were always smaller than $\chi^2_{(0.95;2)} = 5.99$ (red line) for all markers. Using Method II, as shown by Lara et al. (2014), 17 of the 59 markers identified QTLs for resistance to white mold (above the red line) according to the Wald test (Figure 8).

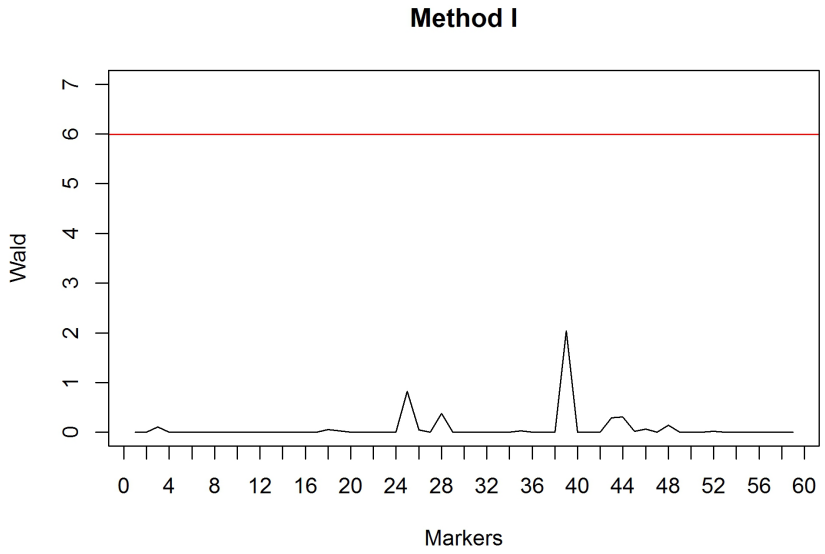

Figure 7. Wald test with 59 markers using Method I. The red line represents $\chi^2_{(0.95;2)} = 5.99$.

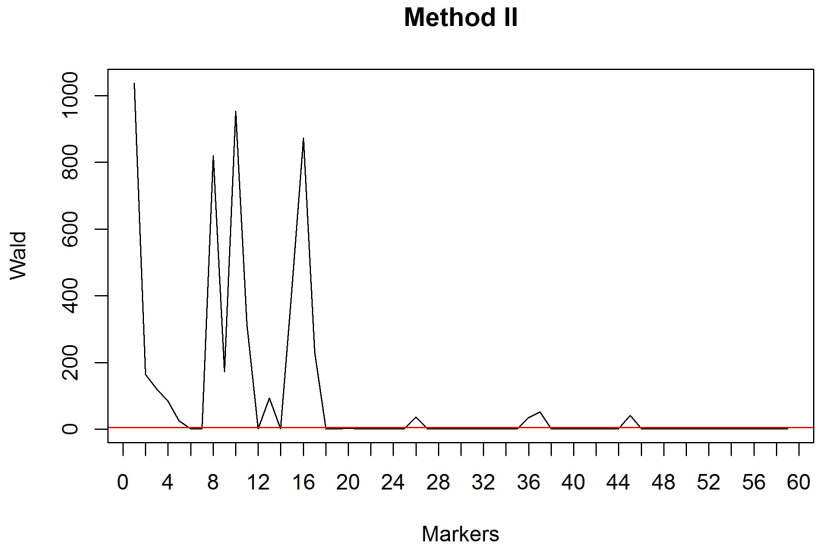

Figure 8. Wald test with 59 markers using Method II. The red line represents $\chi^2_{(0.95;2)} = 5.99$.

DISCUSSION

Method I assumes that most markers have small effects (Xu, 2003), and the inverse of the variance is used as a penalty coefficient. Markers that exhibit distinct effects have a penalty function with smaller values than those with small effects (Xu, 2003). Using the method of Wang et al. (2005), each interval between adjacent markers should have QTLs that are analyzed simultaneously. When many intervals have no QTLs, their estimates are shrunk to zero by the Bayesian methodology. Doerge et al. (1997) used the MAFM technique, which is another method of single-marker mapping that evaluates whether or not the marker is associated with a QTL (Wu et al., 2007).

We followed this approach, but simultaneously used all of the markers by adopting the methods of Wang et al. (2005) and Xu (2003), in order to use markers to identify QTLs rather than the fixed intervals created by adjacent markers. We considered each marker as a pivot for the QTL search at intervals designated by a range of recombination fractions. As the marker moves away (recombination fraction increases) from each position (determined by a random walk within those intervals) we tested for the presence or absence of a QTL.

If the genome is poorly saturated, adjacent markers are very far from each other. Therefore, the recombination fraction used to define the interval required in the method presented in this study may be higher than that shown here, to ensure covering the entire genome during the search process. With a saturated genome, a smaller recombination fraction could be used and avoid such overlapping intervals. The first interval tested in this study was [0; 0.5] to test whether or not the marker was linked to a QTL. Overfitting occurred, because the distance to search for QTLs was great and all of the markers identified a QTL. Therefore, we changed to a smaller range [0; 0.2]. Despite this, Method II identified many QTLs; it is possible that multiple markers detected a single QTL. This result deserves further analysis that is beyond the scope of this study.

In our simulation study, almost all the markers found QTLs within the intervals, and only very distant markers did not. This suggests that the intervals were affected by the presence of the real QTL, resulting in “ghost” detection as they moved away from the true QTL region. However, this was not a problem, because the detection frequency for these QTLs was low. A high detection frequency often indicated that the marker was close to the real QTL region.

We found a pattern when using Method II in the simulation study: markers tended to zero-out the distance estimation the closer they were to the real QTL. This cannot happen using Method I, because effects are located at the markers. Even without a linkage map, this pattern suggests that they could be reordered and consequently the linkage groups (or chromosomes) could be reconstructed, as Method II shows which markers are far from and which ones are close to the real QTL. This pattern was less evident for large losses (80% of the markers), but was still present.

Comparing Figures 1 and 4, we can see that the detection power of both methods was lower for a greater level of loss (80%). Furthermore, Method I was less powerful than Method II for detecting QTLs when the genome was poorly saturated (loss of 80% of the markers). On chromosomes 1, 5, and 8, Method II was better than Method I (Figure 4). On chromosome 7, Method I was better, but it identified only one of the simulated QTLs, i.e., it did not produce satisfactory results. On chromosome 9, Method I identified QTLs with a “cleaner” detection power, and Method II had problems between markers 8 and 12.

Analyzing the real data, we found that the mapping of multiple regression markers proposed by Xu (2003) did not detect QTLs, because of low marker saturation. However, the Bayesian version of MAFM proposed in this study detected QTLs, and it should be used in poorly saturated genomes.

Future work should compare this new technique with other methods of QTL mapping, and use other experimental designs, populations, as well as investigate the potential of the technique for genome-wide association studies, in which models of low dimensionality can be fitted with very dense genotype matrices.

ACKNOWLEDGMENTS

Research supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq, Coordenação de Aperfeiçoamento de Pessoal de Nível Superior-CAPES, and Fundação de Amparo à Pesquisa do Estado de Minas Gerais-FAPEMIG.

REFERENCES

Balestre M, Pinho RGV, Sousa Junior CL and Bueno Filho JSS (2012). Bayesian mapping of multiple traits in maize: the importance of pleiotropic effects in studying the inheritance of quantitative traits. *Theor. Appl. Genet*. 125: 4479-493.

Clopper CJ and Pearson ES (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26: 404-413.

Doerge RW, Zeng ZB and Weir BS (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Stat. Sci*. 12: 195-219.

Edwards MD, Stuber CW and Wendel JF (1987). Molecular-marker-facilitated investigations of quantitative-trait loci in maize: I. Numbers, genomic distribution and types of gene action. *Genetics* 116: 113-125.

Haldane JBS (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *J. Genet*. 8: 299-309.

Hastings WK (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57: 97-109.

Jansen RC (1993). Interval mapping of multiple quantitative trait loci. *Genetics* 135: 205-211.

Joehanes R and Nelson JC (2008). QGene 4.0, an extensible Java QTL-analysis platform. *Bioinformatics* 24: 2788-2789.

Kao CH, Zeng ZB and Teasdale RD (1999). Multiple interval mapping for quantitative trait loci. *Genetics* 152: 1203-1216.

Kosambi DD (1944). The estimation of map distances from recombination values. *Ann. Eugen*. 12: 172-175.

Lander ES and Botstein D (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 121: 185-199.

Lara LAC, Santos JB, Veloso JS, Balestre M, et al. (2014). Identification of QTLs for resistance to *Sclerotinia sclerotiorum* in Carioca common bean by the moving away method. *ISRN Mol. Biol*. Doi: 10.1155/2014/828102.

Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller AH, et al. (1953). Equations of state calculations by fast computing machines. *J. Chem. Phys*. 21: 1087-1092.

R Core Team (2014). *R*: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available at [http://www.R-project.org/]. Accessed May 10, 2014.

Wang H, Zhang YM, Li X, Masinde GL, et al. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* 170: 465-480.

Wu R, Ma CX and Casella G (2007). Statistical genetics of quantitative traits: linkage, maps and QTL. Springer, Berlin.

Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* 163: 789-801.

Xu C, Wang X, Li Z and Xu S (2009). Mapping QTL for multiple traits using Bayesian statistics. *Genet. Res*. 91: 23-37.

Yang R and Xu S (2007). Bayesian shrinkage analysis of quantitative trait loci for dynamic traits. *Genetics* 176: 1169-1185.

Zeng ZB (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proc. Natl. Acad. Sci. U. S. A*. 90: 10972-10976.

Zeng ZB (1994). Precision mapping of quantitative trait loci. *Genetics* 136: 1457-1468.