MODELO HIERÁRQUICO BAYESIANO NA DETERMINAÇÃO DE ASSOCIAÇÃO ENTRE MARCADORES E QTL EM UMA POPULAÇÃO F2

Renato Nunes PEREIRA[1]

Roseli Aparecida LEANDRO [2]

Antonio Augusto Franco GARCIA [3]

Cláudio Lopes SOUZA JUNIOR [3]

Iuri Emmanuel de Paula FERREIRA[4]

[1]Universidade Federal Rural do Rio de Janeiro - UFRRJ, Departamento de Matemática, CEP: 23890-000, Seropédica, RJ, Brasil. E-mail: *rnpmoc@gmail.com*

[2]Universidade de São Paulo - USP, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Ciências Exatas, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *raleandr@usp.br*

[3]Universidade de São Paulo - USP, Escola Superior de Agricultura "Luiz de Queiroz", Departamento de Genética, CEP: 13418-900, Piracicaba, SP, Brasil. E-mail: *augusto.garcia@usp.br; clsouza@usp.br*

[4]Universidade Federal de São Carlos - UFSCAR, Campus Lagoa do Sino, Centro de Ciências da Natureza, CEP: 18290-000, Buri, SP, Brasil. E-mail: *ferreira.iep@gmail.com*

▪ RESUMO: O objetivo do mapeamento de QTL (*Q*uantitative Trait Loci) é identificar a posição de locos que controlam caracteres quantitativos no genoma, isto é, identificar os cromossomos e as localizações nestes em que os QTL se encontram e, além disso, estimar os seus efeitos aditivo e de dominância. Para isso, utiliza-se um grande número de marcadores moleculares espalhados pelos cromossomos que possam estar ligados a estes QTL e, portanto, possam estar associados a características fenotípicas. Devido a isso, os modelos estatísticos possuem elevado número de parâmetros a serem estimados. No entanto, é esperado que muitos destes marcadores não estejam ligados a QTL e, assim, parte destes parâmetros não serão significativos. A proposta deste trabalho é utilizar distribuições *a priori* que permitam a incorporação destas não associações QTL-Marcadores no modelo, que conjuntamente com a informação dos dados conduzam à atualização da informação da ligação marcadores-QTL. Dois modelos foram utilizados: o primeiro que utiliza a distribuição (*a priori*) de encolhimento Lasso Bayesiano e o segundo Estimador de Horseshoe. Para verificar o desempenho dos modelos foram realizadas 1.000 simulações referentes a 10 cenários, nos quais houve variação no número de indivíduos, número de marcadores e níveis de herdabilidade. Foi observado que os modelos propostos têm capacidade de selecionar os marcadores associados aos QTL em todos os cenários. Os modelos foram ajustados à dados de produção de grãos oriundos de progênies de uma população de milho, analisados anteriormente por outras metodologias, para possibilitar comparações de metodologias. A implementação computacional dos algoritmos foi feita utilizando a linguagem C e executada no pacote estatístico R.

▪ PALAVRAS-CHAVE: Lasso; estimador horseshoe; associação; marcadores; QTL.

1 Introdução

Os enormes avanços computacionais das últimas décadas têm permitido o progresso e armazenamento de grandes volumes de informações em diversas áreas de conhecimento. A área biológica, por exemplo, está se beneficiando desses avanços de forma significativa. Do ponto de vista biológico, pode-se observar o desenvolvimento em técnicas moleculares para mapeamento de locos que controlam características quantitativas, denominados de QTL (***Q****uantitative* ***T****rait* ***L****oci*). O objetivo do mapeamento de QTL é identificar sua posição no genoma, determinar o número de QTL que influenciam a característica fenotípica de interesse, assim como identificar em qual cromossomo ele está, qual a sua localização nesse cromossomo, bem como estimar seus efeitos genéticos (LYNCH e WALCH, 1998). Uma vez que as localizações dos QTL não são conhecidas a *priori*, marcadores moleculares são usados frequentemente para auxiliar no seu mapeamento.

Se o mapa genético é suficientemente denso, a maioria dos QTL é potencialmente detectável por causa da ligação estreita com os marcadores moleculares. Como cada marcador é considerado um candidato a QTL, todos eles são incluídos na análise e o modelo estatístico é um modelo supersaturado. Considerando-se o fato de que a maioria deles pode não estar diretamente ligado a QTL, espera-se que muitos dos parâmetros do modelo sejam não significativos. Para

agregar essa informação, vários métodos têm sido propostos na literatura (XU, 2003; WANG *et al.*, 2005; SUN *et al.*, 2009), dentre os quais, um comumente utilizado é o encolhimento bayesiano. Na inferência bayesiana essa situação é tratada especificando distribuições *a priori* especiais para os parâmetros de interesse, isto é, aos efeitos dos marcadores associados aos QTL.

O método encolhimento bayesiano proposto por Xu (2003) é uma alternativa que permite a modelagem associação do marcador com o QTL. Esse método consiste em um modelo hierárquico baseado em um modelo linear aditivo, em que cada coeficiente de regressão segue uma distribuição normal com média zero e uma variância específica para cada coeficiente de regressão. Outra alternativa é o LASSO (Least Absolute Shrinkage and Selector Operator), que é um método de encolhimento que tem sido amplamente usado em análise de regressão para modelos de grandes dimensões (TIBSHIRANI, 1996). O LASSO minimiza a soma de quadrados residuais, restringindo a soma dos valores absolutos dos coeficientes de regressão, e essa restrição permite que algumas estimativas dos coeficientes de regressão seja exatamente zero. Recentemente Park e Casella (2008) propuseram uma formulação bayesiana para o método LASSO de Tibshirani (1996), denominado de LASSO bayesiano. De acordo com Tibshirani (1996) e Park e Casella (2008) a estimativa dos parâmetros obtida pelo método LASSO pode ser interpretada como a moda da distribuição a *posteriori* em um contexto bayesiano, no qual a distribuição a *priori* Laplace (Exponencial Dupla) é atribuída para os parâmetros de regressão.

Um outro método que utiliza distribuições a *priori* de encolhimento é o estimador Horseshoe, proposto por Carvalho *et al.* (2010). Esses autores propuseram uma distribuição a *priori* hierárquica, denominada de *priori* Horseshoe para estimar coeficientes de regressão. O modelo proposto por eles consiste na mistura do parâmetro de escala da distribuição normal com uma distribuição Half-Cauchy padrão nos reais positivos. O estimador Horseshoe é um excelente método de estimação, bem como seleção de preditores no modelo proposto (CARVALHO e POLSON, 2010).

Park e Casella (2008), Yi e Xu (2008) propuseram alguns modelos hierárquicos bayesianos para o mapeamento de múltiplos QTL, que estimam e ajustam os possíveis efeitos genéticos associados aos marcadores ao longo do genoma. Park e Casella (2008), Yi e Xu (2008) utilizaram distribuições a *priori* para os efeitos genéticos que são misturas escalares de distribuições Normais com média zero e uma variância específica para cada efeito, porém desconhecida, sendo necessário atribuir uma distribuição a *priori*. Eles consideraram dois tipos de distribuições a *priori* para a variância, as distribuições Exponencial e a χ^2 - escalonada invertida, em que resulta em uma versão bayesiana do modelo LASSO e o difundido modelo $Student's\ t$, respectivamente. Nesse trabalho cada marcador foi considerado como um potencial QTL com α_j representando o efeito aditivo do QTL e δ_j o seu efeito dominante. Nos dois modelos propostos foram inseridos todos os marcadores moleculares simultaneamente.

O objetivo deste trabalho foi desenvolver modelos hierárquicos bayesianos para verificação de associação entre marcadores e QTL em uma população F_2,

na qual o efeito do marcador será decomposto em efeito aditivo e de dominância. Serão exploradas as distribuições Exponencial e Half-Cauchy como distribuições *a priori* para as quantidades desconhecidas do modelo associadas aos efeitos dos QTL e ajustar os modelos desenvolvidos a um conjunto de dados de uma população de milho tropical. Uma vez que os dados de grãos de milho tropical já foram analisados por diferentes metodologias, isso viabilizará a comparação dos resultados. O objetivo principal em comparar os resultados entres esses trabalhos é apenas para verificar a concordância dos mesmos. A metodologia desenvolvida antes de ser aplicada a dados reais foi aplicada a dados de simulação.

2 Material e métodos

A capacidade de detectar um QTL depende da magnitude do seu efeito sobre a característica, do tamanho da população segregante avaliada, da frequência de recombinação entre marcador e QTL, bem como da herdabilidade da característica analisada. Evidentemente, quanto maiores o efeito, o tamanho da população e a herdabilidade, e mais próximo o marcador do QTL, mais fácil será sua detecção (FERREIRA e GRATTAPAGLA, 1998; LANZA *et al.*, 2000). Muitas pesquisas têm sido realizadas no processo de detecção de QTL e as análises estatísticas, em geral, não são triviais. Nesta seção são ajustados dois modelos para o estudo de associação entre marcadores e QTL utilizando dados simulados e dados experimentais.

2.1 Material

2.1.1 Dados simulados

Para avaliar o desempenho dos dois modelos propostos para determinação de associação entre marcadores e QTL, foi realizado um estudo de simulação. Foram simulados 1.000 conjuntos de dados de experimentos de QTL no software QTLcart. Esse software é inteiramente gratuito para universidades e pode ser obtido por meio da página $ftp://statgen.ncsu.edu/pub/qtlcart/$. É possível simular por meio deste software o mapa genético e a população de mapeamento para o experimento de QTL. Para simular os conjuntos de dados foram construídos dois mapas de ligação, denotados por Mapa I e Mapa II, com as seguintes características:

> **Mapa I:** com cinco cromossomos de comprimento igual a 200 cM, cada um contendo 20 marcadores moleculares dispostos em posições igualmente espaçadas.
> **Mapa II:** com cinco cromossomos de comprimento igual a 200 cM, cada um contendo 40 marcadores moleculares dispostos em posições igualmente espaçadas.

Ao longo de cada mapa simulado foram inseridos cinco QTL: um QTL no cromossomo 1, um QTL no cromossomo 3 e três no cromossomo 5, com efeitos aditivo, dominante e posição descritos nas Tabelas 2 e 3. Foram consideradas

amostras de 200, 400 e 800 indivíduos com dois níveis de herdabilidade global para a característica fenotípica: baixa (0,25) e alta (0,80). Na Tabela 1 são apresentados os 10 cenários considerados no estudo de simulação.

Tabela 1 - Cenários do estudo de simulação

Cenário	Herdabilidade	Nº de marcadores	Nº de indivíduos
1	baixa	100	200
2	baixa	100	400
3	baixa	100	800
4	baixa	200	200
5	baixa	200	400
6	alta	100	200
7	alta	100	400
8	alta	100	800
9	alta	200	200
10	alta	200	400

Tabela 2 - Valores dos efeitos aditivo e dominante dos 5 QTL e seus respectivos marcadores flanqueadores, considerando 100 marcadores

QTL	Cromossomo	Marcador	Aditivo	Dominante
1	C3	M07 – M08	0,0336	0,5512
2	C5	M11 – M12	0,1586	0,6343
3	C5	M02 – M03	0,5940	0,5020
4	C5	M18 – M19	0,2801	−0,3596
5	C1	M08 – M09	0,4576	−0,4410

Tabela 3 - Valores dos efeitos aditivo e dominante dos 5 QTL e seus respectivos marcadores flanqueadores, considerando 200 marcadores

QTL	Cromossomo	Marcador	Aditivo	Dominante
1	C3	M15 – M16	0,0336	0,5512
2	C5	M23 – M24	0,1586	0,6343
3	C5	M05 – M06	0,5940	0,5020
4	C5	M37 – M38	0,2801	−0,3596
5	C1	M17 – M18	0,4576	−0,4410

2.1.2 Dados de uma população de milho tropical

Os dados utilizados nesta subseção estão apresentados em detalhes em Sibov *et al.* (2003a, 2003b) e Sabadin *et al.* (2008). Resumidamente, foi obtida uma

população entre o cruzamento das linhagens endogâmicas L-08-05F e L-14-4B, que possuem comportamentos contrastantes para a produção de grãos. Do cruzamento das linhagens endogâmicas, obteve-se a progênie F_1, sendo que a partir de quatro plantas foram geradas 400 plantas F_2, das quais foram obtidas 400 progênies $F_{2:3}$. As progênies foram avaliadas em dois locais diferentes, no ano de 1999, e em três locais diferentes no ano de 2000, localizados no município de Piracicaba, Estado de São Paulo, Brasil.

As 400 progênies foram divididas em quatro conjuntos com 100 progênies cada e cada grupo foi avaliado em um delineamento látice 10×10, com duas repetições cada. Neste experimento foram avaliados vários caracteres, sendo que no presente trabalho foram utilizados os seguintes caracteres: Produção de Grãos (PG) em Mg por hectare; Altura da Espiga (AE) em cm e Altura da Planta (AP) em cm. O mapa de ligação usado para a localização dos QTL contém 117 locos de marcadores do tipo microssatélites distribuídos em dez grupos de ligação. O comprimento do mapa foi de 1634 cM.

2.2 Métodos

2.2.1 Modelo linear

Seja y_i o valor fenotípico de uma característica quantitativa associado ao indivíduo i, $i = 1, ..., n$, em que n é o número de indivíduos em uma população F_2, e suponha que o indivíduo i é genotipado para p marcadores, os quais são distribuídos ao longo do genoma. Considere o modelo linear descrevendo a relação entre o fenótipo e o genótipo dos marcadores:

$$y_i = \mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j + e_i = \boldsymbol{\mu} + \boldsymbol{X_i\beta} + \boldsymbol{e}, \tag{1}$$

em que μ é uma constante inerente a todas as observações, α_j e δ_j são efeitos aditivos e dominantes, respectivamente, referentes ao marcador j; e_i é o erro aleatório com distribuição normal com média zero e variância σ^2. As variáveis x_{ij1} e x_{ij2} são definidas por meio do modelo epistático de Cockerham (1954) (KAO e ZENG, 2002) e são dadas por:

$$x_{ij1} = z_{ij} - 1 \text{ e } x_{ij2} = (1 + x_{ij1})(1 - x_{ij1}) - 0,5 \tag{2}$$

em que z_{ij} é o número de alelos dominantes do genótipo do j-ésimo marcador para o i-ésimo indivíduo. Assim, para uma população F_2, x_{ij1} e x_{ij2}, considerando-se marcadores co-dominantes, são dadas por:

$$x_{ij1} = \begin{cases} +1 & \text{para} \quad AA; \\ 0 & \text{para} \quad Aa; \\ -1 & \text{para} \quad aa; \end{cases} \quad \text{e} \quad x_{ij2} = \begin{cases} -1/2 & \text{para} \quad AA; \\ 1/2 & \text{para} \quad Aa; \\ -1/2 & \text{para} \quad aa. \end{cases} \tag{3}$$

Na abordagem bayesiana (paramétrica) para a construção da distribuição conjunta a *posteriori* para os parâmetros, é necessário: (1) obter a função

de verossimilhança; (2) incorporar a incerteza relativa dos parâmetros a serem estimados (BOX e TIAO, 1992). Os parâmetros necessários para a descrição do modelo são: μ, σ^2 e os vetores dos coeficientes de regressão $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ..., \alpha_p]^{'}$ e $\boldsymbol{\delta} = [\delta_1, \delta_2, ..., \delta_p]^{'}$ de dimensão p.

2.2.2 Função de verossimilhança

Considerando o modelo descrito pela equação 1, a parametrização de Cockerham dada em 2 e x_{ij1} e x_{ij2} assumindo os valores dados em 3, e assumindo que cada observação Y_i tem distribuição $Y_i \sim N(\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j, \sigma^2)$, a função de verossimilhança dos parâmetros é dada por

$$L(\boldsymbol{\theta}|y) = \prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2}[y_i - (\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j)]^2 \right\} \right\} \quad (4)$$

em que $\boldsymbol{\theta} = (\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta})$.

O modelo especificado pela equação 1 contém um número elevado de parâmetros quando todos os marcadores são incluídos na análise de QTL (modelo supersaturado). Espera-se, no entanto, que muitos destes parâmetros sejam não significativos, necessitando de um tratamento especial. Utilizando a abordagem bayesiana, essa informação importante poderá ser incorporada no ajuste do modelo por meio de distribuições *a priori* apropriadas para os parâmetros. Sendo assim, especifica-se uma distribuição *a priori* que descreva a incerteza que todos os parâmetros relativos a regressão sejam não significativos no modelo e deixe aos dados a "incumbência" de escolher quais dos coeficientes de regressão são necessários para descrever a característica fenotípica. Ressalta-se que a forma da distribuição é importante na especificação da distribuição *a priori* adotada. Um parâmetro não significativo é, naturalmente, especificado através de uma distribuição que coloque um peso maior no zero com muita precisão. Essa especificação é conhecida como encolhimento bayesiano e é proveniente da denominação Bayesian shrinkage.

Neste trabalho utiliza-se duas especificações de distribuições *a priori* para descrever o encolhimento a fim de estimar o efeito dos marcadores associados ao QTL. A primeira especificação é dada por meio do uso da distribuição *a priori* de Laplace, também denonimada, Exponencial Dupla; e a segunda especificação se dá por meio do uso da distribuição *a priori* Horseshoe. Essas duas descrições para a incerteza dos parâmetros de regressão dão origem a dois modelos diferentes, que a partir deste momento são denominados por Modelo I e Modelo II, respectivamente.

2.2.3 Modelo I

A especificação completa do Modelo I é feita considerando o modelo descrito pela equação 1, assumindo-se independência entre os parâmetros do modelo e as

seguintes distribuições *a priori* para os parâmetros:

$$\begin{aligned}
&(\text{ i }) \ \mu \sim N(0, \sigma_u^2);\\
&(\text{ ii }) \ \sigma^2 \sim \text{GI}(a, b);\\
&(\text{iii }) \ \alpha_j \sim \text{ED}(0, \lambda),\ j = 1, \ldots, p;\\
&(\text{ iv }) \ \delta_j \sim \text{ED}(0, \lambda_1),\ j = 1, \ldots, p;
\end{aligned}$$

em que $a > 0$, $b > 0$, λ, λ_1 e σ_u^2 são hiperparâmetros e GI a notação da distribuição Gama Inversa. Ainda, devido à independência entre os parâmetros, tem-se que:

$$\boldsymbol{\alpha} \sim \prod_{j=1}^{p} \text{ED}(0, \lambda) \text{ e } \boldsymbol{\delta} \sim \prod_{j=1}^{p} \text{ED}(0, \lambda_1)$$

em que $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_p]^{'}$ e $\boldsymbol{\delta} = [\delta_1, ..., \delta_p]^{'}$.

A exponencial dupla tem algumas vantagens, dentre elas é que pode ser expressa como um modelo hierárquico de dois níveis (ANDREWS e MALLOWS, 1974). Esse modelo hierárquico de dois níveis, quando comparado com a forma original da exponencial dupla, é mais facilmente tratável tanto do ponto de vista analítico quanto computacional.

Segundo Park e Casella (2008) e Yi e Xu (2008), no modelo hierárquico de dois níveis, o primeiro nível assume que os coeficientes α_j e δ_j seguem distribuições normais independentes com média zero e variâncias desconhecidas υ_j^2 e τ_j^2, respectivamente

$$\boldsymbol{\alpha}|\upsilon_j^2 \sim \prod_{j=1}^{p} N(\alpha_j|0, \upsilon_j^2) \tag{5}$$

e

$$\boldsymbol{\delta}|\tau_j^2 \sim \prod_{j=1}^{p} N(\delta_j|0, \tau_j^2). \tag{6}$$

No segundo nível, assume-se que as variâncias υ_j^2 e τ_j^2 seguem distribuições exponenciais independentes, como especificadas a seguir:

$$\boldsymbol{\upsilon^2}|\lambda^2 \sim \prod_{j=1}^{p} \text{Exp}(\upsilon_j^2|\lambda^2) \tag{7}$$

e

$$\boldsymbol{\tau^2}|\lambda_1^2 \sim \prod_{j=1}^{p} \text{Exp}(\tau_j^2|\lambda_1^2). \tag{8}$$

Em 7 e 8 as expressões $\text{Exp}(\upsilon_j^2|\lambda^2)$ e $\text{Exp}(\tau_j^2|\lambda_1^2)$ são densidades exponenciais, $\boldsymbol{\upsilon^2} = [\upsilon_1^2, ..., \upsilon_p^2]^{'}$ e $\boldsymbol{\tau^2} = [\tau_1^2, ..., \tau_p^2]^{'}$. Ao invés de fixar um valor para os hiperparâmetros

λ^2 e λ_1^2, são atribuídas distribuições *a priori* para esses hiperparâmetros e eles serão estimados juntamente com os outros parâmetros do modelo. As distribuições *a priori* atribuídas a esses dois parâmetros foram respectivamente Gama(a_1, b_1) e Gama(a_2, b_2) com $a_1 > 0$, $a_2 > 0$, $b_1 > 0$ e $b_2 > 0$.

Dado o modelo 1 e as distribuições *a priori* para os parâmetros de interesse, a distribuição conjunta *a posteriori* é dada por:

$$\begin{aligned}
\pi(\theta|\mathbf{y}) \propto & \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\texttt{exp}\left\{-\frac{1}{2\sigma^2}[y_i-(\mu+\sum_{j=1}^{p}x_{ij1}\alpha_j+\sum_{j=1}^{p}x_{ij2}\delta_j)]^2\right\}\right. \\
& \times\frac{1}{\sqrt{2\pi\sigma_\mu^2}}\texttt{exp}\left\{-\frac{1}{2\sigma_\mu^2}\mu^2\right\} \\
& \times(\sigma^2)^{-(1+a)}\texttt{exp}\left\{-\frac{b}{\sigma^2}\right\}\times\prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi\upsilon_j^2}}\texttt{exp}\left\{-\frac{1}{2\upsilon_j^2}\alpha_j^2\right\}\right. \\
& \left.\left.\times\frac{\lambda^2}{2}\texttt{exp}\left\{\frac{\lambda^2\upsilon_j^2}{2}\right\}\times\frac{1}{\sqrt{2\pi\tau_j^2}}\texttt{exp}\left\{-\frac{1}{2\tau_j^2}\delta_j^2\right\}\times\frac{\lambda_1^2}{2}\texttt{exp}\left\{\frac{\lambda_1^2\tau_j^2}{2}\right\}\right\}\right. \\
& \times(\lambda^2)^{a_1-1}\texttt{exp}\{-b_1\lambda^2\}\times(\lambda_1^2)^{a_2-1}\texttt{exp}\{-b_2\lambda_1^2\}
\end{aligned} \tag{9}$$

Visto que a obtenção das distribuições marginais *a posteriori* para os parâmetros de interesse é analiticamente intratável, o algoritmo Gibbs Sampling foi utilizado para a obtenção de uma amostra da distribuição conjunta *a posteriori* e a partir dessa amostra é possível fazer inferências sobre os parâmetros de interesse, ou seja, resumos tais como média, mediana e intervalos de credibilidade para os efeitos aditivos, dominâncias e os demais parâmetros. A descrição da obtenção da amostra da distribuição conjunta *a posteriori* encontra-se no APÊNDICE A.

2.2.4 Modelo II

A especificação completa do Modelo II é feita considerando o modelo descrito pela equação 1, assumindo-se independência entre os parâmetros do modelo e as seguintes distribuições *a priori* para os parâmetros:

$$\begin{aligned}
&(\text{ i }) \ \alpha_j|\upsilon_j^2 \sim N(0,\upsilon_j^2);\\
&(\text{ ii }) \ \upsilon_j^2|\lambda_j \sim \text{C}^+(0,\lambda_j);\\
&(\text{iii }) \ \lambda_j \sim \text{C}^+(0,\phi);\\
&(\text{ iv }) \ \delta_j|\tau_j^2 \sim N(0,\tau_j^2);\\
&(\text{v}) \ \tau_j^2|\lambda_{1_j} \sim \text{C}^+(0,\lambda_{1_j});\\
&(\text{vi}) \ \lambda_{1_j} \sim \text{C}^+(0,\phi_1);
\end{aligned}$$

em que $j = 1, \ldots, p$; $\phi > 0$; $\phi_1 > 0$; $\mathrm{C}^+(0, \lambda_j)$ e $\mathrm{C}^+(0, \lambda_{1_j})$ são distribuições Half-Cauchy padrão nos reais positivos com hiperparâmetros de escala λ_j e λ_{1_j} e eles serão estimados juntamente com os outros parâmetros do modelo. Nesse modelo, diferentemente do Modelo I, a cada υ_j^2 e τ_j^2 está associado um λ_j e λ_{1_j}, respectivamente. Essa medida é necessária para facilitar o processo de convergência dos demais parâmetros do modelo.

Dado o modelo 1 e as distribuições a *priori* para os parâmetros de interesse, a distribuição conjunta *a posteriori* é dada por:

$$\begin{aligned}
\pi(\theta|\mathbf{y}) \propto & \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{1}{2\sigma^2}[y_i - (\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j)]^2\right\}\right. \\
& \times \frac{1}{\sqrt{2\pi\sigma_\mu^2}}\exp\left\{-\frac{1}{2\sigma_\mu^2}\mu^2\right\} \\
& \times (\sigma^2)^{-s+1}\exp\left\{-\frac{t}{\sigma^2}\right\} \times \prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi\upsilon_j^2}}\exp\left\{-\frac{1}{2\upsilon_j^2}\alpha_j^2\right\}\right. \\
& \times \frac{2}{\pi\lambda_j\left[1 + \left(\frac{\upsilon_j^2}{\lambda_j}\right)^2\right]} \times \frac{1}{\sqrt{2\pi\tau_j^2}}\exp\left\{-\frac{1}{2\tau_j^2}\delta_j^2\right\} \\
& \left.\left. \times \frac{2}{\pi\lambda_{1_j}\left[1 + \left(\frac{\tau_j^2}{\lambda_{1_j}}\right)^2\right]} \times \frac{2}{\pi\phi\left[1 + \left(\frac{\lambda_j}{\phi}\right)^2\right]} \times \frac{2}{\pi\phi_1\left[1 + \left(\frac{\lambda_{1_j}}{\phi_1}\right)^2\right]}\right\}\right.
\end{aligned} \tag{10}$$

Visto que a obtenção das distribuições marginais *a posteriori* para os parâmetros de interesse é analiticamente intratável, os algoritmos Metropolis-Hastings e Gibbs Sampling foram utilizados para a obtenção de uma amostra da distribuição conjunta *a posteriori* e a partir dessa amostra é possível fazer inferências sobre os parâmetros de interesse, ou seja, resumos tais como média, mediana, intervalos de credibilidade para os efeitos aditivos, dominâncias e os demais parâmetros. A descrição da obtenção da amostra da distribuição conjunta *a posteriori* encontra-se no APÊNDICE B.

Por meio das distribuições conjuntas *a posteriori* 9 e 10 é possível obter os resumos *a posteriori* para todos os parâmetros de interesse, tais como: média, mediana e intervalos de credibilidade. Os intervalos de credibilidade foram considerados na avaliação da significância do efeito associado a cada marcador, visto que o verdadeiro valor do parâmetro tem probabilidade $100(1-\alpha)\%$ de pertencer a ele. Sendo assim, se após a estimação dos parâmetros e dos intervalos de credibilidade, o valor zero pertencer à um destes intervalo implicará que o efeito do marcador não é significativo no modelo. Utilizando as amostras *a posteriori* de α_j e δ_j e os valores observados x_{ij1} e x_{ij2}, pode-se calcular a proporção da variância fenotípica explicada por cada efeito (herdabilidade), $h_j^2 = v_j\alpha_j^2/\sigma_y^2$ (aditivo) e $H_j^2 = V_j\delta_j^2/\sigma_y^2$ (dominante), em que σ_y^2 é a variância de Y e v_j e V_j são as variâncias

da amostra de $(x_{ij1}; i = 1, ..., n)$ e $(x_{ij2}; i = 1, ..., n)$, respectivamente (YI e XU, 2008).

3 Resultados e discussão

Para facilitar a apresentação e discussão dos resultados, inicialmente descreve-se os resultados do estudo de simulação e em seguida apresentam-se os resultados da análise de dados referentes a uma população de milho tropical.

Para a análise dos dados, foram utilizados os Modelos I e II propostos. Para o ajuste dos Modelos I e II foi implementado um programa utilizando a linguagem C e executado no pacote estatístico R (R CORE TEAM, 2017). A implementação do código C destes modelos podem ser obtidas mediante solicitação ao primeiro autor. A construção da amostra da distribuição conjunta *a posteriori* para os parâmetros foi feita utilizando métodos MCMC, mais especificamente, os algoritmos Gibbs Sampler e Metropolis-Hastings. Foram geradas cadeias com 60.000 iterações, as primeiras 20.000 iterações foram descartadas (*burn-in*) e um espaçamento (*thinning*) de tamanho 27 iterações foi considerado com a finalidade de diminuir a correlação existente entre os valores amostrados, gerando assim uma amostra final de tamanho 1.482. As cadeias resultantes dos algoritmos Gibbs Sampler e Metropolis-Hastings necessitam ter sua convergência diagnosticada. O critério escolhido para verificar a convergência das cadeias foi o critério de Gelman e Rubin (1992), o qual encontra-se disponível, por exemplo, no pacote BOA (*Bayesian Output Analysis*) do programa estatístico *R*, auxiliada por gráficos tais como: gráficos de histórico da cadeias, traço da cadeia, densidade, correlação.

Devido às diferentes especificações das distribuições *a priori* utilizadas para descrever a incerteza *a priori* dos efeitos dos QTL associados aos marcadores, os Modelos I e II diferenciaram-se no número de parâmetros a serem estimados. Na Tabela 4 está disposto o número de parâmetros estimados de acordo com cada modelo nos diferentes conjuntos de dados analisados.

Tabela 4 - Número de parâmetros estimados para cada modelo

Dados	Nº de marcadores	Nº de parâmetros	
		Modelo I	Modelo II
Simulados	100	404	602
	200	804	1202
Milho Tropical	117	472	704

3.1 Dados simulados

Para cada um dos cenários apresentados na seção 2 (página 5) foram gerados 50 conjuntos de dados que foram analisados utilizando-se os Modelos I e II, produzindo assim um total de 1000 análises. O tempo computacional para cada análise em

um computador com a configuracão: Intel (R) Core (TM) i7-2630 QM CPU @ 2.00 GHz e 6 GB de RAM, variou de 15 minutos para o cenário 1 a 2 horas e 45 minutos para o Cenário 10. Embora o número de parâmetros a serem estimados pelo Modelo II seja maior que o Modelo I, não houve uma diferença significativa entre o tempo computacional dos dois modelos. Isso se deve, provavelmente a linguagem de programação utilizada. Os parâmetros υ_j^2, τ_j^2, λ_j e λ_{1j} foram atualizados no Modelo II utilizando o algoritmo Metropolis-Hastings e o Gibbs Sampler no Modelo I. Os demais parâmetros em ambos os modelos foram atualizados utilizando-se o Gibbs Sampler. A taxa de aceitação para os parâmetros em que foi utilizado o Metropolis-Hastings variou de 30% a 46%. Os resultados obtidos utilizando os dois modelos são apresentados nas subseções seguintes por meio de tabelas.

3.1.1 Comparação por meio de simulação entre os resultados do Modelo I e Modelo II

Foi realizada uma comparação entre os resultados do ajuste dos dois modelos propostos neste trabalho para a seleção de marcadores associados a QTL utilizando o ajuste dos dados simulados. O objetivo da comparação é o de avaliar se o Modelo I e o Modelo II estão selecionando os marcadores associados a QTL adequadamente, bem como se são equivalentes, ou seja, se concordam em relação às seleções dos marcadores. Os passos utilizados para a comparação dos modelos são:

1) para cada um dos cenários apresentados na Tabela 1(seção 2, página 5) foram ajustados os Modelos I e II.

2) A ausência do valor zero em pelo menos um dos intervalos de credibilidade, referentes aos efeitos genéticos, indica que existe evidências de que efeito é significativo e que, existem evidências de que o marcador está associado ao QTL.

3) utilizando as informações das Tabelas 2 e 3 e os marcadores selecionados, como descrito no item 2, foi possível identificar quais destes marcadores teriam que ser selecionados.

4) baseado nesta seleção de marcadores foi possível calcular o percentual de acerto de cada modelo, bem como o percentual de falsos positivos.

5) o item 4 foi realizado para cada uma das 50 análises de cada cenário. Para resumir os resultados de cada cenário, trabalhou-se com o percentual médio de acerto das 50 análises, assim como dos falsos positivos.

Os resultados obtidos utilizando os passos descritos nos itens de 1 a 5 estão apresentados nas Tabelas 5 a 7. Como os dois modelos estudados realizam uma estimação pontual, espera-se que no estudo de associação seja selecionado pelo menos um dos marcadores flanqueadores do QTL, não necessariamente têm que selecionar os dois marcadores, mas diversas vezes, além de selecionar os marcadores flanqueadores, foram selecionados outros que estavam próximos destes, formando

assim uma concentração de marcadores, aumentando as evidências de um QTL naquele local. Em algumas situações, ao invés de selecionar os marcadores flanqueadores, selecionou-se marcadores próximos. Mediante isso, nas Tabelas 5 a 7 foram apresentados os resultados considerando o acerto exato (selecionou-se um dos marcadores flanqueadores) e por região (o marcador selecionado estava próximo aos marcadores flanqueadores).

Tabela 5 - Percentual de acertos e de falsos positivos de cada modelo, considerando-se 200 indivíduos

Herdabilidade	Número de Marcadores	Modelo	Efeito	Percentual de Acertos (%)		Falsos Positivos (%)
				Marcador Exato	Por Região	
baixa	100	I	aditivo	21,2	26	1,76
			dominante	36,4	42,4	3,10
		II	aditivo	19,6	22,4	0,82
			dominante	33,6	36,0	0,92
	200	I	aditivo	24	26	2,40
			dominante	32,4	36,0	2,60
		II	aditivo	20,4	22,4	1,45
			dominante	32	40,4	1,55
alta	100	I	aditivo	49,2	50,8	2,58
			dominante	72,4	73,3	4,26
		II	aditivo	46,8	48,8	0,92
			dominante	74,0	74,4	1,04
	200	I	aditivo	49	49,8	3,31
			dominante	71,2	74,2	3,63
		II	aditivo	48,1	49,4	1,56
			dominante	72,2	72,6	1,44

Concordâncias entre os resultados, obtidos por ambos os modelos, indica a equivalência entre os mesmos e, portanto, um ou outro, pode ser escolhido sem grande impacto no resultado final, porém observando nas Tabelas 5 a 7 o percentual médio de acerto do Modelo I foi em geral superior ao Modelo II. Em contrapartida, é possível observar que o número de falsos positivos provenientes do Modelo II em todos os cenários considerados foi inferior ao Modelo I. Observa-se nestas tabelas que os resultados melhoraram a medida em que aumentou o número de indivíduos na população. Assim como variou o número de indivíduos, houve uma variação no número de marcadores, mas os resultados não mudaram e foram semelhantes quando considerado 100 ou 200 marcadores.

3.2 Aplicação a dados experimentais: produção de grãos de milho tropical

Para os dados, considerando o fenótipo produção de grãos, foram ajustados os Modelos I e II. O tempo computacional para a análise em um computador com a

Tabela 6 - Percentual de acertos e de falsos positivos de cada modelo, considerando-se 400 indivíduos

Herdabilidade	Número de Marcadores	Modelo	Efeito	Percentual de Acertos (%)		Falsos Positivos (%)
				Marcador Exato	Por Região	
baixa	100	I	aditivo	40,80	45,20	2,70
			dominante	74,00	77,20	3,42
		II	aditivo	36,40	39,80	1,20
			dominante	56,8	60,4	1,60
	200	I	aditivo	41,95	45,36	2,78
			dominante	73,17	77,00	3,96
		II	aditivo	39,02	40,98	1,08
			dominante	56,58	59,51	1,12
alta	100	I	aditivo	75,20	76,80	2,92
			dominante	97,60	97,60	4,28
		II	aditivo	70,80	72,00	1,38
			dominante	98,00	98,00	1,40
	200	I	aditivo	74,8	77,20	4,65
			dominante	98,4	98,4	5,34
		II	aditivo	71,60	74,00	1,63
			dominante	98,4	98,4	2,17

Tabela 7 - Percentual de acertos e de falsos positivos de cada modelo, considerando-se 800 indivíduos

Herdabilidade	Modelo	Efeito	Percentual de Acertos (%)		Falsos Positivos (%)
			Marcador Exato	Por Região	
baixa	I	aditivo	49,25	54,35	3,82
		dominante	90,25	92,80	4,82
	II	aditivo	44	45,5	1,80
		dominante	80,5	84,0	1,90
alta	I	aditivo	80,58	86,25	5,65
		dominante	98,33	98,33	6,36
	II	aditivo	81,66	82,91	1,98
		dominante	97,91	97,91	3,83

configuracão: Intel (R) Core (TM) i7-2630 QM CPU @ 2.00 GHz e 6 GB de RAM, foi de aproximadamente 56 minutos para cada um dos Modelos. Os resultados obtidos utilizando os dois modelos foram comparados entre si e serão apresentados nas subseções que seguem.

O tempo computacional para a análise da variável em estudo em um computador com a configuracão: Intel (R) Core (TM) i7-2630 QM CPU @ 2.00 GHz e 6 GB de RAM, foi de aproximadamente 56 minutos. Os resultados obtidos

utilizando os dois modelos foram comparados entre si e serão apresentados a seguir. Observa-se nas Figuras 1 e 2 a mediana *a posteriori* dos efeitos aditivo e dominante estimados por meio dos Modelos I e II e à proporção da variância fenotípica explicada por cada efeito para os marcadores ao longo dos cromossomos, e nas Figuras de 3 e 4 os intervalos de credibilidade das estimativas dos parâmetros associados aos marcadores com evidências de associação com QTL.

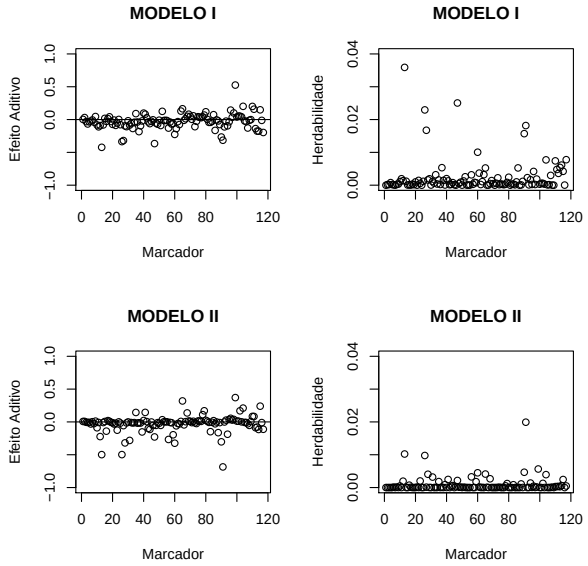

Figura 1 - Mediana *a posteriori* para o efeito aditivo de cada marcador e da herdabilidade.

Observa-se na Tabela 8 a lista completa dos marcadores selecionados por meio dos Modelos I e II e os cromossomos em que se encontram esses marcadores. Na Tabela 8 e nas Figuras 3 e 4 é possível observar que os resultados foram semelhantes para os Modelos I e II.

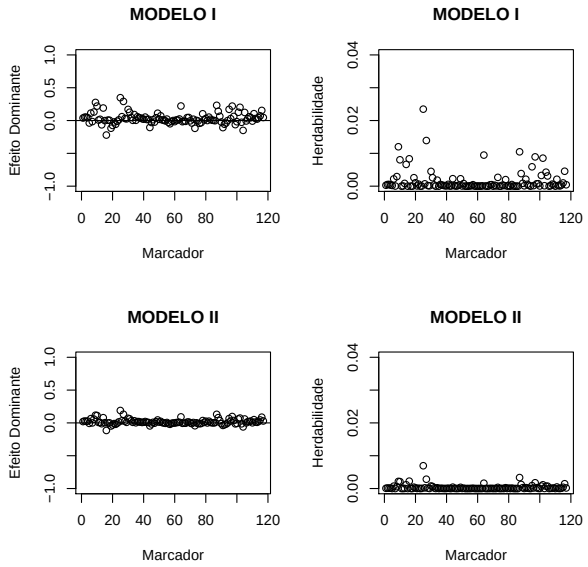

Figura 2 - Mediana *a posteriori* para o efeito dominante de cada marcador e da herdabilidade.

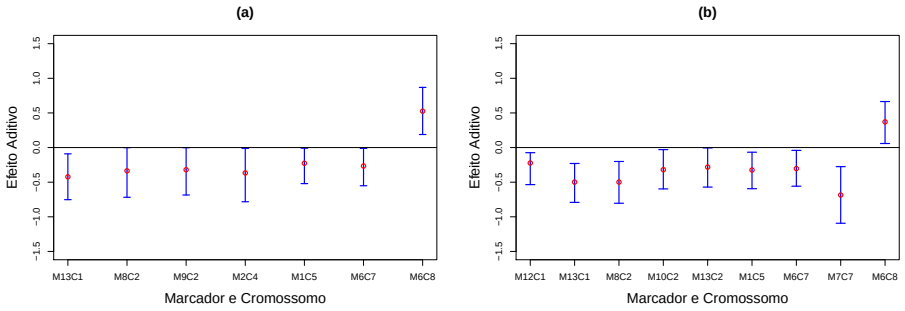

Figura 3 - Intervalo de Credibilidade 95% para o efeito aditivo dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b).

Combinando os efeitos aditivos e dominantes, foram identificados oito marcadores com evidências de associações a QTL utilizando o Modelo I. Esse número de marcadores foi igual a 10 pelo Modelo II. Nos dois modelos, esses marcadores estão localizados nos mesmos cromossomos (1,2,5,7,8), com exceção do marcador umc1652, que pode ser visto na Tabela 8, está localizado no cromossomo 4, identificado apenas pelo Modelo I. Observa-se uma diferença entre o número de marcadores identificados pelos modelos, mas a diferença ocorre porque para

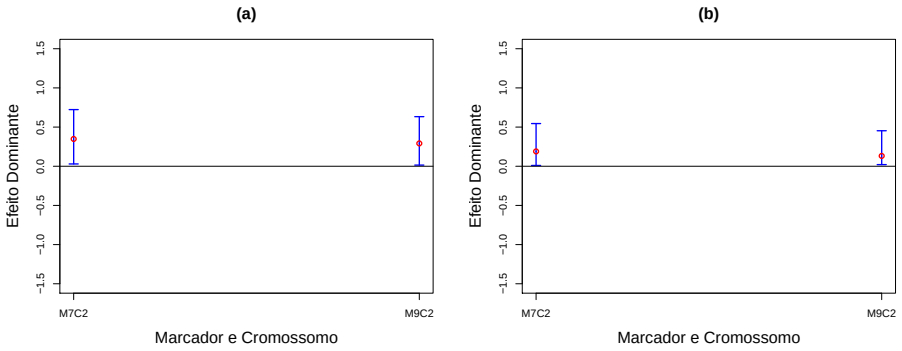

Figura 4 - Intervalo de Credibilidade 95% para o efeito dominante dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b).

Tabela 8 - Marcadores com possíveis associações com QTL de acordo com os Modelos I e II

			Efeito do marcador	
Modelo	Marcador/Crom.[1]	Marcador	Aditivo	Dominância
I	13/1	bnlg0615	-0,4229	-0,0650*
	7/2	bnlg0125	-0,0780*	0,3484
	8/2	umc1845	-0,3358	0,0589*
	9/2	bnlg0166	-0,3198	0,2916
	2/4	umc1652	-0,3661	-0,0236*
	1/5	bnlg1006	-0,2263	-0,0161*
	6/7	dupssr13	-0,2661	-0,0230*
	6/8	bnlg1176	0,5256	-0,0623*
II	12/1	umc1035	-0,2231	0,0056*
	13/1	bnlg0615	-0,4984	-0,0117*
	7/2	bnlg0125	-0,0180*	0,1895
	8/2	umc1845	-0,4980	0,0321*
	10/2	dupssr21	-0,3194	0,0264*
	13/2	umc1633	-0,2818	0,0621*
	1/5	bnlg1006	-0,3243	-0,0022*
	6/7	dupssr13	-0,3033	-0,0109*
	7/7	umc1154	-0,6856	-0,0381*
	6/8	bnlg1176	0,3708	-0,0141*

[1]Cromossomo onde está localizado o marcador.

*Não Significativo.

o Modelo II, ao invés de detectar apenas um marcador associado a um possível QTL, foram identificados vários em sequência, formando uma região, aumentando

as evidências de um QTL nesta região. Isso pode ser visto claramente com os marcadores 7, 8, 10 do cromossomo 2 na Tabela 8, considerando os efeitos aditivos e dominantes simultaneamente. Identificar marcadores em sequência, tornando-os como possíveis marcadores flanqueadores de um determinado QTL, não é exclusividade do Modelo II, embora, isso tenha ocorrido com maior frequência neste modelo.

Algumas análises com metodologias específicas foram realizadas anteriormente a este trabalho, utilizando-se o conjunto de dados milho tropical, fato que possibilita fazer comparações com os resultados obtidos para o fenótipo produção de grãos. A comparação entre os diferentes métodos tem como objetivo principal analisar a concordância entre os resultados e não discriminar entre melhor ou pior método. Na Tabela 9 são apresentados os resultados para os cromossomos com evidências de QTL, obtidos:
1) pelo modelo CIM (Mapeamento por Intervalo Composto) apresentados em Sibov et al. (2003);
2) pela metodologia bayesiana para mapear QTL no caso em que efeitos de epistasia são incorporados no modelo, apresentado por Meyer et al. (2009);
3) e no presente trabalho.

Tabela 9 - Cromossomos com forte evidência da existência de QTL, considerando 4 modelos diferentes

Modelo	QTL/Crom.[1]	Posição do QTL	Intervalo
CIM	1/2	52,88	umc1845-bnlg0166
	2/7	125,20	dupssr13-umc1154
	3/8	57,94	phi0115-bnlg1176
	4/8	74,70	bnlg1176-bnlg1607
bayesiano	1/2	56,69	umc1845-bnlg0166
	2/7	125,47	dupssr13-umc1154
	3/8	77,01	bnlg1176-bnlg1607
Modelo	Marcador/Crom.[2]	Posição do Marcador	Marcador
II	8/2	33,7	umc1845
	9/2	71	bnlg0166
	6/7	103,1	dupssr13
	6/8	63,8	bnlg1176
II	8/2	33,7	umc1845
	6/7	103,1	dupssr13
	7/7	136,8	umc1154
	6/8	63,8	bnlg1176

[1] Cromossomo onde está localizado o QTL.
[2]Cromossomo onde está localizado o marcador.

Na Tabela 9 são apresentados apenas os marcadores dos Modelos I e II que coincidiram com os resultados destes dois trabalhos realizados anteriormente. Como se pode observar, uma parte dos marcadores selecionados neste trabalho está relacionada com QTL de acordo com as metodologias utilizadas por Sibov *et al.* (2003) e Meyer *et al.* (2009), como pode ser visto na Tabela 9. Comparando os resultados dos diferentes modelos, nota-se que há uma concordância na evidência de QTL nos cromossomos 2, 7 e 8, sendo que os dois modelos utilizados neste trabalho identificaram evidências de QTL em outros cromossomos (4,5).

3.3 Discussão

Neste trabalho, foram explorados dois modelos hierárquicos bayesianos: Modelo I baseado em Tibshirani (1996), Park e Casela (2007) e Yi e Xu (2008) e Modelo II baseado em Carvalho e Polson (2010). Os modelos foram utilizados com o objetivo de estimar os possíveis efeitos genéticos associados com todos os marcadores no cromossomo. Para avaliar o desempenho dos modelos na determinação da associação entre marcadores e QTL, realizou-se um estudo de simulação no qual considerou-se uma população F_2. Foi possível notar que tais modelos, devido à estrutura de distribuições *a priori* utilizadas, têm capacidade de selecionar os marcadores com associações aos QTL. Porém foi observado que há a possibilidade, ainda que pequena, de marcadores que não possuam associação com QTL serem selecionados, sendo assim considerados como falsos positivos.

Os marcadores, em geral, que não estavam ligados a QTL no estudo de simulação, tiveram os efeitos estimados reduzidos para um valor o mais próximo possível de zero. Essa redução foi mais acentuada no Modelo II. Acredita-se que isso tenha ocorrido porque para este modelo considerou-se como distribuição a *priori* para as quantidades desconhecidas associadas aos efeitos do QTL a distribuição a *priori* Horseshoe, que é extremamente concentrada no zero.

Observou-se na análise dos dados simulados que nem sempre os dois marcadores flanqueadores ao QTL são detectados, mas simplesmente um deles. Quando foi detectado apenas um marcador, foi exatamente o marcador em que a distância entre ele e o QTL era menor. Tomando como base as análises pelos modelos CIM e Bayesiano, isso se repetiu com os dados de milho tropical para a maioria dos casos. Observa-se na Tabela 9 que, de acordo com os modelos CIM e Bayesiano os marcadores bnlg1176 e bnlg1607 são marcadores flanqueadores de um QTL. Utilizando as posições do QTL informada na Tabela 9 e recorrendo ao mapa com a posição dos marcadores apresentado por Sibov *et al.* (2003), percebe-se que o marcador que teria que ser identificado, caso não identificasse os dois, teria que ser o bnlg1176. De fato, foi o que ocorreu com os Modelos I e II.

Neste estudo, quando considerado o fator herdabilidade da característica fenotípica, percebeu-se que os resultados referentes aos modelos estudados foram melhores para herdabilidade classificada como alta. Esse desempenho foi nítido para os dois efeitos (aditivo e dominante), implicando que os modelos estão sujeitos a conseguirem menos associações quando considerado herdabilidade baixa.

No Modelo II foi necessário considerar para o parâmetro τ_j^2 da distribuição $N(0, \tau_j^2)$ uma distribuição a *priori* Half-Cauchy com parâmetros de escala λ_j específico para cada τ_j^2. Isso foi necessário porque se considerar apenas um único λ para todos os τ_j^2 como no Modelo I, o λ final gerado seria zero ou próximo de zero, dificultando a convergência dos outros parâmetros. Pelo estudo de simulação, percebeu-se que o valor assumido pelo efeito do QTL é extremamente importante e implica na seleção ou não do marcador no processo de associação. Se o efeito for muito pequeno, provavelmente não será detectado, daí a importância de trabalhar com os efeitos aditivos e de dominância, pois quando o efeito aditivo for pequeno e o de dominância grande, ou vice-versa, o maior deles poderá ser detectado por um dos marcadores, o que ocorreu diversas vezes nas análises dos dados simulados.

Os resultados da associação obtidos para o fenótipo produção de grãos utilizando os Modelos I e II foram comparados com trabalhos realizados anteriormente, os modelos CIM e bayesiano apresentados em Sibov *et al.* (2003b) e Meyer *et al.* (2009), respectivamente. Há uma concordância em boa parte dos resultados, sendo que, considerando os modelos atuais haveria uma existência maior de QTL.

4 Conclusões

Neste trabalho foram propostos dois modelos para verificar a associação entre marcadores e QTL. Um estudo de simulação foi realizado para verificar o desempenho dos dois modelos, concluindo-se que os modelos são equivalentes. Foi realizada a análise de um conjunto de dados de produção de grãos oriundos de progênies de uma população de milho utilizando os dois modelos propostos, os resultados foram comparados com os resultados obtidos em análises previamente realizadas na literatura utilizando-se o método CIM (Mapeamento por Intervalo Composto) e bayesiano apresentados em Sibov *et al.* (2003b) e Meyer *et al.* (2009), respectivamente. Há uma concordância em boa parte dos resultados, sendo que, considerando os modelos atuais haveria uma existência maior de QTL.

5 Agradecimentos

Ao CNPq pela bolsa de doutorado concedida ao primeiro autor e aos revisores e editores pelos comentários e sugestões.

PEREIRA, R, N.; LEANDRO, R. A.; GARCIA, A. A. F.; SOUSA JUNIOR, C. L. Bayesian hierarchical model in determining association between markers and QTL in an F2 population. *Rev. Bras. Biom.,* Lavras, v.36, n.2, p.413-437, 2018.

■ *ABSTRACT: The purpose of QTL mapping is to identify the position of loci controlling quantitative traits in the genome, that is, to identify the chromosomes and the locations in which the QTLs meet and, moreover, to estimate Its additive and dominance effects. To do this, a large number of molecular markers spread across the chromosomes that may be linked to these QTLs are used and, therefore, may be associated with phenotypic characteristics. Due to this, the statistical models have a high number of parameters to be estimated. However, it is expected that many of these markers will not be bound to QTL and thus some of these parameters will not be significant. The purpose of this work is to use a priori distribution that allow the incorporation of these non-QTL-Markers associations into the model, which together with the data information lead to the updating of the QTL-binding information. Two models were used: the first using the Lasso Bayesian shrinkage distribution (a priori) and the second Horseshoe Estimator. To verify the performance of the models were performed 1000 simulations referring to 10 scenarios, in which there was variation in the number of individuals, number of markers and heritability levels. It was observed that the proposed models have the ability to select the markers associated with QTL in all scenarios. The models were adjusted to grain yield data from progenies of a maize population, previously analyzed by other methodologies, to allow comparisons of methodologies. The computational implementation of the algorithms was done using the C language and executed in the statistical package R.*

■ *KEYWORDS: Lasso; horseshoe estimator; association; markers; QTL.*

Referências

ANDREWS, D. F.; MALLOWS, C. L. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society, Ser. B*, v.36, p.99-102, 1974.

BOX, G. E.; TIAO, G. C. *Bayesian inference in statistical analysis*. New York: Wiley, 1992. 588p

CARVALHO, C. M.; POLSON, N. G. The Horseshoe estimator for sparse signals. *Biometrika*, v.97, n.2, p.465-480, 2010.

COCKERHAN, C. C. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, v.39, p.859-882, 1954.

FERREIRA, M. E.; GRATTAPAGLIA, D. *Introdução ao uso de marcadores moleculares em análise genética*. 3.ed. Brasília: EMBRAPA, CENARGEN, 1998. 220 p.

KAO, C., H.; ZENG, Z. B. Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. *Genetics*, v.160, p.1243-1261, 2002.

LANZA, M. A.; GUIMARÃES, C. T.; SHUSTER, I. *Aplicação de marcadores moleculares no melhoramento genético*. Belo Horizonte, v.21, p.97-108, 2000.

LYNCH, M.; WALSH, B. *Genetics and analysis of quantitative traits*. Massachusetts: Sinauer Sunderland, 1998. 980p.

MEYER, A. S. *Uma abordagem bayesiana para mapeamento de QTLs em populações experimentais*. 2009, 129p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2009.

PARK, T.; CASELLA, G. The Bayesian Lasso. *Journal of the American Statistical Association* v.103, p.681-686, 2008.

R CORE TEAM. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. Disponível em: https://www.R-project.org/ Vienna, Austria, 2017. Acesso em: 10 Abril 2017.

SABADIN, P. K.; SOUZA JÚNIOR, C. L.; GARCIA, A. A. F. QTl mapping for yeield components in tropical maize population using microsatellite markers. *Hereditas*, v.145, p.194-203, 2008.

SIBOV, S. T.; SOUZA JÚNIOR, C. L.; GARCIA, A. A. F.; GARCIA, A. F.; SILVA, A. R.; MANGOLIN, C.A.; BENCHIMOL, L.L.; SOUZA, A.P. Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 1 . Map construction and localization of loci showing distorted segregation. *Hereditas*, v.139, p.96-106, 2003a.

SIBOV, S. T.; SOUZA JÚNIOR, C. L.; GARCIA, A. A. F.; GARCIA, A. F.; SILVA, A. R.; MANGOLIN, C. A.; BENCHIMOL, L. L.; SOUZA, A. P. Molecular mapping in tropical maize (Zea mays L.) using microsatellite markers. 2. Quantitative trait loci (QTL) for grain yield, plant height, ear height and grain moisture. *Hereditas*, v.139, p.107-115, 2003b.

SUN, W.; IBRRAHIM, J. G.; and ZOU, F. *"Variable Selection by Bayesian Adaptive Lasso and Iterative Adaptive Lasso, with Application for Genome-wide Multiple Loci Mapping"(March 2009)*. The University of North Carolina at Chapel Hill Department of Biostatistics Technical Report Series. Working Paper 10.

TIBSHIRANI, R. Regression shrinkage and selection via the Lasso. *Journal of the American Statistical Association, Ser. B*, v.58, p.267-288, 1996.

WANG, H.; ZHANG, Y, M.; LI, X. Bayesian Shrinkage estimation of quantitative trait loci parameters. *Genetics*, v.170, n.1, p.465-480, 2005.

XU, S. Estimation polygenic effects using markers of the entire genome. *Genetics*, v.163, n.2, p.789-801, 2003.

YI, N.; XU, S. Bayesian LASSO for Quantitative Trait Loci Mapping. *Genetics*, v.179, p.1045-1055, 2008.

Recebido em 13.10.2016.

Aprovado após revisão em 14.07.2017.

APÊNDICE A - ATUALIZAÇÃO DOS PARÂMETROS DO MODELO I

A atualização dos parâmetros foi feita da seguinte forma:

(i) Atualização de $\boldsymbol{\Theta} = (\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2)$: os parâmetros $\mu, \sigma^2, \alpha_j, \delta_j, \upsilon_j^2, \tau_j^2, \lambda^2, \lambda_1^2$ possuem formas conhecidas para suas distribuições condicionais completas e são dadas por:
(a) distribuição condicional completa *a posteriori* para μ:

$$\mu|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{n} S_i}{\sigma^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_\mu^2}\right)}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_\mu^2}}\right), \tag{11}$$

em que $S_i = y_i - \sum_{j=1}^{p} x_{ij1}\alpha_j - \sum_{j=1}^{p} x_{ij2}\delta_j$.
(b) distribuição condicional completa *a posteriori* para σ^2:

$$\sigma^2|\mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim GI\left(\delta + \frac{n}{2}, \frac{1}{\frac{\sum_{i=1}^{n} d_i}{2} + \frac{1}{\gamma}}\right), \tag{12}$$

em que $d_i = \left[y_i - \left(\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j\right)\right]^2$.
(c) distribuição condicional completa *a posteriori* para α_j:

$$\alpha_j|\mu, \sigma^2, \boldsymbol{\alpha_{j^-}}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{n} x_{ij1} h_i}{\sum_{i=1}^{n} x_{ij1}^2 + V_j}, \frac{\sigma^2}{\sum_{i=1}^{n} x_{ij1}^2 + V_j}\right), \tag{13}$$

em que $h_i = y_i - \left(\mu + \sum_{k\neq j}^{p} x_{ik1}\alpha_k + \sum_{j=1}^{p} x_{ij2}\delta_j\right)$, $\boldsymbol{\alpha_{j^-}}$ representa todos os elementos de $\boldsymbol{\alpha}$ exceto α_j e $V_j = \frac{\sigma^2}{\upsilon_j^2}$.
(d) distribuição condicional completa *a posteriori* para δ_j:

$$\delta_j|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta_{j^-}}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{n} x_{ij2} k_i}{\sum_{i=1}^{n} x_{ij2}^2 + U_j}, \frac{\sigma^2}{\sum_{i=1}^{n} x_{ij2}^2 + U_j}\right), \tag{14}$$

em que $k_i = y_i - \left(\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{k\neq j}^{p} x_{ik2}\delta_k\right)$, $\boldsymbol{\delta_{j^-}}$ representa todos os elementos de $\boldsymbol{\delta}$ exceto δ_j e $U_j = \frac{\sigma^2}{\tau_j^2}$.
(e) distribuição condicional completa *a posteriori* para υ_j^2:

$$\upsilon_j^2|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2_{j^-}}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim \text{InvGauss}\left(\sqrt{\frac{\lambda}{\alpha_j^2}}, \lambda^2\right), \tag{15}$$

em que $\boldsymbol{\upsilon}^{\mathbf{2}}_{\boldsymbol{j}-}$ representa todos os elementos de $\boldsymbol{\upsilon}^{\mathbf{2}}$ exceto υ_j^2 e InvGauss é a notação para Inverse Gaussian (Inversa Gaussiana).
(f) distribuição condicional completa *a posteriori* para τ_j^2:

$$\tau_j^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon}^{\mathbf{2}},\boldsymbol{\tau}^{\mathbf{2}}_{\boldsymbol{j}-},\lambda^2,\lambda_1^2,\mathbf{y}\sim \text{InvGauss}\left(\sqrt{\frac{\lambda_1}{\delta_j^2}},\lambda_1^2\right), \tag{16}$$

em que $\boldsymbol{\tau}^{\mathbf{2}}_{\boldsymbol{j}-}$ representa todos os elementos de $\boldsymbol{\tau}^{\mathbf{2}}$ exceto τ_j^2.
(g) distribuição condicional completa *a posteriori* para λ^2:

$$\lambda^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon}^{\mathbf{2}},\boldsymbol{\tau}^{\mathbf{2}}_{\boldsymbol{j}},\lambda_1^2,\mathbf{y}\sim \text{G}\left(p+a_1,\sum_{j=1}^{p}\frac{\upsilon_j^2}{2}+b_1\right). \tag{17}$$

(h) distribuição condicional completa *a posteriori* para λ_1^2:

$$\lambda_1^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon}^{\mathbf{2}},\boldsymbol{\tau}^{\mathbf{2}}_{\boldsymbol{j}},\lambda^2,\mathbf{y}\sim \text{G}\left(p+a_2,\sum_{j=1}^{p}\frac{\tau_j^2}{2}+b_2\right), \tag{18}$$

em que $G(a,b)$ representa a função de densidade de probabildiade gama com $E(X)=a/b$ e $Var(X)=a/b^2$.

APÊNDICE B - ATUALIZAÇÃO DOS PARÂMETROS DO MODELO II

A atualização dos parâmetros $\mu, \sigma^2, \boldsymbol{\alpha}$ e $\boldsymbol{\delta}$ foi feita como no Modelo I e descreve-se no item (i) o procedimento de atualização necessário para obtenção da amostra da distribuição conjunta *a posteriori* para os demais parâmetros.

(i) Atualização de $\boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \boldsymbol{\lambda} = (\lambda_1, ...\lambda_p)$ e $\boldsymbol{\lambda_1} = (\lambda_{1_1}, ..., \lambda_{1_p})$: as distribuições condicionais completas *a posteriori* de cada um dos parâmetros $\upsilon_j^2, \tau_j^2, \lambda_j$ e λ_{1_j} foram obtidas a partir da expressão da distribuição conjunta 10 e são apresentadas a seguir.

(a) a distribuição condicional completa *a posteriori* para $\boldsymbol{\upsilon^2}$ foi obtida individualmente para cada υ_j^2 e é especificada pela expressão:

$$\upsilon_j^2|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2_{j^-}}, \boldsymbol{\tau^2}, \boldsymbol{\lambda}, \boldsymbol{\lambda_1}, \mathbf{y} \propto \frac{1}{\sqrt{\upsilon_j^2}\left[\lambda_j^2 + (\upsilon_j^2)^2\right]} \times \exp\left\{-\frac{1}{2\upsilon_j^2}\alpha_j^2\right\}, \qquad (19)$$

em que $\boldsymbol{\upsilon^2_{j^-}}$ representa todos os elementos de $\boldsymbol{\upsilon^2}$ exceto υ_j^2.

(b) a distribuição condicional completa *a posteriori* para $\boldsymbol{\tau^2}$ foi obtida individualmente para cada τ_j^2 e é especificada pela expressão:

$$\tau_j^2|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2_{j^-}}, \boldsymbol{\lambda}, \boldsymbol{\lambda_1}, \mathbf{y} \propto \frac{1}{\sqrt{\tau_j^2}\left[\lambda_{1_j}^2 + (\tau_j^2)^2\right]} \times \exp\left\{-\frac{1}{2\tau_j^2}\delta_j^2\right\}, \qquad (20)$$

em que $\boldsymbol{\tau^2_{j^-}}$ representa todos os elementos de $\boldsymbol{\tau^2}$ exceto τ_j^2.

(c) a distribuição condicional completa *a posteriori* para $\boldsymbol{\lambda}$ foi obtida individualmente para cada $\boldsymbol{\lambda_j}$ e é especificada pela expressão:

$$\lambda_j|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau_j^2}, \lambda_{j^-}, \lambda_{1_j}, \mathbf{y} \propto \frac{1}{(\phi^2 + \lambda_j^2)} \times \frac{\lambda_j}{\lambda_j^2 + (\upsilon_j^2)^2} \qquad (21)$$

em que λ_{j^-} representa todos os elementos de $\boldsymbol{\lambda}$ exceto λ_j

(d) a distribuição condicional completa *a posteriori* para $\boldsymbol{\lambda_1}$ foi obtida individualmente para cada λ_{1_j} e é especificada pela expressão:

$$\lambda_{1_j}|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau_j^2}, \lambda_j, \lambda_{1_{j^-}}, \mathbf{y} \propto \frac{1}{(\phi_1^2 + \lambda_{1_j}^2)} \times \frac{\lambda_{1_j}}{\lambda_{1_j}^2 + (\tau_j^2)^2} \qquad (22)$$

em que $\lambda_{1_{j^-}}$ representa todos os elementos de $\boldsymbol{\lambda_1}$ exceto λ_{1_j}

Para os parâmetros $\boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \boldsymbol{\lambda}$ e $\boldsymbol{\lambda_1}$, utilizou-se $G(a, b)$ como distribuição candidata, em que $G(a, b)$ representa a função de densidade de probabildiade gama com $E(X) = a/b$ e $Var(X) = a/b^2$. Os valores atribuídos para a e b quando foram gerados os valores propostos para os parâmetros $\boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}$ foram 0.1 e 10, respectivamente. Já para os parâmetros $\boldsymbol{\lambda}$ e $\boldsymbol{\lambda_1}$, $a = 10$ e $b = 0.1$