Universidade de São Paulo
Escola Superior de Agricultura “Luiz de Queiroz”

**Uma abordagem bayesiana para mapeamento de QTLs em
populações experimentais**

Andréia da Silva Meyer

Tese apresentada para obtenção do título de Doutor em
Agronomia. Área de concentração: Estatística e Experi-
mentação Agronômica

Piracicaba
2009

Andréia da Silva Meyer
Licenciada em Matemática

**Uma abordagem bayesiana para mapeamento de QTLs em
populações experimentais**

Orientadora:

Profa. Dra. **ROSELI APARECIDA LEANDRO**

Tese apresentada para obtenção do título de Doutor em
Agronomia. Área de concentração: Estatística e Experi-
mentação Agrônômica

Piracicaba
2009

**Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP**

Meyer, Andréia da Silva
Uma abordagem bayesiana para mapeamento de QTLs em populações experimentais /
Andréia da Silva Meyer. - - Piracicaba, 2009.
129 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura Luiz de Queiroz, 2009.
Bibliografia.

1. Genética estatística 2. Inferência bayesiana 3. Mapeamento genético 4. Método de
Monte Carlo I. Título

CDD 519.542
M612a

“Permitida a cópia total ou parcial deste documento, desde que citada a fonte – O autor”

DEDICATÓRIA

Aos meus pais,
Gerson e Gersina, pelo exemplo de ser humano que são
e por tudo que me ensinaram.

As minhas irmãs,
Claudia e Luciane, por me ensinarem o sentido real da
palavra amor.

Aos meus príncipes,
André, Vinícius e Vítor e a minha princesa **Vitória**, por
encherem minha vida de brilho.

AGRADECIMENTOS

A Deus, pela força para a realização desse trabalho.

A Profa. Dra. Roseli Aparecida Leandro pela orientação dedicada, competente e paciente; pela amizade e apoio durante todo o curso e principalmente pela confiança em mim depositada.

Ao Prof. Dr. Antonio Augusto Franco Garcia, pelo apoio e pelas valiosíssimas contribuições para o desenvolvimento deste trabalho.

Aos Professores Doutores Antonio Augusto Franco Garcia, Júlio Sílvio de Sousa Bueno Filho e Sílvio Sandoval Zocchi pelas importantes sugestões dadas no meu exame de qualificação.

Ao Conselho Nacional de Desenvolvimento Científico- CNPq, pela concessão de bolsa de estudos.

Aos meus amados pais Gerson e Gersina, irmãs Claudia e Luciane, cunhados, André e Nicéio e sobrinhos, André, Vinícius, Vítor e Vitória por todo apoio, amor, incentivo, alegrias e paciência com a minha ausência neste período.

Ao Carlos Creppe Jr. pelo carinho, incentivo e paciência nas horas difíceis e também com a minha ausência.

A querida amiga Giovana Oliveira Silva, pelo companheirismo em todas as fases do doutorado, por ter sempre acreditado em mim e principalmente pela amizade tão preciosa.

As minhas queridas amigas Elizabeth Hashimoto, Joseane Padilha e Juliana Fachini pelo convívio embasado em companheirismo, pela amizade incondicional e suporte emocional.

A minha querida amiga Solange Sabadin pelo carinho, apoio, constantes incentivos e por ter me acolhido em sua casa assim que cheguei à Piracicaba.

A minha família, em especial a minha prima Paula e aos meus tios José, Eunice e Quitéria pelos constantes incentivos.

As minhas amigas Alice, Denise, Fernanda e Silvana pelo apoio e constantes incentivos.

Ao Prof. Dr. José Cláudio Faria, da Universidade Estadual Santa Cruz - UESC e ao técnico de laboratório do Departamento de Ciências Exatas da ESALQ/USP Jorge Alexandre Wiendl pelas valiosas sugestões no uso do programa R.

Ao Marcelo Mollinari pelo auxílio e pela gentil cessão do programa para a simulação dos dados.

Aos professores do Departamento de Ciência Exatas da ESALQ/USP pela amizade e formação, em especialmente à Profa. Dra. Clarice Garcia Borges Demétrio por todos os ensinamentos.

Aos funcionários do Departamento de Ciência Exatas da ESALQ/USP, em especial a Solange Sabadin, Luciane Brajão, Jorge Wiendl e Eduardo Bonilha, pelos auxílios permanentes.

Aos amigos dos cursos de mestrado e doutorado do Departamento de Ciência Exatas da ESALQ/USP, em especial a Afrânio, César, Júlio, Osmar, Sandra, Ana Maria, Idemauro, Ana Paula a atenção, a compreensão e a amizade.

A Escola Superior de Agricultura "Luiz de Queiroz", pela oportunidade de realização do meu doutorado.

A todos que de alguma forma contribuíram para a realização deste trabalho.

SUMÁRIO

RESUMO	8
ABSTRACT	9
1 INTRODUÇÃO	10
2 REVISÃO BIBLIOGRÁFICA	13
2.1 Noções de inferência bayesiana	13
2.1.1 Distribuição a priori	14
2.1.2 Monte Carlo com Cadeias de Markov	15
2.1.3 Monte Carlo com Cadeias de Markov e Saltos Reversíveis	17
2.1.4 Espaço composto e MCMC para seleção de variáveis	18
2.1.5 Método de seleção de modelos	19
2.2 Fundamentos teóricos do mapeamento genético	20
2.2.1 Considerações iniciais	20
2.2.2 Populações utilizadas no mapeamento genético	21
2.2.3 Fração de recombinação e distância genética	21
2.3 Métodos bayesianos para mapear QTLs	22
2.3.1 Modelo QTL	22
2.3.2 Construção da função de verossimilhança	23
2.3.3 Prioris	25
2.3.4 Estimação dos parâmetros	26
Referências	30
3 UMA ABORDAGEM BAYESIANA PARA O MAPEAMENTO DE QTLs	35
Resumo	35
Abstract	35
3.1 Introdução	36
3.2 Metodologia	38
3.2.1 Modelo QTL	38
3.2.2 Estimação dos parâmetros	39
3.3 Estudo de simulação	45
3.3.1 Resultados do estudo de simulação	45
3.4 Discussão	73
3.5 Conclusão	74
Referências	74
4 MAPEAMENTO DE QTLs EM UMA POPULAÇÃO DE MILHO TROPICAL USANDO A ABORDAGEM BAYESIANA	77
Resumo	77

Abstract	77
4.1 Introdução	78
4.2 Material e Métodos	80
4.2.1 Dados fenotípicos e moleculares	80
4.2.2 Mapeamento de QTLs	80
4.3 Resultados	83
4.3.1 Produção de grãos	83
4.3.2 Altura da espiga	90
4.3.3 Altura da planta	98
4.4 Discussão	106
4.5 Conclusão	107
Referências	107
APÊNDICES	111

RESUMO

Uma abordagem bayesiana para mapeamento de QTLs em populações experimentais

Muitos caracteres em plantas e animais são de natureza quantitativa, influenciados por múltiplos genes. Com o advento de novas técnicas moleculares tem sido possível mapear os locos que controlam os caracteres quantitativos, denominados QTLs (*Quantitative Trait Loci*). Mapear um QTL significa identificar sua posição no genoma, bem como, estimar seus efeitos genéticos. A maior dificuldade para realizar o mapeamento de QTLs, se deve ao fato de que o número de QTLs é desconhecido. Métodos bayesianos juntamente com método Monte Carlo com Cadeias de Markov (MCMC), têm sido implementados para inferir conjuntamente o número de QTLs, suas posições no genoma e os efeitos genéticos. O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado desconhecido e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. No presente trabalho foi implementado, utilizando-se o programa estatístico R uma abordagem bayesiana para mapear QTLs em que múltiplos QTLs e os efeitos de epistasia são considerados no modelo. Para tanto foram ajustados modelos com números crescentes de QTLs e o fator de Bayes foi utilizado para selecionar o modelo mais adequado e conseqüentemente, estimar o número de QTLs que controlam os fenótipos de interesse. Para investigar a eficiência da metodologia implementada foi feito um estudo de simulação em que foram considerados duas diferentes populações experimentais: retrocruzamento e F_2 , sendo que para ambas as populações foi feito o estudo de simulação considerando modelos com e sem epistasia. A abordagem implementada mostrou-se muito eficiente, sendo que para todas as situações consideradas o modelo selecionado foi o modelo contendo o número verdadeiro de QTLs considerado na simulação dos dados. Além disso, foi feito o mapeamento de QTLs de três fenótipos de milho tropical: altura da planta (AP), altura da espiga (AE) e produção de grãos utilizando a metodologia implementada e os resultados obtidos foram comparados com os resultados encontrados pelo método CIM.

Palavras-chave: Mapeamento de QTLs; Inferência bayesiana; Fator de Bayes; MCMC

ABSTRACT

A Bayesian approach for mapping QTL in experimental populations

Many traits in plants and animals have quantitative nature, influenced by multiple genes. With the new molecular techniques, it has been possible to map the loci, which control the quantitative traits, called QTL (Quantitative Trait Loci). Mapping a QTL means to identify its position in the genome, as well as to estimate its genetics effects. The great difficulty of mapping QTL relates to the fact that the number of QTL is unknown. Bayesian approaches used with Markov Chain Monte Carlo method (MCMC) have been applied to infer QTL number, their positions in the genome and their genetic effects. The challenge is to obtain the sample from the joined distribution posterior of these parameters, since the number of QTL may be considered unknown and hence the dimension of the parametric space changes according to the number of QTL in the model. In this study, a Bayesian approach was applied, using the statistical program R, in order to map QTL, considering multiples QTL and epistasis effects in the model. Models were adjusted with the crescent number of QTL and Bayes factor was used to select the most suitable model and, consequently, to estimate the number of QTL that control interesting phenotype. To evaluate the efficiency of the applied methodology, a simulation study was done, considering two different experimental populations: backcross and F2, accomplishing the simulation study for both populations, considering models with and without epistasis. The applied approach resulted to be very efficient, considering that for all the used situations, the selected model was the one containing the real number of QTL used in the data simulation. Moreover, the QTL mapping of three phenotypes of tropical corn was done: plant height, corn-cob height and grain production, using the applied methodology and the results were compared to the results found by the CIM method.

Key words: Bayes factor, Bayesian inference, MCMC, QTL mapping

1 INTRODUÇÃO

Muitos caracteres em plantas e animais são de natureza quantitativa, influenciados por múltiplos genes e, conseqüentemente, pela interação entre eles, a qual é denominada epistasia. Com o advento de novas técnicas moleculares tem sido possível mapear os locos que controlam os caracteres quantitativos, denominados QTLs (*Quantitative Trait Loci*). Mapear um QTL significa identificar sua posição no genoma, isto é, identificar em qual cromossomo está e qual sua localização nesse cromossomo, bem como, estimar seus efeitos genéticos (LYNCH; WALSH, 1998). Dessa forma, é possível um melhor entendimento sobre os caracteres quantitativos, que são de fundamental importância nos programas de melhoramento genético, uma vez que estão associados à produção de grãos, altura de plantas, volume, entre outros, que são de grande importância econômica.

Vários modelos tem sido desenvolvidos para o mapeamento de QTLs, variando desde simples modelos de regressão linear e análise de variância até métodos sofisticados, que modelam todo o genoma simultaneamente. O método mais utilizado no momento pelos geneticistas e melhoristas para mapear QTLs é o Mapeamento por Intervalo Composto (CIM) que foi proposto, de forma independente, por Jansen e Stam (1994) e Zeng (1994), este método combina mapeamento por intervalo e regressão linear múltipla e foi proposto com o objetivo de evitar que um QTL presente em qualquer posição do genoma interfira no mapeamento do QTL no intervalo considerado. Isto é feito incluindo no modelo de regressão múltipla a informação de marcadores adjacentes ao intervalo considerado e de marcadores presentes em outros grupos de ligação sob a forma de cofatores. Contudo, o método CIM não permite o mapeamento simultâneo de múltiplos QTLs e conseqüentemente também não permite a análise de epistasias. Visando poder realizar o mapeamento simultâneo de múltiplos QTLs duas alternativas são encontradas na literatura: o Mapeamento por Intervalos Múltiplos (MIM), proposto por Kao e Zeng (1997) e Kao et al. (1999) e as metodologias baseadas em inferência bayesiana por meio de métodos MCMC (*Markov Chain Monte Carlo*) (SATAGOPAN et al., 1996; HEATH, 1997; SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998; SEN; CHURCHILL, 2001; GAFFNEY, 2001; WANG et al., 2005; YI, 2004; YI et al., 2005, YI et al., 2007). O Mapeamento por Intervalos Múltiplos (MIM) combina a possibilidade de mapeamento com o estudo da arquitetura genética dos caracteres quantitativos, a partir de um algoritmo de busca do número, posição, efeitos e interações entre os QTLs.

Os métodos bayesianos utilizam a distribuição conjunta a posteriori para fazer inferências sobre os parâmetros de interesse. Esta distribuição é obtida por meio do teorema de Bayes, que combina a função de verossimilhança com a incerteza relativa aos parâmetros. Assim, tal abordagem tem a vantagem de permitir incorporar na análise informações adicionais baseadas em experimentos anteriores ou conhecimentos de pesquisadores, por meio da atribuição de uma distribuição a pri-

ori para cada parâmetro de interesse. Um dos trabalhos pioneiros utilizando essa abordagem foi apresentado por Satagopan et al. (1996) e desde então a abordagem bayesiana vem ganhando espaço nas descobertas de novas metodologias para realizar o mapeamento de QTLs. O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado como desconhecido e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. Satagopan et al. (1996) ajustou modelos com diferentes números de QTLs e, posteriormente, fez a seleção do modelo mais adequado usando como critério de seleção o fator de Bayes. Trabalhos posteriores (SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998) tratam o número de QTLs como uma quantidade desconhecida e usam o método MCMC com saltos reversíveis apresentado por Green (1995), o qual permite a movimentação entre modelos de dimensões diferentes. O algoritmo MCMC com saltos reversíveis é uma técnica bastante poderosa e geral para explorar as distribuições a posteriori quando o espaço paramétrico não possui dimensão fixa. Contudo, essa habilidade de movimento entre modelos de dimensões diferentes requer a construção de distribuições propostas complexas, o que dificulta a implementação desse algoritmo. Além disso, a complexidade desse algoritmo aumenta a demanda computacional, o que pode comprometer seu desempenho. Trabalhos recentes têm aplicado o espaço composto, proposto por Godsil (2001), no contexto de mapeamento bayesiano de QTLs (YI, 2004; YI et al., 2005; YI et al., 2007). O espaço composto considera a presença de um número máximo de QTLs (S) e utiliza uma variável aleatória binária, γ_{jk} , para indicar inclusão do k -ésimo efeito genético do j -ésimo QTL, ($\gamma_{jk} = 1$) ou exclusão ($\gamma_{jk} = 0$) no modelo, em que $j = 1, \dots, S$ e $k = 1, \dots, K$ e K representa o número de efeitos genéticos. Dessa forma, o espaço paramétrico permanece fixo e a obtenção da amostra da distribuição a posteriori conjunta pode ser realizada utilizando os métodos MCMC tradicionais (YI, 2004).

Outra tarefa desafiadora é a inclusão de efeitos de epistasia no modelo pois, sua inclusão, implica em aumento ou diminuição considerável da dimensão do espaço paramétrico. Com isso, aumenta a complexidade da implementação dos algoritmos utilizados na construção das amostras da distribuição conjunta a posteriori para os parâmetros de interesse. Particularmente, para os algoritmos MCMC com saltos reversíveis, em que cada movimentação entre modelos de dimensões diferentes implica na retirada ou inclusão de muitos parâmetros. Apesar dos desafios da implementação, o algoritmo MCMC com saltos reversíveis tem sido empregado para estudar epistasia (YI; XU, 2002, YI et al., 2003, NARITA; SASAKI, 2004). Contudo, esse método de simulação tem apresentado uma convergência muito lenta e uma demanda computacional intensa, tornando seu emprego não muito recomendável (YI et al., 2005). Assim, se faz necessário o estudo de metodologias bayesianas para mapear QTLs que seja ao mesmo tempo eficiente e de fácil entendimento, principalmente, quando são considerados modelos mais complexos, como é o caso dos

modelos que incluem termos de epistasias. Além disso, existe a necessidade de implementar computacionalmente esses algoritmos de maneira que sejam acessíveis a todos os pesquisadores que tenham interesse no assunto, para que assim haja uma maior difusão da metodologia bayesiana no mapeamento de QTLs.

Diante do exposto, o presente trabalho teve como objetivo a implementação de um programa acessível a todos os pesquisadores, utilizando a metodologia bayesiana para mapear QTLs no caso em que efeitos de epistasia são incorporados ao modelo. Para tanto, para inclusão de efeito de epistasia no modelo utilizou-se o modelo epistático de Cockerham, como apresentado em Kao e Zeng, (2002). O número de QTLs foi considerado como conhecido e o fator de Bayes foi utilizado para selecionar o modelo mais adequado. Para verificar a eficiência da metodologia implementada foi feito um estudo de simulação considerando-se dois tipos de populações experimentais: retrocruzamento e F_2 . Além disso, foi feito o mapeamento de QTLs considerando-se três fenótipos de milho tropical: altura da planta (AP), altura da espiga (AE) e produção de grãos utilizando a metodologia implementada e os resultados obtidos foram comparados com os resultados encontrados pelo método CIM apresentado em Sibov et al. (2003b).

2 REVISÃO BIBLIOGRÁFICA

2.1 Noções de inferência bayesiana

A inferência bayesiana, assim como a freqüentista, trabalha na presença de observações $\mathbf{y}$, que podem ser descritas por uma distribuição de probabilidades $f(\mathbf{y}|\boldsymbol{\theta})$, sendo $\boldsymbol{\theta}$ uma quantidade desconhecida e necessária para descrever $\mathbf{y}$, ou seja, essa quantidade $\boldsymbol{\theta}$, denominada vetor de parâmetros, pode ser vista como um indexador para a possível família de distribuições $\mathfrak{F}$, que descreve de forma adequada o processo que gera as observações $\mathbf{y}$. Pode-se dizer que um dos principais objetivos da estatística é fazer inferências sobre $\boldsymbol{\theta}$. É bastante provável que o pesquisador tenha algum conhecimento a priori sobre esta quantidade, e é então possível, que esse conhecimento seja formalmente incorporado na análise. Neste ponto, os métodos bayesianos e freqüentista se diferenciam. Os freqüentistas não admitem a incorporação dessas informações a priori na análise, por elas não terem sido observadas, e com isso não podem ser verificadas empiricamente, já as metodologias bayesianas incorporam as informações a priori na análise selecionando uma distribuição de probabilidade para $\boldsymbol{\theta}$. Isso pode ser visto como uma das vantagens desta abordagem, porém, também é um ponto de grandes controvérsias entre os estatísticos (GAMERMAN; LOPES, 2006).

Portanto, a abordagem bayesiana requer duas fontes de informação para fazer inferência sobre o parâmetro: a informação sobre os dados, a qual é incorporada na análise através da função de verossimilhança, e uma distribuição a priori para os parâmetros de interesse. Essas duas informações resultam em uma distribuição de probabilidade a posteriori, a qual é usada para fazer inferência sobre os parâmetros do modelo, bem como, para prever novas observações. (CARLIN; LOUIS, 1996; GAMERMAN; LOPES, 2006). A distribuição a posteriori $\pi(\boldsymbol{\theta}|\mathbf{y})$ é obtida por meio do teorema de Bayes que fornece uma regra de atualização de probabilidades, que parte de $\pi(\boldsymbol{\theta})$ e resulta na posteriori $\pi(\boldsymbol{\theta}|\mathbf{y})$ da seguinte forma:

$$\pi(\boldsymbol{\theta}|\mathbf{y}) = \frac{L(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta})}{f(\mathbf{y})}, \quad (2.1)$$

em que

$$f(\mathbf{y}) = \int_{\boldsymbol{\theta}} f(\mathbf{y}, \boldsymbol{\theta}) d\boldsymbol{\theta} = \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}, \text{ para o caso contínuo}$$

$$f(\mathbf{y}) = \sum_{\boldsymbol{\theta}} f(\mathbf{y}, \boldsymbol{\theta}) = \sum_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y})\pi(\boldsymbol{\theta}), \text{ para o caso discreto}$$

e $L(\boldsymbol{\theta}|\mathbf{y})$ é a função de verossimilhança.

Nota-se que $f(\mathbf{y})$ não depende de θ e, portanto, para a determinação da distribuição a posteriori representa apenas uma constante. Assim, da equação (1), observa-se que $\pi(\theta|\mathbf{y})$ é proporcional à função de verossimilhança multiplicada pela priori:

$$\pi(\theta|\mathbf{y}) \propto L(\theta|\mathbf{y})\pi(\theta).$$

A proporcionalidade pode ser usada pelo fato de que quando se multiplica a função de verossimilhança por uma constante não se altera a informação relativa ao parâmetro θ e, assim, a distribuição a posteriori não será alterada (BOX; TIAO, 1992).

2.1.1 Distribuição a priori

As informações sobre o vetor de parâmetros de interesse, θ , que são obtidas antes da realização de um experimento podem ser representadas probabilisticamente através das chamadas distribuições a priori. Portanto, essas distribuições expressam o estado de conhecimento ou ignorância sobre o vetor de parâmetros θ antes de os dados serem obtidos (BOX; TIAO, 1992). Serão discutidas, a seguir, algumas formas de especificação de distribuição a priori.

Distribuições a priori subjetivas

A distribuição a priori subjetiva é usada quando não existe nenhuma base teórica objetiva para o ajuste do modelo. Nessas situações, a distribuição a priori representa a expectativa subjetiva do pesquisador a respeito do parâmetro de interesse. Com isso, a distribuição a posteriori para o parâmetro, obtida combinando essa distribuição a priori e a função de verossimilhança, mostra como o conhecimento pessoal do pesquisador é modificado pela informação dos dados (BOX; TIAO, 1992). O problema de como quantificar essa informação subjetiva e transformá-la em uma distribuição a priori que possa ser usada na abordagem bayesiana, tem sido discutido abundantemente na literatura. Uma interação entre o pesquisador e o estatístico é fundamental para a determinação de tal priori.

Distribuições a priori conjugadas

A idéia de distribuição conjugada é que as distribuições a priori e a posteriori pertençam a mesma classe de distribuições e, com isso, a atualização do conhecimento que se tem sobre o parâmetro θ envolve apenas uma mudança nos parâmetros indexadores da família de distribuições a priori, denominados hiperparâmetros, que diferem do parâmetro θ .

Seja $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$ um vetor de observações que são independentes, identicamente distribuídas e supondo-se que a distribuição de y_i pertença a família exponencial, a distribuição con-

junta para $\mathbf{y}$ é dada por:

$$f(\mathbf{y}|\theta) = \prod_{i=1}^n \exp(a(\theta)b(y_i) + c(\theta) + d(y_i))$$

e a correspondente função de verossimilhança por:

$$L(\theta|\mathbf{y}) \propto \exp\{a(\theta) \sum_{i=1}^n b(y_i) + nc(\theta)\}.$$

Supondo a distribuição a priori conjugada para θ dada por:

$$\pi(\theta; k_1, k_2) \propto \exp\{k_1 a(\theta) + k_2 c(\theta)\},$$

obtém-se a seguinte distribuição a posteriori:

$$\pi(\theta|\mathbf{y}) \propto \exp\{a(\theta) [\sum_{i=1}^n b(y_i) + k_1] + c(\theta) [n + k_2]\}.$$

Gamerman e Lopes (2006) ressaltam que a distribuição a priori conjugada deve ser usada com cuidado, pois sua utilização está muitas vezes associada às facilidades analíticas e nem sempre é uma representação adequada do conhecimento a priori do parâmetro.

Distribuições a priori não informativas

Existem situações em que não existem informações a priori sobre o parâmetro de interesse, ou que a informação disponível é pouco significativa. Este problema pode ser contornado usando as distribuições a priori não-informativas e com isso, a informação resultante da distribuição a posteriori será apenas referente aos dados (CARLIN; LOUIS, 1996).

Uma forma de atribuir distribuição a priori não-informativa é atribuir distribuições de forma que, todos os possíveis valores para um dado parâmetro, tenham a mesma chance de ocorrer. A idéia inicial é utilizar a distribuição uniforme para representar estas situações, e assim, $\pi(\theta) \propto \text{constante}$. Contudo, alguns problemas podem ocorrer com esta escolha, $\pi(\theta)$ é imprópria, isto é, $\int \pi(\theta) \rightarrow \infty$. Jeffreys (1961) propõe uma classe de prioris não informativas invariantes, porém, eventualmente imprópria, a qual se baseia na informação de Fisher.

2.1.2 Monte Carlo com Cadeias de Markov

Na abordagem bayesiana as informações referentes aos parâmetros de interesse são obtidas por meio da distribuição marginal a posteriori. A obtenção da distribuição marginal a posteriori, na maioria das vezes, envolvem resoluções de integrais complexas. Além disso, cálculos dos momentos da distribuição marginal a posteriori, construção de intervalos de credibilidade e regiões

HPD também, na maioria da vezes, envolvem resoluções de integrais complexas. Nestes casos, é possível recorrer aos métodos de Monte Carlo com Cadeias de Markov.

Os métodos Monte Carlo com Cadeias de Markov (MCMC) são uma importante ferramenta computacional dentro da estatística bayesiana. Estes métodos são baseados na simulação de Cadeias de Markov cuja distribuição estacionária $g(\boldsymbol{\theta})$ converge para a distribuição a posteriori de interesse $\pi(\boldsymbol{\theta}|\mathbf{y})$, da qual não é possível gerar diretamente uma amostra (CARLIN; LOUIS, 1996).

Assim, tem-se que os métodos MCMC fazem a associação entre os algoritmos para simulação de distribuições e o método Monte Carlo para aproximações de integrais. São apresentados a seguir, dois algoritmos que permitem gerar Cadeias de Markov cuja distribuição de equilíbrio seja $\pi(\boldsymbol{\theta}|\mathbf{y})$: Metropolis-Hastings e um caso particular deste: Gibbs sampling.

Metropolis-Hastings

O algoritmo Metropolis-Hastings (CHIB; GREENBERG, 1995) é um dos algoritmos usados para gerar amostras da distribuição conjunta a posteriori $\pi(\boldsymbol{\theta}|\mathbf{y})$. Tal algoritmo, é empregado, principalmente, em situações nas quais a distribuição condicional completa a posteriori do parâmetro não possuem forma fechada. Para implementação deste método é necessário considerar uma distribuição $q(\boldsymbol{\theta}, \boldsymbol{\phi})$

Seja a Cadeia de Markov $(\theta^0, \theta^1, \dots, \theta^t, \dots)$, e seja t o estado atual da cadeia. O algoritmo de Metropolis-Hastings gera um valor candidato para o próximo estado da cadeia $(t+1)$, denominado ϕ , utilizando uma distribuição proposta $q(\phi|\theta^t)$. Daí tem-se que este candidato pode ser aceito com probabilidade $\alpha(\theta^t, \phi)$ ou rejeitado com probabilidade $1 - \alpha(\theta^t, \phi)$, em que:

$$\alpha(\theta^t, \phi) = \min \left\{ 1, \frac{\pi(\phi|\mathbf{y})q(\theta^t|\phi)}{\pi(\theta^t|\mathbf{y})q(\phi|\theta^t)} \right\}.$$

Se o valor candidato ϕ foi aceito, então o próximo estado da cadeia θ^{t+1} será igual a ϕ , caso contrário a cadeia não se moverá, ou seja, $\theta^{t+1} = \theta^t$

Em termos práticos, o algoritmo de Metropolis-Hastings pode ser especificado pelos seguintes passos:

- i) inicie o contador de iterações em $t = 0$;
- ii) especifique um valor inicial para θ^0 ;
- iii) gere um valor candidato, ϕ , ao próximo estado a partir da distribuição proposta $q(\phi|\theta^t)$;
- iv) calcule a probabilidade de aceitação $\alpha(\theta^t, \phi)$;
- v) gere u a partir de uma distribuição $U(0, 1)$;

- vi) se $u < \alpha$, então, aceite o valor candidato, e conseqüentemente, faça $\theta^{t+1} = \phi$, caso contrário $\theta^{t+1} = \theta^t$;
- vii) incremente o contador de t para $t + 1$ e volte ao passo (iii).

Gibbs Sampling

Este algoritmo foi introduzido como uma ferramenta MCMC, no contexto de restauração de imagem por Geman e Geman (1984). Ele permite gerar amostras de uma distribuição conjunta a posteriori $\pi(\theta_1, \theta_2, \dots, \theta_p | \mathbf{y})$ a partir das distribuições de cada parâmetro condicionada aos demais parâmetros do modelo, que são denominadas distribuições condicionais completas a posteriori $\pi(\theta_i | \theta_1, \dots, \theta_{i-1}, \dots, \theta_{i+1}, \dots, \theta_p, \mathbf{y})$, desde que essas distribuições condicionais completas possuam formas fechadas (CASELLA; GEORGE, 1992; GELFAND, 2000). O procedimento geral para a execução deste algoritmo se dá da seguinte forma:

- i) iniciar o contador de iterações da cadeia, $t = 0$;
- ii) atribuir valores iniciais para $\theta = (\theta_1^0, \theta_2^0, \dots, \theta_p^0)$;
- iii) obter um novo vetor $\theta^t = (\theta_1^t, \dots, \theta_p^t)$ a partir de θ^{t-1}

$$\theta_1^t \sim \pi(\theta_1 | \theta_2^{t-1}, \theta_3^{t-1}, \dots, \theta_p^{t-1}, \mathbf{y}),$$

$$\theta_2^t \sim \pi(\theta_2 | \theta_1^t, \theta_3^{t-1}, \dots, \theta_p^{t-1}, \mathbf{y}),$$

...

$$\theta_p^t \sim \pi(\theta_p | \theta_1^t, \theta_2^t, \dots, \theta_{p-1}^t, \mathbf{y})$$

- iv) incrementar o contador de t para $t + 1$ e retornar ao passo iii.

Um fator importante a ser considerado, é que os vetores θ^t , que vão sendo gerados, são correlacionados. Assim, surge o problema em como utilizar as realizações da cadeia para fazer inferências sobre θ . Uma alternativa é gerar uma cadeia longa e utilizar vetores espaçados de modo a anular o efeito da correlação. Este espaçamento é denominado de “thin”.

Nota-se que o Gibbs sampling é um caso especial do Metropolis-Hastings, em que distribuição condicional completa é usada como distribuição proposta e, assim, a probabilidade de aceitação é sempre igual a um.

2.1.3 Monte Carlo com Cadeias de Markov e Saltos Reversíveis

Em muitas situações é razoável assumir que existem incertezas em relação ao modelo que melhor se ajuste aos dados. Do ponto de vista bayesiano, o modelo mais adequado pode ser visto, como mais um parâmetro a ser estimado. Para isso cria-se uma variável discreta k que funciona como indicadora do modelo e atribui-se probabilidade a priori $\pi(k)$ para cada modelo k . Para cada k

existe um vetor de parâmetro associado $\theta_k \in \mathbf{R}^{n_k}$, em que a dimensão n_k varia do modelo k para o modelo k' , sendo $k \neq k'$. Assim, dada a função de verossimilhança $L(k, \theta_k | \mathbf{y})$ e a distribuição a priori para θ_k , a distribuição a posteriori para os parâmetros dadas as observações pode ser escrita como:

$$\pi(k, \theta_k | \mathbf{y}) \propto L(k, \theta_k | \mathbf{y}) \pi(\theta_k | k) \pi(k).$$

Como a dimensão de θ_k varia de acordo com k , os algoritmos de Metropolis-Hastings e Gibbs sampling não podem ser usados para gerar amostras da distribuição conjunta a posteriori, pois esses algoritmos são definidos apenas para os casos em que as dimensões são fixas.

Uma alternativa nesse caso é usar o algoritmo MCMC com saltos reversíveis proposto por Green (1995), que pode ser visto como uma generalização do algoritmo de Metropolis-Hastings, o qual introduz movimentos entre espaços paramétricos de diferentes dimensionalidades, enquanto mantém a equação de equilíbrio detalhada, suficiente para garantir a irreversibilidade da Cadeia de Markov (LOPES, 2002).

Seja $J(k \rightarrow k')$ a probabilidade de propor um movimento de um espaço paramétrico de dimensão k para outro de dimensão k' e denotando ϕ como o vetor de parâmetros que estão presentes e tem o mesmo significado em qualquer dimensionalidade, M_k como sendo o modelo corrente de dimensão k , um possível esquema do algoritmo é dado a seguir:

- i) Propor um movimento ao modelo $M_{k'}$ com probabilidade $J(k \rightarrow k')$;
- ii) amostrar $\mathbf{u}$ de uma densidade proposta $q(\mathbf{u} | \theta_k, k, k')$;
- iii) Definir $(\theta_{k'}, \mathbf{u}') = g_{k,k'}(\theta_k, \mathbf{u})$, em que $g_{k,k'}(\cdot)$ é uma bijeção entre $(\theta_k, \mathbf{u})$ e $(\theta_{k'}, \mathbf{u}')$, e $\mathbf{u}$ e $\mathbf{u}'$ têm o papel de emparelhar as dimensões dos vetores θ_k e $\theta_{k'}$.
- iv) A probabilidade de aceitação do novo modelo $(k', \theta_{k'})$ é dada por $\alpha = \min(1, A)$ sendo:

$$A = \frac{\pi(k', \theta_{k'} | \phi) q(u' | \theta_{k'}, k', k) J(k' \rightarrow k)}{\pi(k, \theta_k | \phi) q(u | \theta_k, k, k') J(k \rightarrow k')} \left| \frac{\partial g_{k,k'}(\theta_k, u)}{\partial(\theta_k, u)} \right|$$

O termo referente ao Jacobiano possibilita a mudança entre $(\theta_k, \mathbf{u})$ e $(\theta_{k'}, \mathbf{u}')$.

2.1.4 Espaço composto e MCMC para seleção de variáveis

Outra alternativa para as situações em que existem incertezas em relação a dimensão do espaço paramétrico é a utilização do espaço composto que foi introduzido por Godsill (2001). Este método é uma modificação direta do produto dos espaços, apresentado por Carlin e Chib (1995), o qual considera um espaço que contemple todos possíveis parâmetros envolvidos no modelo e inclui uma variável aleatória para indicar quais parâmetros estão presentes num dado modelo k . Os parâmetros

que foram incluídos no modelo k determinam a função de verossimilhança, e os parâmetros que não foram usados entram na distribuição conjunta a posteriori para os parâmetros em forma de pseudo-prioris, conforme será mostrado seguir.

Dado os K possíveis parâmetros para um determinado modelo, $\theta = (\theta_1, \theta_2, \dots, \theta_K)$, tal que θ_i tem suporte Θ_i . Um conjunto de índices é introduzido para cada modelo, indicando quais parâmetros θ_i são usados pelo modelo k , ou seja, $\mathcal{I}(k) = \{i_1, i_2, \dots, i_k\} \subseteq \{1, 2, \dots, K\}$ e conseqüentemente, estes índices definem a função de verossimilhança, ou seja,

$$L(k, \theta | \mathbf{Y}) = L(k, \theta_{\mathcal{I}(k)} | \mathbf{Y}).$$

em que $\theta_{\mathcal{I}(k)}$ são os parâmetros usados pelo modelo k . Dada as distribuições a priori para os parâmetros $\theta_{i_1}, \theta_{i_2}, \dots, \theta_{i_k}$, tem-se a distribuição conjunta a posteriori:

$$\pi(k, \theta | \mathbf{Y}) = \frac{L(k, \theta_{\mathcal{I}(k)} | \mathbf{Y}) \pi(\theta_{\mathcal{I}(k)} | k) \pi(\theta_{-\mathcal{I}(k)} | \theta_{\mathcal{I}(k)}, k) \pi(k)}{f(\mathbf{Y})},$$

em que $\theta_{-\mathcal{I}(k)}$ denota os parâmetros não usados pelo modelo k . O termo $\pi(\theta_{-\mathcal{I}(k)} | \theta_{\mathcal{I}(k)}, k)$ é denominado como pseudo-priori. O aspecto chave do modelo espaço composto é que o dimensionamento do espaço paramétrico permanece fixo, mesmo quando o indexador do modelo k muda. Assim, o método MCMC padrão, sob as condições usuais de convergência, pode ser aplicado (GODSILL, 2001).

2.1.5 Método de seleção de modelos

Uma forma direta de utilizar os métodos MCMC nos casos de incertezas de modelo é estimar a probabilidade a posteriori para todos os possíveis modelos e, posteriormente, usar um critério para selecionar o melhor modelo. Na literatura, existem diferentes critérios para selecionar o modelo mais adequado. Dentre eles, se destaca o fator de Bayes, que é a razão das verossimilhanças marginais entre dois modelos.

Considerando um determinado modelo M , a verossimilhança marginal é definida por:

$$f(\mathbf{y} | M) = \int L(\boldsymbol{\theta} | \mathbf{y}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}$$

Assim, dados dois modelos, M_1 e M_2 , o fator de Bayes é definido como a razão das verossimilhanças marginais desses dois modelos:

$$B_{12} = \frac{f(\mathbf{y} | M_1)}{f(\mathbf{y} | M_2)}.$$

A interpretação para o Fator de Bayes segue classificação proposta por Jeffreys (1961), como mostra a Tabela 2.1.

Tabela 2.1 – Interpretação do Fator de Bayes Segundo Jeffreys (1961)

B_{12}	$\log(B_{12})$	Evidência a favor do modelo M_1
1 a 3,2	< 0,5	Muito Fraca
3,2 a 10	0,5 a 1	Fraca
10 a 100	1 a 2	Forte
>100	>2	Muito Forte

2.2 Fundamentos teóricos do mapeamento genético

2.2.1 Considerações iniciais

O primeiro artigo a mapear QTLs foi apresentado por Sax (1923), usando como marcador morfológico a cor do tegumento de grãos de feijão. Nesse trabalho, foi encontrada ligação entre genes controlando um caráter qualitativo (cor do tegumento) e a variação quantitativa no peso dos grãos. Posteriormente, Thoday (1961) apresentou os fundamentos sobre como usar informações de locos marcadores ao longo do genoma para o mapeamento de QTLs. Outros trabalhos pioneiros relevantes são os de Elston e Stewart (1973), Soller; Brody e Genizi (1976) e Edwards; Stuber e Wendel (1987). Estes estudos tiveram um grande avanço após o advento dos marcadores moleculares os quais fornecem informações diretamente sobre o material genético (DNA) dos indivíduos em estudo, ou seja, é possível detectar polimorfismo genético diretamente no DNA.

Um marcador de DNA é tipicamente uma pequena região do DNA mostrando polimorfismo (ou diferença) entre indivíduos (LIU, 1998). Dois enfoques metodológicos são usados para detectar tais diferenças: hibridação e amplificação. No primeiro, uma pequena sequência é marcada radioativamente e ligada ao DNA dos indivíduos, usando princípios de pareamento das bases do DNA. Nesse caso, enquadram-se, por exemplo, os marcadores do tipo RFLP (“restriction fragment length polymorphism”). Os marcadores baseados em amplificação têm como princípio o emprego de reações para aumento de sequências específicas, sendo os principais desse tipo o SSR (ou microsátélites), o RAPD (“random amplified polymorphism DNA”) e o AFLP (“amplified fragment length polymorphism”).

Uma vez que os dados de marcadores estão disponíveis, é possível construir um mapa genético. Este mapa genético consiste na representação do cromossomo sobre o qual os marcadores e QTLs estão localizados (DOERGE, 2002). A construção de um mapa genético envolve o ordenamento dos locos dentro de um grupo de ligação, bem como o cálculo das distâncias entre eles (LYNCH; WALSH, 1998).

Os marcadores que estão ligados fazem parte do mesmo grupo de ligação, e assim, é possível localizar QTLs entre estes marcadores. A teoria sobre a construção dos mapas genéticos são bem consolidadas e altamente refinadas (LYNCH; WALSH, 1998). Na prática, o mapa genético não é

denso e os marcadores moleculares não são contínuos, ou seja, tem-se as informações dos genótipos em pontos discretos ao longo do genoma (GAFFNEY, 2001).

2.2.2 Populações utilizadas no mapeamento genético

Um passo muito importante para o mapeamento de QTLs é a seleção de uma população de mapeamento adequada (DOERGE, 2002). As populações derivadas de cruzamentos controlados são interessantes devido à flexibilidade na escolha dos parentais na determinação do esquema de cruzamento, permitindo assim, que a variabilidade genética do caráter seja considerada (TOLEDO et al., 2008). Usualmente, populações de mapeamento são construídas a partir de duas linhagens parentais homozigóticas e contrastantes, resultando assim, em uma geração F1 que será heterozigótica em todos os marcadores e QTLs. Nos indivíduos da geração F1 o desequilíbrio de ligação é máximo e as populações derivadas a partir deles procuram explorar esse desequilíbrio. Alguns dos tipos de populações mais comumente utilizadas são:

- i) população F_2 que é obtida por autofecundação das plantas da geração F1;
- ii) população de retrocruzamento que é obtida pelo cruzamento de plantas F1 com um dos parentais;
- iii) Linhagens puras recombinantes que são derivadas por autofecundação a partir de uma população F_2 .

2.2.3 Fração de recombinação e distância genética

A construção do mapa genético envolve o cálculo da distância entre os marcadores. Tal cálculo é feito em termos de fração de recombinação. O evento de recombinação ocorre quando os genes que estão localizados em um mesmo cromossomo não segregam juntos, formando assim gametas não parentais ou recombinantes. Na ausência de recombinação cada cromossomo é igual a um dos parentais. A probabilidade de ocorrer recombinação entre dois locos, ou fração de recombinação é denotada por r . A porcentagem máxima esperada de eventos recombinantes é de 50%, consequentemente, $r \leq 0,5$.

De acordo com Lynch e Wash (1998), um problema da fração de recombinação é o fato dela não ser aditiva. Para contornar o problema da falta de aditividade da fração de recombinação, surgiram as funções de mapeamento, que são usadas para converter as frações de recombinação em medidas de distâncias (d) com propriedades mais interessantes para o ordenamento dos locos. As funções de mapeamento mais conhecidas são:

- i) a função de Haldane (HALDANE, 1919), que produz uma distância aditiva, d , em termos de fração de recombinação, a qual é dada por:

$$d = -\frac{\ln(1 - 2r)}{2};$$

ii) a função de mapeamento de Kosambi (KOSAMBI, 1944), que considera interferência das permutas nos intervalos adjacentes, a qual é dada por:

$$d = \frac{1}{4} \ln \left(\frac{1 + 2r}{1 - 2r} \right).$$

As distâncias são usualmente expressas em unidades de Morgan ou centiMorgan (cM).

2.3 Métodos bayesianos para mapear QTLs

A abordagem bayesiana tem contribuído no desenvolvimento de metodologias para mapear múltiplos QTLs (SATAGOPAN et al., 1996; HEATH, 1997; SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998; SEN; CHURCHILL, 2001; GAFFNEY, 2001; WANG et al., 2005; YI, 2004; YI et al., 2005, YI et al., 2007).

Os métodos bayesianos utilizam a distribuição conjunta a posteriori para fazer inferências sobre os parâmetros de interesse. Para construir a distribuição conjunta a posteriori para os parâmetros é necessário especificar o relacionamento entre os parâmetros e os dados observados através de modelos estatísticos. Os dados observados incluem as localizações e genótipos dos marcadores e os dados fenotípicos. Os parâmetros de interesse são: número de QTLs (S), localização destes QTLs no genoma (posição e cromossomo), efeitos genéticos (aditivo, dominância e interações). Nesta seção será discutida a construção da distribuição conjunta a posteriori para tais parâmetros.

2.3.1 Modelo QTL

Considerando-se uma população de mapeamento derivada de cruzamentos controlados, e supondo que a característica quantitativa de interesse é afetada por S QTLs, o vetor de valores fenotípicos observados $\mathbf{y} = (y_1, y_2, \dots, y_n)$, pode ser descrito por meio do seguinte modelo de regressão linear:

$$\mathbf{y} = \mathbf{1}\mu + X\boldsymbol{\beta} + \mathbf{e}, \quad (2.2)$$

em que μ é uma constante, $\boldsymbol{\beta}$ é o vetor dos efeitos genéticos, X é a matriz de delineamento, $\mathbf{e}$ o vetor de erros aleatórios e $\mathbf{e} \sim N(0, I\sigma^2)$. A definição da matriz X e do vetor de parâmetros $\boldsymbol{\beta}$ dependem da população de mapeamento. Considerando-se uma população de retrocruzamento os possíveis efeitos genéticos são os efeitos aditivos dos S QTLs, denotado por α_j , $j = 1, \dots, S$, e todas as possíveis interações, $\alpha_j\alpha_{j'}$, entre os efeitos aditivo do QTL j e j' , $j \neq j'$; para uma população F_2 os efeitos genéticos são: aditivos e de dominância dos S QTLs, denotados, respectivamente, por α_j e δ_j , $j = 1, \dots, S$, e todas as possíveis interações: entre os efeitos aditivo; entre os efeitos aditivo e de dominância; entre os efeitos de dominância e aditivo e entre os efeitos de dominância e de dominância do QTL j e j' , denotados por : $\alpha_j\alpha_{j'}$, $\alpha_j\delta_{j'}$, $\delta_{j'}\alpha_j$ e $\delta_j\delta_{j'}$.

A definição dos elementos da matriz X depende dos genótipos Q_{ij} , dos possíveis QTLs, que na prática, não são observáveis. Assim, tem-se que:

$$y_{ij}|Q_{ij} \sim N(\mu + X_{ij}\beta_j, \sigma^2), i = 1, \dots, n; j = 1, \dots, S.$$

Na abordagem bayesiana para a construção da distribuição a posteriori é necessário obter a função de verossimilhança e modelar a incerteza das quantidades desconhecidas no modelo.

2.3.2 Construção da função de verossimilhança

Tem-se que a distribuição marginal, $f(\mathbf{y})$ é dada por:

$$f(\mathbf{y}) = \sum f(\mathbf{y}, Q) = \sum f(\mathbf{y} | Q)f(Q),$$

dessa forma é necessário construir uma distribuição de probabilidades para cada Q_{ij} , visto que $y_{ij}|Q_{ij} \sim N(\mu + X_{ij}\beta_j, \sigma^2)$. Esta distribuição de probabilidade é obtida por meio da fração de recombinação (SATAGOPAN et al., 1996). Para construir esta distribuição de probabilidade é necessário conhecer a posição do suposto j -ésimo QTL, as distâncias entre os marcadores flancueadores a esse suposto QTL e os genótipos dos mesmos. Assim, supondo que a localização, λ_j , do j -ésimo QTL está entre os marcadores k_j e k_{j+1} e conhecendo as distâncias $D_{k_j} \leq \lambda_j < D_{k_{j+1}}$, sendo D_k a distância entre o marcador 1 e k e sabendo-se que os genótipos dos marcadores M_{ik_j} e $M_{ik_{j+1}}$ do i -ésimo indivíduo e considerando-se que S é o número total de QTLs incluídos no modelo e $\lambda = (\lambda_1, \dots, \lambda_S)$ é o vetor de localizações compostos pelas distâncias λ_j , a distribuição de probabilidade para os genótipos dos QTLs para o i -ésimo indivíduo é dada por:

$$f(Q_i|\lambda, M_i, D) = \prod_{j=1}^S f(Q_{ij}|\lambda, M_{ik_j}, M_{ik_{j+1}}, D_{k_j}, D_{k_{j+1}})$$

em que $f(Q_{ij}|\lambda_j, M_i, D_j)$ é calculada utilizando-se as probabilidades apresentadas nas Tabelas 2.2 e 2.3. Nas tabelas mencionadas r_a , r_b e r , são as frações de recombinação obtidas a partir do mapa genético, podendo ser interpretada como a fração de recombinação r_a entre o marcador k e o QTL, a fração de recombinação r_b entre o QTL e o marcador $k + 1$ e $r = r_a + r_b - 2r_ar_b$, a fração de recombinação entre o marcador k e $k + 1$. As frações de recombinação são obtidas utilizando-se uma função de mapeamento, tal como, a função de mapeamento de Haldane (HALDANE, 1919) ou Kosambi (KOSAMBI, 1944) e o vetor de localizações λ .

Assim, a função de verossimilhança para λ e $\theta = (\mu, \beta, \sigma^2)$ pode ser escrita da seguinte forma:

$$L(\lambda, \theta|y) = \prod_{i=1}^n \sum_{q_i} f(y_i|Q_i = q_i, \theta) \pi(Q_i = q_i|\lambda) \quad (2.3)$$

Pode-se observar que $L(\lambda, \theta|y)$, (eq. 2.3), é uma mistura finita de densidades normais. A maximização desta função de verossimilhança é uma tarefa muito difícil quando existem muitos QTLs.

Tabela 2.2 – Probabilidade condicionais do Genótipo do QTL dados os Marcadores Flanqueadores, M_k e M_{k+1} em uma População F_2

M_k	M_{k+1}	$\pi(Q_i = QQ M_k, M_{k+1})$	$\pi(Q_i = Qq M_k, M_{k+1})$	$\pi(Q_i = qq M_k, M_{k+1})$
MM	MM	$\frac{(1-r_a)^2(1-r_b)^2}{(1-r)^2}$	$\frac{2r_a(1-r_a)r_b(1-r_b)}{(1-r)^2}$	$\frac{r_a^2r_b^2}{(1-r)^2}$
MM	Mm	$\frac{(1-r_a)^2r_b(1-r_b)^2}{r(r-1)}$	$\frac{r_a(1-r_a)[r_b^2+(1-r_b)^2]}{r(r-1)}$	$\frac{r_a^2r_b(1-r_b)}{r(r-1)}$
MM	mm	$\frac{(1-r_a)^2r_b^2}{r^2}$	$\frac{2r_a(1-r_a)r_b(1-r_b)}{r^2}$	$\frac{r_a^2(1-r_b)^2}{r^2}$
Mm	MM	$\frac{r_a(1-r_a)(1-r_b)^2}{r(1-r)}$	$\frac{[r_a^2+(1-r_a)^2]r_b(1-r_b)}{r(1-r)}$	$\frac{r_a(1-r_a)r_b^2}{r(1-r)}$
Mm	Mm	$\frac{2r_a(1-r_a)r_b(1-r_b)}{r^2+(1-r)^2}$	$\frac{[r_a^2+(1-r_a)^2][r_b^2+(1-r_b)^2]}{r^2+(1-r)^2}$	$\frac{2r_a(1-r_a)r_b(1-r_b)}{r^2+(1-r)^2}$
Mm	mm	$\frac{r_a(1-r_a)r_b^2}{r(1-r)}$	$\frac{[r_a^2+(1-r_a)^2]r_b(1-r_b)}{r(1-r)}$	$\frac{r_a(1-r_a)(1-r_b)^2}{r(1-r)}$
mm	MM	$\frac{r_a^2(1-r_b)^2}{r^2}$	$\frac{2r_a(1-r_a)r_b(1-r_b)}{r^2}$	$\frac{(1-r_a)^2r_b^2}{r^2}$
mm	Mm	$\frac{r_a^2r_b(1-r_b)}{r(1-r)}$	$\frac{r_a(1-r_a)[r_b^2+(1-r_b)^2]}{r(1-r)}$	$\frac{(1-r_a)^2r_b(1-r_b)}{r(1-r)}$
mm	mm	$\frac{r_a^2r_b^2}{(1-r)^2}$	$\frac{2r_a(1-r_a)r_b(1-r_b)}{(1-r)^2}$	$\frac{(1-r_a)^2(1-r_b)^2}{(1-r)^2}$

Tabela 2.3 – Probabilidade condicionais associadas aos Genótipo do QTL dados os Marcadores Flanqueadores, M_k e M_{k+1} em uma população de retrocruzamento

M_k	M_{k+1}	$\pi(Q_i = Qq M_k, M_{k+1})$	$\pi(Q_i = qq M_k, M_{k+1})$
Mm	Mm	$\frac{(1-r_a)(1-r_b)}{(1-r)}$	$\frac{r_ar_b}{(1-r)}$
Mm	mm	$\frac{(1-r_a)r_b}{r}$	$\frac{r_a(1-r_b)}{(1-r_a)r_b}$
mm	Mm	$\frac{r_a(1-r_b)}{r}$	$\frac{(1-r_a)r}{(1-r_a)r_b}$
mm	mm	$\frac{r_ar_b}{(1-r)}$	$\frac{(1-r_a)(1-r_b)}{(1-r)}$

A abordagem bayesiana contorna esse problema tratando os genótipos dos QTLs como quantidades desconhecidas.

Assim, denotando $Q_i = [Q_{ij}]_{j=1}^S$ e $Q = [Q_j]_{i=1}^n$ como o genótipo para os n indivíduos, a função de verossimilhança para λ , θ e Q pode ser escrita como:

$$L(\lambda, \theta, Q|y) = \prod_{i=1}^n f(y_i|Q_i = q_i, \theta) \pi(Q_i = q_i|\lambda). \quad (2.4)$$

Incorporando-se as incertezas relativas as quantidades desconhecidas do modelo λ , θ e Q e assumindo-se independência entre elas tem-se a distribuição a posteriori dada por:

$$\pi(\lambda, \theta, Q | \mathbf{y}) \propto \prod_{i=1}^n f(y_i|Q_i = q_i, \theta) \pi(Q_i = q_i|\lambda) \pi(Q|\lambda) \pi(\lambda) \pi(\theta) \quad (2.5)$$

em que $\pi(Q|\theta) = \prod_{i=1}^n \pi(Q_j|\lambda)$ e $\pi(\lambda, \theta)$ é densidade conjunta a priori para os parâmetros genéticos.

A seguir, será discutido a especificação de distribuição a priori para os parâmetros envolvidos no modelo.

2.3.3 Prioris

A especificação de distribuição a priori para os parâmetros envolvidos no modelo é uma parte importante da análise bayesiana, pois ela especifica o quanto se tem de conhecimento prévio a respeito dos parâmetros (BOX; TIAO, 1992). Assim, serão apresentadas a seguir, algumas alternativas encontradas na literatura para a escolha de distribuições a priori para os parâmetros envolvidos no mapeamento de QTLs.

Número total de QTLs

A distribuição a priori para o número total de QTLs é, freqüentemente, assumida como uma Poisson com média κ (STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998). Contudo, Green (1995), Gaffney (2001) e Silva (2006) chamam a atenção para o fato de que a escolha de κ tem grande influência sobre a estimativa a posteriori, devendo, portanto, ser feita de forma cuidadosa. Stephens e Fisch (1998) atribuíram uma distribuição de Poisson com média de quatro QTLs quando analisaram seus dados simulados contendo oito QTLs. Sillanpää e Arjas (1998) atribuíram uma distribuição Poisson com média de dois QTL por cromossomo.

Posição do QTL

É comum atribuir, para a posição do j -ésimo QTL, λ_j , distribuição a priori uniforme no intervalo $[a, b]$ (SATAGOPAN et al., 1996; HEATH, 1997; STEPHENS; FISCH, 1998). Gaffney (2001) assume que os QTLs são distribuídos ao longo de todo o genoma. Assim, dado que o comprimento genoma é L , $\lambda_j \sim U[0, L]$.

Satagopan et al. (1996) assumem como distribuição a priori para a constante, μ , uma distribuição normal com média $\eta = 0$ e variância $\tau^2 = 10$. Raftery et al. (1997) sugerem escolher um valor para η que dependa dos dados, por exemplo, $\eta = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Para a variância pode-se escolher $\tau^2 = s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$.

Efeitos genéticos

Satagopan et al. (1996) assume uma distribuição a priori normal tanto para os efeitos aditivos α_j , $j = 1, \dots, S$ como para os de dominâncias δ_j , $j = 1, \dots, S$, com os seguintes hiperparâmetros: $\alpha_j \sim N(\mu_{\alpha_j}, \sigma_{\alpha}^2)$ e $\delta_j \sim N(\mu_{\delta_j}, \sigma_{\delta}^2)$. Gaffney (2001) chama a atenção para o fato de que a escolha de σ_{α}^2 e σ_{δ}^2 pode afetar a seleção do melhor modelo, tornando-se, assim, necessário que esta escolha seja feita com cautela.

Para os efeitos de epistasia: aditivo versus aditivo, aditivo versus dominância, dominância versus aditivo e dominância versus dominância, são atribuídas distribuições a priori conjugadas com hiperparâmetros pré-fixados (YI; XU, 2002; NARITA; SASAKI, 2004)

Variância residual

Freqüentemente, a distribuição a priori atribuída para a variância residual, σ^2 , é a distribuição gama-inversa com hiperparâmetros $\tilde{a}$ e $\tilde{b}$ (SATAGOPAN et al., 1996; GAFFNEY, 2001). Satagopan et al. (1996); Heath (1997) considera $\tilde{a} = \tilde{b} = 1$. Gaffney (2001) sugere $\tilde{a} = 3$ e $\tilde{b} = s_y^2$.

2.3.4 Estimação dos parâmetros

Na abordagem bayesiana as inferências são realizadas a partir da distribuição marginal a posteriori dos parâmetros de interesse, as quais podem ser obtidas integrando a eq. (2.5). Na maioria da vezes não é possível obter uma solução analítica; uma forma de contornar esse problema é usar o método de Monte Carlo com Cadeias de Markov (MCMC) para gerar uma amostra da distribuição conjunta a posteriori. Contudo, o método MCMC tradicional pode ser usado somente quando o espaço paramétrico é fixo e na eq. (2.5), pode-se observar que a dimensão do espaço paramétrico varia de acordo com o número S de QTLs presente no modelo. A seguir será apresentado as formas de estimação dos parâmetros considerando o número de QTLs como conhecido e como desconhecido.

Estimação considerando o número de QTLs conhecido

Satagopan et al. (1996) consideraram o número de QTLs, S , como uma quantidade conhecida. Assim, as estimativas a posteriori para os efeitos genéticos, localização e genótipos dos QTLs são baseadas no método MCMC tradicional. Para a obtenção de uma amostra da distribuição conjunta a posteriori para os parâmetros de interesse é utilizado o algoritmo de Gibbs com passos de Metropolis-Hastings. O fator de Bayes é utilizado para selecionar o modelo mais adequado e, conseqüentemente, estimar o número de QTLs presente no modelo. Um possível esquema de atualização é dado a seguir:

- i) Atualização de λ : O algoritmo Metropolis-Hastings é usado para atualizar cada elemento do vetor de localizações λ , pois a distribuição condicional completa para cada λ_j não tem forma fechada. O desafio é propor uma distribuição $u(\lambda_j, \lambda_j^*)$ para gerar o valor candidato ao próximo estado da cadeia. Satagopan et al. (1996) propuseram gerar um valor candidato λ_j^* a partir de uma distribuição uniforme no intervalo $[\max(\lambda_{j-1}, \lambda - d), \min(\lambda_{j+1}, \lambda + d)]$, sendo $\lambda_0 = 0$ e $\lambda_{S+1} = L$, em que L é o comprimento total do cromossomo. Essa proposta apresenta alguns desafios para sua implementação, pois considera as posições de forma ordenada. Gaffney (2001) propõe gerar um valor λ^{**} a partir de uma distribuição uniforme no intervalo $[\lambda - \varepsilon, \lambda + \varepsilon]$ e o valor candidato $\lambda^* = \min(|\lambda^{**}|, 2L - |\lambda^{**}|)$, em que L é o comprimento do cromossomo. Esta proposta garante que as localizações propostas estejam

no intervalo $[0, L]$, além disso, tem a vantagem de que $u(\lambda_j, \lambda_j^*) = u(\lambda_j^*, \lambda_j)$. Dado λ_j^* , ele será aceito ou não, dependendo da seguinte probabilidade de aceitação:

$$\alpha(\lambda_j, \lambda_j^*) = \frac{\pi(\lambda_j^* | \lambda_{-j}, Q, \theta, y) u(\lambda_j^*, \lambda_j)}{\pi(\lambda_j | \lambda_{-j}, Q, \theta, y) u(\lambda_j, \lambda_j^*)}$$

- ii) Atualização de Q : Cada elemento Q_{ij} é atualizado separadamente a partir da sua distribuição condicional completa a posteriori $\pi(Q_{ij} | \lambda, Q_{i(-j)}, \theta, y)$ (SATAGOPAN et al., 1996; GAFFNEY, 2001; YI, 2004).
- iii) Atualização de θ : Quando as priors atribuídas aos componentes de θ são conjugadas, as distribuições condicionais completas a posteriori tem forma fechada e assim, estes parâmetros podem ser atualizados usando o algoritmo de Gibbs Sampling (SATAGOPAN et al., 1996; GAFFNEY, 2001).

Estimação considerando o número de QTLs como uma quantidade desconhecida

Tratar o número de QTLs como uma quantidade desconhecida implica em incorporar a incerteza a este parâmetro e na mudança na dimensão do espaço paramétrico. Os métodos MCMC tradicionais não podem ser usados. Uma proposta, que tem sido utilizada no mapeamento de QTLs é o uso do algoritmo MCMC com saltos reversíveis introduzido por Green (1995). Este método permite a simulação das distribuições a posteriori sobre espaço de dimensão que variam. O uso desse método assegura que modelos que são mais prováveis sejam visitados com maior frequências (GODSILL, 2001).

No contexto de mapeamento de QTLs os movimentos entre modelos são vistos como passo de nascimento de um QTL ($S \rightarrow S + 1$), com probabilidade p_n , passo morte de QTL ($S \rightarrow S - 1$), com probabilidade p_m ou passo de permanência ($S \rightarrow S$), com probabilidade $p_p = 1 - p_n - p_m$ (SILLANPÄÄ; ARJAS, 1998). Gaffney (2001) ressalta que não existe restrição quanto ao número de QTLs a serem adicionados ou removidos do modelo, contudo, o nascimento ou morte de mais de um QTL pode afetar a razão de aceitação e com isso comprometer o desempenho do algoritmo. A aceitação da proposta de nascimento, morte, ou permanência de número QTLs no modelo depende um probabilidade de aceitação, conforme será mostrado nos itens a seguir.

i) Nascimento de um QTL

Considerando-se o modelo dado pela eq. (3) e supondo-se que existam S QTLs no estado atual, $t - 1$, da cadeia. Propõe-se o movimento do modelo contendo S QTLs para um modelo com $S + 1$ QTLs com probabilidade p_n . Para fazer esse movimento entre modelos, ou seja, para o número de QTLs passar de S para $S + 1$, é preciso propor para este novo QTL, uma localização, genótipos

para os n indivíduos e efeitos genéticos associados. Assim, a probabilidade de aceitar o movimento de S para $S + 1$ QTLs é dada por $\alpha = \min(1, A)$, sendo,

$$A = \frac{\pi(S + 1, \boldsymbol{\lambda}^+, \boldsymbol{\theta}^+, Q^+ | y)}{\pi(S, \boldsymbol{\lambda}, \boldsymbol{\theta}, Q | y)} \times \frac{q(u^+ | S + 1, \boldsymbol{\lambda}^+, \boldsymbol{\theta}^+, Q^+)}{q(u | S, \boldsymbol{\lambda}, \boldsymbol{\theta}, Q)} \times \frac{p_m}{p_n} \times |J|.$$

Em que $\pi(S + 1, \boldsymbol{\lambda}^+, \boldsymbol{\theta}^+, Q^+ | y)$ é a distribuição conjunta a posteriori para os parâmetros de interesse considerando $S + 1$ QTLs, $\boldsymbol{\lambda}^+ = (\boldsymbol{\lambda}, \lambda_{S+1})$, $\boldsymbol{\theta}^+ = (\boldsymbol{\theta}, \theta_{S+1})$, $Q^+ = (Q, Q_{S+1})$ e $q(u^+ | S + 1, \boldsymbol{\lambda}^+, \boldsymbol{\theta}^+, Q^+)$ é a distribuição proposta para adicionar um QTL.

Se proposta do nascimento de um QTL é aceita, o número de QTLs no próximo estado, t , da Cadeia de Markov torna-se, $S^{(t)} = S^{(t-1)} + 1$, e a nova localização, com seus correspondentes genótipos e parâmetros de regressão são aceitos simultaneamente; caso contrário, $S^{(t)} = S^{(t-1)}$ (SILLANPÄÄ; ARJAS, 1998; SILVA, 2006).

ii) Morte de um QTL

Considerando-se o modelo dado pela eq. (3) e supondo-se que existam S QTLs no estado atual, $t - 1$, da cadeia. Propõe-se o movimento do modelo contendo S QTLs para um modelo com $S - 1$ QTLs com probabilidade p_m . Para fazer esse movimento entre modelos, ou seja, para o número de QTLs passar de S para $S - 1$, é preciso selecionar um QTL para ser retirado do modelo, ou seja, é necessário retirar uma localização, os genótipos e efeitos associados a tal QTL. A probabilidade de aceitação é calculada de forma similar ao passo de nascimento e é dada por:

$$\alpha = \min(1, \frac{1}{A}).$$

Se a exclusão do QTL proposto é aceita, o próximo estado da Cadeia de Markov, $S^{(t)} = S^{(t-1)} - 1$, caso contrário, $S^{(t)} = S^{(t-1)}$.

iii) Permanência do Número de QTL

No caso de permanência do número de QTLs, o número de QTLs para novo estado, t , da Cadeia de Markov permanece igual ao número de QTLs no estado $t - 1$. Os demais parâmetros são atualizados como descrito no item (2.4.2.1).

Embora o algoritmo MCMC com saltos reversíveis ofereça a flexibilidade de poder mover entre modelos de dimensões diferentes sua implementação apresenta alguns desafios, tal como, a construção de distribuições propostas complexas. Além disso, quando considerado um modelo mais complexo, por exemplo, incluindo termos de epistasia, a complexidade do algoritmo MCMC com saltos reversíveis aumenta a demanda computacional e pode comprometer seu desempenho (YI et al., 2005). Apesar, dos desafios enfrentados pelos pesquisadores para a implementação do algoritmo MCMC com Saltos Reversíveis, este método tem sido usado em mapeamento de QTLs.

Contudo, observa-se que este método tem sido empregado para modelos mais simples, ou seja, sem considerar termos de interação (HEATH, 1997; SILLANPÄÄ; ARJAS, 1998; STEPHENS; FISCH, 1998; GAFFNEY, 2001; SILVA, 2006).

Mapeamento de QTLs usando espaço composto

Trabalhos recentes têm aplicado o espaço composto, proposto por Godsil (2001), no contexto de mapeamento bayesiano de QTLs (YI, 2004; YI et al., 2005; YI et al., 2007). O espaço composto considera a presença de um número máximo de QTLs (S) e utiliza uma variável aleatória, binária, γ_{jk} , para indicar inclusão do k -ésimo efeito genético do QTL j ($\gamma_{jk} = 1$) ou exclusão ($\gamma_{jk} = 0$) no modelo, em que $j = 1, \dots, S$ e $k = 1, \dots, K$ e K representa o número de efeitos genéticos. Dessa forma, o espaço paramétrico permanece fixo e a obtenção da amostra aleatória da distribuição a posteriori conjunta pode ser realizada utilizando os métodos MCMC tradicionais (YI, 2004).

Dada uma população de mapeamento derivada de cruzamentos controlados e supondo que a característica quantitativa de interesse y é afetada por S QTLs. O valor fenotípico observado, $\mathbf{Y} = (y_1, y_2, \dots, y_n)^T$, é descrito utilizando o seguinte modelo linear:

$$\mathbf{Y} = \mathbf{1}\mu + \mathbf{X}\Gamma\boldsymbol{\beta} + \mathbf{e} \quad (2.6)$$

em que $\Gamma = \text{diag}(\boldsymbol{\gamma})$, μ é a uma constante, $\mathbf{X}$ é a matriz de delineamento, $\boldsymbol{\beta}$ é o vetor contendo todos os possíveis efeitos genéticos e $\mathbf{e}$ é um vetor de erros, tal que $\mathbf{e} \sim N(0, I\sigma^2)$. O número de QTLs é determinado pelo vetor $\boldsymbol{\gamma}$. Portanto, $\boldsymbol{\gamma}$ pode ser visto como o indexador do modelo. Denotando $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mu, \sigma^2)$, $\boldsymbol{\lambda}$ como sendo o vetor contendo as localizações para os possíveis QTLs e $\mathbf{X}$ que é não observável. A idéia chave do espaço composto é particionar o espaço paramétrico $(\boldsymbol{\lambda}, \mathbf{X}, \boldsymbol{\theta})$ da seguinte forma: $(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}})$ e $(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{X}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}})$, em que $(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}})$ representa os parâmetros que foram incluídos no modelo ($\boldsymbol{\gamma} = 1$) e $(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{X}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}})$ representa os parâmetros que não foram incluídos no modelo ($\boldsymbol{\gamma} = 0$)

A função de verossimilhança para um particular $\boldsymbol{\gamma}$, depende somente dos parâmetros usados no modelo, isto é,

$$L(\boldsymbol{\gamma}, \mathbf{X}, \boldsymbol{\theta} | \mathbf{Y}) = L(\boldsymbol{\gamma}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}} | \mathbf{Y})$$

A distribuição a priori para $(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{X}, \boldsymbol{\theta})$ pode ser fatorada como:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{X}, \boldsymbol{\theta}) = \pi(\boldsymbol{\gamma})\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X} | \boldsymbol{\gamma}) = \pi(\boldsymbol{\gamma})\pi(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma})\pi(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{X}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}} | \boldsymbol{\gamma}, \boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}})$$

A distribuição conjunta a posteriori para os parâmetros é expressa por:

$$\pi(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{X}, \boldsymbol{\theta} | \mathbf{Y}) = L(\mathbf{Y} | \boldsymbol{\gamma}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{X} | \boldsymbol{\gamma})\pi(\boldsymbol{\gamma})\pi(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{X}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}} | \boldsymbol{\gamma})$$

$$\pi(\lambda_{-\gamma}, \mathbf{X}_{-\gamma}, \theta_{-\gamma} | \gamma, \lambda_{\gamma}, \mathbf{X}_{\gamma}, \theta_{\gamma})$$

Sendo $\pi(\gamma)$ a distribuição a priori para o modelo indexado, $\pi(\lambda_{\gamma}, \mathbf{X}_{\gamma}, \theta_{\gamma} | \gamma)$ a distribuição conjunta a priori para os parâmetros que estão sendo incluídos no modelo e $\pi(\lambda_{-\gamma}, \mathbf{X}_{-\gamma}, \theta_{-\gamma} | \gamma, \lambda_{\gamma}, \mathbf{X}_{\gamma}, \theta_{\gamma})$ a distribuição conjunta a priori para os parâmetros que não estão sendo incluídos no modelo, a qual é chamada de “pseudo-priori”.

Pode-se observar que mesmo mudando o número de QTLs a serem incluídos no modelo, o dimensionamento do espaço paramétrico permanece o mesmo. Portanto, simulações de $\pi(\gamma, \lambda, \mathbf{X}, \theta | \mathbf{Y})$ podem, então, ser feitas pelos algoritmos MCMC tradicionais.

Referências

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

CARLIN, B.P.; CHIB, S. Bayesian model choice via Markov Chain Monte Carlo. **Journal of the Royal Statistical Society**, London, v. 57, p. 473-484, 1995.

CARLIN, B.P.; LOUIS, T.A. **Bayes and empirical Bayes methods for data analysis**. London: Chapman & Hall, 1996. 399 p.

CASELLA, G.; GEORGE, E.I. Explaining the Gibbs Sampler. **The American Statistician**, New York, v. 46, p. 167-174, 1992.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hasting algorithm. **The American Statistician**, Washington, v. 49, p. 326-335, 1995.

DOERGE, R.W. Mapping and analysis of quantitative trait loci in experimental populations. **Nature Reviews Genetics**, New York, v. 31, p. 43-52, 2002.

EDWARDS, M.D.; STUBER, C.W.; WENDEL, J.F. Molecular-marker-facilitated investigation of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. **Genetics**, Austin, v. 116, p. 113-125, 1987.

ELSTON, R.C.; STEWART, J. The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. **Genetics**, Austin, v. 73, p. 695-711, 1973.

GAFFNEY, P.J. **An efficient reversible jump Markov Chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses.** 2001, 174p. Thesis (Philosophy in Statistics Doctor) - University of Wisconsin, Madison, 2001.

GAMERMAN, D.; LOPES, H.F. **Markov chain Monte Carlo:** stochastic simulation for Bayes inference. London: Chapman & Hall, 2006. 323 p.

GEMAN, S.; GEMAN, D. Stochastic Relaxation, Gibbs Distributions and Bayesian Restoration of Images. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, Los Alamitos v. 6, p. 721-741, 1984.

GELFAND, A.E. Gibbs Sampling. **Journal of the American Statistical Association**, Alexandria, v. 95, p. 1300-1304, 2000.

GODSILL, S.J. On the relationship between MCMC model uncertainty methods. **Journal of computational Graphical Statistics**, Alexandria, v. 10, p. 230-248, 2001.

GREEN, P.J.; Reversible jump markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, Lausanne, v. 82, p. 711-732, 1995.

HALDANE, J.B.S. The combination of linkage value and the calculation of distances between the loci of linkage factors. **Journal of Genetics**, London, v. 8, p. 299-309, 1919.

HEATH, S.C. Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. **The American Journal of Human Genetics**, Boston, v. 61, p. 748-760, 1997.

JANSEN, R.C.; STAM, P. High resolution of quantitative traits into multiple loci via interval mapping. **Genetics**, Austin, v. 136, p. 1447-1455, 1994.

JEFFREY, H. **Theory of probability.** Oxford: Clarendon press, 1961. 447 p.

KAO, C.H.; ZENG, Z-B.; TEASDALE, R.D. Multiple interval mapping for quantitative trait loci. **Genetics**, Austin, v. 152, p. 1203-1216, 1999.

KAO, C.H.; ZENG, Z-B. General formulas for obtaining the MLEs and asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. **Biometrics**, Washington, v. 53, p. 359-371, 1997.

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

KOSAMBI, D.D. The estimation of map distances from recombination values. **Annual Eugenics**, New York, v. 12, p. 172-175, 1944.

LIU, B.H. **Statistical genomics: linkage, mapping and QTL analysis**. Boca Raton: CRC Press, Boca Raton, 1998.

LOPES, H.F. **Bayesian Model Selection**. Disponível em: <http://eagle.ufrj.br/~hedibert/modelselection.pdf>, 2002. Acesso em: 9 fev. de 2009.

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

NARITA, A.; SASAKI, Y. Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. **Genetics Selection Evolution**, Versailles, v. 36, p. 415-433, 2004.

RAFTERY, A.E.; MADIGAN, D.; HOETING, Bayesian model averaging for linear regression models. **Journal of the American Statistical Association**, Alexandria, v. 92, p. 179-191, 1997.

SATAGOPAN, J. M.; YANDELL, B.S. Estimating the number of quantitative trait Loci via Bayesian model determination. **Special contributed paper session on genetic analysis of quantitative trait and complex disease, Biometric section, Joint Statistical Meetings**, Chicago, 1998.

SATAGOPAN, J. M.; YANDELL, B.S. NEWTON M.A.; OSBORN, T.C. A Bayesian approach to detect Quantitative trait Loci using Markov chain Monte Carlo. **Genetics**, Austin, v. 144, p. 805-816, 1996.

SAX, K. The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. **Genetics**, Austin, v. 8, p. 552-560, 1923.

SEN, S.; CHURCHILL A statistical framework for quantitative trait mapping . **Genetics**, Austin, v. 159, p. 371-387, 2001.

SIBOV, S.T.; de SOUZA JÚNIOR, C.L.; GARCIA, A.A.F.; GARCIA, A.F.; SILVA, A.R.; MANGOLIM, 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**, Lund, v. 139, n. 2, p. 107 - 115, 2003b.

SILLANPÄÄ, M.J.; ARJAS, E. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data . **Genetics**, Austin, v. 148, p. 1373-1388, 1998.

SILVA, J.P. **Uma abordagem Bayesiana para o mapeamento de QTLs utilizando o método MCMC com saltos reversíveis** 2006, 80p. Tese (Mestrado em Estatística e Experimentação Agrônômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2006.

SOLLER, M.; BRODY, T.; GENIZI, A. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. **Theoretical and Applied Genetics**, Berlin, v. 47, p. 35-39, 1976.

STEPHENS, D.A.; FISCH, R.D. Bayesian analysis of quantitative trait locus data using reversible jump Markov Chain Monte Carlo **Biometrics**, Washington, v. 54, p. 1334-1347, 1998.

THODAY, J.M. Location of polygenes. **Nature**, New York, v. 191, p. 368-370, 1961.

TOLEDO, E.R. ; LEANDRO, R.A. ; SOUZA JUNIOR, C.L. ; SOUZA, A. P. Mapeamento de QTLs: Uma abordagem bayesiana. **Revista de Matemática e Estatística**, Jaboticabal, v. 26, p. 107-114, 2008.

WANG, H.; ZHANG, Y.M.; MASINDE, G.L.; MOHAN, S; BAYLINK, D.J.; XU, S. Bayesian shrinkage estimation of quantitative trait loci parameters. **Genetics**, Austin, v. 170, p. 465-480, 2005.

YI, N. A unified markov chain monte carlo framework for mapping multiple quantitative trait loci. **Genetics**, Austin, v. 167, p. 967-975, 2004.

YI, N.; XU, S. Mapping quantitative trait loci with epistatic effects. **Genetical Research**, Cambridge, v. 79, p. 185-198, 2002.

YI, N.; ALLISON, D.B.; XU, S. Bayesian model choice and search strategies for mapping multiple epistatic quantitative trait loci. **Genetics**, Austin, v. 165, p. 867-883, 2003.

YI, N.; YANDELL, B.S.; CHURCILL, G.A.; ALLISON, D.B.; EISEN, E.J.; POMP, D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. **Genetics**, Austin, v. 170, p. 1333-1344, 2005.

YI, N.; SHRINER, D.; BANERJEE, S.; MEHTA, T.; POMP, D. YANDELL, B.S. An efficient Bayesian model selection approach for interacting QTL models with many effects. **Genetics**, Austin, v.176 , p. 1865-1877, 2007.

ZENG, Z-B. Theoretical basis for separation of multiple linked genes effects in mapping quantitative trait loci. **Proceedings of the National Academy Sciences of the United States of America**, Washington, v. 90, p. 10972-10976, 1993.

ZENG, Z-B. Precision mapping of quantitative trait loci. **Genetics**, Austin, v. 136, p. 1457-1468, 1994.

3 UMA ABORDAGEM BAYESIANA PARA O MAPEAMENTO DE QTLs

Resumo

Muitos caracteres em plantas e animais são de natureza quantitativa, influenciados por múltiplos genes. Com o advento de novas técnicas moleculares tem sido possível mapear os locos que controlam os caracteres quantitativos, denominados QTLs (*Quantitative Trait Loci*). Mapear um QTL significa identificar sua posição no genoma, bem como, estimar seus efeitos genéticos. A maior dificuldade para realizar o mapeamento de QTLs, se deve ao fato de que o número de QTLs é desconhecido e, conseqüentemente, a dimensão do espaço paramétrico também é desconhecida. Métodos Bayesianos juntamente com método Monte Carlo com Cadeias de Markov (MCMC), têm sido implementados para inferir conjuntamente o número de QTLs, suas posições no genoma e os efeitos genéticos. O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado desconhecido e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. No presente trabalho foi implementado, utilizando-se o programa estatístico R uma abordagem bayesiana para mapear QTLs em que múltiplos QTLs e os efeitos de epistasia são considerados no modelo. Para tanto, o número de QTLs foi tratado como conhecido, foram ajustados modelos com números crescentes de QTLs e o fator de Bayes foi utilizado para selecionar o modelo mais adequado e conseqüentemente, para estimar o número de QTLs que controlam os fenótipos de interesse. Para investigar a eficiência da metodologia implementada foi feito um estudo de simulação em que foram considerados duas diferentes populações experimentais: retrocruzamento e F_2 . Sendo que para ambas as populações foi feito o estudo de simulação considerando modelos com e sem epistasia. A abordagem implementada mostrou-se muito eficiente, sendo que para todas as situações consideradas o modelo selecionado foi o modelo contendo o número verdadeiro de QTLs considerado na simulação dos dados.

Palavras-chave: Mapeamento de QTLs; Inferência bayesiana; Fator de Bayes; MCMC

Abstract

Many traits in plants and animals have quantitative nature, influenced by multiple genes. With the new molecular techniques, it has been possible to map the loci, which control the quantitative traits, called QTL (*Quantitative Trait Loci*). Mapping a QTL means to identify its position in the genome, as well as to estimate its genetics effects. The great difficulty of mapping QTL relates to the fact that the number of QTL is unknown and hence the dimension of the parametric space is also unknown. Bayesian approaches used with Markov Chain Monte Carlo method (MCMC) have been applied to infer QTL number, their positions in the genome and their genetic effects. The challenge is to obtain the sample from the joined distribution posterior of these parameters, since the number of QTL may be considered unknown and hence the dimension of the parametric space changes according to the number of QTL in the model. In this study, a Bayesian approach was applied, using the statistical program R, in order to map QTL, considering multiples QTL and

epistasis effects in the model. Then, the number of QTL was considered as known, models were adjusted with the crescent number of QTL and Bayes factor was used to select the most suitable model and, consequently, to estimate the number of QTL that control interesting phenotypes. To evaluate the efficiency of the applied methodology, a simulation study was done, considering two different experimental populations: backcross and F_2 , accomplishing the simulation study for both populations, considering models with and without epistasis. The applied approach resulted to be very efficient, considering that, for all the used situations, the selected model was the one containing the real number of QTL used in the data simulation.

Keywords: Bayes factor, Bayesian inference, MCMC, QTL mapping

3.1 Introdução

Muitos caracteres medidos em plantas e animais são de natureza quantitativa, influenciados por múltiplos genes e, conseqüentemente, pela interação entre eles, a qual é denominada epistasia. Com o advento de novas técnicas moleculares tem sido possível mapear os locos que controlam os caracteres quantitativos, denominados QTLs (*Quantitative Trait Loci*). Mapear um QTL significa identificar sua posição no genoma, isto é, identificar em qual cromossomo está e qual sua localização nesse cromossomo, bem como, estimar seus efeitos genéticos. Dessa forma, é possível um melhor entendimento sobre os caracteres quantitativos, que são de fundamental importância nos programas de melhoramento genético, uma vez que estão associados à produção de grãos, altura de plantas, volume, entre outros, que são de grande importância econômica.

A maior dificuldade para realizar o mapeamento de QTLs, se deve ao fato de que o número de QTLs é desconhecido e, conseqüentemente, a dimensão do espaço paramétrico também é desconhecida. Métodos Bayesianos juntamente com método Monte Carlo com Cadeias de Markov (MCMC), têm sido implementados para inferir conjuntamente o número de QTLs, suas posições no genoma e os efeitos genéticos (SATAGOPAN et al., 1996; HEATH, 1997; SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998; SEN; CHURCHILL, 2001; GAFFNEY, 2001; WANG et al., 2005; YI; XU, 2002; YI; ALLISON; XU, 2003; NARITA; SASAKI, 2004; YI, 2004; YI et al., 2005; YI et al., 2007). O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado desconhecido e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. Satagopan et al. (1996) considerou o número de QTLs como uma quantidade conhecida, ajustou modelos com diferentes números de QTLs e, posteriormente, fez a seleção do modelo mais adequado usando como critério de seleção o fator de Bayes. Trabalhos posteriores (SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998) contornaram o problema de não conhecer a dimensão do espaço paramétrico tratando o número de QTLs como uma quantidade desconhecida e usando o método MCMC com Saltos Reversíveis apresentado por Green (1995), o qual permite a movimentação entre modelos de dimensões

diferentes. O algoritmo MCMC com saltos reversíveis é uma técnica bastante poderosa e geral para explorar as distribuições a posteriori quando o espaço paramétrico não possui dimensão fixa. Contudo, essa habilidade de movimento entre modelos de dimensões diferentes requer a construção de distribuições propostas complexas, o que dificulta a implementação desse algoritmo. Além disso, a complexidade desse algoritmo aumenta a demanda computacional, o que pode comprometer seu desempenho. Trabalhos recentes têm aplicado o espaço composto, proposto por Godsil (2001), no contexto de mapeamento Bayesiano de QTLs (YI, 2004; YI et al., 2005; YI et al., 2007). O espaço composto considera a presença de um número máximo de QTLs (S) e utiliza uma variável aleatória, binária, γ_{jk} , para indicar inclusão do k -ésimo efeito genético do QTL j ($\gamma_{jk} = 1$) ou exclusão ($\gamma_{jk} = 0$) no modelo, em que $j = 1, \dots, S$ e $k = 1, \dots, K$ e K representa o número de efeitos genéticos. Dessa forma, o espaço paramétrico permanece fixo e a obtenção da amostra aleatória da distribuição a posteriori conjunta pode ser realizada utilizando os métodos MCMC tradicionais (YI, 2004).

Outra tarefa desafiadora é a inclusão de efeitos de epistasia no modelo pois, sua inclusão, implica aumento ou diminuição considerável da dimensão do espaço paramétrico. Com isso, aumenta a complexidade da implementação dos algoritmos utilizados na construção das amostras da distribuição conjunta a posteriori para os parâmetros de interesse. Particularmente, para os algoritmos MCMC com saltos reversíveis, em que cada movimentação entre modelos de dimensões diferentes implica na retirada ou inclusão de muitos parâmetros. Apesar dos desafios da implementação, o algoritmo MCMC com saltos reversíveis tem sido empregado para estudar epistasia (YI; XU, 2002, YI et al., 2003, NARITA; SASAKI, 2004). Contudo, esse método de simulação tem apresentado uma convergência muito lenta e uma demanda computacional intensa, tornando seu emprego não muito recomendável (YI et al., 2005). Assim, se faz necessário o estudo de metodologias bayesianas mais eficientes para mapear QTLs, principalmente, quando são considerados modelos mais complexos, como é o caso dos modelos que incluem termos de epistasias. Além disso, existe a necessidade de implementar esses algoritmos de maneira que sejam acessíveis a todos os pesquisadores que tenham interesse no assunto, para que assim haja uma maior difusão da metodologia bayesiana no mapeamento de QTLs.

Diante do exposto, o presente trabalho teve como objetivo implementar, um programa, acessível a todos os pesquisadores, utilizando a metodologia bayesiana para mapear QTLs no caso em que efeitos de epistasia são incorporados ao modelo. Para inclusão de efeito de epistasia no modelo utilizou-se o modelo epistático de Cockerham, como apresentado por Kao e Zeng, (2002). O número de QTLs foram considerados como conhecido e o fator de Bayes foi utilizado para selecionar o modelo mais adequado como proposto por Satagopan et al. (1996). Para verificar a eficiência da metodologia implementada foi feito um estudo de simulação para dois tipos de população experimentais: retrocruzamento e F_2 .

3.2 Metodologia

3.2.1 Modelo QTL

Considerando-se uma população de mapeamento derivada de cruzamentos controlados, e supondo que a característica quantitativa de interesse é afetada por S QTLs o vetor de valores fenotípicos observados $\mathbf{y} = (y_1, y_2, \dots, y_n)$, foi descrito por meio do seguinte modelo de regressão linear:

$$\mathbf{y} = \mathbf{1}\mu + X\boldsymbol{\beta} + \mathbf{e}, \quad (3.1)$$

em que μ é uma constante, $\boldsymbol{\beta}$ é o vetor dos efeitos genéticos, X é a matriz de delineamento e $\mathbf{e}$ o vetor de erros aleatórios, $\mathbf{e} \sim N(0, I\sigma^2)$.

Para ambas populações, X e $\boldsymbol{\beta}$ foram definidos por meio do modelo epistático de Cockerham, como apresentado por Kao e Zeng (2002). Para a população de retrocruzamento, utilizada no presente trabalho, os possíveis genótipos segregantes são qq e Qq e assim, os elementos da matriz X foram definidos como segue:

$$x_{ij1} = z_{ij} - 0,5,$$

$$x_{ijj^l1} = x_{ij1}x_{ij^l1},$$

em que z_{ij} é o número de alelos dominantes do genótipo do j -ésimo QTL para o i -ésimo indivíduo, $j = 1, \dots, S$ e $j \neq j^l$. Para a população F_2 , os elementos da matriz X foram definidos como segue:

$$\begin{aligned} x_{ij1} &= z_{ij} - 1, \\ x_{ij2} &= (1 + x_{ij1})(1 - x_{ij1}) - 0,5, \\ x_{ijj^lk} &= \begin{cases} x_{ij1}x_{ij^l1} & k = 1 \\ x_{ij1}x_{ij^l2} & k = 2 \\ x_{ij2}x_{ij^l1} & k = 3 \\ x_{ij2}x_{ij^l2} & k = 4. \end{cases} \end{aligned}$$

Para o modelo de Cockerham, $\beta_{j1} = \alpha_j$ e $\beta_{j2} = \delta_j$ correspondem, respectivamente, aos efeitos aditivos e de dominâncias do j -ésimo QTL e $\beta_{jj^l1} = \alpha_j\alpha_{j^l}$, $\beta_{jj^l2} = \alpha_j\delta_{j^l}$, $\beta_{jj^l3} = \delta_j\alpha_{j^l}$ e $\beta_{jj^l4} = \delta_j\delta_{j^l}$ são os efeitos epistáticos entre o loco j e j^l que são definidos respectivamente por: aditivo *versus* aditivo, aditivo *versus* dominância, dominância *versus* aditivo e dominância *versus* dominância (KAO; ZENG, 2002).

3.2.2 Estimação dos parâmetros

A inferência sobre os parâmetros de interesses, número de QTLs S , as localizações destes no cromossomo $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_S)$ e $\theta = (\mu, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma^2)$, foram feitas usando a abordagem bayesiana, sendo o número de QTLs, S , considerado como conhecido, como proposto por Satagopan et al. (1996). Foi assumida independência a priori para os parâmetros do modelo e as distribuições a priori para os mesmos foram especificadas da seguinte maneira: para as localizações foram atribuídas distribuição uniforme no intervalo $[0, L]$, sendo L o comprimento total do cromossomo; para a constante μ foi atribuída uma distribuição $N(0, \tau^2)$, sendo $\tau^2 = 10$, para os efeitos aditivos e de dominância foram atribuídas, respectivamente, $N(0, \sigma_\alpha^2)$ e $N(0, \sigma_\delta^2)$ em que $\sigma_\alpha^2 = 10$ e $\sigma_\delta^2 = 10$, para os efeitos de epistasia foram atribuídas, respectivamente, $N(0, \sigma_{\alpha\alpha}^2)$, $N(0, \sigma_{\alpha\delta}^2)$, $N(0, \sigma_{\delta\alpha}^2)$ e $N(0, \sigma_{\delta\delta}^2)$, sendo $\sigma_{\alpha\alpha}^2 = \sigma_{\alpha\delta}^2 = \sigma_{\delta\alpha}^2 = \sigma_{\delta\delta}^2 = 10$; para σ^2 foi atribuída distribuição gama-inversa com hiperparâmetros $a = 8$ e $b = 8$, como proposto por Raftery et al. (2007).

Assim, dado o modelo (3.1) e as distribuições a priori para os parâmetros de interesse a distribuição conjunta a posteriori é dada por:

$$\begin{aligned} \pi(\lambda, Q, \theta|y) &\propto \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-1}{2\sigma^2} \left(y_i - \mu - \sum_{j=1}^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'1} - \right. \right. \right. \\ &\quad \left. \left. \left. \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'2} \right)^2 \right\} \times \prod_{j=1}^S \pi(Q_{ij}|\lambda_j, M) \right\} \times \left(\frac{1}{L} \right)^S \times \\ &\quad \frac{1}{2\pi\tau^2} \exp \left\{ \frac{-1}{2\tau^2} (\mu)^2 \right\} \times \prod_{j=1}^S \left(\frac{1}{2\pi\sigma_\alpha^2} \exp \left\{ \frac{-1}{2\sigma_\alpha^2} (\alpha_j)^2 \right\} \times \frac{1}{2\pi\sigma_\delta^2} \exp \left\{ \frac{-1}{2\sigma_\delta^2} (\delta_j)^2 \right\} \right) \times \\ &\quad \prod_{j=1}^k \left(\frac{1}{2\pi\sigma_{\alpha\alpha}^2} \exp \left\{ \frac{-1}{2\sigma_{\alpha\alpha}^2} (\alpha_j \alpha_{j'})^2 \right\} \times \frac{1}{2\pi\sigma_{\alpha\delta}^2} \exp \left\{ \frac{-1}{2\sigma_{\alpha\delta}^2} (\alpha_j \delta_{j'})^2 \right\} \frac{1}{2\pi\sigma_{\delta\alpha}^2} \exp \left\{ \frac{-1}{2\sigma_{\delta\alpha}^2} (\delta_j \alpha_{j'})^2 \right\} \times \right. \\ &\quad \left. \frac{1}{2\pi\sigma_{\delta\delta}^2} \exp \left\{ \frac{-1}{2\sigma_{\delta\delta}^2} (\delta_j \delta_{j'})^2 \right\} \right) \times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp \left\{ \frac{-b}{\sigma^2} \right\}. \end{aligned}$$

Neste caso a obtenção da distribuição marginal a posteriori para os parâmetros de interesse são analiticamente intratáveis e os métodos MCMC, foram utilizados para a obtenção de uma amostra da distribuição conjunta a posteriori para os parâmetros de interesse, isso foi feito utilizando o algoritmo de Gibbs com passos Metropolis-Hastings. A atualização dos parâmetros foram feitas da seguinte forma:

- i) Atualização das posições (λ): cada λ_j , $j = 1, \dots, S$ foi atualizado usando o algoritmo de Metropolis-Hastings. O valor proposto para o próximo estado da cadeia foi gerado de acordo com Galfney (2001): para o j -ésimo QTL, um valor λ_j^{**} foi gerado de uma distribuição $U[\lambda_j - d, \lambda_j + d]$, em que d é uma constante para estabilizar a taxa de aceitação, e o valor proposto para o próximo estado da cadeia dado por $\lambda_j^* = \min(|\lambda_j^{**}|, 2L - |\lambda_j^{**}|)$, em que L é o comprimento total do cromossomo. Esta proposta é denotada por $q(\lambda_j^*|\lambda_j)$ e tem a vantagem de que $q(\lambda_j^*|\lambda_j) = q(\lambda_j|\lambda_j^*)$. A razão de aceitação é dada por:

$$\alpha(\lambda_j, \lambda_j^*) = \frac{\pi(\lambda_j^*|\lambda_{-j}, Q, \theta, \mathbf{y})}{\pi(\lambda_j|\lambda_{-j}, Q, \theta, \mathbf{y})}$$

Se a proposta para o próximo estado da cadeia, t , for aceita $\lambda_j^t = \lambda_j^*$, caso contrário $\lambda_j^t = \lambda_j^{t-1}$.

- ii) Atualização de Q_{ij} : a distribuição condicional completa a posteriori para cada Q_{ij} , é uma distribuição discreta com as seguinte probabilidade:

$$p_{ij} = \pi(Q_{ij} = q|\lambda, Q_{i(-j)}, \theta, \mathbf{y}) = \frac{\pi(Q_{ij} = q|\lambda_j)\pi(y_i|\theta, Q_i, Q_{ij} = q)}{\sum_q \pi(Q_{ij} = q|\lambda_j)\pi(y_i|\theta, Q_i, Q_{ij} = q)}$$

sendo $q = 0, 1, 2$ para a população F_2 e $q = 0, 1$ para a população de retrocruzamento que foi utilizada neste trabalho. Assim Q_{ij} pode ser atualizado usando o algoritmo Gibbs sampling.

- iii) Atualização de $\theta = (\mu, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma^2)$: como para os parâmetros $\mu, \alpha_j, \delta_j, \alpha_j\alpha_{j'}, \alpha_j\delta_{j'}, \delta_j\alpha_{j'}, \delta_j\delta_{j'}$ e σ^2 foram atribuídas distribuições a priori conjugadas, suas correspondentes distribuições condicionais completas a posteriori possuem formas fechadas, são dadas por:

- a) distribuição condicional completa a posteriori para μ :

$$\mu|\lambda, Q, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^n(mi)}{\sigma^2(\frac{1}{\tau^2} + \frac{n}{\sigma^2})}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}}\right) \quad (3.2)$$

em que:

$$mi = (y_i - \sum_{j=1}^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'4})$$

b) distribuição condicional completa a posteriori para α_{j^*} :

$$\alpha_{j^*} | \boldsymbol{\lambda}, Q, \mu, [\boldsymbol{\alpha}]_{j \neq j^*}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \sigma_2, y \sim N \left(\frac{\sum_{i=1}^n x_{ij^*1} a_i}{\sigma^2 \left(\frac{1}{\sigma_\alpha^2} + \frac{\sum_{i=1}^n x_{ij^*1}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_\alpha^2} + \frac{\sum_{i=1}^n Q_{ij^*}^2}{\sigma^2}} \right) \quad (3.3)$$

em que:

$$a_i = (y_i - \mu - \sum_{j \neq j^*}^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

c) distribuição condicional completa a posteriori para δ_{j^*} :

$$\delta_{j^*} | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, [\boldsymbol{\delta}]_{j \neq j^*}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \sigma_2, y \sim N \left(\frac{\sum_{i=1}^n x_{ij^*2} d_i}{\sigma^2 \left(\frac{1}{\sigma_\delta^2} + \frac{\sum_{i=1}^n x_{ij^*2}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_\delta^2} + \frac{\sum_{i=1}^n x_{ij^*2}^2}{\sigma^2}} \right) \quad (3.4)$$

em que:

$$d_i = (y_i - \mu - \sum_{j=1}^S \alpha_j x_{ij1} - \sum_{j \neq j^*}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

d) distribuição condicional completa a posteriori para σ^2 :

$$\sigma^2 | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \mathbf{y} \sim IG \left(a + \frac{n}{2}, \frac{1}{\frac{\sum_{i=1}^n S_i}{2} + \frac{1}{b}} \right) \quad (3.5)$$

$$S_i = (y_i - \mu - \sum_{j=1}^S \alpha_j x_{ijx} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})^2$$

d) distribuição condicional completa a posteriori para $\alpha\alpha_{j^*}$:

$$\alpha\alpha_{j^*} | \lambda, Q, \mu, \alpha, \delta, [\alpha\alpha]_{j \neq j^*}, \alpha\delta, \delta\alpha, \delta\delta, \sigma_2, y \sim N\left(\frac{\sum_{i=1}^n x_{ij^*j^*l_1} E1_i}{\sigma^2\left(\frac{1}{\sigma_{\gamma_1}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_1}^2}{\sigma^2}\right)}, \frac{1}{\frac{1}{\sigma_{\gamma_1}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_1}^2}{\sigma^2}}\right) \quad (3.6)$$

em que:

$$E1_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j \neq j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

e) distribuição condicional completa a posteriori para $\alpha\delta_{j^*}$:

$$\alpha\delta_{j^*} | \lambda, Q, \mu, \alpha, \delta, \alpha\alpha, [\alpha\delta]_{j \neq j^*}, \delta\alpha, \delta\delta, \sigma_2, y \sim N\left(\frac{\sum_{i=1}^n x_{ij^*j^*l_2} E2_i}{\sigma^2\left(\frac{1}{\sigma_{\gamma_2}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_2}^2}{\sigma^2}\right)}, \frac{1}{\frac{1}{\sigma_{\gamma_2}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_2}^2}{\sigma^2}}\right) \quad (3.7)$$

em que:

$$E2_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j \neq j^*}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

f) distribuição condicional completa a posteriori para $\delta\alpha_{j^*}$:

$$\delta\alpha_{j^*} | \lambda, Q, \mu, \alpha, \delta, \alpha\alpha, \alpha\delta, [\delta\alpha]_{3j \neq j^*}, \delta\delta, \sigma_2, y \sim N\left(\frac{\sum_{i=1}^n x_{ij^*j^*l_3} E3_i}{\sigma^2\left(\frac{1}{\sigma_{\gamma_3}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_3}^2}{\sigma^2}\right)}, \frac{1}{\frac{1}{\sigma_{\gamma_3}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_3}^2}{\sigma^2}}\right) \quad (3.8)$$

em que:

$$E3_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j \neq j^*}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

g) distribuição condicional completa a posteriori para $\delta\delta_{j^*}$:

$$\gamma_{4j^*} | \lambda, Q, \mu, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, [\delta\delta]_{j \neq j^*}, \sigma^2, y \sim N \left(\frac{\sum_{i=1}^n x_{ij^*j^*l_4} E4_i}{\sigma^2 \left(\frac{1}{\sigma_{\gamma_4}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_4}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_{\gamma_4}^2} + \frac{\sum_{i=1}^n x_{ij^*j^*l_4}^2}{\sigma^2}} \right) \quad (3.9)$$

em que:

$$E4_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j \neq j^*}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

Assim, tem-se que os parâmetros $\mu, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma^2$ podem ser atualizados usando o algoritmo Gibbs sampling.

Para fazer a seleção do modelo mais adequado adotou-se o fator de Bayes. Pelo fato de, neste caso, $f(\mathbf{y}|M)$ não possuir, neste caso, solução analítica utilizou-se os métodos de MCMC para estimar esta quantidade. Foram utilizados dois estimadores: estimador da média harmônica, proposto por Newton e Raftery (1994), dado por:

$$\hat{\pi}_{HM}(\mathbf{y}|M) = \left[\frac{1}{T} \sum_{t=1}^T \frac{1}{L(\mathbf{y}|M, \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\theta}^t)} \right]^{-1},$$

e o estimador estabilizado da média harmônica, como sugerido em Raftery et al. (2007), o qual consiste em integrar o estimador da média Harmônica em relação à σ^2 .

$$\hat{\pi}_{SHM}(\mathbf{y}|M) = \left[\frac{1}{T} \sum_{t=1}^T \frac{1}{L(\mathbf{y}|M, \boldsymbol{\lambda}^t, Q^t, h(\boldsymbol{\theta}^t))} \right]^{-1}, \quad (3.10)$$

em que $h(\boldsymbol{\theta}^t) = (\mu, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta)$. Em (3.10), $\hat{\pi}_{SHM}(\mathbf{y}|M)$ é a densidade de uma distribuição t escalonada, $St_n(\mathbf{y}|X\boldsymbol{\beta}, I, 2a_1)$.

Os passos para implementação do programa, utilizando-se o *Software R*, para construir uma amostra da distribuição conjunta a posteriori para os parâmetros de interesse são os seguintes:

1. Entrada dos dados fenotípicos, dos genótipos dos marcadores moleculares e das distância entre os marcadores moleculares;
2. Construção do genoma: Neste passo faz-se a concatenação dos cromossomos, de forma que todos os cromossomos possam ser considerados simultaneamente;

3. Atribuição dos valores iniciais para todos os parâmetros;
4. Atualização dos parâmetros:
 - 4.1 Inicia o contador de iterações em $t = 1$
 - 4.2 Atualização de $\sigma^2 | \boldsymbol{\lambda}^{t-1}, Q^{t-1}, \mu^{t-1}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, \mathbf{y}$, utilizando a equação (5);
 - 4.3 Atualização de $\lambda_j | \boldsymbol{\lambda}_{-j}^{t-1}, Q^{t-1}, \mu^{t-1}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \gamma_3^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$, conforme descrito no item *i*);
 - 4.3 Atualização de $Q_{ij} | \boldsymbol{\lambda}_{-j}^t, \mu^{t-1}, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$, conforme descrito no item *ii*);
 - 4.4 Atualização de $\mu | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$ utilizando a equação (2);
 - 4.5 Atualização de $\alpha_j | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}_{-j}^{t-1}, \boldsymbol{\delta}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^1, \mathbf{y}$ utilizando a equação (3);
 - 4.6 Atualização de $\delta_j | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}^t, \boldsymbol{\delta}_{-j}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$ utilizando a equação (4);
 - 4.7 Atualização de $\alpha\alpha_j | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}^t, \boldsymbol{\delta}^t, \boldsymbol{\alpha}\boldsymbol{\alpha}_{-j}^{t-1}, \boldsymbol{\alpha}\boldsymbol{\delta}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$ utilizando a equação (6);
 - 4.8 Atualização de $\alpha\delta_j | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}^t, \boldsymbol{\delta}^t, \boldsymbol{\alpha}\boldsymbol{\alpha}^t, \boldsymbol{\alpha}\boldsymbol{\delta}_{-j}^{t-1}, \boldsymbol{\delta}\boldsymbol{\alpha}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$ utilizando a equação (7);
 - 4.9 Atualização de $\delta\alpha_j | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}^t, \boldsymbol{\delta}^t, \boldsymbol{\alpha}\boldsymbol{\alpha}^t, \boldsymbol{\alpha}\boldsymbol{\delta}_2^t, \boldsymbol{\delta}\boldsymbol{\alpha}_{-j}^{t-1}, \boldsymbol{\delta}\boldsymbol{\delta}^{t-1}, (\sigma^2)^t, \mathbf{y}$ utilizando a equação (8);
 - 4.10 Atualização de $\delta\delta_j | \boldsymbol{\lambda}^t, Q^t, \boldsymbol{\alpha}^t, \boldsymbol{\delta}^t, \boldsymbol{\alpha}\boldsymbol{\alpha}^t, \boldsymbol{\alpha}\boldsymbol{\delta}_2^t, \boldsymbol{\delta}\boldsymbol{\alpha}^t, \boldsymbol{\delta}\boldsymbol{\delta}_{-j}^{t-1}, (\sigma^2)^t, \mathbf{y}$ utilizando a equação (9);
 - 4.10 Após atualizar todos os parâmetros calculou-se os valores das densidades $L(\boldsymbol{\lambda}^t, Q^t, \boldsymbol{\theta}^t | \mathbf{y})$ e $L(\boldsymbol{\lambda}^t, Q^t, h(\boldsymbol{\theta}^t) | \mathbf{y})$, para posteriormente realizar o cálculo do fator de Bayes.
 - 4.11 Feita as atualizações dos parâmetros, incrementar o contador de iterações t e voltar ao passo 4.2;
5. Repetir o passo 4.11 M vezes, considerar o período de aquecimento da cadeia “burn-in” e o espaçamento entre os pontos amostrados denominado de “thin”.

3.3 Estudo de simulação

Para investigar a eficiência da metodologia foi realizado um estudo de simulação. No estudo de simulação foram gerados conjuntos de dados de experimentos para mapear QTLs em populações de retrocruzamento e em populações F_2 considerando-se a presença e a ausência de epistasia. No estudo de simulação foram considerados 200 indivíduos, três cromossomos de comprimento igual a 1,3 Morgan, cada um contendo quatorze marcadores moleculares, igualmente espaçados ao longo do cromossomo. Foram inseridos três QTLs e a herdabilidade considerada para a característica fenotípica foi de 80%. Os valores para os parâmetros considerados na simulação são apresentados na Tabela 3.1.

Tabela 3.1 – Valores das posições (λ) e dos efeitos genéticos: (α), (δ), ($\alpha\alpha$), ($\alpha\delta$), ($\delta\alpha$) e ($\delta\delta$) dos três QTLs adotados na simulação dos conjuntos de dados

População	Cromossomo	λ	α	δ	$\alpha\alpha$	$\alpha\delta$	$\delta\alpha$	$\delta\delta$
Retrocruzamento (cenário 1)	1	0,6500	1,0000	—				
	1	1,0500	2,0000	—				
	3	3,2500	3,0000	—				
F_2 (cenário 2)	1	0,6500	1,3359	0,2821				
	1	1,0500	2,0056	—0,6240				
	3	3,2500	2,7559	—0,4530				
Retrocruzamento (cenário 3)	1	0,6500	1,0000	—	0,0000	—	—	—
	1	1,0500	2,0000	—	1,0000	—	—	—
	3	3,2500	3,0000	—	0,0000	—	—	—
F_2 (cenário 4)	1	0,6500	1,3359	0,2821	0,0000	0,0000	0,0000	2,0000
	1	1,0500	2,0056	—0,6240	0,0000	0,0000	0,0000	0,0000
	3	3,2500	2,7559	—0,4530	—1,0000	0,5000	0,0000	0,0000

Considerando-se o modelo 3.1 e as informações dos parâmetros dadas na Tabela 3.1 foram gerados 50 conjuntos de dados para cada um dos cenários a seguir:

- 1) Modelo Sem Epistasia - População de Retrocruzamento (MSE-PR);
- 2) Modelo Com Epistasia - População de Retrocruzamento (MCE-PR);
- 3) Modelo Sem Epistasia - População F_2 (MSE-PF2);
- 4) Modelo Com Epistasia - População F_2 (MCE-PF2).

3.3.1 Resultados do estudo de simulação

Para o estudo de simulação foi implementado um programa (APÊNDICE A), utilizando a abordagem bayesiana para mapear QTLs. A implementação computacional foi realizada utilizando-se

o programa estatístico R (R DEVELOPMENT CORE TEAM, 2008). A construção da amostra da distribuição conjunta a posteriori foi feita utilizando os métodos MCMC. Foram geradas cadeias com 120.000 iterações, sendo que as 1000 primeiras iterações foram descartadas para o período de aquecimento da cadeia (“burn-in”) e foi utilizado espaçamento de 50 iterações, com a finalidade de diminuir a correlação entre os valores amostrados, resultando em uma amostra final de tamanho 2380. A indicação de convergência de cada cadeia foi verificada visualmente e por meio do critério Geweke, implementado no pacote CODA do programa estatístico R.

Resultados do estudo de simulação: MSE-PR

Na análise de cada um dos 50 conjuntos de dados simulados considerando-se o modelo MSE-PR foram ajustados modelos contendo um, dois, três, quatro e cinco QTLs. O tempo computacional de cada um dos conjuntos de dados variou de 4 horas, para o ajuste do modelo com um QTL à 10 horas para o ajuste do modelo com cinco QTLs. O algoritmo Metropolis-Hastings foi utilizado para atualizar a localização dos QTLs, os demais parâmetros foram atualizados utilizando-se o algoritmo Gibbs Sampling. A taxa de aceitação para as localizações dos QTLs variou entre 33% e 50%. Inicialmente foi verificado a convergência das cadeias para os parâmetros envolvidos nos modelos. Em 20% das análises em que foram ajustados o modelo com um QTL, verificou-se que a convergência da cadeia, para pelo menos um dos parâmetros, não foi diagnosticada. O mesmo foi observado quando foram ajustados o modelo com dois QTLs. Em 30% das análises em que foram ajustados o modelo com três QTL, verificou-se que a convergência da cadeia, para pelo menos um dos parâmetros, não foi diagnosticada. A convergência também não foi verificada para pelo menos um dos parâmetros do modelo quando foram ajustados modelos com quatro e cinco QTLs em 100% das análises. Sendo assim, a seleção do modelo foi feita entre os modelos contendo um, dois e três QTLs, e para os conjuntos de dados, 31 no total, nos quais a convergência foi observada para todos os parâmetros envolvidos no modelo. As estimativas do fator de Bayes foram obtidas considerando-se o estimador média harmônica (MH) e estimador estabilizador da média harmônica (EMH). Na Tabela 3.2 apresenta-se a porcentagem de classificação, segundo Jeffreys (1961), do fator de Bayes.

Tabela 3.2 – Classificação do fator de Bayes

Classificação	B_{21}		B_{32}	
	MH(%)	EMH(%)	MH(%)	EMH(%)
Muito Fraca	0	0	0	0
Fraca	3	0	0	0
Forte	10	3	0	0
Muito Forte	87	97	100	100

De acordo com a classificação do fator de Bayes, apresentada na Tabela 3.2, no estudo de simulação o modelo com 2 QTLs se ajusta melhor aos dados que o modelo com 1 QTL, e o modelo com três QTLs se ajusta melhor aos dados que o modelo com 2 QTLs. Sendo assim, o modelo que melhor se ajusta aos dados é o modelo contendo três QTLs, ou seja o modelo contendo o número verdadeiro de QTLs considerados na simulação dos dados. Visto que em 100% das análises o modelo 3 foi escolhido quando comparado ao modelo 2. Assim, o fator de Bayes, para ambos estimadores, foi bastante eficiente para fazer a seleção do modelo.

A Tabela 3.3 apresenta um resumo a posteriori para cada parâmetro obtido na análise de dois conjuntos de dados, selecionados aleatoriamente, no estudo de simulação. Observa-se que os valores estimados estão próximos dos valores verdadeiros. Além disso, observa-se que a região HPD cobre os valores verdadeiros dos parâmetros.

Tabela 3.3 – Resumo a posteriori e a região HPD de 95% (HPD 95%) para os parâmetros do modelo contendo três QTLs, considerando dois diferentes conjuntos de dados simulados a partir do modelo MSE-PR

Conjunto	Par.	Verd	Média	d.p.	Mediana	HPD(2,5%)	HPD(97,5%)
I	μ	0,0000	0,0053	0,0705	0,0100	-0,1329	0,1425
	σ^2	0,9283	0,8941	0,0978	0,8900	0,7220	1,0896
	α_1	1,0000	0,9455	0,1747	0,9500	0,5696	1,2656
	α_2	2,0000	1,8647	0,1664	1,8600	1,5424	2,1860
	α_3	3,0000	3,2183	0,1396	3,2200	2,9477	3,4800
	λ_1	0,6500	0,6793	0,0310	0,6800	0,6249	0,7444
	λ_2	1,0500	1,0523	0,0188	1,0500	1,0158	1,0854
	λ_3	3,2500	3,2402	0,0122	3,2400	3,2172	3,2641
II	μ	0,0000	-0,0353	0,0696	-0,0400	-0,1668	0,0994
	σ^2	0,9283	0,8917	0,0933	0,8900	0,7220	1,0867
	α_1	1,0000	1,0771	0,1570	1,0800	0,7716	1,3980
	α_2	2,0000	2,0755	0,1604	2,0700	1,7651	2,3838
	α_3	3,0000	2,9314	0,1395	2,9300	2,6388	3,1820
	λ_1	0,6500	0,6214	0,0398	0,6200	0,5445	0,6903
	λ_2	1,0500	1,0553	0,0153	1,0600	1,0263	1,0850
	λ_3	3,2500	3,2361	0,0112	3,2400	3,2162	3,2582

As Figuras de 3.1 a 3.6, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para os conjuntos de dados I e II. A linha vertical vermelha, apresentada nos histogramas, representa o valor verdadeiro do parâmetro.

Figura 3.1 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , e para os efeitos aditivos α_1 , α_2 e α_3 , considerando o conjunto de dados I

Figura 3.2 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs, λ_1 , λ_2 e λ_3 , e para os efeitos aditivos α_1 , α_2 e α_3 , considerando o conjunto de dado I

Figura 3.3 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , e para os efeitos aditivos α_1 , α_2 e α_3 , considerando o conjunto de dado I

Figura 3.4 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , e para os efeitos aditivos α_1 , α_2 e α_3 , considerando o conjunto de dados II

Figura 3.5 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 e para os efeitos aditivos α_1 , α_2 e α_3 , considerando o conjunto de dado II

Figura 3.6 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 e para os efeitos aditivos α_1 , α_2 e α_3 , considerando o conjunto de dado II

A Figura 3.7 apresenta um gráfico dos intervalos de credibilidade obtidos na análise dos 31 conjuntos de dados considerando-se o modelo com três QTLs. A linha vertical vermelha representa o valor verdadeiro do parâmetro. Nota-se que os resultados foram muito satisfatórios, tendo em vista que a grande maioria dos intervalos de credibilidade cobriram os verdadeiros valores para todos os parâmetros.

Figura 3.7 – Intervalos de credibilidade para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 e para os efeitos aditivos α_1 , α_2 e α_3 do modelo contendo três QTLs, considerando os 31 conjuntos de dados simulados utilizando o modelo MSE-PR

Resultados do estudo de simulação: MSE-PF2

Na análise de cada um dos 50 conjuntos de dados simulados considerando-se o modelo MSE-PF2, assim como na análise dos dados considerando-se o modelo MSE-PR, foram ajustados modelos contendo um, dois, três, quatro e cinco QTLs. O tempo computacional de cada um dos conjuntos de dados variou de 5 horas, para o ajuste do modelo com um QTL à 17 horas para o ajuste do modelo com cinco QTLs. A taxa de aceitação do algoritmo Metropolis-Hastings para as atualizações das localizações dos QTLs variou de 31% a 51%. Foi verificada a convergência das cadeias para os parâmetros envolvidos nos modelos. Em 18% das análises em que foram ajustados o modelo com um QTL, verificou-se que a convergência da cadeia, para pelo menos um dos parâmetros, não foi diagnosticada. Em 32% das análises em que foram ajustados o modelo com dois QTL, verificou-se que a convergência da cadeia, para pelo menos um dos parâmetros, não foi diagnosticada. Em 18% das análises em que foram ajustados o modelo com dois QTL, verificou-se que a convergência da

cadeia, para pelo menos um dos parâmetros, não foi diagnosticada. Novamente, a convergência não foi verificada para pelo menos um dos parâmetros do modelo quando foram ajustados modelos com quatro e cinco QTLs em 100% das análises. Sendo assim, a seleção do modelo foi feita entre os modelos contendo um, dois e três QTLs, e para os conjuntos de dados, 30 no total, nos quais a convergência foi observada para todos os parâmetros envolvidos no modelo. Na Tabela 3.4 apresenta-se a porcentagem de classificação, segundo Jeffreys (1961), do fator de Bayes.

Tabela 3.4 – Classificação do fator de Bayes

Classificação	B_{21}		B_{32}	
	MH (%)	EMH (%)	MH (%)	EMH (%)
Muito Fraca	7	7	0	0
Fraca	4	0	0	0
Forte	15	19	0	0
Muito Forte	74	74	100	100

De acordo com a classificação do fator de Bayes, apresentada na Tabela 3.4, no estudo de simulação o modelo com 2 QTLs se ajusta melhor aos dados que o modelo com 1 QTL, e o modelo com três QTLs se ajusta melhor aos dados que o modelo 2 QTLs. Assim, no estudo de simulação o modelo que melhor se ajusta aos dados é o modelo com três QTLs, ou seja o modelo contendo o número verdadeiro de QTLs considerados na simulação dos dados. Assim, novamente, o fator de Bayes, para ambos estimadores, foi bastante eficiente para fazer a seleção do modelo.

A Tabela 3.5 apresenta um resumo a posteriori para cada parâmetro obtido na análise de dois conjuntos de dados, selecionados aleatoriamente, no estudo de simulação. Observa-se que os valores estimados estão próximos dos valores verdadeiros. Além disso, observa-se que a região HPD cobre os valores verdadeiros dos parâmetros.

As Figuras de 3.8 a 3.13, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para os conjuntos de dados I e II. A linha vertical vermelha, apresentada nos histogramas, representa o valor verdadeiro do parâmetro.

A Figura 3.14 apresenta os intervalos de credibilidade obtidos no estudo de simulação para o modelo contendo 3 QTLs, sendo que a linha vertical representa o valor verdadeiro. Novamente, os resultados foram satisfatórios, tendo em vista que a grande maioria dos intervalos de credibilidade cobriram os verdadeiros valores para todos os parâmetros.

Tabela 3.5 – Resumo a posteriori e a região HPD de 95% (HPD 95%) para os parâmetros dos modelos contendo três QTLs, considerando dois conjuntos de dados simulados a partir do modelo MSE-PF2

Conjunto	Par.	Verd	Média	d.p.	Mediana	HPD(2, 5)	HPD(97, 5)
I	μ	0,0000	0,0947	0,1030	0,1000	-0,1156	0,2824
	σ^2	1,7129	1,6110	0,2041	1,5900	1,2352	2,0220
	α_1	1,3359	1,2998	0,1667	1,3000	1,0056	1,6555
	α_2	2,0056	2,0003	0,1680	2,0100	1,6784	2,3331
	α_3	2,7559	2,6735	0,1459	2,6800	2,3728	2,9412
	δ_1	0,2821	0,3621	0,2133	0,3600	-0,0712	0,7728
	δ_2	-0,6240	-0,8275	0,2111	-0,8300	-1,2324	-0,4204
	δ_3	-0,4530	-0,3876	0,2022	-0,3800	-0,7532	0,0270
	λ_1	0,6500	0,6496	0,0202	0,6500	0,6139	0,6878
	λ_2	1,0500	1,0366	0,0144	1,0400	1,0094	1,0655
	λ_3	3,2500	3,2639	0,0105	3,2600	3,2428	3,2831
II	μ	0,0000	-0,0662	0,1020	-0,0700	-0,2744	0,1268
	σ^2	1,7129	1,6265	0,2025	1,6100	1,2605	2,0426
	α_1	1,3359	1,2302	0,1649	1,2300	0,9128	1,5651
	α_2	2,0056	2,0965	0,1721	2,1000	1,7332	2,4026
	α_3	2,7559	2,6705	0,1529	2,6700	2,3633	2,9666
	δ_1	0,2821	0,3759	0,2163	0,3800	-0,0641	0,7803
	δ_2	-0,6240	-0,6207	0,2081	-0,6200	-1,0585	-0,2531
	δ_3	-0,4530	-0,4079	0,2060	-0,4100	-0,8005	-0,0036
	λ_1	0,6500	0,6502	0,0208	0,6500	0,6128	0,6908
	λ_2	1,0500	1,0258	0,0131	1,0300	1,0016	1,0510
	λ_3	3,2500	3,2603	0,0102	3,2600	3,2399	3,2797

Figura 3.8 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dados I

Figura 3.9 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dados I

Figura 3.10 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado I

Figura 3.11 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dados II

Figura 3.12 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado II

Figura 3.13 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado II

Figura 3.14 – Intervalos de credibilidade para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e de dominância δ_1 , δ_2 e δ_3 , do modelo contendo três QTLs considerando os 30 conjuntos de dados simulados utilizando o modelo MSE-PF2

Resultados do estudo de simulação: MCE-PR

Na análise de cada um dos 50 conjuntos de dados simulados considerando-se o modelo MCE-PR foram ajustados modelos contendo três QTLs com e sem efeitos de epistasia. O tempo computacional de cada um dos conjuntos de dados variou de 7 horas, para o ajuste do modelo com três QTLs e sem efeitos de epistasia à 11 horas para o ajuste do modelo com três QTLs e todos os possíveis efeitos de epistasia. A taxa de aceitação do algoritmo Metropolis-Hasting para atualizar as localizações dos QTLs variou entre 32% e 46%. Inicialmente foi verificado a convergência das cadeias para os parâmetros envolvidos nos modelos. Em 20% das análises em que foram ajustados o modelo com três QTL com e sem epistasia, verificou-se que a convergência da cadeia, para pelo menos um dos parâmetros, não foi diagnosticada. Assim, a seleção do modelo foi feita entre os modelos com três QTLs com e sem efeitos de epistasia, e para os conjuntos de dados, 40 no total, nos quais a convergência foi observada para todos os parâmetros envolvidos no modelo. Na Tabela 3.6 apresenta-se a porcentagem de classificações das estimativas do fator de Bayes que foi obtidas por meio do estimador média harmônica (MH) e estimador estabilizador da média harmônica (EMH).

Tabela 3.6 – Classificação do fator de Bayes

Classificação	B_{3+3} (1)	
	MH (%)	EMH (%)
Nenhuma evidência	26	28
Muito Fraca	3	3
Fraca	5	8
Forte	18	13
Muito Forte	49	49

(1) Fator de Bayes considerando o modelo com três QTLs e todos os possíveis efeitos de epistasia (3_{++}) versus o modelo com três QTLs e sem efeitos de epistasia

De acordo com a classificação do fator de Bayes, apresentada na Tabela 3.6, no estudo de simulação o modelo com 3 QTLs e com todos os possíveis efeitos de epistasia se ajusta melhor aos dados que o modelo com 3 QTLs e sem efeitos de epistasia, visto que, tal modelo é selecionado em 74% das análises considerando-se o estimador média hamônica (MH) e em 72% das análises considerando-se o estimador estabilizador da média harmônica.

A Tabela 3.7 apresenta um resumo a posteriori para cada um dos parâmetros obtido na análise de dois conjuntos de dados, selecionados aleatoriamente, no estudo de simulação. Observa-se que os valores estimados estão próximos dos valores verdadeiros. Além disso, observa-se que a região HPD cobre os valores verdadeiros dos parâmetros.

Tabela 3.7 – Resumo a posteriori e a região HPD 95% (HPD 95%) para os parâmetros do modelos contendo três QTLs e todos os possíveis efeitos epistáticos, considerando os conjuntos de dados I e II simulados a partir do modelo MCE-PR

Conjunto	Par.	Verd	Média	d.p.	Mediana	HPD(2, 5%)	IC(97, 5%)
I	μ	2,0000	2,0468	0,0900	2,0500	1,8715	2,2168
	σ^2	0,9481	0,9461	0,1119	0,9400	0,7429	1,1744
	α_1	1,0000	0,8233	0,1958	0,8200	0,4487	1,2062
	α_2	2,0000	1,9400	0,2022	1,9500	1,5606	2,3243
	α_3	3,0000	2,8792	0,1458	2,8800	2,5797	3,1681
	λ_1	0,6500	0,7026	0,0608	0,6800	0,6210	0,8346
	λ_2	1,0500	1,0387	0,0143	1,0400	1,0121	1,0643
	λ_3	3,2500	3,2576	0,0127	3,2600	3,2357	3,2844
	$\alpha_1\alpha_2$	0,0000	0,2161	0,3706	0,2100	-0,5312	0,8900
	$\alpha_1\alpha_3$	1,0000	1,3519	0,3616	1,3500	0,6721	2,0525
	$\alpha_2\alpha_3$	0,0000	-0,1329	0,3649	-0,1100	-0,8739	0,5115
II	μ	2,0000	2,0216	0,0930	2,0200	1,8476	2,2118
	σ^2	0,9481	1,1000	0,1266	1,1000	0,8571	1,3437
	α_1	1,0000	0,7536	0,2104	0,7600	0,3486	1,1620
	α_2	2,0000	1,9243	0,2007	1,9300	1,4913	2,2850
	α_3	3,0000	2,9111	0,1536	2,9100	2,6097	3,1998
	λ_1	0,6500	0,6603	0,0257	0,6600	0,6062	0,7065
	λ_2	1,0500	1,0594	0,0184	1,0600	1,0234	1,0933
	λ_3	3,2500	3,2550	0,0126	3,2600	3,2306	3,2786
	$\alpha_1\alpha_2$	0,0000	0,0414	0,3912	0,0400	-0,7372	0,7921
	$\alpha_1\alpha_3$	1,0000	1,1046	0,3889	1,1100	0,3379	1,8512
	$\alpha_2\alpha_3$	0,0000	0,2803	0,3894	0,2800	-0,4343	1,0952

As Figuras de 3.15 a 3.20, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para os conjuntos simulados a partir do modelo MCE-PR. A linha vertical vermelha, apresentada nos histogramas, representa o valor verdadeiro do parâmetro.

A Figura 3.21 apresenta os intervalos de credibilidade obtidos no estudo de simulação para o modelo contendo 3 QTLs e todos os possíveis efeitos de epistasia, sendo que a linha vertical representa o valor verdadeiro dos parâmetros.

Figura 3.15 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasias $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, considerando o conjunto de dados I

Figura 3.16 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasias $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, considerando o conjunto de dados I

Figura 3.17 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, considerando o conjunto de dado I

Figura 3.18 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, considerando o conjunto de dados II

Figura 3.19 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, considerando o conjunto de dado II

Figura 3.20 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, considerando o conjunto de dado II

Figura 3.21 – Intervalo de Credibilidade para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$ e $\alpha_2\alpha_3$, do modelo contendo três QTLs considerando os 40 conjuntos de dados simulados utilizando o modelo MCE-PR

Resultados do estudo de simulação: MCE-PF2

Na análise de cada um dos 50 conjuntos de dados simulados considerando-se o modelo MCE-PF2 foram ajustados modelos contendo três QTLs com e sem efeitos de epistasia. O tempo computacional de cada um dos conjuntos de dados variou de 9 horas, para o ajuste do modelo com três QTLs e sem efeitos de epistasia à 15 horas para o ajuste do modelo com três QTLs e todos os possíveis efeitos de epistasia. A taxa de aceitação do algoritmo Metropolis-Hasting para atualizar as localizações dos QTLs variou entre 35% e 53%. Inicialmente foi verificado a convergência das cadeias para os parâmetros envolvidos nos modelos. Em 20% e 24% das análises em que foram ajustados, respectivamente, o modelo com três QTL sem e com efeito de epistasia, verificou-se que a convergência da cadeia não foi obtida, para pelo menos um dos parâmetros. Assim, a seleção do modelo foi feita entre os modelos com três QTLs com e sem efeitos de epistasia, e para os conjuntos de dados, 33 no total, nos quais a convergência foi observada para todos os parâmetros envolvidos no modelo. Na Tabela 3.8 apresenta-se a porcentagem de classificações das estimativas do fator de Bayes que foi obtidas por meio do estimador média harmônica (MH) e estimador estabilizador da média harmônica (EMH).

Tabela 3.8 – Classificação do fator de Bayes

Classificação	B_{3+3}	
	MH (%)	EMH (%)
Nenhuma evidência	39	33
Muito Fraca	9	12
Fraca	12	12
Forte	6	3
Muito Forte	34	40

⁽¹⁾ Fator de Bayes considerando o modelo com três QTLs e todos os possíveis efeitos de epistasia (3_{+}) versus o modelo com três QTLs e sem efeitos de epistasia

De acordo com a classificação do fator de Bayes, apresentada na Tabela 3.8, no estudo de simulação o modelo 3 com todos os possíveis efeitos de epistasia se ajusta melhor aos dados que o modelo 3 sem efeitos de epistasia, visto que tal modelo foi selecionado em 61% das análises considerando-se o estimador média hamônica (MH) e em 67% das análises considerando-se o estimador Estabilizador da média harmônica.

A Tabela 3.9 apresenta um resumo a posteriori para cada parâmetro obtido na análise de dois conjuntos de dados, selecionados aleatoriamente, no estudo de simulação. Observa-se que os valores estimados estão próximos dos valores verdadeiros. Além disso, observa-se que a região HPD cobre os valores verdadeiros dos parâmetros.

As Figuras de 3.22 a 3.33, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para os conjuntos de dados I e II simulados a partir do modelo MCE-PF2. A linha vertical vermelha, apresentada nos histogramas, representa o valor verdadeiro do parâmetro.

As Figuras 3.34 e 3.35 apresentam os intervalos de credibilidade obtidos no estudo de simulação para o modelo contendo 3 QTLs e todos os possíveis efeitos de epistasia, sendo que a linha vertical representa o valor verdadeiro dos parâmetros. Os resultados das análises foram muito satisfatórios, tendo em vista que a grande maioria dos intervalos de credibilidade cobriram os verdadeiros valores para todos os parâmetros.

Tabela 3.9 – Resumo a posteriori e a região HPD 95% (HPD 95%) para os parâmetros do modelos contendo três QTLs e todos os possíveis efeitos epistáticos, considerando os conjuntos de dados I e II simulados a partir do modelo MCE-PF2

Conjunto	Par.	Verd	Média	d.p.	Mediana	HPD(2, 5%)	HPD(97, 5%)
I	μ	2,0000	2,1915	0,1546	2,1900	1,8826	2,4837
	σ^2	2,2958	2,3137	0,2877	2,2900	1,7986	2,9212
	α_1	1,3359	1,3965	0,2484	1,3900	0,9348	1,8805
	α_2	2,0056	2,4928	0,2632	2,5000	1,9453	2,9870
	α_3	2,7559	2,7241	0,1820	2,7200	2,3930	3,0945
	δ_1	0,2821	0,2469	0,3240	0,2500	-0,3832	0,8899
	δ_2	-0,6240	-0,8989	0,3174	-0,9000	-1,5031	-0,2965
	δ_3	-0,4530	-0,2060	0,2474	-0,2000	-0,6623	0,2852
	λ_1	0,6500	0,6329	0,0278	0,6300	0,5671	0,6772
	λ_2	1,0500	1,0449	0,0141	1,0400	1,0190	1,0734
	λ_3	3,2500	3,2341	0,0121	3,2300	3,2106	3,2567
	$\alpha_1\alpha_2$	0,0000	0,1998	0,4113	0,1900	-0,6662	0,9486
	$\alpha_1\alpha_3$	0,0000	0,0458	0,2744	0,0500	-0,4897	0,5893
	$\alpha_2\alpha_3$	-1,0000	-1,1215	0,2992	-1,1200	-1,6661	-0,5215
	$\alpha_1\delta_2$	0,0000	0,4263	0,4819	0,4200	-0,5330	1,3671
	$\alpha_1\delta_3$	0,5000	0,8044	0,3925	0,7900	0,0274	1,5582
	$\alpha_2\delta_3$	0,0000	-0,4534	0,4317	-0,4500	-1,3804	0,3358
	$\delta_1\alpha_2$	0,0000	0,2654	0,5094	0,2600	-0,7533	1,2407
	$\delta_1\alpha_3$	0,0000	-0,2543	0,3766	-0,2500	-1,0258	0,4224
	$\delta_2\alpha_3$	0,0000	-0,1456	0,3646	-0,1500	-0,8361	0,5605
	$\delta_1\delta_2$	2,0000	0,8206	0,6398	0,8200	-0,3990	2,0834
	$\delta_1\delta_3$	0,0000	0,0794	0,5083	0,0800	-0,9008	1,1049
	$\delta_3\delta_3$	0,0000	-0,3406	0,4993	-0,3400	-1,3016	0,6384
II	μ	2,0000	2,1540	0,1793	2,1600	1,7988	2,4837
	σ^2	2,2958	2,5384	0,3228	2,5100	1,9162	2,9212
	α_1	1,3359	1,6562	0,3071	1,6500	1,0999	1,8805
	α_2	2,0056	1,8230	0,3285	1,8400	1,1750	2,9870
	α_3	2,7559	2,4417	0,1977	2,4400	2,0534	3,0945
	δ_1	0,2821	0,3841	0,3844	0,3900	-0,3805	0,8899
	δ_2	-0,6240	-0,8898	0,3969	-0,9000	-1,6130	-0,2965
	δ_3	-0,4530	-0,5947	0,2771	-0,5900	-1,1551	0,2852
	λ_1	0,6500	0,6521	0,0282	0,6500	0,6067	0,6772
	λ_2	1,0500	1,0475	0,0193	1,0500	1,0109	1,0734
	λ_3	3,2500	3,2342	0,0158	3,2300	3,2043	3,2567
	$\alpha_1\alpha_2$	0,0000	0,1391	0,5042	0,1500	-0,9280	0,9486
	$\alpha_1\alpha_3$	0,0000	-0,1952	0,3140	-0,1900	-0,8320	0,5893
	$\alpha_2\alpha_3$	-1,0000	-0,6271	0,3587	-0,6400	-1,3754	-0,5215
	$\alpha_1\delta_2$	0,0000	-0,6030	0,5550	-0,6100	-1,7271	1,3671
	$\alpha_1\delta_3$	0,5000	0,7643	0,4371	0,7700	-0,0756	1,5582
	$\alpha_2\delta_3$	0,0000	-0,5522	0,4808	-0,5400	-1,4429	0,3358
	$\delta_1\alpha_2$	0,0000	0,3074	0,6010	0,3000	-0,8476	1,2407
	$\delta_1\alpha_3$	0,0000	0,2583	0,4114	0,2600	-0,5760	0,4224
	$\delta_2\alpha_3$	0,0000	0,3335	0,4058	0,3400	-0,4704	0,5605
	$\delta_1\delta_2$	2,0000	1,2560	0,7094	1,2500	-0,1590	2,0834
	$\delta_1\delta_3$	0,0000	0,2133	0,5439	0,2200	-0,8959	1,1049
	$\delta_3\delta_3$	0,0000	-0,2965	0,5547	-0,3000	-1,3525	0,6384

Figura 3.22 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3) e para os efeitos aditivos α_1 , α_2 e α_3 e para os efeitos de os efeitos de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dados I

Figura 3.23 – Histogramas dos valores gerados para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, considerando o conjunto de dados I

Figura 3.24 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para os efeitos de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado I

Figura 3.25 – Trajetória das cadeias geradas para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, considerando o conjunto de dado I

Figura 3.26 – Autocorrelação das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para os efeitos de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado I

Figura 3.27 – Autocorrelação das cadeias geradas para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, considerando o conjunto de dado I

Figura 3.28 – Histogramas dos valores gerados para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para os efeitos de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dados II

Figura 3.29 – Histogramas dos valores gerados para as epistasias $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, considerando o conjunto de dados II

Figura 3.30 – Trajetória das cadeias geradas para a constante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para os efeitos de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado II

Figura 3.31 – Trajetória das cadeias geradas para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, considerando o conjunto de dado II

Figura 3.32 – Autocorrelação das cadeias geradas para a constante do modelo (μ), para σ^2 , para as localizações dos QTLs (λ_1 , λ_2 e λ_3 , para os efeitos aditivos α_1 , α_2 e α_3 e para os efeitos de dominância δ_1 , δ_2 e δ_3 , considerando o conjunto de dado II

Figura 3.33 – Autocorrelação das cadeias geradas para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, considerando o conjunto de dado II

Figura 3.34 – Intervalo de Credibilidade para a constante do modelo (μ), para σ^2 , para as localizações dos QTLs (λ_1 , λ_2 e λ_3), para os efeitos aditivos (α_1 , α_2 e α_3) e de dominância (δ_1, δ_2 e δ_3) do modelo ajustado com três QTLs considerando os 33 conjuntos de dados simulados utilizando o modelo MCE-PF2

Figura 3.35 – Intervalo de Credibilidade para as epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$, do modelo ajustado com três QTLs considerando os 33 conjuntos de dados simulados utilizando o modelo MCE-PF2

3.4 Discussão

No presente trabalho foi implementado, utilizando-se o programa estatístico R (R DEVELOPMENT CORE TEAM, 2008), uma abordagem bayesiana para mapear QTLs em que múltiplos QTLs e os efeitos de epistasia são considerados no modelo. Para tanto, o número de QTLs foi tratado como conhecido, foram ajustados modelos com números crescentes de QTLs e o fator de Bayes foi utilizado para selecionar o modelo mais adequado e conseqüentemente, para estimar o número de QTLs que controlam os fenótipos de interesse. O fator Bayes foi estimado por meio de dois diferentes estimadores: média harmônica e estabilizador da média harmônica. Para poder modelar múltiplos QTLs e os efeitos de epistasia foi feita a concatenação dos grupos de ligação e assim todos os cromossomos puderam ser considerados simultaneamente.

Para investigar a eficiência da metodologia foi feito um estudo de simulação em que foram considerados duas populações experimentais: retrocruzamento e F_2 . Sendo que para ambas as populações foi feito o estudo de simulação considerando modelos com e sem epistasia. Nos quatro casos o número verdadeiro de QTLs considerado na simulação dos dados foi de três QTLs. Para ambas populações experimentais: retrocruzamento e F_2 nos casos em que não foram considerados efeitos de epistasia no modelo foram ajustados modelos com 1, 2, 3, 4 e 5 QTLs e observou-se que para as análises em que o número de QTLs foi maior que o número verdadeiro de QTLs considerado na simulação, a convergência da cadeia não foi verificada para pelo menos um dos parâmetros, sendo assim a seleção de modelos foi feita apenas entre os modelos com 1, 2 e 3 QTLs e somente para as amostras em que todos os parâmetros apresentam convergência. Observou-se que o fator de Bayes foi muito eficiente para selecionar o melhor modelo, pois para ambas as populações foi selecionado o modelo com o número verdadeiro de QTL. Para os casos em que foram considerados efeitos de epistasia no modelo foram ajustados modelos com o número de verdadeiro de QTLs com e sem efeitos de epistasia, e novamente observou-se que a metodologia implementada foi bastante eficiente tendo em vista que na grande maioria das amostras o fator de Bayes selecionou o modelo contendo o número verdadeiro de QTLs e efeitos de epistasia.

Assim, diante do exposto a abordagem bayesiana para mapear QTLs em que o número de QTLs é considerado como conhecido e o fator de Bayes é utilizado para determinar o número de QTLs mostrou-se bastante e eficiente, nas quatro situações consideradas: para as populações experimentais de retrocruzamento e F_2 com e sem efeitos de epistasia. Devido a limitação de tempo, os efeitos individuais das epistasia ainda não foram testados, ficando assim como proposta de estudos futuros, além disso pretende-se otimizar o programa implementado neste trabalho visando a diminuição do tempo computacional para a realização das análises.

3.5 Conclusão

Os resultados obtidos nesta pesquisa levaram as seguintes conclusões:

- I) A utilização da abordagem bayesiana considerando o número de QTLs como conhecido mostrou-se uma excelente alternativa para mapear QTLs quando múltiplos QTLs e efeitos de epistasia são considerados simultaneamente no modelo para duas populações experimentais: retrocruzamento e F_2 .
- II) A utilização do fator de Bayes para realizar a seleção de modelos se mostrou muito eficiente para ambos estimadores: média harmônica e estabilizador da média harmônica.

Referências

- CASELLA, G.; GEORGE, E.I. Explaining the Gibbs Sampler. **The American Statistician**, New York, v. 46, p. 167-174, 1992.
- GAFFNEY, P.J. **An efficient reversible jump Markov Chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses**. 2001, 174p. Thesis (Philosophy in Statistics Doctor) - University of Wisconsin, Madison, 2001.
- GODSILL, S.J. On the relationship between MCMC model uncertainty methods. **Journal of computational Graphical Statistics**, Alexandria, v. 10, p. 230-248, 2001.
- GREEN, P.J.; Reversible jump markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, Lausanne, v. 82, p. 711-732, 1995.
- HEATH, S.C. Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. **The American Journal of Human Genetics**, Boston, v. 61, p. 748-760, 1997.
- JEFFREY, H. **Theory of probability**. Oxford: Clarendon press, 1961. 447 p.
- KAO, C.H.; ZENG, Z-B. Modeling Epistasis of Quantitative Trait Loci Using Cockerham's Model. **Genetics**, Austin, v. 160, p. 1243-1261, 2002.

NARITA, A.; SASAKI, Y. Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. **Genetics Selection Evolution**, Versailles, v. 36, p. 415-433, 2004.

NEWTON, M.A.; RAFTERY, A.E. Approximate Bayesian inference by the weighted likelihood bootstrap. **Journal Of Royal Statistical Society**, London, v. 56, p. 1-48, 1994.

RAFTERY, A.E.; NEWTON, M.A.; SATAGOPAN, J. M.; KRIVITSKY, P.N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: BERNARDO, J.M.; BAYARRI, M.J.; BERGER, J.O.; DAWID, A.P.; HECKERMAN, A.F.; SMITH, A.F.M; WEST, M. **Bayesian statistics**. Oxford: Oxford University Press, 2007, 8, p.1-45.

SATAGOPAN, J. M.; YANDELL, B.S. Estimating the number of quantitative trait Loci via Bayesian model determination. **Special contributed paper session on genetic analysis of quantitative trait and complex disease, Biometric section, Joint Statistical Meetings**, Chicago, 1998.

SATAGOPAN, J. M.; YANDELL, B.S. NEWTON M.A.; OSBORN, T.C. A Bayesian approach to detect Quantitative trait Loci using Markov chain Monte Carlo. **Genetics**, Austin, v. 144, p. 805-816, 1996.

SEN, S.; CHURCHILL A statistical framework for quantitative trait mapping . **Genetics**, Austin, v. 159, p. 371-387, 2001.

SILLANPÄÄ, M.J.; ARJAS, E. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data . **Genetics**, Austin, v. 148, p. 1373-1388, 1998.

STEPHENS, D.A.; FISCH, R.D. Bayesian analysis of quantitative trait locus data using reversible jump Markov Chain Monte Carlo **Biometrics**, Washington, v. 54, p. 1334-1347, 1998.

WANG, H.; ZHANG, Y.M.; MASINDE, G.L.; MOHAN, S; BAYLINK, D.J.; XU, S. Bayesian shrinkage estimation of quantitative trait loci parameters. **Genetics**, Austin, v. 170, p. 465-480, 2005.

YI, N. A unified markov chain monte carlo framework for mapping multiple quantitative trait loci. **Genetics**, Austin, v. 167, p. 967-975, 2004.

YI, N.; ALLISON, D.B.; XU, S. Bayesian model choice and search strategies for mapping multiple epistatic quantitative trait loci. **Genetics**, Austin, v. 165, p. 867-883, 2003.

YI, N.; XU, S. Mapping quantitative trait loci with epistatic effects. **Genetics Research** , Cambridge, v. 79, p. 185-198, 2002.

YI, N.; YANDELL, B.S.; CHURCILL, G.A.; ALLISON, D.B.; EISEN, E.J.; POMP, D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. **Genetics**, Austin, v. 170, p. 1333-1344, 2005.

YI, N.; SHRINER, D.; BANERJEE, S.; MEHTA, T.; POMP, D YANDELL, B.S. An efficient Bayesian model selection approach for interacting QTL models with many effects. **Genetics**, Austin, v.176 , p. 1865-1877, 2007.

4 MAPEAMENTO DE QTLs EM UMA POPULAÇÃO DE MILHO TROPICAL USANDO A ABORDAGEM BAYESIANA

Resumo

Muitos caracteres medidos em plantas e animais são de natureza quantitativa, ou seja, influenciados por múltiplos genes. Com o advento de novas técnicas moleculares tem sido possível mapear os locos que controlam tais caracteres, denominados QTLs (*Quantitative Trait Loci*). Mapear um QTL significa identificar sua posição no genoma, bem como, estimar seus efeitos genéticos e interações. A maior dificuldade para realizar o mapeamento de QTLs deve-se ao fato de que seu número é desconhecido e, conseqüentemente, a dimensão do espaço paramétrico também é desconhecida. Métodos Bayesianos juntamente com método Monte Carlo com Cadeias de Markov (MCMC), têm sido implementados para inferir conjuntamente o número de QTLs, suas posições no genoma e os seus efeitos genéticos. O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado desconhecido e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. No presente trabalho foi utilizada uma abordagem bayesiana para mapear QTLs para caracteres numa população de milho tropical: altura de planta (AP), altura de espiga (AE) e produção de grãos. Para tanto, foram considerados múltiplos QTLs e efeitos de epistasia no modelo e o número de QTLs foi tratado como conhecido. Os métodos MCMC foram utilizados para gerar uma amostra da distribuição conjunta a posteriori dos parâmetros. Foram ajustados modelos com números crescentes de QTLs e o fator de Bayes foi utilizado para selecionar o modelo mais adequado e conseqüentemente, para estimar o número de QTLs que controlam os caracteres de interesse. Os resultados obtidos no presente trabalho foram comparados com os resultados obtidos pelo método frequentista para mapear QTLs.

Palavras-chave: Mapeamento de QTLs; Inferência Bayesiana; Mapeamento por Intervalo Composto; Fator de Bayes; MCMC

Abstract

Many traits in plants and animals have quantitative nature, influenced by multiple genes. With the new molecular techniques, it has been possible to map the loci, which control the quantitative traits, called QTL (*Quantitative Trait Loci*). Mapping a QTL means to identify its position in the genome, as well as to estimate its genetics effects and interactions. The great difficulty of mapping QTL relates to the fact that the number of QTL is unknown and hence the dimension of the parametric space is also unknown. Bayesian approaches used with Markov Chain Monte Carlo method (MCMC) have been applied to infer QTL number, their positions in the genome and their genetic effects. The challenge is to obtain the sample from the joint posterior distribution of these parameters, since the number of QTL may be considered unknown and hence the dimension of the parametric space changes according to the number of QTL in the model. In this study, a Bayesian approach was applied, using a code implemented in the statistical program R, in order to map QTL for traits in a population of tropical corn: plant height, corn-cob height and grain

production. So, multiples QTL and epistasis effects were considered in the model and the number of QTL was considered as known. The MCMC methods were used to create a sample from the joint posterior distribution of the parameters. Models were adjusted with the crescent number of QTL and Bayes factor was used to select the most suitable model and, consequently, to estimate the number of QTL that control interesting phenotype. The results from this study were compared to the results obtained using the frequentist approach for mapping QTL. Keywords: Bayes Factor; Bayesian Inference; Composite Interval Mapping; MCMC QTL Mapping

4.1 Introdução

Muitos caracteres em plantas e animais são de natureza quantitativa, sendo controlados por múltiplos genes e pela interação entre eles (epistasia). Com o advento de novas técnicas moleculares tem sido possível mapear os locos que controlam os caracteres quantitativos, denominados QTLs (“quantitative trait loci”). Mapear um QTL significa identificar suas posições no genoma, isto é, identificar em qual cromossomo estão e qual sua localização nesses cromossomos, bem como estimar seus efeitos genéticos e interações. Dessa forma, é possível um melhor entendimento sobre os caracteres quantitativos, que são de fundamental importância nos programas de melhoramento genético, uma vez que estão associados as características mais importantes, como produção de grãos, altura de plantas, volume, entre outros.

Vários modelos estatísticos tem sido propostos para o mapeamento de QTLs, variando desde simples modelos de regressão linear e análise de variância até métodos sofisticados, que modelam todo o genoma simultaneamente. O método mais utilizado no momento pelos geneticistas e melhoristas é o Mapeamento por Intervalo Composto (CIM) que foi proposto, de forma independente, por Jansen e Stam (1994) e Zeng (1994), tal modelo considera a procura por QTLs num dado intervalo usando como cofatores marcadores fora desse intervalo, com a finalidade de contornar a variação causada por outros QTLs. Contudo, o método CIM não permite o mapeamento simultâneo de múltiplos QTLs e não permite portanto estimar os efeitos de epistasias.

Visando realizar o mapeamento simultâneo de múltiplos QTLs, duas alternativas são encontradas na literatura com maior destaque: o mapeamento por Múltiplos Intervalos (MIM), proposto por Kao e Zeng (1997) e Kao et al. (1999) e as metodologias baseadas em inferência bayesiana por meio de métodos MCMC (Markov Chain Monte Carlo) (SATAGOPAN et al., 1996; HEATH, 1997; SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998; SEN; CHURCHILL, 2001; GAFFNEY, 2001; WANG et al., 2005; YI, 2004; YI et al., 2005, YI et al., 2007). O Mapeamento de Múltiplos Intervalos (MIM) combina a possibilidade de mapeamento com o estudo da arquitetura genética dos caracteres quantitativos, a partir de um algoritmo de busca do número, posição, efeitos e interações entre os QTLs. Os métodos bayesianos utilizam a distribuição conjunta a posteriori para fazer inferências sobre os parâmetros de interesse. Esta

distribuição é obtida por meio do teorema de Bayes, que combina a função de verossimilhança com a incerteza relativa aos parâmetros. Assim, tal abordagem tem a vantagem de permitir incorporar na análise informações adicionais baseadas em experimentos anteriores ou conhecimentos de pesquisadores, por meio da atribuição de uma distribuição a priori para cada parâmetro de interesse.

Um dos trabalhos pioneiros utilizando essa abordagem foi apresentado por Satagopan et al. (1996) e desde então a abordagem bayesiana vem ganhando espaço nas descobertas de novas metodologias para realizar o mapeamento de QTLs. O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado como desconhecido e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. Satagopan et al. (1996) considerou o número de QTLs como uma quantidade fixa, ajustou modelos com diferentes números de QTLs e, posteriormente, fez a seleção do modelo mais adequado usando como critério de seleção o Fator de Bayes. Trabalhos posteriores (SATAGOPAN; YANDELL, 1998; STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998) contornaram o problema de não conhecer a dimensão do espaço paramétrico tratando o número de QTLs como uma quantidade desconhecida e usando o método MCMC com Saltos Reversíveis (GREEN, 1995), o qual permite a movimentação entre modelos de dimensões diferentes. O algoritmo MCMC com saltos reversíveis é uma técnica bastante poderosa e geral para explorar as distribuições a posteriori quando o espaço paramétrico não possui dimensão fixa. Contudo, essa habilidade de movimento entre modelos de dimensões diferentes requer a construção de distribuições propostas complexas, o que dificulta a sua implementação desse algoritmo. Além disso, a complexidade desse algoritmo aumenta a demanda computacional, o que pode comprometer seu desempenho.

Outra tarefa desafiadora é a inclusão de efeitos de epistasia no modelo, pois sua inclusão, implica aumento ou diminuição considerável da dimensão do espaço paramétrico, aumentando a complexidade da implementação dos algoritmos utilizados na construção das amostras da distribuição conjunta a posteriori para os parâmetros de interesse. Apesar dos desafios da implementação, o algoritmo MCMC com saltos reversíveis tem sido empregado para estudar epistasia (YI; XU, 2002, YI et al., 2003, NARITA; SASAKI, 2004). Contudo, esse método de simulação tem apresentado uma convergência muito lenta e uma demanda computacional intensa, tornando seu emprego não muito recomendável (YI et al., 2005). Assim faz-se necessário o estudo de metodologias bayesianas para mapear QTLs que sejam ao mesmo tempo eficientes e de fácil entendimento, principalmente quando são considerados modelos mais complexos, como os que incluem termos de epistasia. Além disso, existe a necessidade de implementar esses algoritmos e de maneira que sejam acessíveis a todos os pesquisadores que tenham interesse no assunto, para que haja uma maior difusão da metodologia bayesiana para mapeamento de QTLs.

O presente trabalho teve como objetivo comparar os resultados obtidos pela abordagem bayesiana

para mapear QTLs com os obtidos pelo método CIM. Para tanto, foram usados caracteres dos fenótipos altura de planta (AP), altura de espiga (AE) e produção de grãos apresentado em Sibov et al. (2003b) para milho tropical.

4.2 Material e Métodos

4.2.1 Dados fenotípicos e moleculares

Os dados utilizados neste trabalho são apresentados em detalhes por Sibov et al. 2003a, 2003b) e Sabadin et al. (2008). Resumidamente, foi obtida uma população entre o cruzamento da linhagens endogâmicas L-08-05F e L-14-4B, que possuem comportamento 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 $F_{2:3}$, que foram cruzadas entre si e semeadas em linhas com 20 plantas para aumentar o número de sementes necessárias para as avaliações experimentais. 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 num delineamento látice 10×10 , com duas repetições cada. Neste experimento foram medidos vários caracteres, sendo que no presente trabalho foram utilizados os seguintes caracteres: Produção de Grãos (PG) em Mg por hectare⁻¹; Altura da Planta (AP) em cm e Altura da Espiga (AE) em cm.

O Mapa de ligação usado para a localização dos QTLs contém 117 locos de marcadores do tipo microsatélites, que estão distribuídos em dez grupos de ligação, que foram descritos em Sibov et al.(2003a). O comprimento do mapa foi de aproximadamente 1634 cM. A função de mapeamento utilizada foi a de Kosambi (KOSAMBI, 1944).

4.2.2 Mapeamento de QTLs

Método clássico

Sibov et al. (2003b) realizou o mapeamento de QTLs utilizando o Mapeamento por Intervalo Composto (CIM) proposto por Zeng (1994), usando o QTLCartographer versão 1.15 (BASTEN et al. 1994, 1999). O número de marcadores cofatores foram obtidos a partir do método de seleção de variáveis *stepwise*. O nível crítico genômico (*threshold*) para rejeição de H_0 (ausência de QTL) foi obtido com base em 1000 permutações (CHURCHILL; DOERGE, 1994). Todas as análises foram feitas utilizando o módulo 6 do módulo *Zmapqtl* do QTLCartographer.

Método bayesiano

Supondo que a característica quantitativa de interesse é afetada por S QTLs, o vetor de valores fenotípicos observados $\mathbf{y} = (y_1, y_2, \dots, y_n)$, foi descrito por meio do seguinte modelo de regressão linear:

$$\mathbf{y} = \mathbf{1}\mu + X\boldsymbol{\beta} + \mathbf{e}, \quad (4.1)$$

em que μ é uma constante, $\boldsymbol{\beta}$ é o vetor dos efeitos genéticos, X é a matriz de delineamento e $\mathbf{e}$ o vetor de erros aleatórios, $e \sim N(0, I\sigma^2)$.

A matriz X e o vetor de parâmetros $\boldsymbol{\beta}$ foram definidos por meio do modelo epistático de Cockerham, como apresentado e Kao e Zeng, (2002). Os elementos da matriz X foram definidos por:

$$\begin{aligned} x_{ij1} &= z_{ij} - 1, \\ x_{ij2} &= (1 + x_{ij1})(1 - x_{ij1}) - 0,5, \\ x_{ijj'k} &= \begin{cases} x_{ij1}x_{ij'1} & k = 1 \\ x_{ij1}x_{ij'2} & k = 2 \\ x_{ij2}x_{ij'1} & k = 3 \\ x_{ij2}x_{ij'2} & k = 4 \end{cases} \end{aligned}$$

em que z_{ij} é o número de alelos dominantes do genótipo do j -ésimo QTL para o i -ésimo indivíduo, $j = 1, \dots, S$ e $j \neq j'$. Para o modelo de Cockerham, $\beta_{j1} = \alpha_j$ e $\beta_{j2} = \delta_j$ correspondem, respectivamente, aos efeitos aditivos e de dominâncias do j -ésimo QTL e $\beta_{jj'1} = \alpha_j\alpha_{j'}$, $\beta_{jj'2} = \alpha_j\delta_{j'}$, $\beta_{jj'3} = \delta_j\alpha_{j'}$ e $\beta_{jj'4} = \delta_j\delta_{j'}$ são os efeitos epistáticos entre o loco j e j' : interações aditiva versus aditiva, aditiva versus dominância, dominância versus aditiva e dominância versus dominância (KAO; ZENG, 2002).

A inferência sobre os parâmetros de interesses, número de QTLs S , as localizações destes no cromossomo $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_S)$ e $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \sigma^2)$, foram feitas usando a abordagem Bayesiana, sendo o número de QTLs, S , considerado como conhecido, como proposto por Satagopan et al. (1996). Foi assumida independência a priori para os parâmetros do modelo e as distribuições a priori para os mesmos foram especificadas da seguinte maneira: para as localizações foram atribuídas distribuição Uniforme no intervalo $[0, L]$, sendo L o comprimento total do cromossomo; para a constante μ foi atribuída uma distribuição $N(0, \tau^2)$, sendo $\tau^2 = 10$, para os efeitos aditivos e de dominância foram atribuídas, respectivamente, $N(0, \sigma_\alpha^2)$ e $N(0, \sigma_\delta^2)$ em que $\sigma_\alpha^2 = 10$ e $\sigma_\delta^2 = 10$, para os efeitos de epistasia foram atribuídas, respectivamente, $N(0, \sigma_{\alpha\alpha}^2)$, $N(0, \sigma_{\alpha\delta}^2)$, $N(0, \sigma_{\delta\alpha}^2)$ e $N(0, \sigma_{\delta\delta}^2)$, sendo $\sigma_{\alpha\alpha}^2 = \sigma_{\alpha\delta}^2 = \sigma_{\delta\alpha}^2 = \sigma_{\delta\delta}^2 = 10$; para σ^2 foi atribuída

distribuição gama-inversa com hiperparâmetros $a = 8$ e $b = 8$, como proposto por Raftery et al. (2007).

Assim, dado o modelo (4.1) e as distribuições a priori para os parâmetros de interesse a distribuição conjunta a posteriori é dada por:

$$\begin{aligned} \pi(\boldsymbol{\lambda}, Q, \boldsymbol{\theta} | y) &\propto \prod_{i=1}^n \left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ \frac{-1}{2\sigma^2} \left(y_i - \mu - \sum_{j=1}^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'1} - \right. \right. \right. \\ &\quad \left. \left. \left. \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'2} \right)^2 \right\} \times \prod_{j=1}^S \pi(Q_{ij} | \lambda_j, M) \right\} \times \left(\frac{1}{L} \right)^S \times \\ &\quad \frac{1}{2\pi\tau^2} \exp \left\{ \frac{-1}{2\tau^2} (\mu)^2 \right\} \times \prod_{j=1}^S \left(\frac{1}{2\pi\sigma_\alpha^2} \exp \left\{ \frac{-1}{2\sigma_\alpha^2} (\alpha_j)^2 \right\} \times \frac{1}{2\pi\sigma_\delta^2} \exp \left\{ \frac{-1}{2\sigma_\delta^2} (\delta_j)^2 \right\} \right) \times \\ &\quad \prod_{j=1}^k \left(\frac{1}{2\pi\sigma_{\alpha\alpha}^2} \exp \left\{ \frac{-1}{2\sigma_{\alpha\alpha}^2} (\alpha_j \alpha_{j'})^2 \right\} \times \frac{1}{2\pi\sigma_{\alpha\delta}^2} \exp \left\{ \frac{-1}{2\sigma_{\alpha\delta}^2} (\alpha_j \delta_{j'})^2 \right\} \frac{1}{2\pi\sigma_{\delta\alpha}^2} \exp \left\{ \frac{-1}{2\sigma_{\delta\alpha}^2} (\delta_j \alpha_{j'})^2 \right\} \times \right. \\ &\quad \left. \frac{1}{2\pi\sigma_{\delta\delta}^2} \exp \left\{ \frac{-1}{2\sigma_{\delta\delta}^2} (\delta_j \delta_{j'})^2 \right\} \right) \times \frac{b^a}{\Gamma(a)} (\sigma^2)^{-(a+1)} \exp \left\{ \frac{-b}{\sigma^2} \right\}. \end{aligned}$$

A obtenção da distribuição marginal a posteriori para os parâmetros de interesse é analiticamente muito complexas, e assim métodos computacionalmente intensivos, tais como os métodos MCMC, foram utilizados para a obtenção de uma amostra da distribuição conjunta a posteriori para os parâmetros de interesse. Foi utilizado o algoritmo de Gibbs com passos Metropolis-Hastings. A atualização dos parâmetros foram feitas da seguinte forma:

- i) Atualização das posições ($\boldsymbol{\lambda}$): cada λ_j , $j = 1, \dots, S$ foi atualizado usando o algoritmo de Metropolis-Hastings. O valor proposto para o próximo estado da cadeia foi gerado de acordo com Gaffney (2001): para o j -ésimo QTL, um valor λ_j^{**} foi gerado de uma distribuição $U[\lambda_j - d, \lambda_j + d]$, e o valor proposto para o próximo estado da cadeia dado por $\lambda_j^* = \min(|\lambda_j^{**}|, 2L - |\lambda_j^{**}|)$, em que L é o comprimento total do cromossomo. Esta proposta é denotada por $q(\lambda_j^* | \lambda_j)$ e a razão de aceitação é dada por:

$$\alpha(\lambda_j, \lambda_j^*) = \frac{\pi(\lambda_j^* | \lambda_{-j}, Q, \boldsymbol{\theta}, \mathbf{y})}{\pi(\lambda_j | \lambda_{-j}, Q, \boldsymbol{\theta}, \mathbf{y})}$$

Se a proposta para o próximo estado da cadeia, t , for aceita $\lambda_j^t = \lambda_j^*$, caso contrário $\lambda_j^t = \lambda_j^{t-1}$.

- ii) Atualização de Q_{ij} : a distribuição condicional completa a posteriori para cada Q_{ij} , é uma distribuição discreta com as seguinte probabilidade:

$$p_{ij} = \pi(Q_{ij} = q | \lambda, Q_{i(-j)}, \boldsymbol{\theta}, \mathbf{y}) = \frac{\pi(Q_{ij} = q | \lambda_j) \pi(y_i | \boldsymbol{\theta}, Q_i, Q_{ij} = q)}{\sum_q \pi(Q_{ij} = q | \lambda_j) \pi(y_i | \boldsymbol{\theta}, Q_i, Q_{ij} = q)}$$

sendo $q = 0, 1, 2$. Assim Q_{ij} foi atualizado usando o algoritmo Gibbs sampling.

- iii) Atualização de $\boldsymbol{\theta} = (\mu, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma^2)$: como para os parâmetros $\mu, \alpha_j, \delta_j, \alpha_j\alpha_{j'}, \alpha_j\delta_{j'}, \delta_j\alpha_{j'}, \delta_j\delta_{j'}$ e σ^2 foram atribuídas distribuições a priori conjugadas, suas correspondentes distribuições condicionais completas a posteriori possuem formas fechadas cujas expressões são apresentadas no APÊNDICE B.

A seleção do modelo foi feita utilizando-se o fator de Bayes, usando como estimadores a média harmônica, proposto por Newton e Raftery (1994) e o estimador estabilizado da média harmônica, proposto em Raftery et al. (2007).

4.3 Resultados

A implementação computacional foi realizada utilizando-se o programa estatístico R (R DEVELOPMENT CORE TEAM, 2008). A construção da amostra da distribuição conjunta a posteriori foi feita utilizando os métodos MCMC. Foram geradas cadeias com 120.000 iterações, sendo que as 1000 primeiras iterações foram descartadas para o período de aquecimento da cadeia (*burn-in*) e foi utilizado espaçamento de 50 iterações, com a finalidade de diminuir a correlação entre os valores amostrados, resultando em uma amostra final de tamanho 2380. A indicação de convergência de cada cadeia foi verificada visualmente e por meio do critério Geweke, implementado no pacote CODA do programa estatístico R. Na análise de cada um dos três conjuntos de dados foram ajustados modelos contendo números crescentes de QTLs, iniciando em um QTL e parando no número de QTLs selecionado pelo fator de Bayes.

4.3.1 Produção de grãos

Na análise do conjunto de dados referentes à produção de grãos foram ajustados modelos contendo um, dois, três e quatro QTLs. O tempo computacional variou de 23 horas, para o ajuste do modelo com um QTL à 90 horas para o ajuste do modelo com quatro QTLs. O computador usado tinha a seguinte configuração: Intel(R) Core (TM) 2 DUO, T5500, 1.66 GHz e 1 GB de RAM. A taxa de aceitação utilizada pelo algoritmo Metropolis-Hastings para atualizar as localizações dos QTLs variou entre 41% e 56%. Inicialmente foi verificado a convergência das cadeias para os parâmetros envolvidos nos modelos. Posteriormente, foi feita a seleção do modelo utilizando-se como critério o fator de Bayes. As estimativas do fator de Bayes foram obtidas considerando-se o estimador média

harmônica (MH) e estimador estabilizador da média harmônica (EMH). Na Tabela 4.1 apresenta-se os resultados do fator de Bayes e sua classificação, segundo Jeffreys (1961).

Tabela 4.1 – Fator de Bayes

Modelos(1)	MH	EMH	Classificação
B_{21}	2,1342E+05	5,8402E+03	Muito Forte
B_{32}	3,3101E+06	3,2684E+04	Muito Forte
B_{43}	0,0019	0,0192	Nenhuma Evidência

⁽¹⁾ Fator de Bayes considerando o modelo com $k+1$ e k , $B_{(k+1)k}$

De acordo com a classificação do fator de Bayes, apresentada na Tabela 4.1, existe evidências de que o modelo com 2 QTLs se ajusta melhor aos dados que o modelo com 1 QTL, da mesma forma, existe evidências de que o modelo com 3 QTLs se ajusta melhor aos dados que o modelo com 2 QTLs e o modelo com 3 QTLs se ajusta melhor aos dados que o modelo com 4 QTLs. Sendo assim, o modelo que melhor se ajusta aos dados, segundo o fator de Bayes, é o modelo contendo três QTLs.

As Figuras 4.1 a 4.6, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para o modelo ajustado com três QTLs. Nota-se por meio dos histogramas e das trajetórias da cadeias, Figuras 4.1, 4.2, 4.3, 4.4, que há indícios de convergência para todos os parâmetros. Além disso, nota-se que por meio das Figuras 4.5 e 4.6 que o resultado da função de autocorrelação foi satisfatório para todos os parâmetros.

Figura 4.1 – Histogramas dos valores gerados para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivo α_1 , α_2 e α_3 , de dominância δ_1 , δ_2 e δ_3

Figura 4.2 – Histogramas dos valores gerados para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$

Figura 4.3 – Trajetória das cadeias geradas para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivo α_1 , α_2 e α_3 , de dominância δ_1 , δ_2 e δ_3

Figura 4.4 – Trajetória das cadeias geradas para para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_2\delta_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$

Figura 4.5 – Função de Autocorrelação das cadeias geradas para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 e λ_3 , para os efeitos aditivo α_1 , α_2 e α_3 , de dominância δ_1 , δ_2 e δ_3

Figura 4.6 – Função de autocorrelação das cadeias geradas para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_2\alpha_3$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_2\alpha_3$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_2\delta_3$

A Tabela 4.2 apresenta um resumo a posteriori para os parâmetros do modelo que foi selecionado pelo fator de Bayes, ou seja o modelo ajustado com três QTLs e epistasia. Vale ressaltar que para que todos os cromossomos pudessem ser considerados simultaneamente, foi feita a concatenação dos dez grupos de ligação. Assim, as localizações dos QTLs (λ) são apresentadas considerando os resultados da concatenação dos grupos de ligação. Além disso, as localizações são apresentados em Morgan. Nota-se que as estimativas dos efeitos aditivo, de dominância, das interações aditivo versus aditivo e dominância versus aditivo são em sua maioria positivos; já as estimativas para as estimativas das interações aditivo versus dominância são todas negativas e as estimativas para as interações dominância versus dominância são todas positivas.

Tabela 4.2 – Resumo a posteriori e região HPD de 95% (HPD 95%) para as localizações dos QTLs (λ), para os efeitos aditivos (α), de dominância (δ) e epistasias ($\alpha\alpha$, $\alpha\delta$, $\delta\alpha$, $\delta\delta$) do modelo com três QTLs

Parâmetro	Média	d.p.	Mediana	HPD(2, 5%)	HPD(97, 5%)
μ	1,0936	0,0111	1,0900	1,0713	1,1147
σ^2	0,0355	0,0034	0,0400	0,0291	0,0419
α_1	0,0911	0,0208	0,0208	0,0482	0,1295
α_2	0,0733	0,0186	0,0186	0,0382	0,1099
α_3	-0,0773	0,0209	0,0209	-0,1173	-0,0351
δ_1	-0,0641	0,0324	0,0324	-0,1281	-0,0026
δ_2	0,0492	0,0306	0,0306	-0,01	0,1099
δ_3	0,0128	0,0332	0,0332	-0,054	0,0785
λ_1	2,9446	0,0421	2,95	2,8576	3,0235
λ_2	12,7563	0,0449	12,76	12,6734	12,8471
λ_3	13,7738	0,0477	13,77	13,6801	13,8656
$\alpha_1\alpha_2$	-0,0066	0,0287	0,0287	-0,0643	0,0476
$\alpha_1\alpha_3$	0,042	0,0311	0,0311	-0,0186	0,1038
$\alpha_2\alpha_3$	0,0001	0,0305	0,0305	-0,0539	0,0631
$\alpha_1\delta_2$	-0,0603	0,0429	0,0429	-0,1408	0,0259
$\alpha_1\delta_3$	-0,0055	0,0424	0,0424	-0,0927	0,0742
$\alpha_2\delta_3$	-0,0005	0,0446	0,0446	-0,0801	0,0947
$\delta_1\alpha_2$	0,0286	0,0448	0,0448	-0,0588	0,113
$\delta_1\alpha_3$	-0,0039	0,0447	0,0447	-0,0872	0,087
$\delta_2\alpha_3$	0,002	0,0441	0,0441	-0,0885	0,084
$\delta_1\delta_2$	0,0322	0,0607	0,0607	-0,0855	0,1518
$\delta_1\delta_3$	0,0158	0,0609	0,0609	-0,1022	0,1344
$\delta_3\delta_3$	0,002	0,0592	0,0592	-0,1128	0,1176

A Tabela 4.3 apresenta os resultados para as localizações e efeitos genéticos principais, obtidos pelo modelo CIM apresentados em Sibov, et al. (2003b) e os obtidos no presente trabalho. Para fazer a comparação entre as duas abordagens, a unidade dos efeitos genéticos: aditivo e dominância

apresentados em Sibov, et al. (2003b), foram transformadas para a mesma unidade dos efeitos utilizadas no presente trabalho. Nota-se que o modelo CIM mapeou 4 QTLs enquanto a abordagem bayesiana mapeou 3 QTLs. Nota-se no entanto que o QTL em 4/8 mapeado por Sibov et al. (2003b) explica apenas 7,98% da variância fenotípica, o que indica que o mesmo pode ter baixo efeito ou é um falso positivo.

Tabela 4.3 – Estimativas das localizações e efeitos aditivos e de dominâncias dos QTLs mapeados para produção de grãos, considerando dois diferentes métodos de mapeamentos

Método	Qtl/Crom.(1)	Posição	Intervalo	Efeitos	
				Aditivo	Dominância
CIM	1/2	52,88	umc1845-bnlg0166	0,0855	-0,0968
	2/7	125,20	dupssr13-umc1154	0,0804	0,0614
	3/8	57,94	phi0115-bnlg1176	-0,0776	-0,0065
	4/8	74,70	bnl1176-bnlg1607	-0,0915	0,0293
Bayesiano	1/2	51,66	umc1845-bnlg0166	0,0911	-0,0641
	2/7	125,33	dupssr13-umc1154	0,0733	0,0492
	3/8	78,18	bnl1176-bnlg1607	-0,0773	0,0128

(1) Cromossomo onde está localizado o QTL

Na Figura 4.7 é possível observar que houve boa concordância entre três QTLs mapeados pela abordagem bayesiana e os QTLs mapeados usando o modelo CIM. A única discordância foi observada no cromossomo 8 em que o modelo CIM mapeou um QTL a mais em relação a bayesiana, este fato deve ter ocorrido, provavelmente, pelo fato deste QTL apresentar o menor R^2 entre os mapeados em Sibov et al. (2003b).

Figura 4.7 – Gráfico mostrando a localização de todos os QTLs mapeados no presente trabalho e em Sibov et al. (2003b) para o fenótipo produção de grãos. Os triângulos azuis representam os QTLs mapeados por Sibov et al. (2003b) e os losangos vermelhos representam os QTLs mapeados pelo presente trabalho

4.3.2 Altura da espiga

Na análise do conjunto de dados referentes à altura da espiga foram ajustados modelos contendo um, dois, três, quatro e cinco QTLs. O tempo computacional variou de 23 horas, para o ajuste do modelo com um QTL à 110 horas para o ajuste do modelo com cinco QTLs, usando o mesmo computador mencionado anteriormente. A taxa de aceitação utilizada pelo algoritmo Metropolis-Hastings para atualizar as localizações dos QTLs variou entre 24% e 50%. Após a verificação da convergência das cadeias para os parâmetros envolvidos nos modelos, foi feita a seleção do modelo utilizando-se como critério o fator de Bayes. Na Tabela 4.4 apresenta-se os resultados do fator de Bayes, obtidas considerando-se o estimador média harmônica (MH) e estimador estabilizador da média harmônica (EMH), e sua classificação, segundo Jeffreys (1961).

Tabela 4.4 – Fator de Bayes

Modelos(1)	MH	EMH	Classificação
B_{21}	30,7752	2,4868	Forte/Muito Fraca
B_{32}	25,4537	7,4890	Forte/Fraca
B_{43}	14464,6908	20,8394	Muito Forte/Forte
B_{54}	1,3795E-116	1,5798E-18	Nenhuma Evidência

⁽¹⁾ Fator de Bayes considerando o modelo com $k+1$ e k , $B_{(k+1)k}$

De acordo com a classificação do fator de Bayes, apresentada na Tabela 4.4, o modelo com 2 QTLs se ajusta melhor aos dados que o modelo com 1 QTL, o modelo com 3 QTLs se ajusta melhor aos dados que o modelo com 2 QTLs, o modelo com 4 QTLs se ajusta melhor aos dados que o modelo com 3 QTLs e o modelo com 4 QTLs se ajusta melhor aos dados que o modelo 5 QTLs. Assim, o modelo que melhor se ajusta aos dados é o modelo com quatro QTLs.

As Figuras 4.8 a 4.16, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para o modelo ajustado com três QTLs. Nota-se por meio dos histogramas, das trajetórias da cadeias e pela função de autocorrelação que há indícios de convergência para todos os parâmetros.

Figura 4.8 – Histogramas dos valores gerados para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 , λ_3 e λ_4 , para os efeitos aditivo α_1 , α_2 , α_3 e α_4 de dominância δ_1 , δ_2 , δ_3 e δ_4

Figura 4.9 – Histogramas dos valores gerados para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\alpha_4$, $\alpha_2\alpha_3$, $\alpha_2\alpha_4$, $\alpha_3\alpha_4$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_1\delta_4$, $\alpha_2\delta_3$, $\alpha_2\delta_4$, $\alpha_3\delta_4$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_1\alpha_4$, $\delta_2\alpha_3$

Figura 4.10 – Histogramas dos valores gerados $\delta_2\alpha_4$, $\delta_3\alpha_4$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_1\delta_4$, $\delta_2\delta_3$, $\delta_2\delta_4$, $\delta_3\delta_4$

Figura 4.11 – Trajetória das cadeias geradas para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 , λ_3 e λ_4 , para os efeitos aditivo α_1 , α_2 , α_3 e α_4 de dominância δ_1 , δ_2 , δ_3 e δ_4

Figura 4.12 – Trajetória das cadeias geradas para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\alpha_4$, $\alpha_2\alpha_3$, $\alpha_2\alpha_4$, $\alpha_3\alpha_4$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_1\delta_4$, $\alpha_2\delta_3$, $\alpha_2\delta_4$, $\alpha_3\delta_4$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_1\alpha_4$, $\delta_2\alpha_3$

Figura 4.13 – Trajetória das cadeias geradas para $\delta_2\alpha_4$, $\delta_3\alpha_4$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_1\delta_4$, $\delta_2\delta_3$, $\delta_2\delta_4$, $\delta_3\delta_4$

Figura 4.14 – Função de Autocorrelação das cadeias geradas para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 , λ_3 e λ_4 , para os efeitos aditivo α_1 , α_2 , α_3 e α_4 de dominância δ_1 , δ_2 , δ_3 e δ_4

Figura 4.15 – Função de autocorrelação das cadeias geradas epistasia $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\alpha_4$, $\alpha_2\alpha_3$, $\alpha_2\alpha_4$, $\alpha_3\alpha_4$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_1\delta_4$, $\alpha_2\delta_3$, $\alpha_2\delta_4$, $\alpha_3\delta_4$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_1\alpha_4$, $\delta_2\alpha_3$

Figura 4.16 – Função de autocorrelação das cadeias geradas para $\delta_2\alpha_4$, $\delta_3\alpha_4$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_1\delta_4$, $\delta_2\delta_3$, $\delta_2\delta_4$, $\delta_3\delta_4$

A Tabela 4.5 apresenta um resumo a posteriori para os parâmetros do modelo ajustado com quatro QTLs, que foi selecionado pelo fator de Bayes. Nota-se que as estimativas dos efeitos aditivos são todas positivas e para os efeitos de dominância são em sua maioria negativas, as estimativas das epistasias são equilibradas entre positivas e negativas.

A Tabela 4.6 apresenta os resultados para as localizações e efeitos genéticos principais, obtidos pelo modelo CIM apresentados em Sibov, et al. (2003b) e os obtidos pela abordagem bayesiana. Nota-se que, novamente, a abordagem bayesiana mapeou um QTLs a menos que o modelo CIM, sendo que o modelo CIM mapeou 5 QTLs enquanto a abordagem bayesiana mapeou 4 QTLs. Observa-se que há semelhança entre três QTLs mapeados pela abordagem bayesiana e os QTLs mapeados usando o modelo CIM. As exceções foram observadas no cromossomo 1, em que a bayesiana não mapeou nenhum QTL e o modelo CIM mapeou um QTL e no cromossomo 2 em que a bayesiana mapeou um QTL e o modelo CIM não mapeou QTL nesse cromossomo (Figura 4.17). Isso pode ser devido ao fato dos QTLs terem pequenos efeitos, ou ainda em função da inclusão de efeitos epistáticos no modelo na abordagem bayesiana.

Tabela 4.5 – Resumo a posteriori e região HPD de 95% (HPD 95%) para as localizações dos QTLs (λ), para os efeitos aditivos (α), de dominância (δ) e epistasias ($\alpha\alpha$, $\alpha\delta$, $\delta\alpha$, $\delta\delta$) do modelo com quatro QTLs

Parâmetro	Média	d.p.	Mediana	HPD(2, 5%)	HPD(97, 5%)
μ	0,8549	0,0037	0,8500	0,8476	0,8621
σ^2	0,0041	0,0004	0,0000	0,0034	0,0048
α_1	0,0223	0,0068	0,0200	0,0092	0,0358
α_2	0,0147	0,0066	0,0100	0,0025	0,0282
α_3	0,0184	0,0059	0,0200	0,0072	0,0303
α_4	0,0040	0,0063	0,0000	-0,0086	0,0158
δ_1	-0,0139	0,0097	-0,0100	-0,0335	0,0045
δ_2	-0,0061	0,0095	-0,0100	-0,0249	0,0123
δ_3	0,0001	0,0099	0,0000	-0,0198	0,0185
δ_4	-0,0105	0,0092	-0,0100	-0,0284	0,0067
λ_1	29.186	0,0647	29.200	27.918	30.434
λ_2	116.976	0,0418	117.000	116.127	117.745
λ_3	126.808	0,0724	126.800	125.548	128.351
λ_4	146.586	0,0499	146.600	145.621	147.580
$\alpha_1\alpha_2$	0,0009	0,0100	0,0000	-0,0195	0,0192
$\alpha_1\alpha_3$	-0,0049	0,0094	-0,0100	-0,0229	0,0135
$\alpha_1\alpha_4$	-0,0091	0,0098	-0,0100	-0,0290	0,0094
$\alpha_2\alpha_3$	-0,0001	0,0087	0,0000	-0,0175	0,0166
$\alpha_2\alpha_4$	-0,0041	0,0095	0,0000	-0,0227	0,0146
$\alpha_3\alpha_4$	-0,0041	0,0090	0,0100	-0,0122	0,0230
$\alpha_1\delta_2$	-0,0015	0,0142	0,0000	-0,0290	0,0265
$\alpha_1\delta_3$	-0,0097	0,0147	-0,0100	-0,0380	0,0182
$\alpha_1\delta_4$	-0,0164	0,0139	-0,0200	-0,0425	0,0106
$\alpha_2\delta_3$	-0,0042	0,0133	0,0000	-0,0302	0,0217
$\alpha_2\delta_4$	0,0145	0,0131	0,0100	-0,0103	0,0393
$\alpha_3\delta_4$	-0,0172	0,0128	-0,0200	-0,0431	0,0061
$\delta_1\alpha_2$	0,0121	0,0135	0,0100	-0,0125	0,0399
$\delta_1\alpha_3$	0,0085	0,0138	0,0100	-0,0188	0,0355
$\delta_1\alpha_4$	-0,0051	0,0142	-0,0100	-0,0342	0,0220
$\delta_2\alpha_3$	-0,0013	0,0119	0,0000	-0,0260	0,0199
$\delta_2\alpha_4$	-0,0018	0,0132	0,0000	-0,0286	0,0229
$\delta_3\alpha_4$	0,0310	0,0145	0,0300	0,0037	0,0606
$\delta_1\delta_2$	0,0055	0,0195	0,0100	-0,0338	0,0432
$\delta_1\delta_3$	0,0041	0,0192	0,0000	-0,0344	0,0413
$\delta_1\delta_4$	0,0067	0,0196	0,0100	-0,0315	0,0443
$\delta_2\delta_3$	0,0086	0,0181	0,0100	-0,0278	0,0429
$\delta_2\delta_4$	-0,0050	0,0175	-0,0100	-0,0404	0,0286
$\delta_3\delta_4$	-0,0050	0,0210	0,0300	-0,0123	0,0685

Tabela 4.6 – Estimativas das localizações e efeitos aditivos e de dominâncias dos QTLs mapeados para altura da espiga, considerando dois diferentes métodos de mapeamentos

Método	QTL/Crom. (1)	Posição	Intervalo	Efeitos	
				Aditivo	Dominância
CIM	1/1	150,38	bnlg0615-phi0037	0,0198	0,0056
	2/7	35,70	umc1632-umc1409	0,0169	-0,0111
	3/7	91,70	bnlg0434-dupssr13	0,0204	-0,0041
	4/7	119,20	dupssr13-umc1154	0,0221	0,0044
	5/9	1,01	umc1893-bnlg0430	-0,0084	-0,0242
Bayesiano	1/2	49,06	umc11845-bnlg0166	0,0223	-0,0139
	2/7	19,46	umc1428-umc1632	0,0147	-0,0061
	3/7	117,78	dupssr13-umc1154	0,0184	0,0001
	4/9	11,06	umc1893-bnlg0430	0,0040	-0,0105

(1) Cromossomo onde está localizado o QTL

Figura 4.17 – Gráfico mostrando a localização de todos os QTLs mapeados no presente trabalho e em Sibov et al. (2003b) para o fenótipo altura da espiga. Os triângulos azuis representam os QTLs mapeados por Sibov et al. (2003b) e os losangos vermelhos representam os QTLs mapeados pelo presente trabalho

4.3.3 Altura da planta

Na análise do conjunto de dados referentes à altura da planta foram ajustados modelos contendo um, dois, três, quatro e cinco QTLs. O tempo computacional variou de 23 horas, para o ajuste do modelo com um QTL à 110 horas para o ajuste do modelo com cinco QTLs, no mesmo computador mencionado anteriormente. A taxa de aceitação utilizada pelo algoritmo Metropolis-Hastings para atualizar as localizações dos QTLs variou entre 29% e 66%. Após a verificação da convergência das cadeias para os parâmetros envolvidos nos modelos, foi feita a seleção do modelo utilizando-se como critério o fator de Bayes. Na Tabela 4.7 apresenta-se os resultados do fator de Bayes, obtidos considerando-se o estimador média harmônica (MH) e estimador estabilizador da média harmônica (EMH), e sua classificação, segundo Jeffreys (1961).

Tabela 4.7 – Fator de Bayes

Modelos	MH	EMH	Classificação
B_{21}	164,3134	14,3657	Muito Forte/Forte
B_{32}	0,6183	4,1449	Nenhuma/Fraca
B_{43}	1,1565E+07	129,6249	Muito Forte
B_{54}	0,1221	0,7733	Nenhuma Evidência

⁽¹⁾ Fator de Bayes considerando o modelo com $k+1$ e k , $B_{(k+1)k}$

De acordo com a classificação do fator de Bayes, apresentada na Tabela 4.7, o modelo com 2 QTLs se ajusta melhor aos dados que o modelo com 1 QTL. Quando compara-se o modelo com 3 QTLs com o com 2 QTLs, houve uma contradição entre os resultados dos dois estimadores do fator de Bayes utilizados, sendo que para o estimador média harmônica não existe evidências de que o modelo três se ajuste melhor aos dados que o modelo 2. Já para os estimador estabilizador da média harmônica existe uma fraca evidência de que o modelo com 3 QTLs se ajuste melhor que o modelo com 2 QTLs, sendo assim, foi ajustado o modelo com quatro QTLs e para ambos estimadores do fator de Bayes existe evidências de que o modelo com 4 QTLs se ajusta melhor aos dados que o modelo com 3 QTLs. Na comparação do modelo com 5 QTLs com o modelo com 4 não existe nenhuma evidência de que o modelo com 5 se ajuste melhor aos dados que o modelo com 4 QTLs. Assim, segundo o fator de Bayes, o modelo que melhor se ajusta aos dados é o modelo contendo 4 QTLs.

As Figuras 4.18 a 4.26, apresentam as análise gráfica das cadeias de todos os parâmetros obtidas para o modelo ajustado com três QTLs. Nota-se por meio dos histogramas, das trajetórias da cadeias e pela função de autocorrelação que há indícios de convergência para todos os parâmetros.

Figura 4.18 – Histogramas dos valores gerados para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 , λ_3 e λ_4 , para os efeitos aditivo α_1 , α_2 , α_3 e α_4 de dominância δ_1 , δ_2 , δ_3 e δ_4

Figura 4.19 – Histogramas dos valores gerados para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\alpha_4$, $\alpha_2\alpha_3$, $\alpha_2\alpha_4$, $\alpha_3\alpha_4$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_1\delta_4$, $\alpha_2\delta_3$, $\alpha_2\delta_4$, $\alpha_3\delta_4$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_1\alpha_4$, $\delta_2\alpha_3$

Figura 4.20 – Histogramas dos valores gerados para $\delta_2\alpha_4$, $\delta_3\alpha_4$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_1\delta_4$, $\delta_2\delta_3$, $\delta_2\delta_4$, $\delta_3\delta_4$

Figura 4.21 – Trajetória das cadeias geradas para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 , λ_3 e λ_4 , para os efeitos aditivo α_1 , α_2 , α_3 e α_4 de dominância δ_1 , δ_2 , δ_3 e δ_4

Figura 4.22 – Trajetória das cadeias geradas para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\alpha_4$, $\alpha_2\alpha_3$, $\alpha_2\alpha_4$, $\alpha_3\alpha_4$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_1\delta_4$, $\alpha_2\delta_3$, $\alpha_2\delta_4$, $\alpha_3\delta_4$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_1\alpha_4$, $\delta_2\alpha_3$

Figura 4.23 – Trajetória das cadeias geradas para $\delta_2\alpha_4$, $\delta_3\alpha_4$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_1\delta_4$, $\delta_2\delta_3$, $\delta_2\delta_4$, $\delta_3\delta_4$

Figura 4.24 – Função de Autocorrelação das cadeias geradas para a contante do modelo μ , para σ^2 , para as localizações dos QTLs λ_1 , λ_2 , λ_3 e λ_4 , para os efeitos aditivo α_1 , α_2 , α_3 e α_4 de dominância δ_1 , δ_2 , δ_3 e δ_4

Figura 4.25 – Função de autocorrelação das cadeias geradas para $\alpha_1\alpha_2$, $\alpha_1\alpha_3$, $\alpha_1\alpha_4$, $\alpha_2\alpha_3$, $\alpha_2\alpha_4$, $\alpha_3\alpha_4$, $\alpha_1\delta_2$, $\alpha_1\delta_3$, $\alpha_1\delta_4$, $\alpha_2\delta_3$, $\alpha_2\delta_4$, $\alpha_3\delta_4$, $\delta_1\alpha_2$, $\delta_1\alpha_3$, $\delta_1\alpha_4$, $\delta_2\alpha_3$

Figura 4.26 – Função de autocorrelação das cadeias geradas para $\delta_2\alpha_4$, $\delta_3\alpha_4$, $\delta_1\delta_2$, $\delta_1\delta_3$, $\delta_1\delta_4$, $\delta_2\delta_3$, $\delta_2\delta_4$, $\delta_3\delta_4$

A Tabela 4.8 apresenta um resumo a posteriori para os parâmetros do modelo com QTLs, que foi selecionado pelo fator de Bayes. Nota-se que as estimativas dos efeitos aditivos são todas positivas, as estimativas para os efeitos de dominância foram em sua maioria negativas e as estimativas para efeito de epistasia foram em sua maioria positivas.

A Tabela 4.9 apresenta os resultados para as localizações e efeitos genéticos principais, obtidos pelo modelo CIM apresentados em Sibov, et al. (2003b) e os obtidos pela abordagem bayesiana. Nota-se que para o caráter altura de plantas a abordagem bayesiana e o modelo CIM mapearam o mesmo número de QTLs.

Observa-se na Figura 4.27 que os resultados foram semelhantes para três QTLs mapeados pela abordagem bayesiana e os QTLs mapeados usando o modelo CIM. Observa-se que o modelo CIM mapeou um QTL a mais no cromossomo 1, e abordagem bayesiana mapeou um QTL a mais no cromossomo 7. Estas diferenças, provavelmente, deve-se ao fato da bayesiana considerar múltiplos QTLs simultaneamente no modelo, juntamente com efeitos epistáticos.

Tabela 4.8 – Resumo a posteriori e região HPD de 95% (HPD 95%) para as localizações dos QTLs (λ), para os efeitos aditivos (α), de dominância (δ) e epistasias ($\alpha\alpha$, $\alpha\delta$, $\delta\alpha$, $\delta\delta$) do modelo com quatro QTLs

Parâmetro	Média	d.p.	Mediana	HPD(2, 5%)	HPD(97, 5%)
μ	1,7161	0,0054	1,7200	1,7052	1,7260
σ^2	0,0083	0,0008	0,0100	0,0069	0,0098
α_1	0,0306	0,0087	0,0300	0,0145	0,0488
α_2	0,0297	0,0101	0,0300	0,0104	0,049
α_3	0,0226	0,0088	0,0200	0,0055	0,0395
α_4	0,0258	0,0084	0,0300	0,0095	0,042
δ_1	-0,0052	0,0131	-0,0100	-0,031	0,0193
δ_2	-0,0263	0,0156	-0,0300	-0,0554	0,0053
δ_3	-0,0196	0,0129	-0,0200	-0,0441	0,0057
δ_4	0,0221	0,0149	0,0200	-0,005	0,0532
λ_1	15,368	0,0328	15,400	14,786	16,038
λ_2	29,509	0,0401	29,500	28,691	30,222
λ_3	117,493	0,0560	117,500	116,545	118,681
λ_4	127,207	0,0526	127,200	126,150	128,204
$\alpha_1\alpha_2$	0,0000	0,0149	0,0000	-0,0289	0,0288
$\alpha_1\alpha_3$	-0,0159	0,0132	-0,0200	-0,0417	0,0105
$\alpha_1\alpha_4$	0,0109	0,0125	0,0100	-0,0147	0,0346
$\alpha_2\alpha_3$	0,0033	0,0135	0,0000	-0,0251	0,0276
$\alpha_2\alpha_4$	0,0005	0,0130	0,0000	-0,0224	0,0288
$\alpha_3\alpha_4$	0,0005	0,0133	0,0000	-0,0291	0,0236
$\alpha_1\delta_2$	0,0013	0,0223	0,0000	-0,0417	0,0456
$\alpha_1\delta_3$	0,0307	0,0181	0,0300	-0,0051	0,0633
$\alpha_1\delta_4$	-0,0228	0,0198	-0,0200	-0,0614	0,0158
$\alpha_2\delta_3$	-0,0072	0,0193	-0,0100	-0,0447	0,0306
$\alpha_2\delta_4$	-0,0140	0,0207	-0,0100	-0,0552	0,0255
$\alpha_3\delta_4$	0,0070	0,0220	0,0100	-0,0355	0,0484
$\delta_1\alpha_2$	0,0087	0,0204	0,0100	-0,0322	0,0473
$\delta_1\alpha_3$	-0,0166	0,0186	-0,0200	-0,0503	0,0204
$\delta_1\alpha_4$	-0,0112	0,0175	-0,0100	-0,0448	0,0226
$\delta_2\alpha_3$	0,0096	0,0195	0,0100	-0,029	0,0478
$\delta_2\alpha_4$	0,0021	0,0180	0,0000	-0,032	0,0381
$\delta_3\alpha_4$	0,0073	0,0181	0,0100	-0,029	0,0423
$\delta_1\delta_2$	-0,0590	0,0323	-0,0600	-0,1213	0,0029
$\delta_1\delta_3$	0,0131	0,0263	0,0100	-0,0352	0,0672
$\delta_1\delta_4$	-0,0034	0,0286	0,0000	-0,0568	0,0537
$\delta_2\delta_3$	0,0082	0,0275	0,0100	-0,0445	0,0619
$\delta_2\delta_4$	0,0028	0,0266	0,0000	-0,0497	0,0557
$\delta_3\delta_4$	0,0028	0,0295	0,0300	-0,0336	0,0828

Tabela 4.9 – Estimativas das localizações e efeitos aditivos e de dominâncias dos QTLs mapeados para altura da planta, considerando dois diferentes métodos de mapeamentos

Método	QTL/Crom.(1)	Posição	Intervalo	Efeitos	
				Aditivo	Dominância
CIM	1/1	79,38	bnlg2238-umc2025	-0.0042	-0.0508
	2/1	153,38	bnlg0615-phi0037	0.0325	-0.0054
	3/2	53,88	umc11845-bnlg0166	0.0349	-0.0429
	4/7	117,20	dupssr13-umc1154	0.0286	0.0281
Bayesiano	1/1	153,68	bnlg0615-phi0037	-0.0306	-0.0052
	2/2	53,29	umc11845-bnlg0166	0.0297	-0.0263
	3/7	24,63	umc1428-umc1632	0.0226	-0.0196
	4/7	121,77	dupssr13-umc1154	0.0258	0.0221

(1) Cromossomo onde está localizado o QTL

Figura 4.27 – Gráfico mostrando a localização de todos os QTLs mapeados no presente trabalho e em Sibov et al. (2003b) para o fenótipo altura da planta. Os triângulos azuis representam os QTLs mapeados por Sibov et al. (2003b) e os losangos vermelhos representam os QTLs mapeados pelo presente trabalho

4.4 Discussão

Mapear um QTL significa identificar sua posição no genoma, isto é, identificar em qual cromossomo está e qual sua localização nesse cromossomo, bem como, estimar seus efeitos genéticos (incluindo epistasia). Dessa forma, é possível um melhor entendimento sobre os caracteres quantitativos, que são de fundamental importância nos programas de melhoramento genético, uma vez que estão associados à produção de grãos, altura de plantas, volume, entre outros, que são de grande importância econômica. A maior dificuldade para realizar o mapeamento de QTLs se deve ao fato de que o número de QTLs é desconhecido e, conseqüentemente, a dimensão do espaço paramétrico também é desconhecida.

A abordagem bayesiana juntamente com método Monte Carlo com Cadeias de Markov (MCMC), têm sido implementados para inferir conjuntamente o número de QTLs, suas posições no genoma e seus efeitos genéticos. O desafio está em obter a amostra da distribuição conjunta a posteriori desses parâmetros, uma vez que o número de QTLs pode ser considerado uma variável aleatória e a dimensão do espaço paramétrico muda de acordo com o número de QTLs presente no modelo. Na literatura são encontradas algumas alternativas para contornar este problema tais como método MCMC com Saltos Reversíveis e o espaço composto. Contudo a complexidade de implementação e entendimento destas metodologias acabam comprometendo seus usos. Outra alternativa bastante razoável é considerar o número de QTLs como conhecido e usar um critério para selecionar o modelo mais adequado. Outra tarefa desafiadora é a inclusão de efeitos de epistasia no modelo pois, sua inclusão, implica aumento ou diminuição considerável da dimensão do espaço paramétrico.

No presente trabalho foi utilizada abordagem bayesiana para mapear QTLs para três fenótipos distintos, associados a milho tropical: Produção de Grãos, Altura da Planta e Altura da Espiga. Para tanto, foi utilizado a abordagem bayesiana em que o número de QTLs é considerado como conhecido e a seleção do modelo foi feita utilizando o fator Bayes que foi estimado por meio de dois diferentes estimadores: média harmônica e estabilizador da média harmônica. Foi feita a concatenação dos grupos de ligação e assim todos os cromossomos puderam ser considerados simultaneamente, possibilitando assim a inclusão de efeitos epistáticos no modelo.

Para cada um dos três fenótipos foram ajustados sucessivos modelos com número crescente de QTLs, sendo que a incrementação de QTLs no modelo foi feita até que o fator Bayes selecionasse o melhor modelo. Desta forma, para todos os fenótipos foram ajustados modelos com um QTL a mais do que o modelo que foi selecionado pelo fator de Bayes. De forma geral, os dois estimadores utilizados para obter o fator de Bayes tiveram resultados bastante semelhantes.

Os resultados do mapeamento obtidos para os três diferentes fenótipos pela abordagem bayesiana foi comparado aos resultados obtidos pelo modelo CIM que estão apresentados em Sibov et al. (2003b). Algumas discordância entre número de QTLs e posições no genoma foram observadas.

Para o fenótipo produção de grãos foram mapeados 3 QTLs utilizando a abordagem bayesiana enquanto pelo modelo CIM foram mapeados 4 QTLs. Provavelmente, a abordagem bayesiana mapeou um número menor de QTLs por ter considerado epistasia no modelo ou porque o QTL mapeado pelo CIM tinham baixo efeito o que de certa forma é esperado já que os modelos usados foram diferentes. Nota-se que há concordância entre os três QTLs mapeados pela bayesiana e pelo modelo CIM. Para o fenótipo altura de espiga foram mapeados 4 QTLs utilizando a abordagem bayesiana enquanto pelo modelo CIM foram mapeados 5 QTLs. Novamente, observou-se um número menor de QTLs quando usada a abordagem bayesiana. Para este fenótipo algumas discodâncias foram observadas entre os resultados das duas metodologias. O modelo CIM mapeou um QTL no cromossomos 1, três QTLs no cromossomo 7 e um QTL no cromossomo 9, enquanto que a bayesiana mapeou um QTLs no cromossomo 2, dois QTLs no cromossomo 7 e um QTL no cromossomo 9. Para a altura da planta o número de QTLs mapeados foi o mesmo para as duas metodologias. Para este fenótipo apenas uma discordância foi observada, o modelo CIM mapeou um QTL no cromossomo 1 que a bayesiana não mapeou e a bayesiana mapeou um QTL no cromossomo 7 que o modelo CIM não mapeou. Vale ressaltar que a abordagem bayesiana utilizada neste trabalho considerou todos os QTLs simultaneamente no modelo e ainda incluiu termos de epistasia, assim é esperado que exista algumas discodância com o resultado apresentado pelo modelo CIM. Devido a limitação de tempo os efeitos individuais das epistasia ainda não foram testados, ou seja, é ainda necessário verificar quais efeitos devem ou não permanecer no modelo.

4.5 Conclusão

Os resultados obtidos nesta pesquisa levaram as seguintes conclusões:

- I) Os resultados obtidos no mapeamento de QTLs dos três fenótipos de milho tropical: produção de grão, altura da espiga e altura da planta foram semelhantes aos apresentados pelo modelo CIM.
- II) A abordagem Bayesiana utilizada no presente trabalho tem a grande vantagem de permitir considerar múltiplos QTLs e incluir efeitos de epistasia no modelo, possibilitando assim um melhor entendimento da arquitetura genética dos caracteres estudados.

Referências

BASTEN, C.J.; ZENG, Z.-B. **Windows QTL Cartographer 1.15**. Department of Statistics, North Carolina State University, Raleigh, NC, 1994.

CASELLA, G.; GEORGE, E.I. Explaining the Gibbs Sampler. **The American Statistician**, New York, v. 46, p. 167-174, 1992.

CHURCHILL, G.A; DOERGE, R.W. Empirical threshold values for quantitative trait mapping. **Genetics**, Austin, v. 138, p. 963-971, 1994.

GAFFNEY, P.J. **An efficient reversible jump Markov Chain Monte Carlo approach to detect multiple loci and their effects in inbred crosses.** 2001, 174p. Thesis (Philosophy in Statistics Doctor) - University of Wisconsin, Madison, 2001.

GREEN, P.J.; Reversible jump markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, Lausanne, v. 82, p. 711-732, 1995.

HEATH, S.C. Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. **The American Journal of Human Genetics**, Boston, v. 61, p. 748-760, 1997.

JEFFREY, H. **Theory of probability.** Oxford: Clarendon press, 1961. 447 p.

KAO, C.H.; ZENG, Z-B.; TEASDALE, R.D. Multiple interval mapping for quantitative trait loci. **Genetics**, Austin, v. 152, p. 1203-1216, 1999.

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

KOSAMBI, D.D. The estimation of map distances from recombination values. **Annual Eugenics**, New York, v. 12, p. 172-175, 1944.

NARITA, A.; SASAKI, Y. Detection of multiple QTL with epistatic effects under a mixed inheritance model in an outbred population. **Genetics Selection Evolution**, Versailles, v. 36, p. 415-433, 2004.

NEWTON, M.A.; RAFTERY, A.E. Approximate Bayesian inference by the weighted likelihood bootstrap. **Journal Of Royal Statistical Society**, London, v. 56, p. 1-48, 1994.

RAFTERY, A.E.; NEWTON, M.A.; SATAGOPAN, J. M.; KRIVITSKY, P.N. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. In: BERNARDO, J.M.; BAYARRI, M.J.; BERGER, J.O.; DAWID, A.P.; HECKERMAN, A.F.; SMITH, A.F.M; WEST, M. **Bayesian statistics**. Oxford: Oxford University Press, 2007, 8, p.1-45.

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

SATAGOPAN, J. M.; YANDELL, B.S. Estimating the number of quantitative trait Loci via Bayesian model determination. **Special contributed paper session on genetic analysis of quantitative trait and complex disease, Biometric section, Joint Statistical Meetings**, Chicago, 1998.

SATAGOPAN, J. M.; YANDELL, B.S. NEWTON M.A.; OSBORN, T.C. A Bayesian approach to detect Quantitative trait Loci using Markov chain Monte Carlo. **Genetics**, Austin, v. 144, p. 805-816, 1996.

SEN, S.; CHURCHILL A statistical framework for quantitative trait mapping . **Genetics**, Austin, v. 159, p. 371-387, 2001.

SIBOV, S.T.; de SOUZA JÚNIOR, C.L.; GARCIA, A.A.F.; GARCIA, A.F.; SILVA, A.R.; MANGOLIM, 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**, Lund, v. 139, n. 2, p. 96 - 106, 2003a.

SIBOV, S.T.; de SOUZA JÚNIOR, C.L.; GARCIA, A.A.F.; GARCIA, A.F.; SILVA, A.R.; MANGOLIM, 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**, Lund, v. 139, n. 2, p. 107 - 115, 2003b.

SILLANPÄÄ, M.J.; ARJAS, E. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data . **Genetics**, Austin, v. 148, p. 1373-1388, 1998.

STEPHENS, D.A.; FISCH, R.D. Bayesian analysis of quantitative trait locus data using reversible jump Markov Chain Monte Carlo **Biometrics**, Washington, v. 54, p. 1334-1347, 1998.

WANG, H.; ZHANG, Y.M.; MASINDE, G.L.; MOHAN, S; BAYLINK, D.J.; XU, S. Bayesian shrinkage estimation of quantitative trait loci parameters. **Genetics**, Austin, v. 170, p. 465-480, 2005.

YI, N. A unified markov chain monte carlo framework for mapping multiple quantitative trait loci. **Genetics**, Austin, v. 167, p. 967-975, 2004.

YI, N.; ALLISON, D.B.; XU, S. Bayesian model choice and search strategies for mapping multiple epistatic quantitative trait loci. **Genetics**, Austin, v. 165, p. 867-883, 2003.

YI, N.; XU, S. Mapping quantitative trait loci with epistatic effects. **Genetics Research** , Cambridge, v. 79, p. 185-198, 2002.

YI, N.; YANDELL, B.S.; CHURCILL, G.A.; ALLISON, D.B.; EISEN, E.J.; POMP, D. Bayesian model selection for genome-wide epistatic quantitative trait loci analysis. **Genetics**, Austin, v. 170, p. 1333-1344, 2005.

YI, N.; SHRINER, D.; BANERJEE, S.; MEHTA, T.; POMP, D YANDELL, B.S. An efficient Bayesian model selection approach for interacting QTL models with many effects. **Genetics**, Austin, v.176 , p. 1865-1877, 2007.

ZENG, Z-B. Theoretical basis for separation of multiple linked genes effects in mapping quantitative trait loci. **Procedure of the National Academic Science USA**, Washington, v. 90, p. 10972-10976, 1993.

ZENG, Z-B. Precision mapping of quantitative trait loci. **Genetics**, Austin, v. 136, p. 1457-1468, 1994.

APÊNDICES

APÊNDICE A - Programa em R utilizado

```

s<-3 # s é igual ao número de QTLs
D<-0.035 # D parametro de tuning
nep<-((4*s*(s-1))/2)/4 # número de epistasia
for(iter in 1:niter) {
  # Atualização de  $\sigma^2$ 
  sigma2.draw <- sigma2.draw.f(y, j, Q, Xaa, X12, X21, X22, mu.draw, alpha.draw, delta.draw,
aa.draw, ad.draw, da.draw, dd.draw)
  # Atualização de  $\lambda$ 
  lambdapp<-NULL
  lambdap<-NULL
  nqtl<-1
  marc.flanqp<-matrix(0,s,2)
  dap<-NULL
  dbp<-NULL
  dp<-NULL
  d.mfp<-matrix(0,s,2)
  rap<-NULL
  rbp<-NULL
  rp<-NULL
  mf.probp<-NULL
  A<-NULL
  probb<-matrix(0,1,s)
  fprobp<-matrix(0,n,s)
  while(nqtl<=s) {
    qprob<-matrix(0,16,3)
    lambdap[nqtl]<- geralambda(lambda,D,nqtl,comp.gen)
    res <- length(dist.marc[dist.marc<=lambdap[nqtl]])
    marc.flanqp[nqtl,]<-c(res,res+1)
    d.mfp[nqtl,]<-c(dist.marc[marc.flanqp[nqtl,1]],dist.marc[marc.flanqp[nqtl,2]])
    dap[nqtl]<- lambdap[nqtl]-d.mfp[nqtl,1]
    dbp[nqtl]<- d.mfp[nqtl,2]-lambdap[nqtl]
    dp[nqtl]<- diff(d.mfp[nqtl,1:2])
    rap[nqtl]<-fh.d.r(dap[nqtl])
    rbp[nqtl]<-fh.d.r(dbp[nqtl])
    rp[nqtl]<-rap[nqtl]+rbp[nqtl]-2*rap[nqtl]*rbp[nqtl]
  }
}

```

```

matp<-marc.flanqp
res.mqprobp<-NULL
# Construção da matriz de probabilidades
res.mqprobp<-mqprob(rap[nqtl],rbp[nqtl],rp[nqtl])
g.marc.flanqp<-cbind(qtl[,matp[nqtl,1]], qtl[,matp[nqtl,2]])
# Cálculo das probabilidades
mf.probp[nqtl]<-prod(f.fprob(g.marc.flanqp,Q,res.mqprobp,nqtl))
lambda.atual<-lambda
A[nqtl]<-(mf.probp[nqtl])/(mf.probp[nqtl])
u<-runif(s,0,1)
probb[,nqtl]=min(1, A[nqtl])
if(u[nqtl] <= probb[nqtl])
lambda[nqtl]=lambdap[nqtl]
da[nqtl]= dap[nqtl]
db[nqtl]= dbp[nqtl]
d[nqtl]= dp[nqtl]
ra[nqtl]= rap[nqtl]
rb[nqtl]= rbp[nqtl]
r[nqtl]= rp[nqtl]
mat[nqtl,]= matp[nqtl,]
taxa[nqtl]=taxa[nqtl]+1
} else {
lambda[nqtl]=lambda.atual[nqtl]
da[nqtl]= da[nqtl]
db[nqtl]= db[nqtl]
d[nqtl] = d[nqtl]
ra[nqtl] = ra[nqtl]
rb[nqtl] = rb[nqtl]
r[nqtl] = r[nqtl]
mat[nqtl,] = mat[nqtl,]
}
nqtl=nqtl+1
# Atualização da matriz de genótipos Q
mf.probp<-matrix(0,1,s)
for( j in 1:s) {
g.marc.flanqp<-NULL

```

```

qprob<-matrix(0,16,3)
res.mqprob1<-NULL
res.mqprob1<-mqprob(ra[j],rb[j],r[j])
g.marc.flaq<-cbind(qtl[,mat[j,1]], qtl[,mat[j,2]])
Q[,j]<-gera.Q(g.marc.flaq,res.mqprob1,j,sigma2.draw,alpha.draw,mu.draw,delta.draw,
aa.draw,ad.draw,da.draw,dd.draw)
# Cálculo das probabilidades
mf.prob[,j]<-prod(f.fprob(g.marc.flaq,Q,res.mqprob1,j)) } x1<-Q - 1
x2<-(1+x1)*(1-x1)- 0.5
Xaa<-geraXaa(x1)
X12<-geraX12(x1,x2)
X21<- geraX21(x1,x2)
X22<-geraX22(x2)
# Atualização da contante geral  $\mu$ 
mu.draw <- mu.draw.f(y, j,Q,Xaa,X12,X21,X22,alpha.draw, delta.draw,
sigma2.draw,
aa.draw,ad.draw,da.draw,dd.draw)
# Atualização da contante geral  $\alpha$ 
for(j in 1:s)
alpha [j] <- alpha.j.draw.f(y, j,Q,Xaa,X12,X21,X22, mu.draw, alpha, delta.draw, sigma2.draw,
aa.draw,ad.draw,da.draw,dd.draw)
alpha.draw <- c( alpha)
# Atualização da contante geral  $\delta$ 
for(j in 1:s)
delta[j] <- delta.j.draw.f(y, j,Q,Xaa,X12,X21,X22, mu.draw, alpha.draw, delta, sigma2.draw,
aa.draw,ad.draw,da.draw,dd.draw)
delta.draw <- c(delta)
# Atualização da contante geral  $\alpha\alpha$ 
for(j in 1:nep)
aa[j] <- aa.j.draw.f(y, j,Q,Xaa,X12,X21,X22, mu.draw, alpha.draw, delta.draw, sigma2.draw,
aa,ad.draw,da.draw,dd.draw)
aa.draw <- c(aa)
# Atualização da contante geral  $\alpha\delta$ 
for(j in 1:nep)
ad[j]<-ad.j.draw.f(y, j,Q,Xaa,X12,X21,X22, mu.draw, alpha.draw, delta.draw, sigma2.draw,
aa.draw,ad,da.draw,dd.draw)

```

```

ad.draw<-c(ad)
# Atualização da contante geral  $\delta\alpha$ 
for(j in 1:nep)
da[j]<- da.j.draw.f(y, j,Q,Xaa,X12,X21,X22, mu.draw, alpha.draw,
delta.draw, sigma2.draw,
aa.draw,ad.draw,da,dd.draw)
da.draw<-c(da)
# Atualização da contante geral  $\delta\delta$ 
for(j in 1:nep)
dd[j]<-dd.j.draw.f(y, j,Q,Xaa,X12,X21,X22, mu.draw, alpha.draw,
delta.draw, sigma2.draw,
aa.draw,ad.draw,da.draw,dd)
dd.draw<-c(dd)
# Fator de Bayes - Média Harmônica
bayes1<-postL(y,j,x1,x2,Xaa,X12,X21,X22, mu.draw, alpha.draw,
delta.draw, sigma2.draw,aa.draw,ad.draw,da.draw,dd.draw,mf.prob)
bayes=rbind(bayes,bayes1)
# Fator de Bayes - Estabilizador da Média Harmônica
bayesT1<-postLT(y,j,x1,x2,Xaa,X12,X21,X22, mu.draw, alpha.draw, delta.draw, sigma2.draw,
aa.draw,ad.draw,da.draw,dd.draw)
bayesT=rbind(bayesT,bayesT1)
parametros=rbind(parametros,c((lambda),mu.draw,alpha.draw,delta.draw,
aa.draw,ad.draw,da.draw,dd.draw,sigma2.draw))
}
### Funções utilizadas ###

# Função para converter a distância em fração de recombinação e vice-versa
# I. Função de Haldane
#(I1) Convertendo distância (d) em fração de recombinação (r)
fh.d.r<- function(d) 1/2*(1-exp(-2*(d)))
# Kosambi
fh.d.r<- function(d) 1/2*((exp(4*(d))-1) / (exp(4*(d))+1))
#(I2) Convertendo fração de recombinação (r) em distância (d)
fh.r.d<- function(r) -1/2*log(1-2*r)
# Construção da matriz de probabilidades
mqprob <- function(ra,rb,r)

```

```

ra1 <- 1-ra
rb1 <- 1-rb
r1 <- 1-r
#QQ# qprob[1,1]<-pQQMMMM<-((ra1)^2 * (rb1)^2)/(r1)^2
qprob[2,1]<-pQQMMMm<-((ra1)^2 * rb * (rb1))/(r*(r1))
qprob[3,1]<-pQQMMmm<-((ra1)^2*rb^2)/r^2
qprob[4,1]<-pQQMmMM<-(ra*(ra1)*(rb1)^2)/(r*(r1))
qprob[5,1]<-pQQMmMm<-(2*ra*(ra1)*rb*(rb1))/((r1)^2 + r^2)
qprob[6,1]<-pQQMmmm<-(ra*(ra1)*rb^2)/(r*(r1))
qprob[7,1]<-pQQmmMM<-(ra^2*(rb1)^2)/(r^2)
qprob[8,1]<-pQQmmMm<-(ra^2*rb*(rb1))/(r*(r1))
qprob[9,1]<-pQQmmmm<-(ra^2*rb^2)/(r1^2)
qprob[10,1]<-pQQMMPP<- ra1^2
qprob[11,1]<-pQQMmPP<- ra*ra1
qprob[12,1]<-pQQmmPP<- ra^2
qprob[13,1]<-pQQPPMM<- rb1^2
qprob[14,1]<-pQQPPMm<- rb*rb1
qprob[15,1]<-pQQPPmm <- rb^2
qprob[16,1]<-pQQPPPP <- 1/4
#Qq#
qprob[1,2]<-pQqMMMM<-(2*ra*(ra1)*rb*(rb1))/(r1)^ 2
qprob[2,2]<-pQqMMMm<-(ra*(ra1)*(rb^2+(rb1)^2))/(r*(r1))
qprob[3,2]<-pQqMMmm<-(2*ra*(ra1)*rb*(rb1))/r^2
qprob[4,2]<-pQqMmMM<-(ra^2+(ra1)^2)*rb*(rb1))/(r*(r1))
qprob[5,2]<-pQqMmMm<-(ra^2+(ra1)^2)*(rb^2+(rb1)^2))/((r1)^2 + r^2)
qprob[6,2]<-pQqMmmm<-(ra^2+(ra1)^2)*rb*(rb1))/(r*(r1))
qprob[7,2]<-pQqmmMM<-(2*ra*(ra1)*rb*(rb1))/r^2
qprob[8,2]<-pQqmmMm<-(ra*(ra1)*(rb^2+(rb1)^2))/(r*(r1))
qprob[9,2]<-pQqmmmm<-(2*ra*(ra1)*rb*(rb1))/(r1)^2
qprob[10,2]<-pQqMMPP<- 2*ra*ra1
qprob[11,2]<-pQqMmPP<- ra1^2 + ra^2
qprob[12,2]<-pQqmmPP<- 2* ra*ra1
qprob[13,2]<-pQqPPMM<- 2*rb*rb1
qprob[14,2]<-pQqPPMM<- rb1^2 + rb^2
qprob[15,2]<-pQqPPmm<- 2*rb*rb1
qprob[16,2]<-pQqPPPP <- 1/2

```

```

#qq#
qprob[1,3]<-pqqMMMM<-(ra^2*rb^2)/(r1)^2
qprob[2,3]<-pqqMMMm<-(ra^2*rb*(rb1))/(r*(r1))
qprob[3,3]<-pqqMMmm<-(ra^2*(rb1)^2)/r^2
qprob[4,3]<-pqqMmMM<-(ra*(ra1)*rb^2)/(r*(r1))
qprob[5,3]<-pqqMmMm<-(2*ra*(ra1)*rb*(rb1))/((r1)^2 + r^2)
qprob[6,3]<-pqqMmmm<-(ra*(ra1)*(rb1)^2)/(r*(r1))
qprob[7,3]<-pqqmmMM<-((ra1)^2*rb^2)/r^2
qprob[8,3]<-pqqmmMm<-((ra1)^2 * rb *(rb1))/(r*(r1))
qprob[9,3]<-pqqmmmm<-((ra1)^2*(rb1)^2)/(r1)^2
qprob[10,3]<-pqqMMPP<- ra^2
qprob[11,3]<-pqqMmPP<- ra* ra1
qprob[12,3]<-pqqmmPP<- ra1^2
qprob[13,3]<-pqqPPMM<- rb^2
qprob[14,3]<-pqqPPMm<- rb* rb1
qprob[15,3]<-pqqPPmm<- rb1^2
qprob[16,3]<-pqqPPPP<- 1/4
return(qprob)
}
# Função para gerar os genótipos dos QTLs
gera.Q<-function(g.marc.flanq,res.mqprob,j,sigma2.draw,alpha.draw,mu.draw,delta.draw,
aa.draw,ad.draw,da.draw,dd.draw){
for (i in 1:n) {
qprob<-NULL
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==2)) qprob<-res.mqprob[1,]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==1)) qprob<-res.mqprob[2,]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==0)) qprob<-res.mqprob[3,]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==2)) qprob<-res.mqprob[4,]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==1)) qprob<-res.mqprob[5,]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==0)) qprob<-res.mqprob[6,]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==2)) qprob<-res.mqprob[7,]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==1)) qprob<-res.mqprob[8,]
if((g.marc.flanq[i,1]==0)&& (g.marc.flanq[i,2]==0)) qprob<-res.mqprob[9,]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==-1)) qprob<-res.mqprob[10,]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==-1)) qprob<-res.mqprob[11,]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==-1)) qprob<-res.mqprob[12,]

```

```

if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==2)) qprob<-res.mqprob[13,]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==1)) qprob<-res.mqprob[14,]
if((g.marc.flanq[i,1]==-1) &&(g.marc.flanq[i,2]==0)) qprob<-res.mqprob[15,]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==-1)) qprob<-res.mqprob[16,]
}
Q[i,j]=2
x1[i,j]<- 1
x2[i,j]<- -0.5
if(j==1) {(Xaa[i,1]<-x1[i,1]*x1[i,2])
(Xaa[i,2]<-x1[i,1]*x1[i,3])
(X12[i,1]<-x1[i,1]*x2[i,2])
(X12[i,2]<-x1[i,1]*x2[i,3])
(X21[i,1]<-x2[i,1]*x1[i,2])
(X21[i,2]<-x2[i,1]*x1[i,3])
(X22[i,1]<-x2[i,1]*x2[i,2])
(X22[i,2]<-x2[i,1]*x2[i,3])
}
if(j==2) {(Xaa[i,1]<-x1[i,1]*x1[i,2])
(Xaa[i,3]<-x1[i,2]*x1[i,3])
(X12[i,1]<-x1[i,1]*x2[i,2])
(X12[i,3]<-x1[i,2]*x2[i,3])
(X21[i,1]<-x2[i,1]*x1[i,2])
(X21[i,3]<-x2[i,2]*x1[i,3])
(X22[i,1]<-x2[i,1]*x2[i,2])
(X22[i,3]<-x2[i,2]*x2[i,3])
}
if(j==3) {(Xaa[i,2]<-x1[i,1]*x1[i,3])
(Xaa[i,3]<-x1[i,2]*x1[i,3])
(X12[i,2]<-x1[i,1]*x2[i,3])
(X12[i,3]<-x1[i,2]*x2[i,3])
(X21[i,2]<-x2[i,1]*x1[i,3])
(X21[i,3]<-x2[i,2]*x1[i,3])
(X22[i,2]<-x2[i,1]*x2[i,3])
(X22[i,3]<-x2[i,2]*x2[i,3])
}
prob1<-dnorm(y[i,], mean=(mu.draw + sum(alpha.draw*x1[i,]) + sum(delta.draw*x2[i,])+

```

```

Xaa[i,]%*%aa.draw + X12[i,]%*%ad.draw + X21[i,]%*%da.draw + X22[i,]%*%dd.draw),
sd=sqrt(sigma2.draw), log = FALSE)
Q[i,j]=1
x1[i,j]<- 0
x2[i,j]<- 0.5
if(j==1) {(Xaa[i,1]<-x1[i,1]*x1[i,2])
(Xaa[i,2]<-x1[i,1]*x1[i,3])
(X12[i,1]<-x1[i,1]*x2[i,2])
(X12[i,2]<-x1[i,1]*x2[i,3])
(X21[i,1]<-x2[i,1]*x1[i,2])
(X21[i,2]<-x2[i,1]*x1[i,3])
(X22[i,1]<-x2[i,1]*x2[i,2])
(X22[i,2]<-x2[i,1]*x2[i,3])
}
if(j==2) {(Xaa[i,1]<-x1[i,1]*x1[i,2])
(Xaa[i,3]<-x1[i,2]*x1[i,3])
(X12[i,1]<-x1[i,1]*x2[i,2])
(X12[i,3]<-x1[i,2]*x2[i,3])
(X21[i,1]<-x2[i,1]*x1[i,2])
(X21[i,3]<-x2[i,2]*x1[i,3])
(X22[i,1]<-x2[i,1]*x2[i,2])
(X22[i,3]<-x2[i,2]*x2[i,3])
}
if(j==3) {(Xaa[i,2]<-x1[i,1]*x1[i,3])
(Xaa[i,3]<-x1[i,2]*x1[i,3])
(X12[i,2]<-x1[i,1]*x2[i,3])
(X12[i,3]<-x1[i,2]*x2[i,3])
(X21[i,2]<-x2[i,1]*x1[i,3])
(X21[i,3]<-x2[i,2]*x1[i,3])
(X22[i,2]<-x2[i,1]*x2[i,3])
(X22[i,3]<-x2[i,2]*x2[i,3])
}
prob2<-dnorm(y[i,], mean=(mu.draw + sum(alpha.draw*x1[i,]) + sum(delta.draw*x2[i,])+
Xaa[i,]%*%aa.draw + X12[i,]%*%ad.draw + X21[i,]%*%da.draw + X22[i,]%*%dd.draw),
sd=sqrt(sigma2.draw), log = FALSE)
Q[i,j]=0

```



```

x1[i,j]<- -1
x2[i,j]<- -0.5
if(j==1) {(Xaa[i,1]<-x1[i,1]*x1[i,2])
(Xaa[i,2]<-x1[i,1]*x1[i,3])
(X12[i,1]<-x1[i,1]*x2[i,2])
(X12[i,2]<-x1[i,1]*x2[i,3])
(X21[i,1]<-x2[i,1]*x1[i,2])
(X21[i,2]<-x2[i,1]*x1[i,3])
(X22[i,1]<-x2[i,1]*x2[i,2])
(X22[i,2]<-x2[i,1]*x2[i,3])
}
if(j==2) {(Xaa[i,1]<-x1[i,1]*x1[i,2])
(Xaa[i,3]<-x1[i,2]*x1[i,3])
(X12[i,1]<-x1[i,1]*x2[i,2])
(X12[i,3]<-x1[i,2]*x2[i,3])
(X21[i,1]<-x2[i,1]*x1[i,2])
(X21[i,3]<-x2[i,2]*x1[i,3])
(X22[i,1]<-x2[i,1]*x2[i,2])
(X22[i,3]<-x2[i,2]*x2[i,3])
}
if(j==3) {(Xaa[i,2]<-x1[i,1]*x1[i,3])
(Xaa[i,3]<-x1[i,2]*x1[i,3])
(X12[i,2]<-x1[i,1]*x2[i,3])
(X12[i,3]<-x1[i,2]*x2[i,3])
(X21[i,2]<-x2[i,1]*x1[i,3])
(X21[i,3]<-x2[i,2]*x1[i,3])
(X22[i,2]<-x2[i,1]*x2[i,3])
(X22[i,3]<-x2[i,2]*x2[i,3])
}
prob3<-dnorm(y[i,], mean=(mu.draw + sum(alpha.draw*x1[i,]) + sum(delta.draw*x2[i,])+
Xaa[i,]*%aa.draw + X12[i,]*%ad.draw + X21[i,]*%da.draw + X22[i,]*%dd.draw),
sd=sqrt(sigma2.draw), log = FALSE)
probt<-cbind(prob1,prob2,prob3)
qprob1<-(qprob[1]*prob1)/sum(qprob*probt)
qprob2<-(qprob[2]*prob2)/sum(qprob*probt)
qprob3<-(qprob[3]*prob3)/sum(qprob*probt)

```

```

sum(qprob1+qprob2+qprob3)
qprobf<-cbind(qprob1,qprob2,qprob3)
Q[i,j] <- sample(c(2,1,0),1,replace=TRUE,prob=qprobf)
}
return(Q[,j])
}
#Cálculo das probabilidades#
f.fprob<-function(g.marc.flanq,Q,res.mqprob,j) for (i in 1:n)
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==2) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[1,1] if((g.marc.flanq[i,1]==2) &&
(g.marc.flanq[i,2]==2) && Q[i,j]==1) fprob[i,j]<-res.mqprob[1,2]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==2) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[1,3] if((g.marc.flanq[i,1]==2) &&
(g.marc.flanq[i,2]==1) && Q[i,j]==2) fprob[i,j]<-res.mqprob[2,1]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==1) && Q[i,j]==1)
fprob[i,j]<-res.mqprob[2,2] if((g.marc.flanq[i,1]==2) &&
(g.marc.flanq[i,2]==1) && Q[i,j]==0) fprob[i,j]<-res.mqprob[2,3]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==0) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[3,1] if((g.marc.flanq[i,1]==2) &&
(g.marc.flanq[i,2]==0) && Q[i,j]==1) fprob[i,j]<-res.mqprob[3,2]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==0) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[3,3] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==2) && Q[i,j]==2) fprob[i,j]<-res.mqprob[4,1]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==2) && Q[i,j]==1)
fprob[i,j]<-res.mqprob[4,2] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==2) && Q[i,j]==0) fprob[i,j]<-res.mqprob[4,3]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==1) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[5,1] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==1) && Q[i,j]==1) fprob[i,j]<-res.mqprob[5,2]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==1) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[5,3] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==0) && Q[i,j]==2) fprob[i,j]<-res.mqprob[6,1]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==0) && Q[i,j]==1)
fprob[i,j]<-res.mqprob[6,2] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==0) && Q[i,j]==0) fprob[i,j]<-res.mqprob[6,3]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==2) && Q[i,j]==2)

```

```

fprob[i,j]<-res.mqprob[7,1] if((g.marc.flanq[i,1]==0) &&
(g.marc.flanq[i,2]==2) && Q[i,j]==1) fprob[i,j]<-res.mqprob[7,2]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==2) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[7,3] if((g.marc.flanq[i,1]==0) &&
(g.marc.flanq[i,2]==1) && Q[i,j]==2) fprob[i,j]<-res.mqprob[8,1]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==1) && Q[i,j]==1)
fprob[i,j]<-res.mqprob[8,2] if((g.marc.flanq[i,1]==0) &&
(g.marc.flanq[i,2]==1) && Q[i,j]==0) fprob[i,j]<-res.mqprob[8,3]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==0) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[9,1] if((g.marc.flanq[i,1]==0) &&
(g.marc.flanq[i,2]==0) && Q[i,j]==1) fprob[i,j]<-res.mqprob[9,2]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==0) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[9,3]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==-1) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[10,1] if((g.marc.flanq[i,1]==2) &&
(g.marc.flanq[i,2]==-1) && Q[i,j]==1) fprob[i,j]<-res.mqprob[10,2]
if((g.marc.flanq[i,1]==2) && (g.marc.flanq[i,2]==-1) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[10,3] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==-1) && Q[i,j]==2) fprob[i,j]<-res.mqprob[11,1]
if((g.marc.flanq[i,1]==1) && (g.marc.flanq[i,2]==-1) && Q[i,j]==1)
fprob[i,j]<-res.mqprob[11,2] if((g.marc.flanq[i,1]==1) &&
(g.marc.flanq[i,2]==-1) && Q[i,j]==0) fprob[i,j]<-res.mqprob[11,3]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==-1) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[12,1] if((g.marc.flanq[i,1]==0) &&
(g.marc.flanq[i,2]==-1) && Q[i,j]==1) fprob[i,j]<-res.mqprob[12,2]
if((g.marc.flanq[i,1]==0) && (g.marc.flanq[i,2]==-1) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[12,3]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==2) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[13,1] if((g.marc.flanq[i,1]==-1) &&
(g.marc.flanq[i,2]==2) && Q[i,j]==1) fprob[i,j]<-res.mqprob[13,2]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==2) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[13,3]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==1) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[14,1] if((g.marc.flanq[i,1]==-1) &&
(g.marc.flanq[i,2]==1) && Q[i,j]==1) fprob[i,j]<-res.mqprob[14,2]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==1) && Q[i,j]==0)

```

```

fprob[i,j]<-res.mqprob[14,3] if((g.marc.flanq[i,1]==-1) &&
(g.marc.flanq[i,2]==0) && Q[i,j]==2) fprob[i,j]<-res.mqprob[15,1]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==0) && Q[i,j]==1)
fprob[i,j]<-res.mqprob[15,2] if((g.marc.flanq[i,1]==-1) &&
(g.marc.flanq[i,2]==0) && Q[i,j]==0) fprob[i,j]<-res.mqprob[15,3]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==-1) && Q[i,j]==2)
fprob[i,j]<-res.mqprob[16,1] if((g.marc.flanq[i,1]==-1) &&
(g.marc.flanq[i,2]==-1) && Q[i,j]==1) fprob[i,j]<-res.mqprob[16,2]
if((g.marc.flanq[i,1]==-1) && (g.marc.flanq[i,2]==-1) && Q[i,j]==0)
fprob[i,j]<-res.mqprob[16,3]

}
return(fprob[,j]) }
# Função para gerar  $\mu$ 
mu.draw.f <- function(y, j,Q,Xaa,X12,X21,X22,alpha, delta,
sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
cte1 <- 1/tau2 + n/sigma2
m.mu <- (eta/tau2 + sum( y - Q%*%alpha - QQ%*%delta- Xaa%*%aa -
X12%*%ad - X21%*%da - X22%*%dd)/sigma2)/cte1
var.mu <- 1/cte1
rnorm(1, m.mu, sqrt(var.mu))
}
# Função para gerar  $\alpha$ 
alpha.j.draw.f <- function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
Qj <- as.matrix(Q[,j])
cte2 <- 1/taua + sum( Q[,j]^2 )/sigma2
m.alpha.j <- sum( Q[,j]*(y - mu - Qj%*%alpha[-j] - QQ%*%delta- Xaa%*%aa - X12%*%ad -
X21%*%da - X22%*%dd ) )/ (sigma2*cte2)
var.alpha.j <- 1/cte2
rnorm(1, m.alpha.j, sqrt(var.alpha.j))

```

```

}
# Função para gerar  $\delta$ 
delta.j.draw.f<-function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
QQj <- as.matrix(QQ[,-j])
cte2 <- 1/taud + sum( QQ[,j]^2 )/sigma2
m.delta.j <- sum( QQ[,j]*(y - mu - Q%%alpha - QQj%%delta[-j] - Xaa%%aa - X12%%ad -
X21%%da - X22%%dd ) )/ (sigma2*cte2)
var.delta.j <- 1/cte2
rnorm(1,m.delta.j,sqrt(var.delta.j))
}
# Função para gerar  $\alpha\alpha$ 
aa.j.draw.f <- function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
cte2 <- 1/tauaa + sum( Xaa[,j]^2 )/sigma2
m.aa.j <- sum( Xaa[,j]*(y - mu - Q%%alpha - QQ%%delta - Xaa[,j]%%aa[-j] - X12%%ad -
X21%%da - X22%%dd ) )/ (sigma2*cte2)
var.aa.j <- 1/cte2
rnorm(1, m.aa.j, sqrt(var.aa.j))
}
# Função para gerar  $\alpha\delta$ 
ad.j.draw.f <- function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
cte2 <- 1/tauad + sum( X12[,j]^2 )/sigma2
m.ad.j <- sum( X12[,j]*(y - mu - Q%%alpha - QQ%%delta - Xaa%%aa - X12[,j]%%ad[-j] -
X21%%da - X22%%dd ) )/ (sigma2*cte2)
var.ad.j <- 1/cte2
rnorm(1, m.ad.j, sqrt(var.ad.j))
}
# Função para gerar  $\delta\alpha$ 

```

```

da.j.draw.f <-function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
cte2 <- 1/tauda + sum( X21[,j]^2 )/sigma2
m.da.j <- sum( X21[,j]*(y - mu - Q%%alpha - QQ%%delta - Xaa%%aa - X12%%ad - X21[,-
j]%%da[-j] - X22%%dd ) )/ (sigma2*cte2)
var.da.j <- 1/cte2
rnorm(1, m.da.j, sqrt(var.da.j))
}
# Função para gerar  $\delta\delta$ 
dd.j.draw.f <- function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
cte2 <- 1/taudd + sum( X22[,j]^2 )/sigma2
m.dd.j <- sum( X22[,j]*(y - mu - Q%%alpha - QQ%%delta - Xaa%%aa - X12%%ad - X21%%da
- X22[,j]%%dd[-j] ) )/ (sigma2*cte2)
var.dd.j <- 1/cte2
rnorm(1, m.dd.j, sqrt(var.dd.j))
}
# Função para gerar  $\sigma^2$ 
sigma2.draw.f <- function(y, j,Q,Xaa,X12,X21,X22, mu, alpha, delta,aa,ad,da,dd){
x1<-Q-1
Q<-x1
QQ <- x2<-(1+x1)*(1-x1)- 0.5
a <- n/2 + a1
b <- 1/b1 + sum( (y - mu - Q%%alpha - QQ%%delta - Xaa%%aa - X12%%ad - X21%%da
- X22%%dd)^2 )/2
precision<-rgamma(1, shape=a, scale=1/b)
1/precision
}
# Função para gerar  $\lambda$ 
geralambda<-function(lambda,D,nqtl,comp.gen)
lambdapp[nqtl]<-runif(1,lambda[nqtl] - D,lambda[nqtl] + D)
lambdap[nqtl]=min(abs(lambdapp[nqtl]),(2*comp.gen-abs(lambdapp[nqtl])))

```

```

return(lambdap[nqtl])
}
# Função para gerar  $X_{aa}$ 
geraXaa<- function(x1)
for (j in 1:nep){
if(j==1) (Xaa[,j]<-x1[,1]*x1[,2])
if(j==2) (Xaa[,j]<-x1[,1]*x1[,3])
if(j==3) (Xaa[,j]<-x1[,2]*x1[,3])
}
return(Xaa)
}
# Função para gerar  $X_{12}$ 
geraX12<- function(x1,x2){
for (j in 1:nep){
if(j==1) (X12[,j]<-x1[,1]*x2[,2])
if(j==2) (X12[,j]<-x1[,1]*x2[,3])
if(j==3) (X12[,j]<-x1[,2]*x2[,3])
} return(X12) }
# Função para gerar  $X_{21}$ 
geraX21<- function(x1,x2) {
for (j in 1:nep){
if(j==1) (X21[,j]<-x2[,1]*x1[,2])
if(j==2) (X21[,j]<-x2[,1]*x1[,3])
if(j==3) (X21[,j]<-x2[,2]*x1[,3])
}
return(X21)
}
# Função para gerar  $X_{22}$ 
geraX22<- function(x2){
for (j in 1:nep){
if(j==1) (X22[,j]<-x2[,1]*x2[,2])
if(j==2) (X22[,j]<-x2[,1]*x2[,3])
if(j==3) (X22[,j]<-x2[,2]*x2[,3])
}
return(X22)
}

```

Função para gerar Fator de Bayes - Média Harmônica

```
postL<-function(y, j,x1,x2,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd,mf.prob) {
  prod(dnorm(y, mean=(mu + x1%%alpha + x2%%delta + Xaa%%aa + X12%%ad + X21%%da
  + X22%%dd), sd=sqrt(sigma2), log = FALSE))
}
```

Função para gerar Fator de Bayes - Estabilizador da Média Harmônica

```
postLT<-function(y, j,x1,x2,Xaa,X12,X21,X22, mu, alpha, delta, sigma2,aa,ad,da,dd){
  mm=(mu + x1%%alpha + x2%%delta + Xaa%%aa + X12%%ad + X21%%da + X22%%dd)
  cov1=diag(rep(1, length(y)))
  dmvt(t(y), mm, cov1, df = b1, log = FALSE)
}
```

APÊNDICE B - Distribuições condicional completa a posteriori

a) distribuição condicional completa a posteriori para μ :

$$\mu | \lambda, Q, \alpha, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma_2, \mathbf{y} \sim N \left(\frac{\sum_{i=1}^n (mi)}{\sigma^2 \left(\frac{1}{\tau^2} + \frac{n}{\sigma^2} \right)}, \frac{1}{\frac{1}{\tau^2} + \frac{n}{\sigma^2}} \right) \quad (2)$$

em que:

$$mi = (y_i - \sum_{j=1}^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'4})$$

b) distribuição condicional completa a posteriori para α_{j^*} :

$$\alpha_{j^*} | \lambda, Q, \mu, [\alpha]_{j \neq j^*}, \delta, \alpha\alpha, \alpha\delta, \delta\alpha, \delta\delta, \sigma_2, \mathbf{y} \sim N \left(\frac{\sum_{i=1}^n x_{ij^*1} a_i}{\sigma^2 \left(\frac{1}{\sigma_\alpha^2} + \frac{\sum_{i=1}^n x_{ij^*1}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_\alpha^2} + \frac{\sum_{i=1}^n x_{ij^*1}^2}{\sigma^2}} \right) \quad (3)$$

em que:

$$a_i = (y_i - \mu - \sum_{j \neq j^*}^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'4})$$

c) distribuição condicional completa a posteriori para δ_{j^*} :

$$\delta_{j^*} | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, [\delta]_{j \neq j^*}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \sigma_2, y \sim N \left(\frac{\sum_{i=1}^n x_{ij^*2} d_i}{\sigma^2 \left(\frac{1}{\sigma_\delta^2} + \frac{\sum_{i=1}^n x_{ij^*1}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_\delta^2} + \frac{\sum_{i=1}^n x_{ij^*2}^2}{\sigma^2}} \right) \quad (4)$$

em que:

$$d_i = (y_i - \mu - \sum_{j=1}^S \alpha_j x_{ij1} - \sum_{j \neq j^*}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

d) distribuição condicional completa a posteriori para σ^2 :

$$\sigma^2 | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \mathbf{y} \sim IG \left(a + \frac{n}{2}, \frac{1}{\frac{\sum_{i=1}^n S_i}{2} + \frac{1}{b}} \right) \quad (5)$$

$$S_i = (y_i - \mu - \sum_{j=1}^S \alpha_j x_{ijx} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})^2$$

d) distribuição condicional completa a posteriori para $\alpha \alpha_{j^*}$:

$$\alpha \alpha_{j^*} | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, [\boldsymbol{\alpha\alpha}]_{j \neq j^*}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \sigma_2, y \sim N \left(\frac{\sum_{i=1}^n x_{ij^*j'l_1} E1_i}{\sigma^2 \left(\frac{1}{\sigma_{\gamma_1}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_1}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_{\gamma_1}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_2}^2}{\sigma^2}} \right) \quad (6)$$

em que:

$$E1_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j \neq j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

e) distribuição condicional completa a posteriori para $\alpha \delta_{j^*}$:

$$\gamma_{2j^*} | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, [\boldsymbol{\alpha\delta}]_{j \neq j^*}, \boldsymbol{\delta\alpha}, \boldsymbol{\delta\delta}, \sigma_2, \mathbf{y} \sim N \left(\frac{\sum_{i=1}^n x_{ij^*j'l_2} E2_i}{\sigma^2 \left(\frac{1}{\sigma_{\gamma_2}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_2}^2}{\sigma^2} \right)}, \frac{1}{\frac{1}{\sigma_{\gamma_2}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_2}^2}{\sigma^2}} \right) \quad (7)$$

em que:

$$E2_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j \neq j^*}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

f) distribuição condicional completa a posteriori para $\delta\alpha_{j^*}$:

$$\gamma_{3j^*} | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, [\boldsymbol{\delta\alpha}]_{3j \neq j^*}, \boldsymbol{\delta\delta}, \sigma_2, y \sim N\left(\frac{\sum_{i=1}^n x_{ij^*j'l_3} E3_i}{\sigma^2 \left(\frac{1}{\sigma_{\gamma_3}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_3}^2}{\sigma^2}\right)}, \frac{1}{\frac{1}{\sigma_{\gamma_3}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_3}^2}{\sigma^2}}\right) \quad (8)$$

em que:

$$E3_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j \neq j^*}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j=1}^k \delta_j \delta_{j'} x_{ijj'l_4})$$

g) distribuição condicional completa a posteriori para $\delta\delta_{j^*}$:

$$\gamma_{4j^*} | \boldsymbol{\lambda}, Q, \mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\alpha\alpha}, \boldsymbol{\alpha\delta}, \boldsymbol{\delta\alpha}, [\boldsymbol{\delta\delta}]_{j \neq j^*}, \sigma_2, y \sim N\left(\frac{\sum_{i=1}^n x_{ij^*j'l_4} E4_i}{\sigma^2 \left(\frac{1}{\sigma_{\gamma_4}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_4}^2}{\sigma^2}\right)}, \frac{1}{\frac{1}{\sigma_{\gamma_4}^2} + \frac{\sum_{i=1}^n x_{ij^*j'l_4}^2}{\sigma^2}}\right) \quad (9)$$

em que:

$$E4_i = (y_i - \mu - \sum_j^S \alpha_j x_{ij1} - \sum_{j=1}^S \delta_j x_{ij2} - \sum_{j=1j^*}^k \alpha_j \alpha_{j'} x_{ijj'l_1} - \sum_{j=1}^k \alpha_j \delta_{j'} x_{ijj'l_2} - \sum_{j=1}^k \delta_j \alpha_{j'} x_{ijj'l_3} - \sum_{j \neq j^*}^k \delta_j \delta_{j'} x_{ijj'l_4})$$