Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Modelo hierárquico bayesiano na determinação de associação entre marcadores e QTL em uma população F_2

Renato Nunes Pereira

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2012

Renato Nunes Pereira
Licenciado em Matemática

Modelo hierárquico bayesiano na determinação de associação entre marcadores e QTL em uma população F_2

Orientadora:
Prof^{a}. Dra. **ROSELI APARECIDA LEANDRO**

Tese apresentada para obtenção do título de Doutor em Ciências. Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2012

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

Pereira, Renato Nunes
Modelo hierárquico bayesiano na determinacão de associacão entre marcadores e QTL em uma populacão F2 / Renato Nunes Pereira. - - Piracicaba, 2012.
126 p. : il.

Tese (Doutorado) - - Escola Superior de Agricultura "Luiz de Queiroz", 2012.

1. Distribuições (Probabilidade) 2. Genética estatística 3. Inferência bayesiana 4. Marcador molecular 5. Mapeamento genético 6. Modelos matemáticos 7. Milho 8. Populações vegetais I. Título

CDD 633.15
P436m

Dedicatória

Aos meus pais, **José e Santa**,
que souberam conduzir com
muita sabedoria a minha
formação.

AGRADECIMENTOS

Tentarei expressar nestes poucos parágrafos parte da minha gratidão em relação a algumas pessoas que colaboraram direta ou indiretamente para a conclusão desse trabalho. Ao longo desses quatro anos posso dizer que aprendi na prática o significado de várias palavras.

Aprendi que heróis são pessoas que fizeram o que era necessário fazer, que demonstram amor e compaixão e agem sempre de acordo com esses sentimentos, além de apresentar um caráter consistente. Por isso e por muitas outras coisas eu gostaria de agradecer aos meus heróis, Santa e José, por meus estudos sempre terem sido um sonho para eles e por terem me dado todo apoio necessário para realizá-lo.

Aprendi que verdadeiras amizades continuam a crescer mesmo a longas distâncias e que bons amigos são a família que Deus nos permitiu escolher. Diante disso eu gostaria de agradecer a minha família em Piracicaba. A Rose e Keliny pelo carinho, compreensão e por me fazer perceber que ao lado de bons amigos podemos fazer qualquer coisa, ou nada, e ainda assim termos bons momentos. Obrigado por terem sido a minha luz, o meu discernimento e minha alegria. Ao Tiago e Popó por permitirem que caminhássemos juntos e sermos sempre unidos. Ao Tiago Egídio pelo exemplo de bondade e serenidade. A minha querida sobrinha Ana Paula, pela delicadeza de saber me ouvir e permitir que ao seu lado eu conseguisse trabalhar os valores essenciais para o ser humano.

Aprendi que as pessoas com quem mais nos importamos na vida são tomadas de nós muito rapidamente. Obrigado Jovelino, por ter semeado a alegria, honestidade, humildade e amor na nossa família e principalmente por permitir que aprendêssemos com você, mesmo no momento mais difícil.

Aprendi que nada na vida conquistamos sozinhos. Sempre precisamos de outras pessoas para alcançar os nossos objetivos. Muitas vezes um simples gesto pode mudar a nossa vida e contribuir para o nosso sucesso. Obrigado Professora Dra. Roseli Aparecida Leandro pela sua orientação, pelo acompanhamento do meu trabalho, pela confiança em mim depositada e amizade.

Aprendi principalmente a agradecer. Obrigado,

Aos professores do programa de Pós-graduação em Estatística e Experimentação Agronômica, Dr. César Gonçalves de Lima, Dra. Clarice Garcia Borges Demétrio,

Dr. Décio Barbin, Dr. Edwin Ortega, Dr. Gabriel Adrián Sarriés, Dr. Sílvio Zocchi, Dra. Sônia Maria de Stefano Piedade, Dra Taciana Villela Savian e Dr. Vitor Augusto Ozaki por todos os ensinamentos e amizade.

Aos professores Doutores Andréia da Silva Meyer, Júlio Sílvio de Sousa Bueno Filho e Taciana Villela Savian pelas importantes sugestões dadas no meu exame de qualificação.

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

Ao Conselho Nacional de Desenvolvimento Científico - CNPQ, pela concessão de bolsas de estudos.

À Escola Superior de Agricultura "Luiz de Queiroz" - ESALQ/USP, pela excelência de estrutura pública oferecida.

Às secretárias Luciane Brajão e Solange Sabadin que, indo além de suas funções, foram excelentes amigas.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, Eduardo Bonilha, Jorge Alexandre Wiendl e Rosni Honofre Aparecido Pinto pelos auxílios permanentes.

Aos professores do Departamento de Estatística da UFRRJ, pelo incentivo, amizade e confiança em mim depositada.

Aos amigos Patrícia Ferreira e Braga pelo carinho, companhia e disposição.

Aos amigos dos cursos de mestrado e doutorado do Departamento de Ciências Exatas da ESALQ/USP, em especial a Ana Júlia, Ana Patrícia Bastos Peixoto, Eduardo Monteiro, Epaminondas, Elizabeth Mie Hashimoto, Fabiane de Lima, Fernanda Rizzato, Juliana Fachini, Kuang Hongyu, Luíz Ricardo, Mariana Urbano, Pedro Cerqueira, Renata Alcarde, Ricardo Alves de Olinda, Rodrigo Pescim e Simone Sartório, pelo auxílio em momentos de dúvidas e, principalmente pela amizade.

Ao Iuri Emmanuel pela disponibilidade e auxílio na implementação computacional.

A todos os alunos do curso de Pós-graduação em Estatística e Experimentação Agronômica da ESALQ/USP, com os quais tive o prazer de conviver.

Aos meus irmãos e cunhados pelo carinho, companhia, força, compreensão e torcida em todos os momentos.

Aos meus sobrinhos Ana Paula, João Vitor, Lucas, Marcony e Maria Eduarda pelos momentos de alegria e descontração.

Aos sempre amigos Antônio Silveira, Daniel Soares, Gisely, Ilma, Janaina, José Antônio Silveira, Luciana, Maria, Simone, Tia Nau, Tia Nen, Tia Zilda, Tio Zé, Vinícius e Zeny pela sincera e sólida amizade.

Ao Teodoro Robens Freitas, pelos ensinamentos, incentivo e principalmente por acreditar em mim.

Ao amigo Udi Florião pela força e disposição em sempre me ajudar.

Aos amigos do teatro por me proporcionarem momentos de alegria e principalmente pela experiência em atuar em "Darwin e o Canto dos Canários Cegos".

E a Deus, por ter me dado tudo que eu sempre precisei para que eu pudesse tornar cada sonho realidade.

Que a força do medo que tenho
Não me impeça de ver o que anseio
Que a morte de tudo em que acredito
Não me tape os ouvidos e a boca
Que a música que ouço ao longe
Seja linda ainda que tristeza
Que as palavras que eu falo
Não sejam ouvidas como prece e
nem repetidas com fervor
Apenas respeitadas
Que a arte nos aponte uma resposta
Mesmo que ela não saiba
E que ninguém a tente complicar
Porque é preciso simplicidade pra fazê-la florescer
Porque metade de mim é platéia
E a outra metade é canção.
Trecho da música "Metade"de Oswaldo Montenegro.

SUMÁRIO

RESUMO

Modelo hierárquico bayesiano na determinação de associação entre marcadores e QTL em uma população F_2

O objetivo do mapeamento de QTL (*Quantitative Trait Loci*) é identificar sua posição no genoma, isto é, identificar em qual cromossomo está e qual sua localização nesse cromossomo, bem como estimar seus efeitos genéticos. Uma vez que as localizações dos QTL não são conhecidas *a priori*, marcadores são usados frequentemente para auxiliar no seu mapeamento. Alguns marcadores podem estar altamente ligados a um ou mais QTL e, dessa forma eles podem mostrar uma alta associação com a característica fenotípica. O efeito genético do QTL e os valores fenotípicos de uma característica quantitativa são normalmente descritos por um modelo linear. Uma vez que as localizações dos QTL não são conhecidas *a priori*, marcadores são utilizados para representá-los. Em geral, é utilizado um número grande de marcadores. Esses marcadores são utilizados no modelo linear para proceder ao processo de associação; dessa forma o modelo especificado contém um número elevado de parâmetros a serem estimados. No entanto, é esperado que muitos destes parâmetros sejam não significativos, necessitando de um tratamento especial. Na estimação bayesiana esse problema é tratado por meio da estrutura de distribuições *a priori* utilizada. Um parâmetro que é esperado assumir o valor zero (não significativo) é naturalmente especificado por meio de uma distribuição que coloque um peso maior no zero, encolhimento bayesiano. Neste trabalho é proposta a utilização de dois modelos que utilizam distribuições *a priori* de encolhimento. Um dos modelos está relacionado com o uso da distribuição *a priori* Laplace (Lasso bayesiano) e o outro com a Horseshoe (Estimador Horseshoe). Para avaliar o desempenho dos modelos na determinação da associação entre marcadores e QTL, realizou-se um estudo de simulação. Foi analisada a associação entre marcadores e QTL utilizando três características fenotípicas: produção de grãos, altura da espiga e altura da planta. Comparou-se os resultados obtidos neste trabalho com análises feitas na literatura na detecção dos marcadores associados a essas características. A implementação computacional dos algoritmos foi feita utilizando a linguagem C e executada no pacote estatístico R. O programa implementado na linguaguem C é apresentado e disponibilizado. Devido à interação entre as linguagens de programação C e R, foi possível executar o programa no ambiente R.

Palavras-chave: Lasso, Horseshoe, Associação, Marcadores, QTL

ABSTRACT

Bayesian hierarchical model in the determination of association between markers and QTL in a F_2 population

The objective of the mapping of quantitative trait loci (QTL) is to identify its position in the genome, ie, identify which chromosome is and what is its location in the chromosome, as well as to estimate their genetic effects. Since the location of QTL are not known a *priori*, markers are often used to assist in it mapping. Some markers may be closely linked to one or more QTL, and thus they may show a strong association with the phenotypic trait. The genetic effect of QTL and the phenotypic values of a quantitative trait are usually described by a linear model. Since the QTL locations are not known a *priori*, markers are used to represent them. Generally is used a large number of markers. These markers are used in the linear model to make the process of association and thus the model specified contains a large number of parameters to be estimated. However, it is expected that many of these parameters are not significant, requiring a special treatment. In Bayesian estimation this problem is treated through structure *priori* distribution used. A parameter that is expected to assume the value zero (not significant) is naturally specified by means of a distribution that put more weight at zero, bayesian shrinkage. This paper proposes the use of two models using *priori* distributions to shrinkage. One of the models is related to the use of *priori* distribution Laplace (bayesian Lasso) and the other with Horseshoe (Horseshoe Estimator). To evaluate the performance of the models to determine the association between markers and QTL, we performed a simulation study. We analyzed the association between markers and QTL using three phenotypic traits: grain yield, ear height and plant height. We compared the results obtained in this study with analyzes in the literature on the detection of markers associated with these characteristics. The computational implementation of the algorithms was done using the C language and executed the statistical package R. The program is implemented in C languages presented and made available. Due to the interaction between the programming languages C and R, it was possible execute the program in the environment R.

Keywords: Lasso, Horseshoe, Association, Markers, QTL

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Os enormes avanços tecnológicos das últimas décadas têm permitido o armazenamento de grandes volumes de informações em diversas áreas de conhecimento. A área biológica, por exemplo, está se beneficiando destes avanços de forma significativa. Do ponto de vista biológico, pode-se observar o progresso em diversas técnicas, tais como técnicas moleculares para mapeamento de locos que controlam características quantitativas, denominados de QTL (***Q****uantitative* ***T****rait* ***L****oci*). O objetivo do mapeamento de QTL é identificar sua posição no genoma, isto é, identificar em qual cromossomo está e qual sua localização nesse cromossomo, bem como estimar seus efeitos genéticos (LYNCH; WALCH, 1998). Uma vez que as localizações dos QTL não são conhecidas a *priori*, marcadores são usados frequentemente para auxiliar no seu mapeamento. Alguns marcadores podem estar altamente ligados a um ou mais QTL, e dessa forma eles podem mostrar uma alta associação com a característica fenotípica. A maioria dos marcadores, no entanto, pode não estar diretamente ligada ao QTL e, portanto, nenhuma associação será esperada entre esses marcadores e a característica fenotípica (YI; XU, 2008).

Os conhecimentos recentes da genética quantitativa, aliados à utilização de metodologias estatísticas e técnicas computacionais têm contribuído para progressos genéticos na agricultura, zootecnia, medicina etc. Um grande progresso no mapeamento de QTL surgiu com o Mapeamento por Intervalo (IM do inglês, *Interval Mapping*) proposto por Lander e Botstein (1989) para a população derivada do cruzamento entre linhagens. Daí em diante, muitos métodos têm sido desenvolvidos, incluindo a análise baseada em regressão, máxima verossimilhança e a bayesiana.

O método IM considera um par de marcas adjacentes e tenta inferir a respeito da existência de um QTL em qualquer posição entre essas duas marcas. Esse método permite obter estimativas mais precisas de efeito e posição dos QTL, além de aumentar o poder de detecção em relação à análise de marcas individuais. Nesse método, um modelo linear é usado para associar o fenótipo ao genótipo do QTL e as estimativas dos parâmetros são obtidas pelo método de máxima verossimilhança. O teste da razão de verossimilhanças, ou o LOD score, é usado para verificar a existência de ligação entre os marcadores e o QTL (LYNCH; WALSH, 1998).

Dois métodos que têm sido amplamente utilizados para mapear QTL são: (1) o Mapeamento por Intervalo Composto (CIM, do inglês *Composite-Interval Mapping*)

e (2) o Mapeamento por Intervalos Múltiplos (MIM, do inglês *Multiple-Interval Mapping*). O método CIM foi proposto, independentemente, por Jansen e Stam (1994) e Zeng (1994). Esse método combina o mapeamento por intervalo e regressão linear múltipla e foi construído com o propósito de evitar que um QTL presente em qualquer posição do genoma interfira no mapeamento do QTL no intervalo sob análise. No método CIM, os efeitos dos QTL fora do intervalo considerado são inseridos no modelo sob a forma de cofatores, sendo, assim eliminados do resíduo. A técnica de Mapeamento por Intervalos Múltiplos foi proposta por Kao e Zeng (1997) e Kao et al. (1999), sendo baseada no modelo de Cocherham (1954) para interpretação de parâmetros genéticos e no método de máxima verossimilhança para estimação de parâmetros genéticos. Com o MIM, pode-se obter um aumento na precisão e no poder de detecção de QTL, além de permitir a estimação e análise dos efeitos de interação entre QTL (epistasia). Estes métodos descritos são desenvolvidos no âmbito da máxima verossimilhança, que tem limitação no ajuste de modelo saturado de grandes dimensões. Uma alternativa para ajustar modelos com um grande número de parâmetros é a utilização da metodologia bayesiana.

A aplicação da abordagem bayesiana no mapeamento de QTL surge com o trabalho de Satagopan et al. (1996), em que técnicas de MCMC[1] são aplicadas para simultaneamente identificar múltiplos QTL e a magnitude de seus efeitos. Esses autores ajustaram modelos com diferentes números de QTL e utilizaram o fator de Bayes como critério de escolha do modelo mais adequado. Desde então, a abordagem bayesiana tem sido bastante utilizada (SATAGOPAN; YANDELL, 1998; YI; XU, 2002, MEYER et al., 2009, CHE; XU, 2010) e tem apresentado algumas vantagens quando comparada com os métodos clássicos. Uma das grandes vantagens da inferência bayesiana é a possibilidade de estimar o número de QTL conjuntamente com a estimação dos seus efeitos e posições. Satagopan e Yandell (1998), assim como (STEPHENS; FISCH, 1998; SILLANPÄÄ; ARJAS, 1998) tratam o número de QTL como uma quantidade desconhecida e usam o método MCMC com saltos reversíveis proposto por Green (1995) para realizar a sua estimação. Uma vantagem do método MCMC com saltos reversíveis está na possibilidade de alternar entre modelos com espaços paramétricos de dimensões diferentes. No entanto, a complexidade dos passos deste algoritmo aumenta a demanda computacional e isso pode interferir no desempenho do algoritmo.

[1]Métodos de Monte Carlo via Cadeias de Markov, do inglês *Markov Chain Monte Carlo.*

A inferência bayesiana tem colaborado com o processo de mapeamento de QTL, desde um modelo aditivo simples a modelos com inclusão de efeitos de epistasia. Apesar da inferência bayesiana ter facilitado o processo de estudo de mapeamento de QTL, ainda tem sido um desafio trabalhar com alguns modelos, como por exemplo os que incluem efeitos de epistasia, pois essa inclusão, implica em um aumento considerável da dimensão do espaço paramétrico, o que aumenta a complexidade da implementação dos algoritmos utilizados na obtenção das amostras de distribuição conjunta para os parâmetros de interesse. O algoritmo MCMC com saltos reversíveis tem sido empregado para estudar epistasia e tem alcançado bons resultados (YI; XU, 2002; YI et al., 2003). Dois outros métodos que têm alcançado bons resultados com modelos de grandes dimensões são o encolhimento bayesiano (bayesian shrinkage), em que uma distribuição a *priori* é atribuída para os parâmetros de regressão, os quais neste trabalho, representam os efeitos do QTL associados aos marcadores (WANG et al.,2005, XU, 2003) e a seleção de variáveis por busca estocástica (SSVS, do inglês *Stochastic Search Variable Selection*) (Yi et al., 2005).

De acordo com Che e Xu (2010), os mapeamentos por meio do Shrinkage e SSVS são mais eficientes em termos de avaliação do genoma inteiro, porque são estatisticamente fáceis de entender e também porque proporcionam uma oportunidade melhor de avaliar todo o genoma. Esses dois métodos estão relacionados com o método Lasso (*Least absolute shrinkage and selection operator*) para a análise de regressão (TIBSHIRANI, 1996). Esses métodos utilizam um algoritmo especial para reduzir os efeitos dos QTL, que não estão ligados à característica, a zero ou próximo de zero. Estimar o efeito de QTL como zero é o mesmo que excluí-lo do modelo.

Park e Casella (2008) propuseram uma formulação bayesiana para o método Lasso de Tibshrinani (1996) denominado de Lasso bayesiano. O estimador Lasso pode ser interpretado como uma estimativa da moda *a posteriori* em um contexto bayesiano, utilizando distribuição *a priori* Laplace (Exponencial Dupla) (TIBSHIRANI, 1996; PARK; CASELLA, 2008). Esse método é implementado considerando-se um modelo hierárquico e consiste na mistura do parâmetro de escala da distribuição normal com uma distribuição que seja exponencialmente distribuída. Recentemente Yi e Xu (2008) utilizaram o Lasso bayesiano para associação entre marcadores e QTL em uma população *Backcross*. De acordo com esses autores, esse método pode ser extendido para outros tipos de população. No entanto, poucos trabalhos têm explorado populações do tipo F_2. Na população

F_2 há um efeito dominante associado a cada QTL, o que implica em um aumento considerável do espaço paramétrico em relação a outras populações nas quais esse efeito não é considerado na modelagem.

Este trabalho tem por objetivo:

(i) Utilizar modelos hierárquicos bayesianos na associação entre marcadores e QTL em uma população F_2;

  (a) Explorar as distribuições Exponencial e Half-Cauchy como distribuições *a priori* para as quantidades desconhecidas no modelo;

  (b) Fazer a implementação de um programa eficiente e acessível a todos os pesquisadores para verificação da associação;

  (c) Realizar o ajuste dos modelos desenvolvidos considerando-se dados simulados e diferentes níveis de herdabilidade.

(ii) Ajustar os modelos desenvolvidos a um conjunto de dados de uma população de milho tropical.

1.1 Organização do texto

Este trabalho está dividido em 5 seções, da seguinte forma:

1) Na seção 2 são apresentados, brevemente, os conhecimentos necessários para o entendimento deste trabalho: Conceitos de inferência bayesiana, métodos computacionais, conceitos básicos de genética, o método Lasso, além de ser abordada a formulação bayesiana para o Lasso.

2) Na seção 3 são apresentadas duas propostas de modelos para associação entre marcador e QTL, que estão relacionados ao uso das distribuições *a priori* Exponencial e Half-Cauchy.

3) Na seção 4 são apresentados os resultados obtidos por meio dos Modelos I e II para os dados simulados e posteriormente para os dados de uma população de milho tropical.

4) Na seção 5 são apresentadas as considerações finais.

2 REVISÃO BIBLIOGRÁFICA

O desenvolvimento e a aplicação de técnicas para a análise de dados relacionados a QTL têm ganhado um certo destaque nas últimas décadas. Vários métodos estatísticos têm sido propostos e conseguido um avanço significativo. Esse avanço se deve em boa parte à inferência bayesiana e ao notável avanço computacional. Nesse contexto, com o intuito de estudar modelos para dados relacionados a QTL, faz-se necessária uma breve revisão de conceitos de inferência bayesiana e métodos computacionais, bem como de alguns conceitos básicos de genética.

2.1 Conceitos em inferência bayesiana

Um dos principais objetivos da estatística é fazer inferências ou predições acerca de parâmetros de interesse, por meio de métodos clássicos ou bayesianos. A inferência clássica, assim como a bayesiana, trabalha na presença de observações $\mathbf{y}$, que podem ser descritas por uma distribuição de probabilidades $f(\boldsymbol{y}|\boldsymbol{\theta})$, sendo $\boldsymbol{\theta}$ uma quantidade desconhecida e necessária para descrever a distribuição de $\mathbf{y}$. 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 $\mathbf{y}$. Segundo Paulino et al. (2003), a abordagem clássica foi adotada de forma quase unânime pelos estatísticos durante a primeira metade do século XX. No entanto, nas últimas décadas, renasce uma forte alternativa a esta abordagem: a inferência bayesiana.

A inferência bayesiana modela a incerteza relativa aos parâmetros com base em evidências *a priori*, por meio de distribuições de probabilidade conhecidas como distribuições *a priori*. A abordagem bayesiana requer duas fontes de informações para fazer inferência sobre uma quantidade desconhecida: a informação sobre ela presente nos dados expressa pela função de verossimilhança e seu conhecimento prévio modelado por meio da distribuição *a priori*, $\pi(\boldsymbol{\theta})$. Combinando-se essas duas fontes de informação e utilizando o Teorema de Bayes é obtida a distribuição *a posteriori* para $\boldsymbol{\theta}$, denotado por $\pi(\boldsymbol{\theta}|\boldsymbol{y})$:

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta})}{f(\boldsymbol{y})}, \tag{1}$$

em que

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

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

A função $f(\boldsymbol{y})$, no denominador, não depende de θ e, portanto, para a determinação da distribuição de interesse $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, representa apenas uma constante. Por essa razão a equação 1 reduz-se a :

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) = cL(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) \tag{2}$$

em que $c = \frac{1}{f(\boldsymbol{y})}$ é a constante normalizadora, ou ainda,

$$\pi(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta}|\boldsymbol{y})\pi(\boldsymbol{\theta}) \tag{3}$$

sendo que o símbolo $\propto$ denota proporcionalidade. A proporcionalidade pode ser usada porque quando se multiplica a função de verossimilhança por uma constante não se altera a informação relativa ao parâmetro $\boldsymbol{\theta}$ e, assim, a distribuição a posteriori não será alterada (BOX; TIAO, 1992). O teorema de Bayes fornece a regra para a atualização da probabilidade sobre $\boldsymbol{\theta}$, partindo de $\pi(\boldsymbol{\theta})$ e chegando a $\pi(\boldsymbol{\theta}|\boldsymbol{y})$.

2.1.1 Distribuição *a priori*

A distribuição *a priori* modela a incerteza relativa ao parâmetro antes da observação dos dados, isto é, resume a informação relativa ao parâmetro $\boldsymbol{\theta}$, desconhecido antes da realização do experimento. Tal distribuição desempenha um papel importante na inferência bayesiana, pois pode ser utilizada para representar probabilisticamente o conhecimento prévio ou ignorância relativa sobre a quantidade desconhecida $\boldsymbol{\theta}$ antes dos dados serem obtidos (BOX; TIAO, 1992). A informação *a priori* que se pretende incorporar na análise é a informação fornecida por um especialista ou pesquisador, tais como dados históricos do problema ou de experimentos análogos (PAULINO et al., 2003).

A distribuição *a priori* atribuída para um parâmetro $\boldsymbol{\theta}$ pode diferir de pesquisador para pesquisador. Há casos em que a distribuição *a priori* é informativa, refletindo algum conhecimento prévio sobre o parâmetro desconhecido. Outras vezes, a distribuição *a priori* é não informativa, permitindo que os dados "mostrem" as informações sobre $\boldsymbol{\theta}$.

2.1.2 Distribuições *a priori* conjugadas

Uma família de distribuições *a priori* é conjugada se as distribuições *a priori* e *a posteriori* pertencem à mesma classe de distribuições e, dessa forma, a atualização do

conhecimento que se tem sobre o prâmetro $\boldsymbol{\theta}$ envolve apenas uma mudança nos parâmetros indexadores da família de distribuição a *priori*, denominados hiperparâmetros,que difere dos parâmetros $\boldsymbol{\theta}$. Seja $\boldsymbol{y}^T = (y_1, ..., y_n)$ um vetor de observações de variáveis aleatórias independentes e identicamente distribuídas na família exponencial, a função de distribuição conjunta de $\boldsymbol{Y_1, ..., Y_n}$ é dada por:

$$f(\boldsymbol{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} \exp\{a(\theta)b(y_i) + c(\theta) + d(y_i)\}; \tag{4}$$

e sua função de verossimilhança por:

$$L(\boldsymbol{\theta}|y) \propto \exp\{a(\theta) \sum_{i=1}^{n} b(y_i) + nc(\theta)\}; \tag{5}$$

em que $a(\theta)$ e $c(\theta)$ são funções reais de θ; $b(y_i)$ e $d(y_i)$ são funções reais de $\mathbf{y}$. Supondo a distribuição *a priori* conjugada para $\boldsymbol{\theta}$ dada por:

$$\pi(\boldsymbol{\theta}; k_1, k_2) \propto \exp\{k_1 a(\theta) + k_2 c(\theta)\}; \tag{6}$$

obtém-se a seguinte distribuição *a posteriori*

$$\pi(\boldsymbol{\theta}|y) \propto \exp\left\{ a(\theta) \left[\sum_{i=1}^{n} b(y_i) + k_1 \right] + c(\theta)[n + k_2] \right\}. \tag{7}$$

O uso de distribuições *a priori* conjugadas é muito importante na estatística bayesiana, porém 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 prévio do parâmetro.

2.1.3 Distribuições *a priori* não informativas

A utilização de uma distribuição *a priori* não informativa implica em que a informação contida nos dados é dominante, no sentido de que o conhecimento *a priori* seja vago, ou seja, não há informações *a priori* sobre o parâmetro de interesse, ou que a informação disponível seja pouco expressiva (GELMAN et al., 2000). A princípio uma forma de atribuir distribuição *a priori* não informativa é pensar que todos os possíveis valores para um dado parâmetro tenham a mesma chance de ocorrer, isto é, com uma distribuição *a priori* uniforme, ($\pi(\theta) \propto k$). No entanto, algumas dificuldades podem surgir com esta escolha: $\pi(\theta)$ é imprópria, isto é, $\int \pi(\theta) \to \infty$. Jeffreys (1961) propôs uma classe de distribuições *a priori* não informativas invariantes, porém, eventualmente imprópria, a qual se baseia na informação de Fisher.

2.1.4 Distribuições *a priori* hierárquicas

A ideia de distribuições *a priori* hierárquicas é dividir a especificação da distribuição *a priori* em estágios. A distribuição *a priori* de um parâmetro θ depende dos valores dos hiperparâmetros, por exemplo ϕ. Nesse caso pode-se escrever $\pi(\theta|\phi)$ ao invés de $\pi(\theta)$. Além disso ao invés de fixar valores para os hiperparâmetros, pode-se especificar uma distribuição *a priori* $\pi(\phi)$, completando assim o segundo estágio na hierarquia. Dessa forma a distribuição *a priori* conjunta é simplesmente $\pi(\theta, \phi) = \pi(\theta|\phi)\pi(\phi)$ e a distribuição *a priori* marginal de θ pode ser então obtida por integração como

$$\pi(\theta) = \int \pi(\theta, \phi)d\phi = \int \pi(\theta|\phi)\pi(\phi)d\phi.$$

Nesse caso a distribuição conjunta *a posteriori* pode ser escrita na forma:

$$\pi(\theta, \phi|\boldsymbol{y}) \propto L(y|\theta, \phi)\pi(\theta|\phi)\pi(\phi) \propto L(y|\theta)\pi(\theta|\phi)\pi(\phi)$$

pois a distribuição dos dados depende somente de θ. Em outras palavras, dado θ, $\boldsymbol{y}$ e ϕ são independentes.

Teoricamente, não há limitação quanto ao número de estágios, mas devido à complexidade resultante, as distribuições *a priori* hierárquicas são especificadas em geral em dois ou três estágios. Devido à dificuldade de interpretação dos hiperparâmetros em estágios mais altos, é prática comum especificar distribuições *a priori* não informativas para estes níveis (GELMAN, 2006; EHLERS, 2005).

2.1.5 Métodos computacionais

Para se inferir em relação a qualquer parâmetro unidimensional do vetor $\boldsymbol{\theta}$, a distribuição conjunta *a posteriori* dos parâmetros (multidimensional) deve ser integrada em relação a todos os outros parâmetros que a constituem, ou seja, deve-se procurar obter a distribuição marginal de cada um dos parâmetros (PAULINO et al.,2003; BOX; TIAO, 1992). Em geral, existem dificuldades para a obtenção de uma forma analítica para a distribuição marginal. Esssas dificuldades, em geral, devem-se à complexidade das distribuições conjuntas obtidas ou devido à dimensão do parâmetro $\boldsymbol{\theta}$ em estudo. Assim, a obtenção dos momentos fica comprometida e formas alternativas são necessárias para seu cálculo. Neste trabalho serão utilizados algoritmos especiais para obtenção de uma amostra da distribuição conjunta *a posteriori*, baseados em cadeias de Markov.

2.1.6 Métodos de Monte Carlo via Cadeias de Markov

Uma Cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia, ϕ_{t+1}, depende somente do estado atual, ϕ_t e dos dados, não da história passada da cadeia (GAMERMAN, 2006). As primeiras iterações são influenciadas pelo estado inicial, ϕ_1, e são descartadas. Esse período é conhecido como aquecimento da cadeia (*burn-in*). Além disso, considera-se uma dependência entre as observações subsequentes da cadeia e, para diminuir a alta correlação existente entre os valores amostrais, deve-se considerar um espaçamento entre as iterações armazenadas, digamos k iterações, esse valor é conhecido como *thinning*.

A ideia dos métodos MCMC é obter uma amostra da distribuição conjunta dos parâmetros de interesse, por meio de um processo iterativo. Ao final de cada ciclo de atualizações, os valores gerados são considerados amostras aleatórias da distribuição de probabilidade conjunta. Os algoritmos mais utilizados para gerar amostras em inferência bayesiana são o amostrador de Gibbs (Gibbs Sampler) e o Metropolis-Hastings.

2.1.7 Gibbs-Sampler (Amostrador de Gibbs)

O amostrador de Gibbs foi inicialmente utilizado por Geman e Geman (1984) no contexto de restauração de imagens e, desde então, tem sido muito discutido e utilizado em diversas áreas de pesquisa (GELFAND; SMITH, 1990; GELFAND, 2000). Por meio do Gibbs-Sampler, no contexto da inferência bayesiana, é possível gerar amostras de uma distribuição conjunta *a posteriori* $\pi(\theta_1, \theta_2, ..., \theta_p|\boldsymbol{y})$ a partir das distribuições de cada parâmetro condicionada aos demais parâmetros do modelo, que são denominadas distribuições condicionais completas *a posteriori* $\pi(\theta_i|\theta_1, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_p, \boldsymbol{y})$, desde que essas distribuições condicionais completas possuam formas conhecidas na literatura (CASELLA; GEORGE, 1992; GELFAND, 2000).

Segundo Gamerman e Lopes (2006), o procedimento geral para execução do Amostrador de Gibbs pode ser descrito nos seguintes passos:

(1) definem-se os valores iniciais $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$ para os parâmetros;

(2) amostra-se iterativamente $\theta_1^{(1)}$ de $\pi(\theta_1|\theta_2^{(0)}, \theta_3^{(0)} ..., \theta_p^{(0)}, \boldsymbol{Y})$

$$\theta_2^{(1)} \text{ de } \pi(\theta_2|\theta_1^{(1)}, \theta_3^{(0)} ..., \theta_p^{(0)}, \boldsymbol{Y})$$

$$\theta_3^{(1)} \text{ de } \pi(\theta_3|\theta_1^{(1)}, \theta_2^{(1)}..., \theta_p^{(0)}, \boldsymbol{Y})$$

...

$$\theta_p^{(1)} \text{ de } \pi(\theta_p|\theta_1^{(1)}, \theta_2^{(1)}..., \theta_{p-1}^{(1)}, \boldsymbol{Y})$$

completa-se, assim, uma iteração do esquema e uma transição de

$\boldsymbol{\theta}^0$ para $\boldsymbol{\theta}^{(1)} = (\theta_1^{(1)}, \theta_2^{(1)}, ..., \theta_p^{(1)})$;

(3) repete-se o passo (2) t vezes, ou seja, após um grande número de iterações (t iterações),

obtém-se $\boldsymbol{\theta}^{(t)} = (\theta_1^{(t)}, \theta_2^{(t)}, ..., \theta_p^{(t)})$.

O conjunto dos t valores amostrados representa amostras da distribuição conjunta *a posteriori* de $\boldsymbol{\theta}$. A partir dessa amostra, obtém-se as estimativas pontuais (médias, medianas e moda) e estimativas por região (intervalos de credibilidade e HPD) para os parâmetros amostrados. O HPD é denominado intervalo de credibilidade de máxima densidade *a posteriori*.

2.1.8 Metropolis-Hastings

O algoritmo de Metropolis-Hasting gera uma amostra da distribuição conjunta *a posteriori* $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, a partir das distribuições condicionais completas com formas analíticas não conhecidas na literatura. Tal algoritmo usa a ideia de que um valor é gerado de uma distribuição auxiliar ou candidata e este valor é aceito com uma dada probabilidade (METROPOLIS et al., 1953; HASTINGS, 1970). O algoritmo Metropolis-Hastings está estruturado nos seguintes passos:

(1) Inicialize o contador de iterações $t = 0$ e especifique os valores iniciais

$\boldsymbol{\theta}^0 = (\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)})$;

(2) Gere um valor θ^c da distribuição proposta $q(.|\theta_1)$;

(3) Calcule a probabilidade de aceitação $\alpha(\theta_1, \theta^c)$,

$$\alpha(\theta_1, \theta^c) = \min\left\{1, \frac{\pi(\theta^c|\theta_2, ..., \theta_p)q(\theta_1|\theta^c)}{\pi(\theta_1|\theta_2, ..., \theta_p)q(\theta^c|\theta_1)}\right\};$$

(4) Gere um valor u, a partir de uma distribuição $U(0, 1)$;

(5) Se $u < \alpha$, então aceite o valor candidato e faça $\theta_1^{(t+1)} = \theta^c$.

Caso contrário, rejeite e faça $\theta_1^{(t+1)} = \theta^t$;

(6) Incremente o contador de t para $t + 1$ e volte ao passo (2) até atingir a convergência.

2.2 Aspectos gerais da genética

Para desenvolver a metodologia do mapeamento de QTL, bem como estudo de associação, é necessária a familiarização com alguns termos básicos de genética que são extremamente importantes para o desenvolvimento deste tipo de pesquisa.

Genética é a parte da Biologia que estuda a hereditariedade, ou seja, é a ciência que investiga as razões de semelhanças que se manifestam nos organismos relacionados por descendência. As bases dessa ciência apoiam-se nas experiências do botânico austríaco Gregor Mendel. Mendel iniciava seus trabalhos com um planejamento, isto é, escolhia o material, praticava experiência com ele, fazia observações e formulava hipóteses. Do resultado, tirava conclusões, generalizava e estabelecia leis.

De uma série de experiências, Mendel propôs que a existência de características (tais como a cor) das flores é devida à existência de um par de unidades elementares de hereditariedade, agora conhecidas como genes. Define-se, ainda, o loco como o local, no cromossomo,[1] em que se localiza o gene. Formas alternativas de um gene que podem ocorrer em um determinado loco são chamados de alelos.

Um indivíduo diploide tem dois alelos em um determinado loco. Quando um indivíduo possui alelos idênticos para um dado loco, diz-se que ele é homozigoto; caso apresente alelos diferentes no mesmo loco em cada cromossomo homólogo, dá-se-lhe de heterozigoto. Geralmente, emprega-se letra maiúscula (A) para indicar alelos de genes dominantes e letra minúscula (a) para indicar alelos de genes recessivos, como segue: AA - Homozigoto Dominante, aa- Homozigoto Recessivo e Aa - Heterozigoto. Um indivíduo pode ter, ainda, alelos co-dominantes. Nesse caso, permite-se identificar apenas duas classes **fenotípicas** ($A\_$ e aa).

Valor Fenotípico: no estudo de herança de caracteres quantitativos adota-se o modelo básico:

$$Y = G + E, \tag{8}$$

que define o valor fenotípico (Y), diretamente mensurado nos indivíduos, como resultado da ação do genótipo (G), ou valor genotípico, sob a influencia do ambiente (E).

Os valores Y, G e E podem ser expressos em qualquer unidade que represente

[1]Um cromossomo é uma longa sequência de DNA, que contém vários genes, e outras sequências de nucleótidos (nucleotídeos) com funções específicas nas células dos seres vivos.

uma propriedade biológica que possa ser medida de maneira contínua, tal como peso, altura, produção de grãos etc. É importante salientar que o efeito genético de um gene (de um QTL) pode ser decomposto em dois componentes: efeito genético aditivo e efeito genético de dominância. Em termos da equação 8, tem-se:

$$Y = (EA + ED) + E, \tag{9}$$

em que EA e ED são os efeitos aditivos e dominantes, respectivamente. O efeito genético aditivo é o valor fenotípico que pode ser predito linearmente por meio do número de alelos de um certo tipo que definem o genótipo, enquanto que o efeito de dominância é o valor fenotípico que não pode ser explicado linearmente (resíduo genético devido ao efeito de interação entre os alelos A e a no mesmo loco) (DUARTE, 2007).

2.3 Fundamentos teóricos do mapeamento genético

Metodologias estatísticas e técnicas computacionais têm sido bons aliados na detecção de regiões cromossômicas responsáveis por alguma variação observada em características quantitativas. Características quantitativas são aquelas cuja expressão fenotípica apresenta variações contínuas, atribuídas à segregação simultânea de muitos genes distribuídos pelo genoma, em regiões definidas como QTL (*Quantitative Trait Loci*). Com o auxílio de metodologias estatísticas pode-se realizar o mapeamento de QTL. Mapear um QTL significa identificar quantos QTL estão presentes no genoma, sua posição e estimar seus efeitos genéticos.

Segundo Silva (2001), para mapear um QTL é necessário um mapa de ligação de marcadores polimórficos que cubram todo o genoma, e que exista variabilidade para o caráter quantitativo que se deseja estudar. Apesar de toda a teoria necessária para o estudo de QTL estar disponível desde a década de 20, até meados da década de 60 obteve-se pouco sucesso na identificação de marcadores ligados a QTL, pois marcadores utilizados nesse período eram controlados por genes associados a caracteres morfológicos, em geral, fenótipos de fácil identificação visual, como nanismo, cor de pétalas etc , mas, que são marcadores pouco polimórficos, pois não cobrem todo o genoma, além de não serem seletivamente neutros, no sentido de não afetarem o caráter em questão e nem as características reprodutivas do indivíduo.

Segundo Ferreira e Grattapaglia (1998), a revolução nesse quadro se iniciou com o desenvolvimento de marcadores isoenzimáticos, o que ampliou o número de

marcadores genéticos disponíveis, e com o advento das técnicas modernas de biologia molecular, que permitiram o surgimento de diversos métodos de detecção de polimorfismo ao nível de DNA. Os marcadores genéticos obtidos com base em fenótipos moleculares são denominados marcadores moleculares. Os principais tipos de marcadores moleculares funcionam utilizando o princípio de hibridização ou amplificação de DNA. Entre os identificados por hibridização estão os marcadores RFLP (Restriction Fragment Length Polymorphism) e minissatélites ou locos VNRT (Variable Number of Tandem Repeats). Entre as técnicas que utilizam a amplificação do DNA como base estão: RAPD (Random Amplified Polymorphic DNA), SCAR (Sequence Characterized Amplified Regions), STS (Sequence Tagged Sites), AFLP (Amplified Fragment Length Polymorphism), microssatélite e SNPs (Single Nucleotide Polymorphism). Dentre todas, as técnicas de microssatélites e de SNPs são as mais utilizadas atualmente.

Uma vez que os dados de marcadores estão disponíveis, é possível construir um mapa genético. Esse mapa genético consiste na representação do cromossomo sobre o qual os marcadores e QTL 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 da distância entre eles (LYNCH; WALSH, 1998).

2.4 Seleção de mapeamento

Segundo Silva (2006), a seleção da população de mapeamento envolve a escolha de genitores e a determinação do tipo de cruzamento, sendo essa considerada uma etapa crítica para o sucesso da construção do mapa. Além disso, tal seleção influencia diretamente no mapeamento de QTL, já que esse processo depende dos marcadores genitores espalhados por todo genoma. Para realizar o mapeamento, vários tipos de populações podem ser utilizadas, sendo que as populações mais frequentemente utilizadas são as derivadas de cruzamentos entre linhagens homozigóticas fenotipicamente constrastantes (F_1), tais como as gerações segregantes F_2 e retrocruzamento (LYNCH; WALSCH, 1998).

(i) **População F_2:** é obtida por autofecundação das plantas F_1 resultantes do cruzamento entre genitores homozigóticos. Nessa população, os marcadores co-dominantes segregam na proporção $1 : 2 : 1$ ($AA : Aa : aa$) e os dominantes, na proporção $3 : 1$ ($A\_, aa$). Essa população possui algumas vantagens, dentre as quais a facilidade e rapidez de sua obtenção. No mapeamento de QTL estão disponíveis as in-

formações dos três genótipos (AA, Aa, aa), o que aumenta a precisão da estimativa dos componentes genéticos (Schuster; Cruz, 2004).

(ii) **População de retrocruzamento:** é obtida pelo cruzamento de plantas de população F_1 com um dos genitores. Apenas dois genótipos são obtidos, heterozigotos (Aa) e homozigoto (aa). A rapidez na obtenção da população é uma das grandes vantagens de se trabalhar com população de retrocruzamento.

2.5 Análise de múltiplos marcadores

Análise de marcador é uma espécie de estudo de associação em que cada marcador é considerado um candidato a QTL. Se o mapa dos marcadores é suficientemente denso, a maioria dos QTL são potencialmente detectáveis por causa da ligação estreita com os marcadores. Como cada marcador é considerado um candidato a QTL, todos eles são incluídos na análise de QTL e o modelo pode ser supersaturado. Considerando o fato de que alguns marcadores podem estar intimamente ligados a um ou mais QTL, e que a maioria deles pode não estar diretamente ligado ao QTL, espera-se que muitos dos parâmetros do modelo sejam não significativos, necessitando assim de um procedimento de seleção de variáveis para incluir e excluir marcadores no modelo (SILLANPÄÄ; ARJAS, 1998; YI et al. 2003,2005; YI, 2004; YI; SHRINER, 2008).

Uma abordagem alternativa é a utilização de um método de encolhimento (*shrinkage*) que inclui todas as variáveis no modelo e usa distribuição *a priori* informativa para reduzir os efeitos triviais a zero. O Lasso (*Least absolute shrinkage and selection operator*) é um método de encolhimento que tem sido amplamente usado em análise de regressão para modelos de grandes dimensões (TIBSHIRANI, 1996) e será apresentado na subseção a seguir.

2.5.1 Encolhimento bayesiano (bayesian shrinkage)

O efeito genético do QTL e os valores fenotípicos de uma característica quantitativa são normalmente descritos por um modelo linear. Uma vez que as localizações dos QTL não são conhecidas *a priori*, marcadores são utilizados para representá-los. Em geral, é utilizado um número grande de marcadores. Esses marcadores são utilizados no modelo linear para proceder ao processo de associação e dessa forma o modelo especificado contém um número elevado de parâmetros a serem estimados. No entanto, é esperado que

muitos desses parâmetros sejam não significativos, o que gera a necessidade de um tratamento especial. Na estimação bayesiana esse problema é tratado por meio da estrutura de distribuições *a priori* utilizada. Um parâmetro que é esperado assumir o valor zero (não significativo) é naturalmente especificado por meio de uma distribuição que coloque um peso maior no zero, encolhimento bayesiano. A seguir serão descritos dois métodos que utilizam distribuições *a priori* de encolhimento. Um dos métodos está relacionado com o uso da distribuição *a prior* Laplace (Lasso bayesiano) e o outro com a *Horseshoe* (Estimador Horseshoe).

A estimação utilizando a distribuição *a priori* Laplace, também conhecida como exponencial dupla, está relacionada com o método de estimação conhecido como Lasso (*Least absolute shrinkage and selection operator*), proposto por Tibshirani (1996). De acordo com Park e Casella (2008), o Lasso é normalmente usado para estimar os parâmetros de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ no modelo

$$y_i = \mu + \sum_{j=1}^{p} x_{ij}\beta_j + e_i = \mu + X_i\boldsymbol{\beta} + e_i, \tag{10}$$

em que, y_i é o valor fenotípico do i-ésimo indivíduo ($i = 1, 2, ..., n$); μ é um parâmetro comum a todas as observações; x_{ij} denota o genótipo do marcador j (j = 1,2,...,p) do indivíduo i; o coeficiente β_j representa o efeito do QTL associado ao marcador j; e_i é o erro residual. Assume-se que $e_i \sim N(0, \sigma^2)$, $X_i = (x_{i1}, x_{i2}, ..., x_{ip})$ e $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$.

Sendo $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$, a estimativa Lasso $(\hat{\mu}, \hat{\boldsymbol{\beta}})$ é definida por :

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg\min \left\{ \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2 \right\} \quad \text{sujeito à restrição que} \quad \sum_{j=1}^{p} |\beta_j| \leq t \tag{11}$$

em que $t \geq 0$.

O método de estimação Lasso baseado na minimização da soma de quadrados residuais, como descrito na equação 11, considera que a soma dos valores absolutos dos coeficientes de regressão seja $\sum_{j=1}^{p} |\beta_j| \leq t$ para $t \geq 0$. Essa restrição permite que algumas estimativas dos coeficientes de regressão seja exatamente zero, realizando simultaneamente um procedimento de encolhimento e seleção de modelos.

Segundo Yi e Xu (2008), a estimativa do Lasso dos coeficientes pode ser computada eficientemente utilizando o algoritmo LARS de Efron et al. (2004). O estimador Lasso pode ser interpretado como uma estimativa da moda *a posteriori* quando os

parâmetros de regressão têm distribuições *a priori* Laplace independentes (TIBSHIRANI, 1996; PARK; CASELLA, 2008).

Uma variável aleatória, Y, contínua possui distribuição de Laplace ou Exponencial Dupla, $Y \sim ED(\mu, \lambda)$, se sua função densidade de probabilidade é dada por

$$f(y; \mu, \lambda) = \frac{\lambda}{2} e^{-\lambda|y-\mu|}, \quad -\infty < y < \infty \quad , -\infty < \mu < \infty, \lambda > 0 \tag{12}$$

em que μ é o parâmetro de locação e λ o parâmetro de escala. Encontra-se na Figura 1 os gráficos da distribuição exponencial dupla, com diferentes especificações para seus parâmetros, mostrando que o peso zero é muito maior do que, por exemplo distribuições *a priori* Gaussiana com mesmo fator de escala e o gráfico da distribuição normal padrão.

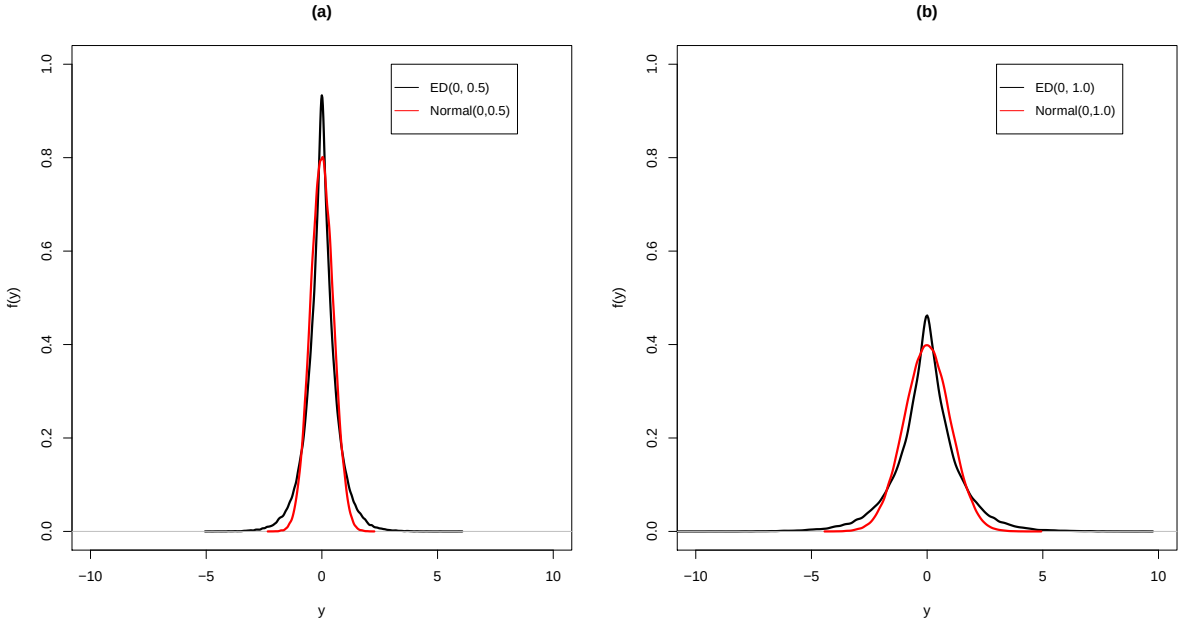

Figura 1 - Comparação entre as distribuições Normal e Exponencial Dupla com parâmetros diferentes

Nessas figuras pode-se observar como a distribuição ED pode ser utilizada para descrever a crença de que um parâmetro pode não ter efeito significativo. A distribuição Exponencial Dupla tem sido utilizada em diversas áreas, como, por exemplo, na modelagem de variáveis financeiras (movimento Browniano de Laplace) (JOHNSON; KOTZ, 1970). Porém há algumas vantagens e dificuldade em se trabalhar com essa distribuição. Dentre as vantagens, é que ela pode ser expressa como uma mistura na escala

de distribuições normais com variâncias que seguem distribuições exponenciais independentes (ANDREWS; MALLOWS, 1974), isto é:

$$\frac{\lambda}{2}e^{-\lambda|\beta_j|} = \int_0^\infty \left[\frac{\exp\left(-\frac{\beta_j^2}{2\tau_j^2}\right)}{\sqrt{2\pi\tau_j^2}}\right]\left[\frac{\lambda^2}{2}\exp\left(-\frac{\lambda^2}{2}\tau_j^2\right)\right]\mathrm{d}\tau_j^2, \tag{13}$$

em que β_j é o efeito desconhecido associado ao j-ésimo marcador e τ_j^2 é a variância associada a β_j, à qual será atribuída uma distribuição *a priori*. Na equação 13 tem-se a variável aleatória $\beta_j|\tau_j^2 \sim N(0, \sigma_j^2)$ combinada com $\tau_j^2 \sim Exp(\frac{\lambda^2}{2})$. Motivados com a possibilidade de trabalhar com essa variação da distribuição exponencial dupla, vários autores propuseram usá-la como distribuição a *priori* em modelos hierárquicos (FIGUEIREDO, 2003; BAE; MALLICK, 2004; YUAN; LIN, 2005; PARK; CASELLA, 2008; YI; XU, 2008).

2.5.2 Formulação hierárquica do modelo Lasso bayesiano e Horseshoe

2.5.2.1 Lasso bayesiano

Recentemente, Park e Casella (2008) propuseram a utilização do amostrador de Gibbs para o Lasso, para a obtenção das estimativas dos parâmetros de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ do modelo 10. Especificamente, eles consideraram uma análise bayesiana utilizando a distribuição a *priori* Laplace para $\boldsymbol{\beta}|\sigma^2$ e uma distribuição a *priori* não informativa para σ^2, a saber,

$$\begin{aligned} \pi(\boldsymbol{\beta}|\sigma^2) &= \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}} \\ \pi(\sigma^2) &= 1/\sigma^2 \end{aligned} \tag{14}$$

De acordo com esses autores, a especificação condicional a σ^2 na expressão 14 é importante, porque garante que a distribuição conjunta $\pi(\beta, \sigma^2)$ seja unimodal. A falta de unimodalidade retarda a convergência do amostrador de Gibbs e produz estimativas pontuais menos significativas.

O amostrador de Gibbs para o Lasso bayesiano explora a representação da distribuição de Laplace da expressão 13. Baseado nessa expressão, Park e Casella (2008),

propuseram a seguinte estrutura hierárquica para o modelo:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n(\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_n), \\
\pi(\boldsymbol{\beta}|\sigma^{^2}, \tau_1^2, ..., \tau_p^2) &\sim N_p(0_p, \sigma^2 \boldsymbol{D}_{\boldsymbol{\tau}}), \\
\boldsymbol{D}_{\boldsymbol{\tau}} &= diag(\tau_1^2, ..., \tau_p^2), \\
\pi(\sigma^2, \tau_1^2, ..., \tau_p^2) &= \pi(\sigma^2) d\sigma^2 \prod_{j=1}^{p} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} d\tau_j^2,
\end{aligned}$$

em que $\sigma^2, \tau_1^2, ..., \tau_p^2 > 0$ e para a constante μ uma distribuição a *priori* não informativa. Os cálculos para se obter a distribuição de Laplace especificada na expressão 14, por meio da equação 13, encontra-se no APÊNDICE A.

2.5.2.2 Horseshoe

Carvalho e Polson (2010) propuseram o estimador Horseshoe para a obtenção das estimativas dos parâmetros de regressão $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_p)^T$ do modelo 10, que consiste na mistura do parâmetro de escala da distribuição normal com uma distribuição Half-Cauchy. O estimador Horseshoe é um excelente método de estimação, bem como seleção de preditores no modelo proposto (CARVALHO; POLSON, 2010). A estimação dos parâmetros $\beta_j^{'}s$ usando misturas de distribuições se dá da seguinte forma:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N_n(\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_n), \\
\beta_j|\tau_j^2 &\sim N(0, \tau_j^2), \\
\tau_j^2|\lambda &\sim \mathrm{C}^+(0, \lambda), \\
\lambda &\sim \mathrm{C}^+(0, \phi),
\end{aligned} \tag{15}$$

em que $\mathrm{C}^+(0, \lambda)$ e $\mathrm{C}^+(0, \phi)$ são distribuições Half-Cauchy padrão nos reais positivos com hiperparâmetros de escala λ e ϕ. Carvalho et al. (2010) denominaram a estrutura hierárquica da expressão 15 de *priori* Horseshoe. A *priori* Horseshoe tem duas características interessantes que a torna particularmente útil como uma distribuição *a priori* de encolhimento para problemas de sinais esparsos, ou seja, muitos das quantidades que estão sendo estimadas são nulas ou próximas de zero. Sua cauda, como a de Cauchy, permite que os sinais fortes, ou marcadores com associações continuem a ser grandes (isto é, não são retraídos, encolhidos) e, no entanto, seu ponto extremamente concentrado na origem prevê retração severa para zero de alguns marcadores.

3 MATERIAL E MÉTODOS

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

3.1 Material

3.1.1 Dados simulados

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

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

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

Ao longo de cada mapa simulado foram inseridos cinco QTL: um, nos cromossomos 1 e 3 e três, no cromossomo 5, com efeitos aditivo e dominante e posições descritos nas Tabelas 2 e 3. Os tamanhos amostrais considerados foram: n= 200, 400 e 800 indivíduos. Para os cenários foram considerados dois níveis de herdabilidade global para a característica

fenotípica, baixa e alta. Combinando-se os diferentes níveis de herdabilidade, números de marcadores e números de indivíduos, tem-se na Tabela 1 os seguintes cenários:

Tabela 1 - Cenários considerados no estudo de simulação

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

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

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

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

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

3.1.2 Dados de uma população de milho tropical

Os dados utilizados nesta subseção estão apresentados em detalhes em Sibov et al. (2003a, 2003b) e Sabadin et al. (2008). Resumidamente, foi obtida uma população entre o cruzamento das linhagens endogâmicas L-08-05F e L-14-4B, que possuem comportamentos contrastantes para a produção de grãos. Do cruzamento das linhagens endogâmicas, obteve-se a progênie F_1, sendo que a partir de quatro plantas foram geradas 400 plantas F_2, das quais foram obtidas 400 progênies $F_{2:3}$, 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 em um 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 Espiga (AE) em cm e Altura da Planta (AP) em cm. O mapa de ligação usado para a localização dos QTL contém 117 locos de marcadores do tipo microssatélites distribuídos em dez grupos de ligação. O comprimento do mapa foi de aproximadamente 1634 cM.

3.2 Métodos

3.2.1 Modelo linear

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

$$y_i = \mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j + e_i = \boldsymbol{\mu} + \boldsymbol{X_i\beta} + \boldsymbol{e}, \tag{16}$$

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

modelo epistático de Cockerham (KAO; ZENG, 2002) e são dadas por:

$$x_{ij1} = z_{ij} - 1 \text{ e } x_{ij2} = (1 + x_{ij1})(1 - x_{ij1}) - 0,5 \tag{17}$$

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

$$x_{ij1} = \begin{cases} +1 & \text{para} \quad AA; \\ 0 & \text{para} \quad Aa; \\ -1 & \text{para} \quad aa; \end{cases} \quad \text{e} \quad x_{ij2} = \begin{cases} -1/2 & \text{para} \quad AA; \\ 1/2 & \text{para} \quad Aa; \\ -1/2 & \text{para} \quad aa. \end{cases} \tag{18}$$

Na abordagem bayesiana (paramétrica) para a construção da distribuição conjunta a *posteriori* para os parâmetros, é necessário: (1) obter a função de verossimilhança; (2) incorporar a incerteza relativa dos parâmetros a serem estimados. Os parâmetros necessários para a descrição do modelo são: μ, σ^2 e os vetores dos coeficientes de regressão $\boldsymbol{\alpha} = \{\alpha_j\}$ e $\boldsymbol{\delta} = \{\delta_j\}$ de dimensão p.

3.3 Função de verossimilhança

Considerando o modelo descrito pela equação 16, a parametrização de Cockerham dada em 17 e x_{ij1} e x_{ij2} assumindo os valores dados em 18, e assumindo que cada observação Y_i tem distribuição $Y_i \sim N(\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j, \sigma^2)$, a função de verossimilhança dos parâmetros é dada por

$$L(\boldsymbol{\theta}|y) \;=\; \prod_{i=1}^{n}\left\{ \frac{1}{\sqrt{2\pi\sigma^2}} \texttt{exp}\left\{ -\frac{1}{2\sigma^2}[y_i - (\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j)]^2 \right\} \right\} \tag{19}$$

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

O modelo especificado pela equação 16 contém um número elevado de parâmetros quando todos os marcadores são incluídos na análise de QTL (modelo supersaturado). Espera-se, no entanto, que muitos destes parâmetros sejam não significativos, necessitando de um tratamento especial. Utilizando a abordagem bayesiana, essa informação importante poderá ser incorporada no ajuste do modelo por meio de distribuições *a priori* apropriadas para os parâmetros. Sendo assim, especifica-se uma distribuição *a priori* que descreva a incerteza que todos os parâmetros relativos a regressão sejam não significativos no modelo e deixe aos dados a "incumbência" de escolher quais dos coeficientes

de regressão são necessários para descrever a característica fenotípica. Ressalta-se que a forma da distribuição é importante na especificação da distribuição *a priori* adotada. Um parâmetro não significativo é, naturalmente, especificado através de uma distribuição que coloque um peso maior no zero com muita precisão. Essa especificação é conhecida como encolhimento bayesiano e é proveniente da denominação bayesian shrinkage.

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

3.3.1 Modelo I

A especificação completa do Modelo I é feita considerando o modelo descrito pela equação 16, assumindo-se independência entre os parâmetros do modelo e as seguintes distribuições *a priori* para os parâmetros:

$$(\text{ i }) \ \mu \sim N(0, \sigma_u^2);$$

$$(\text{ ii }) \ \sigma^2 \sim \text{GI}(a, b);$$

$$(\text{iii }) \ \alpha_j \sim \text{ED}(0, \lambda), \ j = 1, \ldots, p;$$

$$(\text{ iv }) \ \delta_j \sim \text{ED}(0, \lambda_1), \ j = 1, \ldots, p;$$

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

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

em que $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_p)^T$ e $\boldsymbol{\delta} = (\delta_1, ..., \delta_p)^T$.

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

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

$$\boldsymbol{\alpha}|\upsilon_j^2 \sim \prod_{j=1}^{p} N(\alpha_j|0,\upsilon_j^2) \tag{20}$$

e

$$\boldsymbol{\delta}|\tau_j^2 \sim \prod_{j=1}^{p} N(\delta_j|0,\tau_j^2). \tag{21}$$

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

$$\boldsymbol{\upsilon^2}|\lambda^2 \sim \prod_{j=1}^{p} \text{Exp}(\upsilon_j^2|\lambda^2) \tag{22}$$

e

$$\boldsymbol{\tau^2}|\lambda_1^2 \sim \prod_{j=1}^{p} \text{Exp}(\tau_j^2|\lambda_1^2). \tag{23}$$

Em 22 e 23 as expressões $\text{Exp}(\upsilon_j^2|\lambda^2)$ e $\text{Exp}(\tau_j^2|\lambda_1^2)$ são densidades exponenciais, $\boldsymbol{\upsilon^2} = (\upsilon_1^2,...,\upsilon_p^2)$ e $\boldsymbol{\tau^2} = (\tau_1^2,...,\tau_p^2)$. Ao invés de fixar um valor para os hiperparâmetros λ^2 e λ_1^2, são atribuídas distribuições *a priori* para esses hiperparâmetros e eles serão estimados juntamente com os outros parâmetros do modelo. As distribuições *a priori* atribuídas a esses dois parâmetros foram respectivamente $\text{Gama}(a_1,b_1)$ e $\text{Gama}(a_2,b_2)$ com $a_1 > 0$, $a_2 > 0$, $b_1 > 0$ e $b_2 > 0$.

Dado o modelo 16 e as distribuições *a priori* para os parâmetros de interesse, a distribuição conjunta *a posteriori* é dada por:

$$\begin{aligned}
\pi(\theta|\mathbf{y}) \propto & \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\texttt{exp}\left\{-\frac{1}{2\sigma^2}[y_i-(\mu+\sum_{j=1}^{p}x_{ij1}\alpha_j+\sum_{j=1}^{p}x_{ij2}\delta_j)]^2\right\}\right. \\
& \times \frac{1}{\sqrt{2\pi\sigma_\mu^2}}\texttt{exp}\left\{-\frac{1}{2\sigma_\mu^2}\mu^2\right\} \\
& \times (\sigma^2)^{-(1+a)}\texttt{exp}\left\{-\frac{b}{\sigma^2}\right\} \times \prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi\upsilon_j^2}}\texttt{exp}\left\{-\frac{1}{2\upsilon_j^2}\alpha_j^2\right\}\right. \\
& \left.\left. \times \frac{\lambda^2}{2}\texttt{exp}\left\{\frac{\lambda^2\upsilon_j^2}{2}\right\} \times \frac{1}{\sqrt{2\pi\tau_j^2}}\texttt{exp}\left\{-\frac{1}{2\tau_j^2}\delta_j^2\right\} \times \frac{\lambda_1^2}{2}\texttt{exp}\left\{\frac{\lambda_1^2\tau_j^2}{2}\right\}\right\}\right. \\
& \times (\lambda^2)^{a_1-1}\texttt{exp}\{-b_1\lambda^2\} \times (\lambda_1^2)^{a_2-1}\texttt{exp}\{-b_2\lambda_1^2\}
\end{aligned} \tag{24}$$

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

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

(a) distribuição condicional completa *a posteriori* para μ:

$$\mu|\sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{n} S_i}{\sigma^2\left(\frac{n}{\sigma^2} + \frac{1}{\sigma_\mu^2}\right)}, \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\sigma_\mu^2}}\right), \tag{25}$$

em que $S_i = y_i - \sum_{j=1}^{p} x_{ij1}\alpha_j - \sum_{j=1}^{p} x_{ij2}\delta_j$.

(b) distribuição condicional completa *a posteriori* para σ^2:

$$\sigma^2|\mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim GI\left(\delta + \frac{n}{2}, \frac{1}{\frac{\sum_{i=1}^{n} d_i}{2} + \frac{1}{\gamma}}\right), \tag{26}$$

em que $d_i = \left[y_i - \left(\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j\right)\right]^2$.

(c) distribuição condicional completa *a posteriori* para α_j:

$$\alpha_j|\mu, \sigma^2, \boldsymbol{\alpha_{j^-}}, \boldsymbol{\delta}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{n} x_{ij1}h_i}{\sum_{i=1}^{n} x_{ij1}^2 + V_j}, \frac{\sigma^2}{\sum_{i=1}^{n} x_{ij1}^2 + V_j}\right), \tag{27}$$

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

(d) distribuição condicional completa *a posteriori* para δ_j:

$$\delta_j|\mu, \sigma^2, \boldsymbol{\alpha}, \boldsymbol{\delta_{j^-}}, \boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \lambda^2, \lambda_1^2, \mathbf{y} \sim N\left(\frac{\sum_{i=1}^{n} x_{ij2}k_i}{\sum_{i=1}^{n} x_{ij2}^2 + U_j}, \frac{\sigma^2}{\sum_{i=1}^{n} x_{ij2}^2 + U_j}\right), \tag{28}$$

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

(e) distribuição condicional completa *a posteriori* para υ_j^2:

$$\upsilon_j^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2_{j-}},\boldsymbol{\tau^2},\lambda^2,\lambda_1^2,\mathbf{y}\sim \text{InvGauss}\left(\sqrt{\frac{\lambda}{\alpha_j^2}},\lambda^2\right), \tag{29}$$

em que $\boldsymbol{\upsilon^2_{j-}}$ representa todos os elementos de $\boldsymbol{\upsilon^2}$ exceto υ_j^2 e InvGauss é a notação para Inverse Gaussian (Inversa Gaussiana).

(f) distribuição condicional completa *a posteriori* para τ_j^2:

$$\tau_j^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2},\boldsymbol{\tau^2_{j-}},\lambda^2,\lambda_1^2,\mathbf{y}\sim \text{InvGauss}\left(\sqrt{\frac{\lambda_1}{\delta_j^2}},\lambda_1^2\right), \tag{30}$$

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

(g) distribuição condicional completa *a posteriori* para λ^2:

$$\lambda^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2},\boldsymbol{\tau^2_j},\lambda_1^2,\mathbf{y}\sim \text{G}\left(p+a_1,\sum_{j=1}^{p}\frac{\upsilon_j^2}{2}+b_1\right). \tag{31}$$

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

$$\lambda_1^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2},\boldsymbol{\tau^2_j},\lambda^2,\mathbf{y}\sim \text{G}\left(p+a_2,\sum_{j=1}^{p}\frac{\tau_j^2}{2}+b_2\right), \tag{32}$$

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

3.3.2 Modelo II

A especificação completa do Modelo II é feita considerando o modelo descrito pela equação 16, assumindo-se independência entre os parâmetros do modelo e as seguintes distribuições *a priori* para os parâmetros:

$$\text{(i)}\ \ \alpha_j|\upsilon_j^2\sim N(0,\upsilon_j^2);$$

$$\text{(ii)}\ \ \upsilon_j^2|\lambda_j\sim \text{C}^+(0,\lambda_j);$$

$$\text{(iii)}\ \ \lambda_j\sim \text{C}^+(0,\phi);$$

$$\text{(iv)}\ \ \delta_j|\tau_j^2\sim N(0,\tau_j^2);$$

$$\text{(v)}\ \ \tau_j^2|\lambda_{1_j}\sim \text{C}^+(0,\lambda_{1_j});$$

(vi) $\lambda_{1_j} \sim \mathrm{C}^+(0, \phi_1)$;

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

Dado o modelo 16 e as distribuições a *priori* para os parâmetros de interesse, a distribuição conjunta *a posteriori* é dada por:

$$\begin{aligned}
\pi(\theta|\mathbf{y}) \propto & \prod_{i=1}^{n}\left\{\frac{1}{\sqrt{2\pi\sigma^2}}\mathtt{exp}\left\{-\frac{1}{2\sigma^2}[y_i - (\mu + \sum_{j=1}^{p} x_{ij1}\alpha_j + \sum_{j=1}^{p} x_{ij2}\delta_j)]^2\right\}\right. \\
& \times \frac{1}{\sqrt{2\pi\sigma_\mu^2}}\mathtt{exp}\left\{-\frac{1}{2\sigma_\mu^2}\mu^2\right\} \\
& \times (\sigma^2)^{-s+1}\mathtt{exp}\left\{-\frac{t}{\sigma^2}\right\} \times \prod_{j=1}^{p}\left\{\frac{1}{\sqrt{2\pi\upsilon_j^2}}\mathtt{exp}\left\{-\frac{1}{2\upsilon_j^2}\alpha_j^2\right\}\right. \\
& \times \frac{2}{\pi\lambda_j\left[1 + \left(\frac{\upsilon_j^2}{\lambda_j}\right)^2\right]} \times \frac{1}{\sqrt{2\pi\tau_j^2}}\mathtt{exp}\left\{-\frac{1}{2\tau_j^2}\delta_j^2\right\} \\
& \left.\left. \times \frac{2}{\pi\lambda_{1_j}\left[1 + \left(\frac{\tau_j^2}{\lambda_{1_j}}\right)^2\right]} \times \frac{2}{\pi\phi\left[1 + \left(\frac{\lambda_j}{\phi}\right)^2\right]} \times \frac{2}{\pi\phi_1\left[1 + \left(\frac{\lambda_{1_j}}{\phi_1}\right)^2\right]}\right\}\right.
\end{aligned} \tag{33}$$

Visto que a obtenção das distribuições marginais *a posteriori* para os parâmetros de interesse é analiticamente intratável, os algoritmos Metropolis-Hastings e Gibbs Sampling foram utilizados para a obtenção de uma amostra da distribuição conjunta *a posteriori* e a partir dessa amostra é possível fazer inferências sobre os parâmetros de interesse, ou seja, resumos tais como média, mediana, intervalos de credibilidade para os efeitos aditivos, dominâncias e os demais parâmetros.

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

(i) Atualização de $\boldsymbol{\upsilon^2}, \boldsymbol{\tau^2}, \boldsymbol{\lambda} = (\lambda_1, \ldots \lambda_p)$ e $\boldsymbol{\lambda_1} = (\lambda_{1_1}, \ldots, \lambda_{1_p})$: as distribuições condicionais completas *a posteriori* de cada um dos parâmetros $\upsilon_j^2, \tau_j^2, \lambda_j$ e λ_{1_j} foram obtidas a partir da expressão da distribuição conjunta 33 e são apresentadas a seguir.

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

$$\upsilon_j^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2_{j^-}},\boldsymbol{\tau^2},\boldsymbol{\lambda},\boldsymbol{\lambda_1},\mathbf{y} \propto \frac{1}{\sqrt{\upsilon_j^2\left[\lambda_j^2+(\upsilon_j^2)^2\right]}} \times \exp\left\{-\frac{1}{2\upsilon_j^2}\alpha_j^2\right\}, \tag{34}$$

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

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

$$\tau_j^2|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2},\boldsymbol{\tau^2_{j^-}},\boldsymbol{\lambda},\boldsymbol{\lambda_1},\mathbf{y} \propto \frac{1}{\sqrt{\tau_j^2\left[\lambda_{1_j}^2+(\tau_j^2)^2\right]}} \times \exp\left\{-\frac{1}{2\tau_j^2}\delta_j^2\right\}, \tag{35}$$

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

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

$$\lambda_j|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2},\boldsymbol{\tau^2_j},\lambda_{j^-},\lambda_{1_j},\mathbf{y} \propto \frac{1}{(\phi^2+\lambda_j^2)} \times \frac{\lambda_j}{\lambda_j^2+(\upsilon_j^2)^2} \tag{36}$$

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

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

$$\lambda_{1_j}|\mu,\sigma^2,\boldsymbol{\alpha},\boldsymbol{\delta},\boldsymbol{\upsilon^2},\boldsymbol{\tau^2_j},\lambda_j,\lambda_{1_{j^-}},\mathbf{y} \propto \frac{1}{(\phi_1^2+\lambda_{1_j}^2)} \times \frac{\lambda_{1_j}}{\lambda_{1_j}^2+(\tau_j^2)^2} \tag{37}$$

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

Por meio das distribuições conjuntas *a posteriori* 24 e 33 é possível obter os resumos *a posteriori* para todos os parâmetros de interesse, tais como: média, mediana e intervalos de credibilidade. Utilizando as amostras *a posteriori* de α_j e δ_j e os valores observados x_{ij1} e x_{ij2}, pode-se calcular a proporção da variância fenotípica explicada por cada efeito (herdabilidade), $h_j^2 = v_j\alpha_j^2/\sigma_y^2$ (aditivo) e $H_j^2 = V_j\delta_j^2/\sigma_y^2$ (dominante), em que σ_y^2 é a variância de Y e v_j e V_j são as variâncias da amostra de $(x_{ij1}; i = 1, ..., n)$ e $(x_{ij2}; i = 1, ..., n)$, respectivamente (Yi e Xu, 2008). Intervalos de credibilidade foram considerados na avaliação da significância do efeito associado a cada marcador.

4 RESULTADOS E DISCUSSÃO

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

Para a análise dos dados, foram utilizados os Modelos I e II propostos. Para o ajuste desses modelos foi implementado um programa (APÊNDICE C) utilizando a linguagem C e executado no pacote estatístico R (R DEVELOPMENT CORE TEAM, 2011). A construção da amostra da distribuição conjunta *a posteriori* para os parâmetros foi feita utilizando métodos MCMC, mais especificamente, os algoritmos Gibbs Sampler e Metropolis-Hastings. Para as estimativas dos parâmetros dos modelos foram utilizados os mesmos procedimentos em todas as análises realizadas, ou seja, foram geradas cadeias com 60.000 iterações, as primeiras 20.000 iterações foram descartadas (*burn-in*) e um espaçamento (*thin*) de tamanho 27 iterações foi considerado com a finalidade de diminuir a correlação existente entre os valores amostrados, gerando assim uma amostra final de tamanho 1.482.

As cadeias resultantes dos algoritmos Gibbs Sampler e Metropolis-Hastings necessitam ter sua convergência diagnosticada. Os métodos utilizados para diagnosticar a convergência neste trabalho foram: Gelman e Rubin (1992) e Geweke (1992). Os critérios adotados para verificar a convergência nos dados simulados e de milho tropical foram diferentes, sendo o Geweke adotado para os dados simulados e o do Gelman e Rubin para os dados de milho tropical, pois para a aplicação do teste de Gelman e Rubin (1992) é necessária a construção de pelo menos duas cadeias para cada quantidade desconhecida no modelo. Esse fato torna a implementação computacional mais demorada. Como foram realizadas 1000 simulações, optou-se por utilizar na análise dos dados simulados o critério de Geweke (1992). Esses critérios são implementados no pacote BOA (*Bayesian Output Analysis*) do programa estatístico R. Em algumas das análises, verificou-se que a convergência para pelo menos um dos parâmetros não foi diagnosticada.

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

Tabela 4 - Número de parâmetros estimados para cada modelo

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

4.1 Dados simulados

Para cada um dos cenários apresentados na seção 3 (página 42) foram gerados 50 conjuntos de dados que foram analisados utilizando-se os Modelos I e II, produzindo assim um total de 1000 análises. O tempo computacional para cada análise em um computador com a configuracão: Intel (R) Core (TM) i7-2630 QM CPU @ 2.00 GHz e 6 GB de RAM, variou de 15 minutos para o cenário 1 a 2 horas e 45 minutos para o Cenário 10. Embora o número de parâmetros a serem estimados pelo Modelo II seja maior que o Modelo I, não houve uma diferença significativa entre o tempo computacional dos dois modelos. Isso se deve, provavelmente a linguagem de programação utilizada. Os parâmetros υ_j^2, τ_j^2, λ_j e λ_{1j} foram atualizados no Modelo II utilizando o algoritmo Metropolis-Hastings e o Gibbs Sampler no Modelo I. Os demais parâmetros em ambos os modelos foram atualizados utilizando-se o Gibbs Sampler. A taxa de aceitação para os parâmetros em que foi utilizado o Metropolis-Hastings variou de 30% a 46%. Os resultados obtidos utilizando os dois modelos são apresentados nas subseções seguintes por meio de tabelas e representações gráficas.

4.1.1 Herdabilidade baixa

Dentre os conjuntos analisados, escolheu-se um para ilustrar os resultados das estimativas e convergências dos parâmetros. O conjunto escolhido possui 400 indivíduos e 200 marcadores. Nota-se na Tabela 4 que os modelos possuem um número elevado de parâmetros a serem estimados. Devido a grande quantidade de parâmetros, optou-se por apresentar as representações gráficas, tais como histograma, trajetória das cadeias e função de autocorrelação de apenas alguns dos parâmetros envolvidos no modelo.

Essa escolha foi feita aleatoriamente, com exceção da constante do modelo μ e σ^2.

Um resumo gráfico das distribuições marginais *a posteriori* dos parâmetros, baseado no método MCMC, é apresentado nas Figuras 2 a 7.

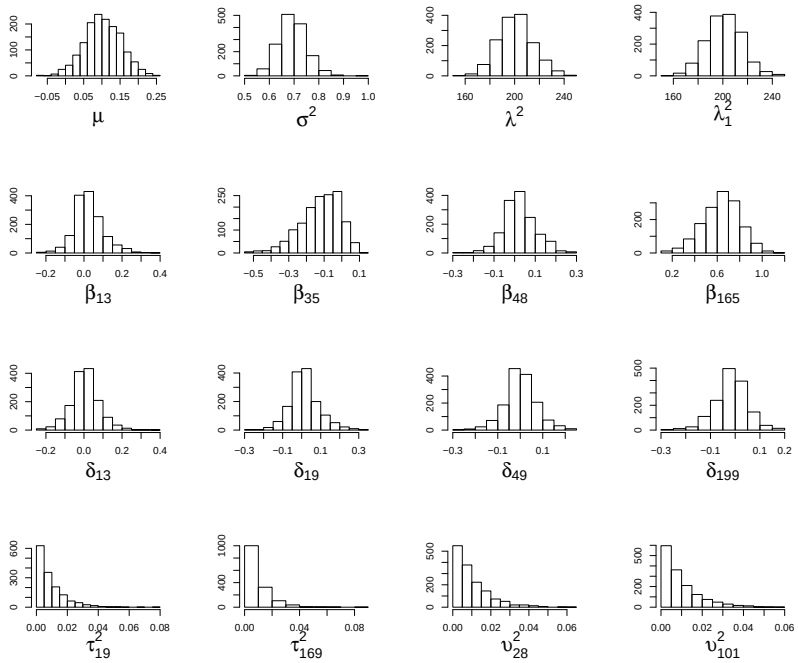

Figura 2 - Histograma dos valores gerados utilizando o Modelo I

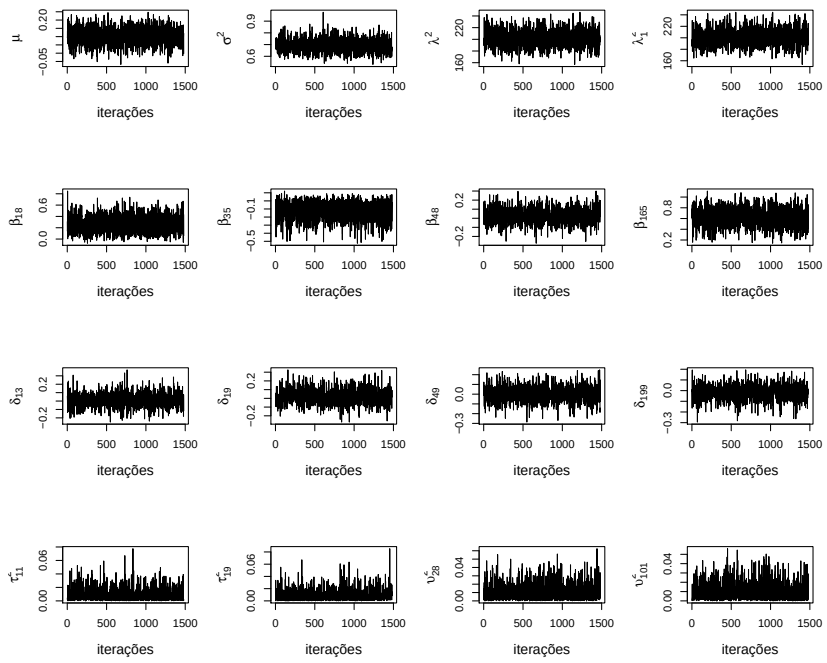

Figura 3 - Trajetória das cadeias geradas utilizando o Modelo I

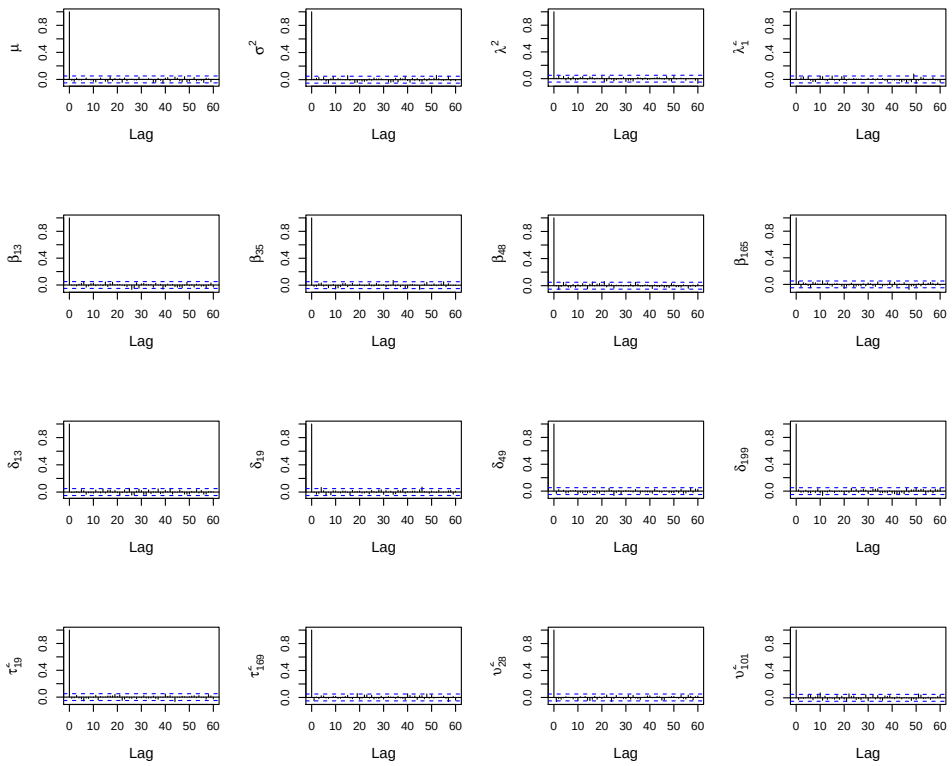

Figura 4 - Função de Autocorrelação das cadeias geradas utilizando o Modelo I

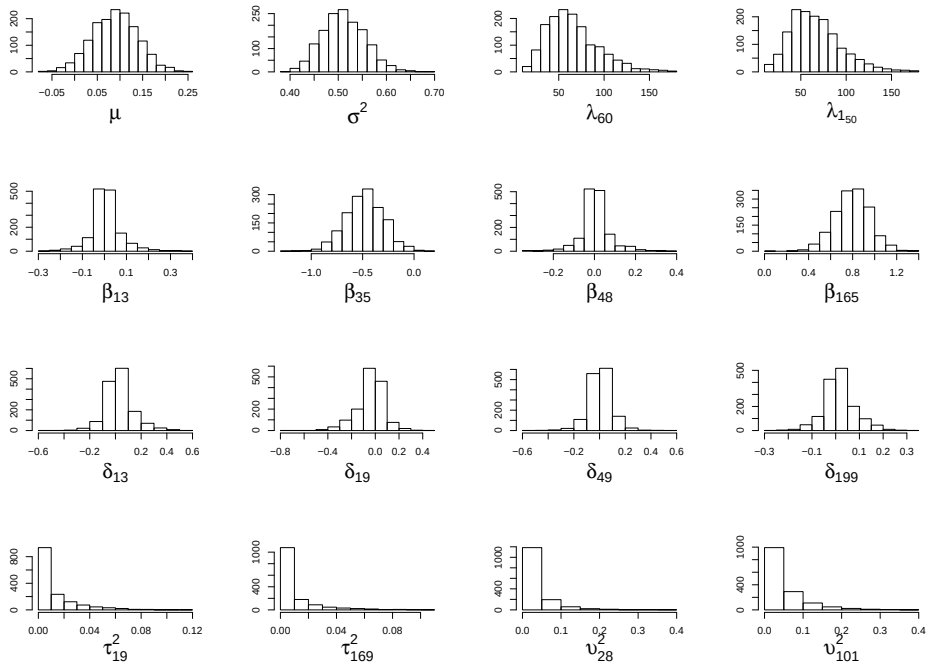

Figura 5 - Histograma dos valores gerados utilizando o Modelo II

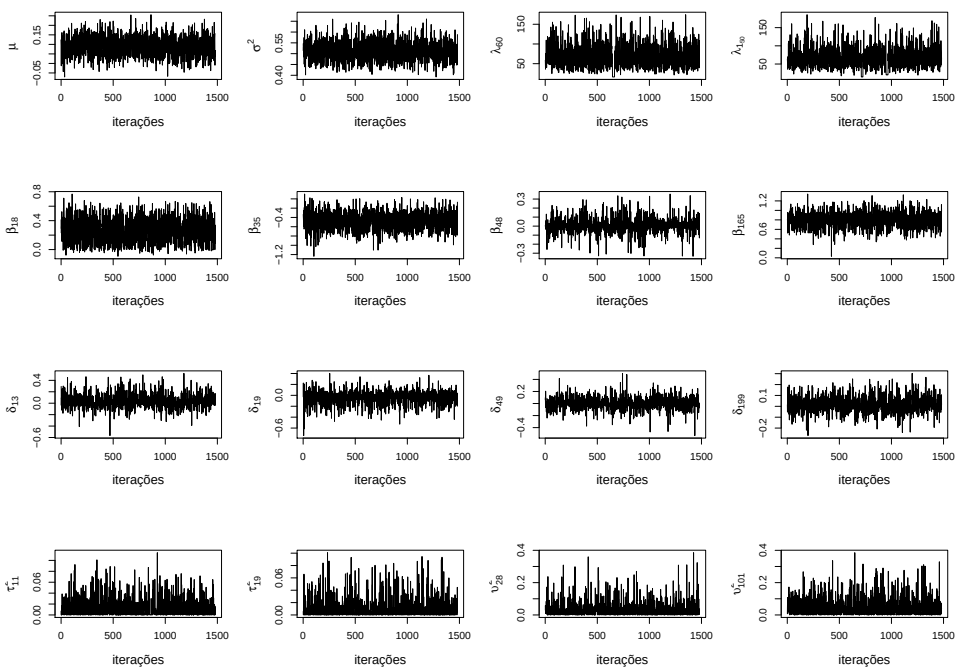

Figura 6 - Trajetória das cadeias geradas utilizando o Modelo II

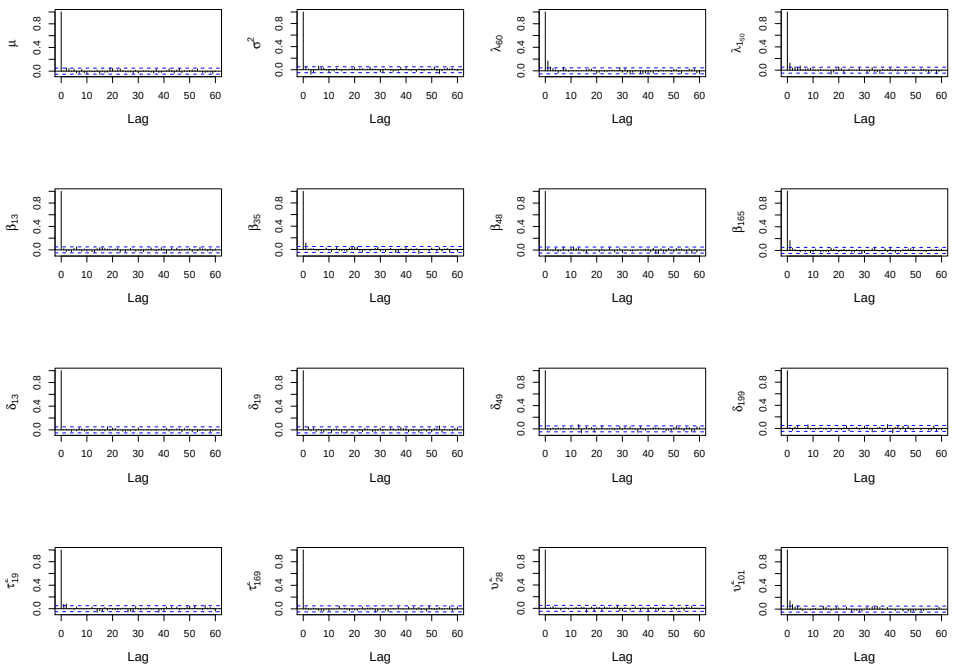

Figura 7 - Função de Autocorrelação das cadeias geradas utilizando o Modelo II

A partir das representações gráficas das cadeias de cada parâmetro geradas por meio do amostrador de Gibbs e Metropolis-Hastings, Figuras 3 e 6, pode-se observar que o processo de geração das amostras não apresentou irregularidade, tais como valores discrepantes ou tendências de estabilizações fora dos limites de convergência do processo iterativo. Os histogramas apresentados por meio das Figuras 2 e 5 sugerem que a distribuição *a posteriori* apresenta uma tendência à simetria para boa parte dos parâmetros. Tem-se nas Figuras 4 e 7 as representações gráficas das funções de autocorrelação. A função de autocorrelação mede o grau de correlação de uma variável, em um dado instante, consigo mesma, em um instante de tempo posterior. Nota-se, Figuras 4 e 7, que com o passar do tempo as autocorrelações vão ficando limitadas, dessa forma a amostra obtida poderá ser considerada como uma amostra "aproximadamente independente"do parâmetro considerado. Adicionalmente, pode-se considerar que o *burn-in* e o *thin* adotados na simulação foram suficientes, no sentido, de oferecerem boas estimativas para os momentos de interesse.

Nas Figuras 8 e 9 são apresentados os gráficos referentes a mediana *a posteriori* dos efeitos aditivo e dominante estimados por meio dos Modelos I e II e à proporção da variância fenotípica explicada por cada efeito (herdabilidade) para os marcadores ao longo dos cromossomos, enquanto que nas Figuras 10 e 11 estão os gráficos com os intervalos de credibilidade das estimativas dos parâmetros associados aos marcadores com evidências de associação com QTL.

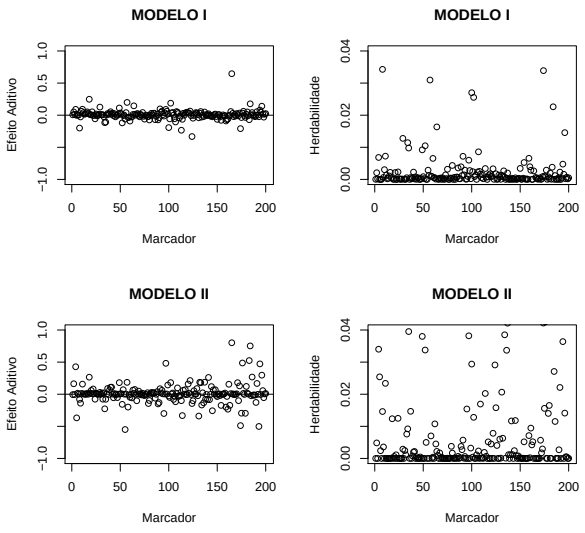

Figura 8 - Mediana *a posteriori* para o efeito aditivo de cada marcador e da herdabilidade, considerando os Modelos I e II

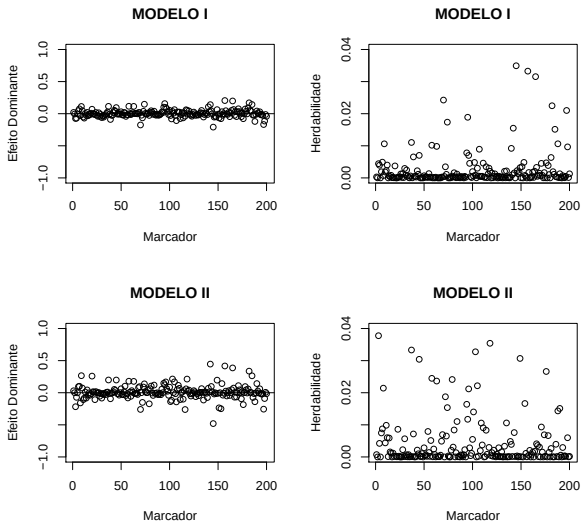

Figura 9 - Mediana *a posteriori* para o efeito dominante de cada marcador e da herdabilidade, considerando os Modelos I e II

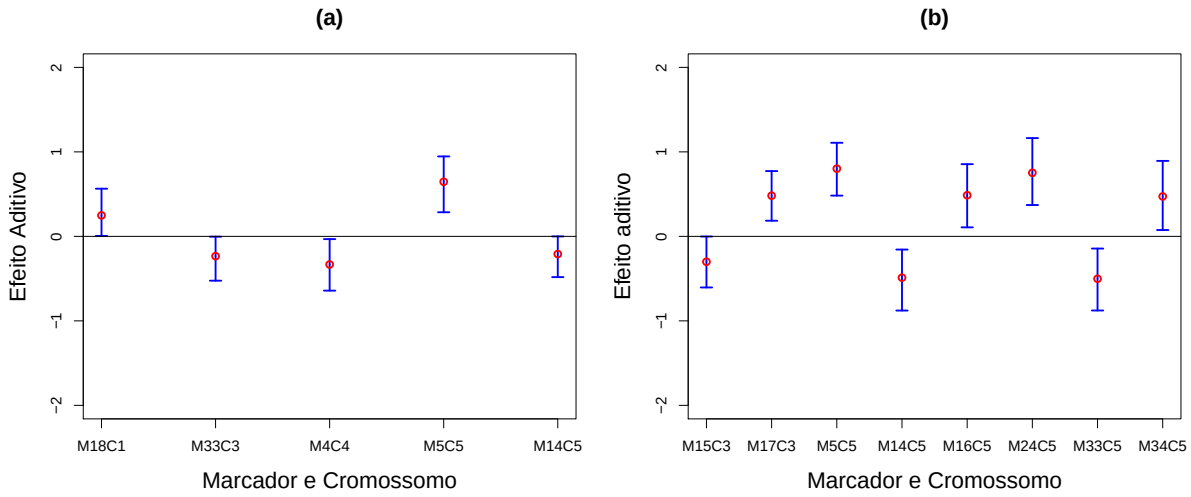

Figura 10 - Intervalo de Credibilidade 95% para o efeito aditivo dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

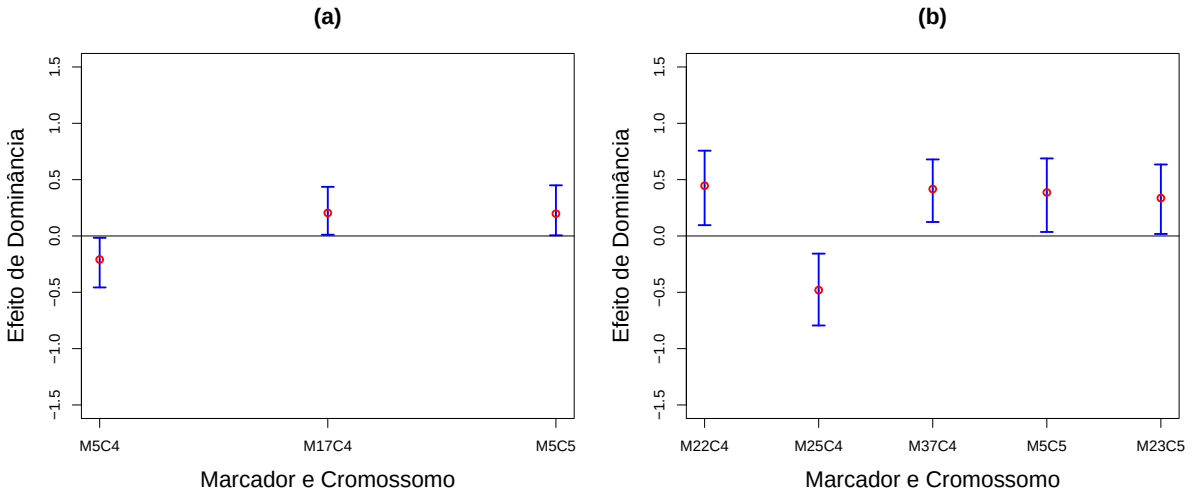

Figura 11 - Intervalo de Credibilidade 95% para o efeito dominante dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

A cada estimativa do efeito do QTL associado aos marcadores das Figuras 8 e 9 foi calculado o seu intervalo de credibilidade. Considerou-se um marcador com evidência de associação a um QTL, quando o valor zero não pertencia ao intervalo de credibilidade. Sendo assim, com base nos resultados expressos nas Figuras 10 e 11, verifica-se que sete marcadores foram selecionados pelo Modelo I e 12 pelo Modelo II. Esse número de marcadores foi determinado pela combinação dos resultados dos efeitos aditivo e dominante, sendo que, para aqueles marcadores que apareceram em ambos os efeitos, foi feita uma única contagem. Dos sete marcadores selecionados por meio do Modelo I, dois (M18C1 e M5C5) estão de fato associados a QTL, conforme informações contidas na Tabela 3. Os outros cinco marcadores selecionados não possuem associação direta, sendo considerados nesse caso como falsos positivos. Dos 12 marcadores selecionados por meio do Modelo II, quatro (M15C3, M5C5, M23C5, M24C5) estão associados a QTL, três (M17C3, M33C5, M34C5) estão próximos aos marcadores flanqueadores dos QTL e os cinco restantes não possuem associação direta, sendo considerados, como no Modelo I, falsos positivos.

Nos dados com herdabilidade baixa, bem como nos outros analisados utilizando os dois modelos propostos neste trabalho, observou-se cada uma das estimativas dos efeitos associadas aos marcadores, independente da evidência da "não signifi-

cância"dos parâmetros associados a esses marcadores. Isso foi necessário para verificar o comportamento de todas as estimativas dos marcadores flanqueadores dos QTL nas análises. Um resultado considerado importante, foi que boa parte dos marcadores flanqueadores das Tabelas 2 e 3 que não foram selecionados com associação a QTL, teve um valor estimado que se afastou consideravelmente do zero, quando comparado com os demais marcadores. A estimativa do marcador 37, que flanqueia um QTL do cromossomo 5 juntamente com o marcador 38 (Tabela 3) ilustra bem essa situação. Observa-se na Tabela 5 a estimativa associada a esse marcador e os LI (Limites Inferiores) e LS (Limites Superiores) do intervalo de credibilidade. Nota-se que o LS do intervalo ultrapassa muito pouco o valor zero. Embora isso tenha ocorrido, esses marcadores não foram computados como associação ao QTL.

Tabela 5 - Efeito e LI e LS dos intervalos para o marcador 37 do cromossomo 5

Modelo	Marcador	Efeito Dominante	LI	LS
I	37	-0,1691	-0,4312	0,0061
II	37	-0,2524	-0,5843	0,0093

Uma observação pertinente em relação aos resultados observados no estudo de simulação, foi que os valores estimados para os marcadores associados ao efeito do QTL foram em geral mais próximos do efeito do QTL simulado quando utilizou-se o Modelo II. Na Tabela 3, tem-se que o efeito dominante do QTL posicionado entre os marcadores 37 e 38 é igual a $0,3596$, e utilizando a Tabela 5 como ilustração, tem-se que o valor estimado pelo Modelo II aproxima-se mais deste efeito.

4.1.2 Herdabilidade alta

Da mesma forma que nos dados de herdabilidade baixa, foram escolhidos alguns parâmetros para apresentar as representações gráficas, tais como histograma, trajetória das cadeias e função de autocorrelação. Observa-se nas Figuras 12 a 17 as representações da trajetória das cadeias, função de autocorrelação e histogramas das densidades marginais de cada parâmetro. A partir dessas representações gráficas é possível verificar que não houve resultados que caracterizassem a não convergência dos parâmetros, como, por exemplo, valores discrepantes ou tendências de estabilização fora dos limites de convergência do processo iterativo. Os histogramas apresentados por meio das Figuras 12 e

15 sugerem que a distribuição a *posteriori* apresenta uma tendência à simetria para boa parte dos parâmetros. Pode-se observar nas Figuras 14 e 17 que o resultado da função de autocorrelação foi satisfatório para todos os parâmetros.

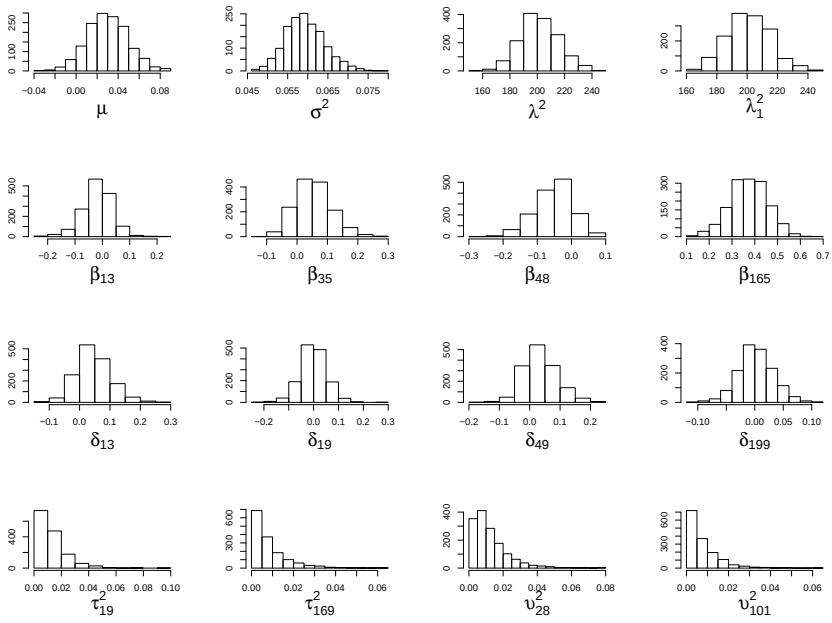

Figura 12 - Histograma dos valores gerados utilizando o Modelo I

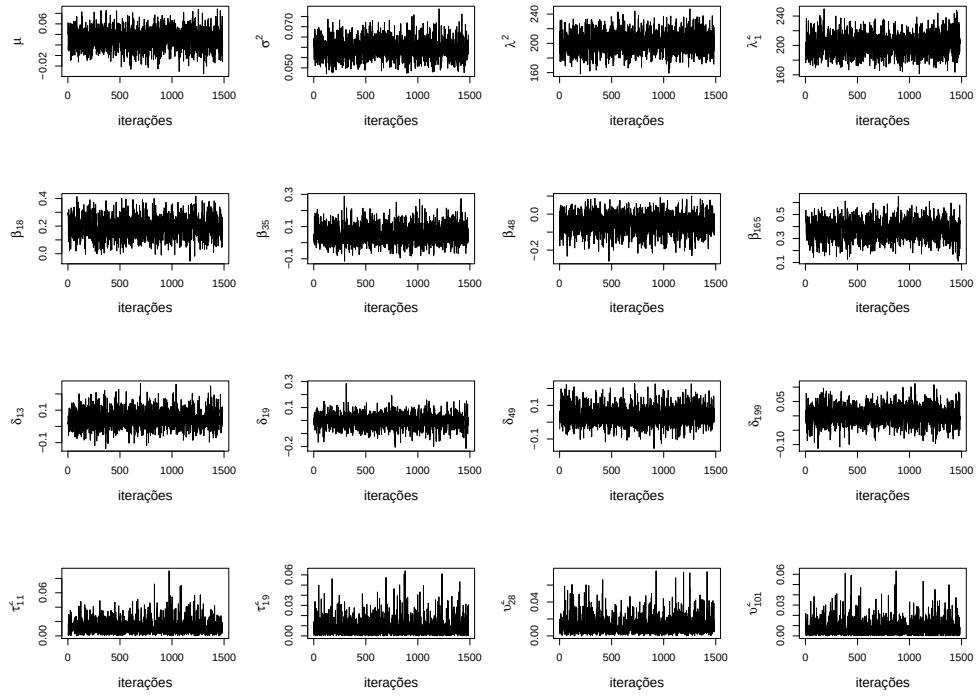

Figura 13 - Trajetória das cadeias geradas utilizando o Modelo I

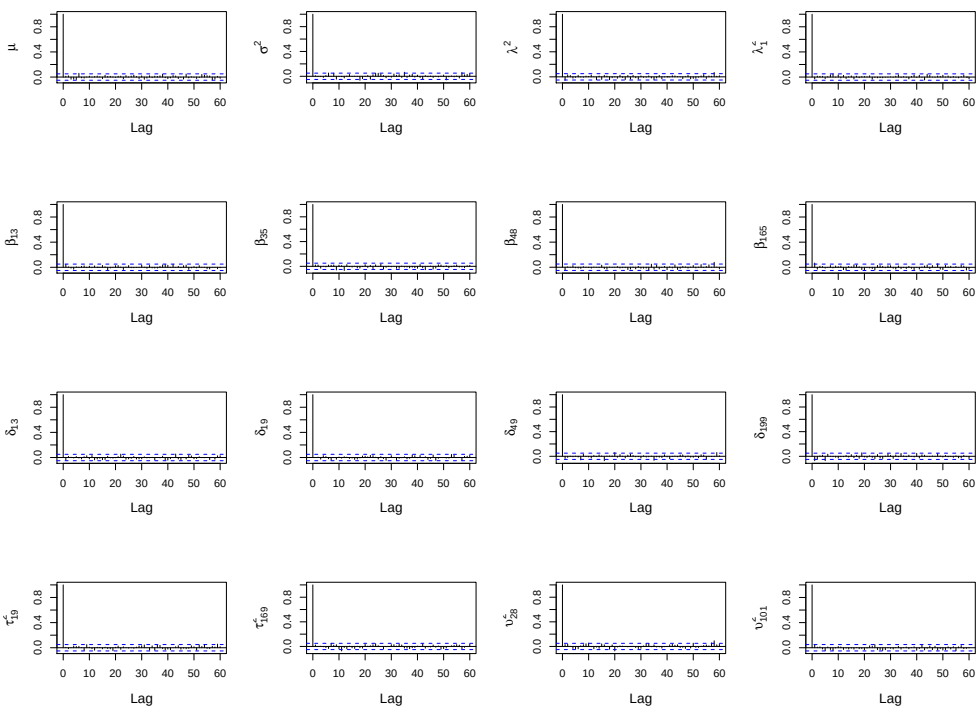

Figura 14 - Função de Autocorrelação das cadeias geradas utilizando o Modelo I

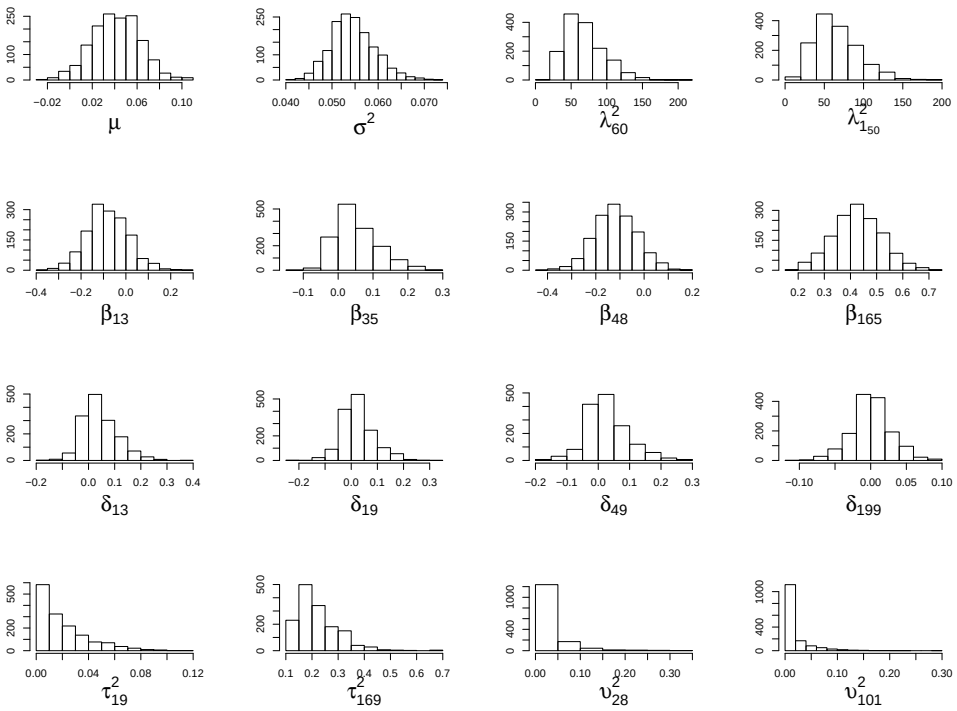

Figura 15 - Histograma dos valores gerados utilizando o Modelo II

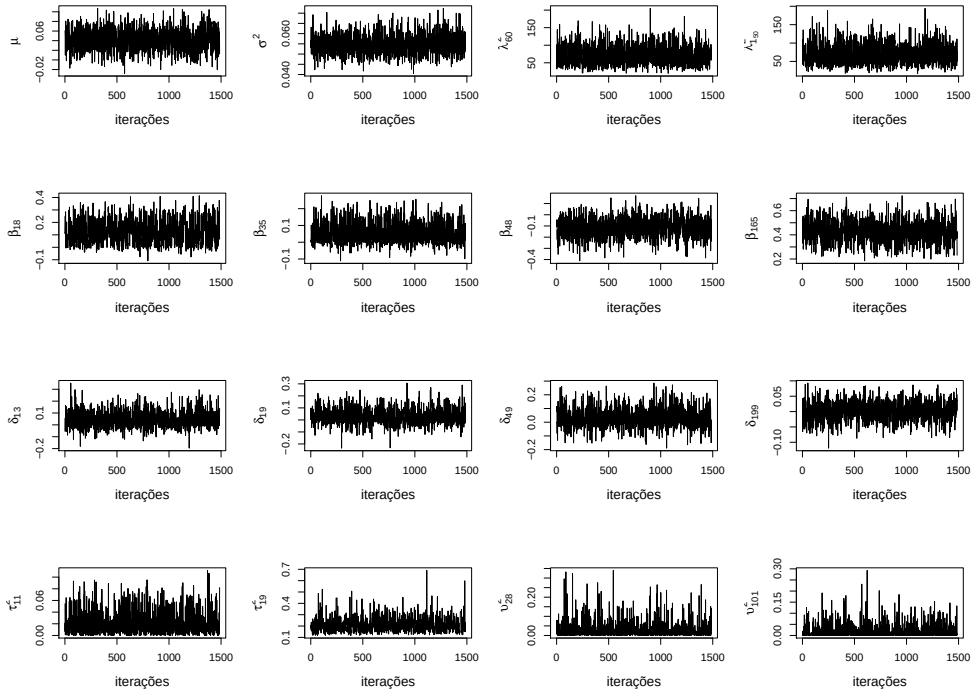

Figura 16 - Trajetória das cadeias geradas utilizando o Modelo II

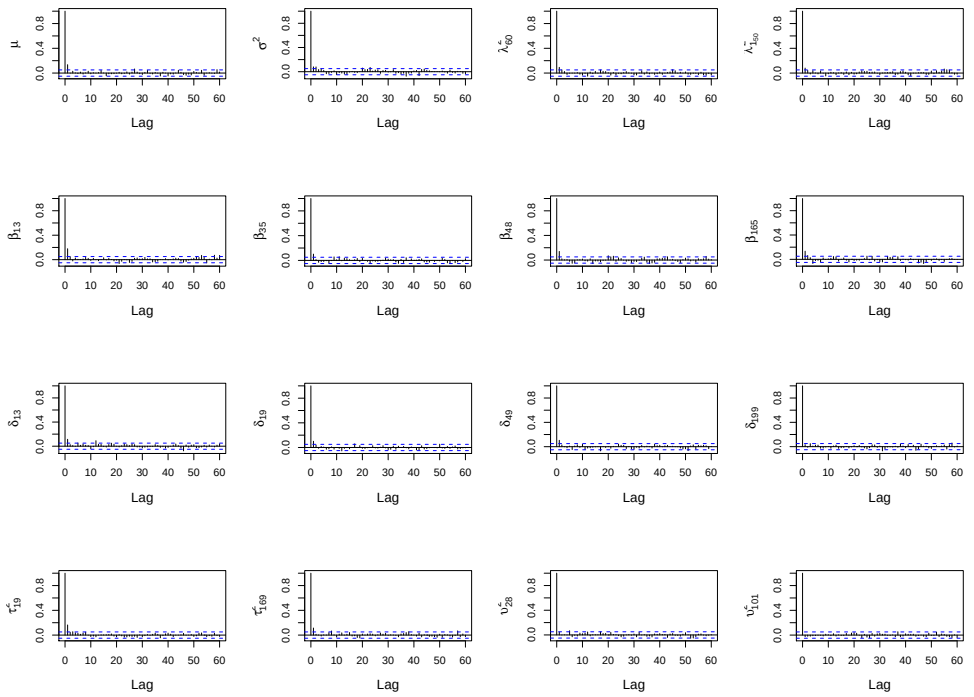

Figura 17 - Função de Autocorrelação das cadeias geradas utilizando o Modelo II

Observa-se por meio das Figuras 18 e 19 a mediana *a posteriori* dos efeitos aditivo e dominante estimados por meio dos Modelos I e II e à proporção da variância

fenotípica explicada por cada efeito para os marcadores ao longo dos cromossomos, e nas Figuras 20 e 21 os gráficos com os intervalos de credibilidade das estimativas dos parâmetros associados aos marcadores com evidências de associação com QTL.

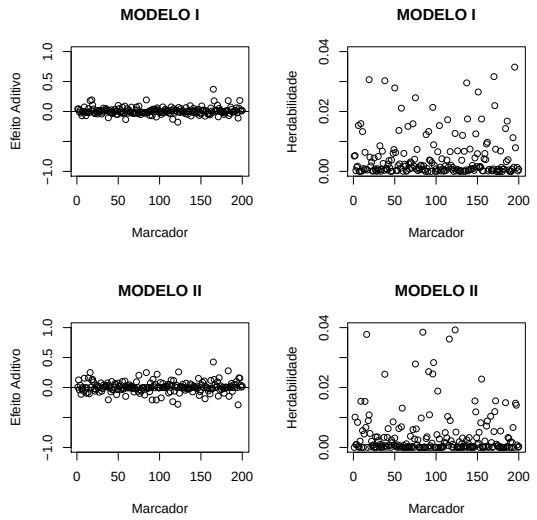

Figura 18 - Mediana *a posteriori* para o efeito aditivo de cada marcador e da herdabilidade, considerando os Modelos I e II

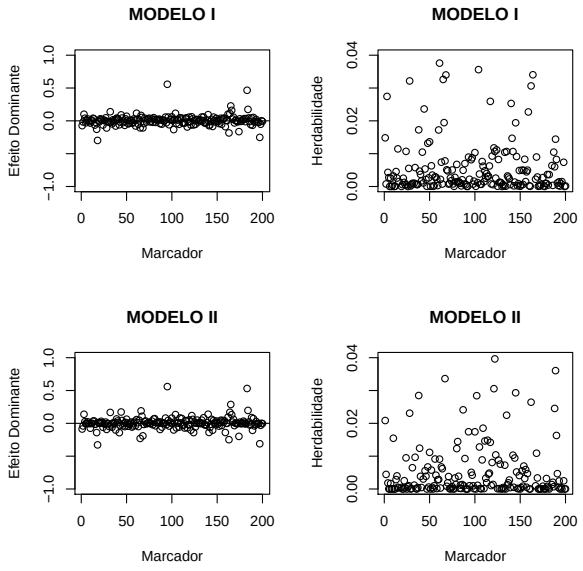

Figura 19 - Mediana *a posteriori* para o efeito dominante de cada marcador e da herdabilidade, considerando os Modelos I e II

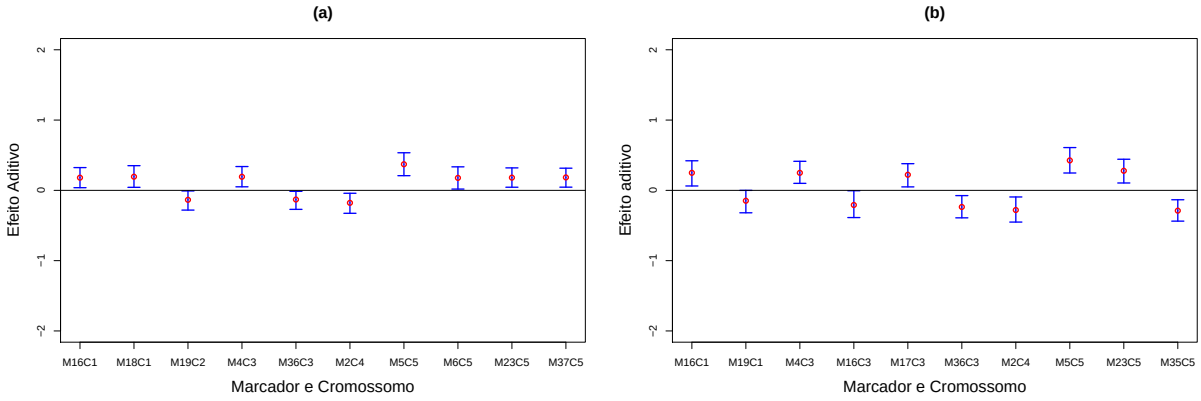

Figura 20 - Intervalo de Credibilidade 95% para o efeito aditivo dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

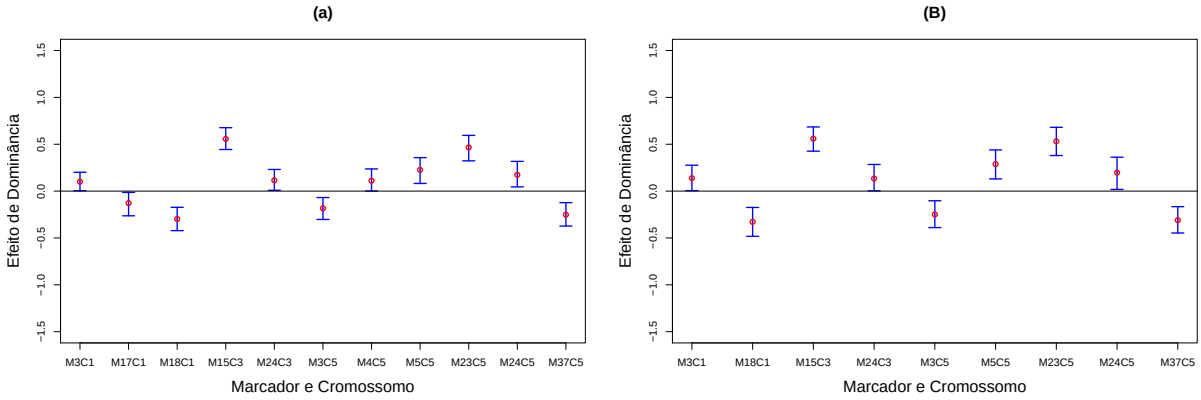

Figura 21 - Intervalo de Credibilidade 95% para o efeito dominante dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

A partir das informações contidas na Tabela 3 (página 42), pode-se observar que são 10 os marcadores associados diretamente com QTL. Combinando os efeitos aditivo e dominante, observa-se nas Figuras 20 e 21 que dezessete marcadores foram selecionados por meio do Modelo I e dezoito por meio do modelo II. Dos 17 marcadores selecionados pelo Modelo I, oito (M17C1, M18C1, M15C3, M5C5, M6C5, M23C5, M24C5, M37C5) estão de fato associados a QTL, três (M16C1, M3C5, M4C5) estão próximos aos marcadores flanqueadores e os seis restantes não possuem nenhuma associação direta, sendo considerados como falsos positivos. Dos 18 marcadores selecionados pelo Modelo II, sete (M18C1, M15C3, M16C3, M5C5, M23C5, M24C5, M37C5) estão de fato associados a QTL, cinco

(M16C1, M19C1, M17C3, M3C5, M35C5) estão próximos aos marcadores flanqueadores e os seis restantes não possuem nenhuma associação direta, sendo considerados como falsos positivos.

Com base nos valores dos efeitos dos QTL expressos na Tabela 3, é possível observar que o QTL que está localizado entre os marcadores 15 e 16 do cromossomo 3 possui um efeito aditivo igual a $0,0336$, efeito este considerado relativamente pequeno em relação aos outros. Pelo gráfico da Figura 20 nota-se que não foi possível conseguir associação entre esses marcadores e o QTL por meio do Modelo I, visto que ele não apareceu na lista dos marcadores com efeitos significativos. Já no Modelo II foi selecionado o marcador 16, porém com uma estimativa negativa. O fato de ter sido estimado um valor negativo é devido, provavelmente, ao valor do efeito do QTL está bem próximo do zero. A distribuição *a priori* atribuída aos efeitos do Modelo II é extremamente concentrada no zero, e dá a esse modelo uma possibilidade maior de capturar esses QTL com efeito pequeno, apesar de que poucas análises conseguiram essa associação. Analisando os resultados das Tabelas 6 a 8, o único resultado em que o Modelo II supera o Modelo I em relação ao efeito aditivo se vê na Tabela 8, na qual ultrapassou o limite de 80% de acertos. Isso pode estar associado ao fato de ter conseguido mais associações ligadas ao QTL localizado no cromossomo 3.

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

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

1) para cada um dos cenários apresentados na Tabela 1(seção 3, página 42) foram ajustados os Modelos I e II.

2) pela análise da média e mediana *a posteriori* dos efeitos aditivo e dominante, foram selecionados os marcadores com evidências de associação a QTL.

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

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

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

Os resultados obtidos utilizando os passos descritos nos itens de 1 a 5 estão apresentados nas Tabelas 6 a 8. Como os dois modelos estudados realizam uma estimação pontual, espera-se que no estudo de associação seja selecionado pelo menos um dos marcadores flanqueadores do QTL, não necessariamente têm que selecionar os dois marcadores, mas diversas vezes, além de selecionar os marcadores flanqueadores, foram selecionados outros que estavam próximos destes, formando assim uma concentração de marcadores, aumentando as evidências de um QTL naquele local. Em algumas situações, ao invés de selecionar os marcadores flanqueadores, selecionou-se marcadores próximos. Mediante isso, nas Tabelas 6 a 8 foram apresentados os resultados considerando o acerto exato (selecionou-se um dos marcadores flanqueadores) e por região (o marcador selecionado estava próximo aos marcadores flanqueadores).

Tabela 6 - Percentual de acertos e de falsos positivos de cada modelo, considerando-se 200 indivíduos

Herdabilidade	Nº de Marcadores	Modelo	Efeito	Percentual de Acertos (%)		Falsos Positivos (%)
				Marcador Exato	Por Região	
baixa	100	I	aditivo	21,2	26	1,76
			dominante	36,4	42,4	3,10
		II	aditivo	19,6	22,4	0,82
			dominante	33,6	36,0	0,92
	200	I	aditivo	24	26	2,40
			dominante	32,4	36,0	2,60
		II	aditivo	20,4	22,4	1,45
			dominante	32	40,4	1,55
alta	100	I	aditivo	49,2	50,8	2,58
			dominante	72,4	73,3	4,26
		II	aditivo	46,8	48,8	0,92
			dominante	74,0	74,4	1,04
	200	I	aditivo	49	49,8	3,31
			dominante	71,2	74,2	3,63
		II	aditivo	48,1	49,4	1,56
			dominante	72,2	72,6	1,44

Concordâncias entre os resultados, obtidos por ambos os modelos, indica a equivalência entre os mesmos e, portanto, um ou outro, pode ser escolhido sem grande impacto no resultado final, porém observando nas Tabelas 6 a 8 o percentual médio de acerto do Modelo I foi em geral superior ao Modelo II. Em contrapartida, é possível observar que o número de falsos positivos provenientes do Modelo II em todos os cenários considerados foi inferior ao Modelo I. Observa-se nestas tabelas que os resultados melhoraram a medida em que aumentou o número de indivíduos na população. Assim como variou o número de indivíduos, houve uma variação no número de marcadores, mas os resultados não mudaram e foram semelhantes quando considerado 100 ou 200 marcadores.

Tabela 7 - Percentual de acertos e de falsos positivos de cada modelo, considerando-se 400 indivíduos

Herdabilidade	Nº de Marcadores	Modelo	Efeito	Percentual de Acertos (%)		Falsos Positivos (%)
				Marcador Exato	Por Região	
baixa	100	I	aditivo	40,80	45,20	2,70
			dominante	74,00	77,20	3,42
		II	aditivo	36,40	39,80	1,20
			dominante	56,8	60,4	1,60
	200	I	aditivo	41,95	45,36	2,78
			dominante	73,17	77,00	3,96
		II	aditivo	39,02	40,98	1,08
			dominante	56,58	59,51	1,12
alta	100	I	aditivo	75,20	76,80	2,92
			dominante	97,60	97,60	4,28
		II	aditivo	70,80	72,00	1,38
			dominante	98,00	98,00	1,40
	200	I	aditivo	74,8	77,20	4,65
			dominante	98,4	98,4	5,34
		II	aditivo	71,60	74,00	1,63
			dominante	98,4	98,4	2,17

Tabela 8 - Percentual de acertos e de falsos positivos de cada modelo, considerando-se 800 indivíduos

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

4.3 População de milho tropical

Para os dados, considerando os fenótipos: Produção de grãos, Altura da planta e Altura da espiga, foram ajustados os Modelos I e II. O tempo computacional para a análise em um computador com a configuracão: Intel (R) Core (TM) i7-2630 QM CPU @ 2.00 GHz e 6 GB de RAM, foi de aproximadamente 56 minutos para cada um dos Modelos. Os resultados obtidos utilizando os dois modelos foram comparados entre si e serão apresentados nas subseções que seguem.

Devido ao grande número de parâmetros nos modelos ajustados envolvendo os diferentes fenótipos, apenas alguns parâmetros terão as suas representações gráficas utilizadas para verificar a convergência, expostos nesta subseção.

4.3.1 Produção de grãos

Observa-se por meio das Figuras 22 a 27 as representações gráficas da trajetória das cadeias, função de autocorrelação e histogramas das densidades marginais *a posteriori* associadas aos efeitos aditivos, dominantes e demais parâmetros. Conforme discutido nos resultados de herdabilidade baixa e alta, nota-se por meio das Figuras 22, 23, 25 e 26 que há indícios de convergência para os parâmetros. Observa-se por meio das Figuras 24 e 27 que o resultado da função de autocorrelação foi satisfatório para todos os parâmetros.

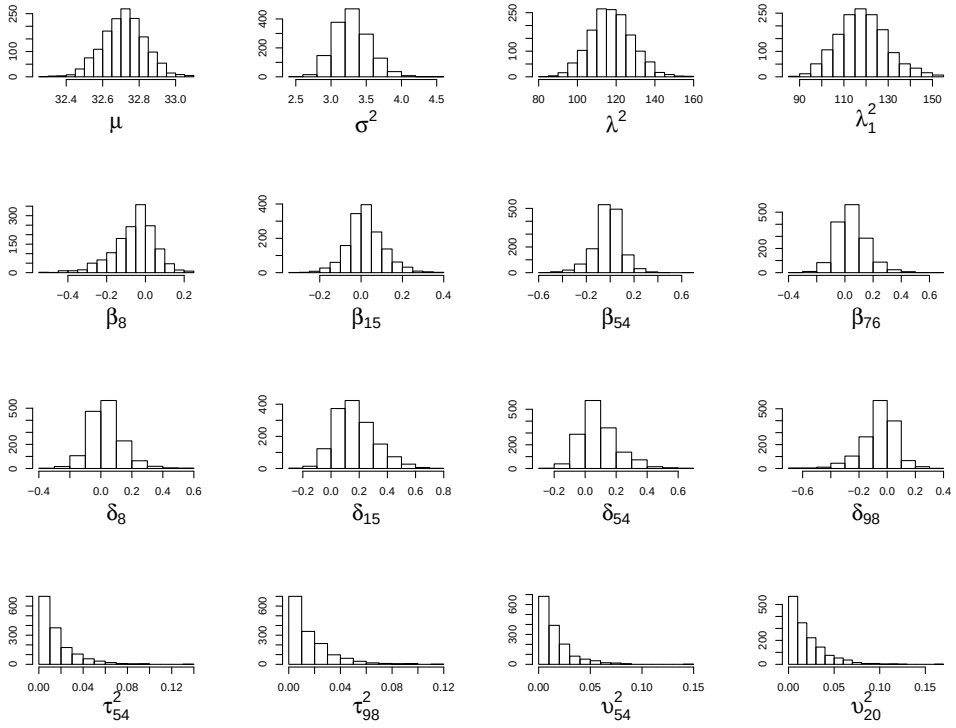

Figura 22 - Histograma dos valores gerados utilizando o Modelo I

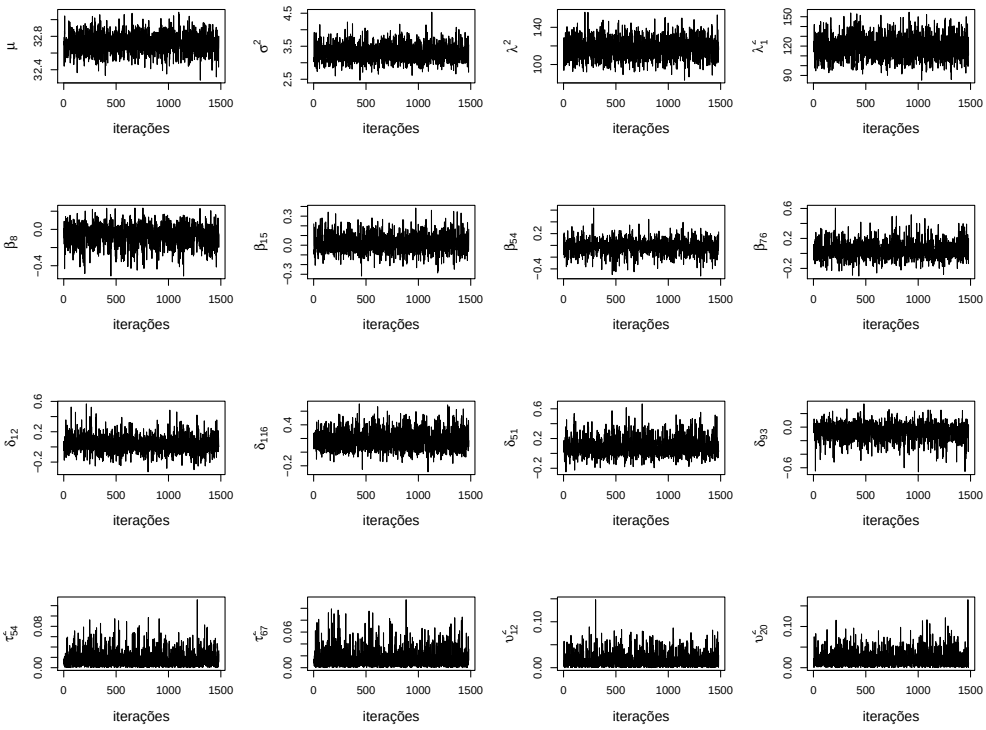

Figura 23 - Trajetória das cadeias geradas utilizando o Modelo I

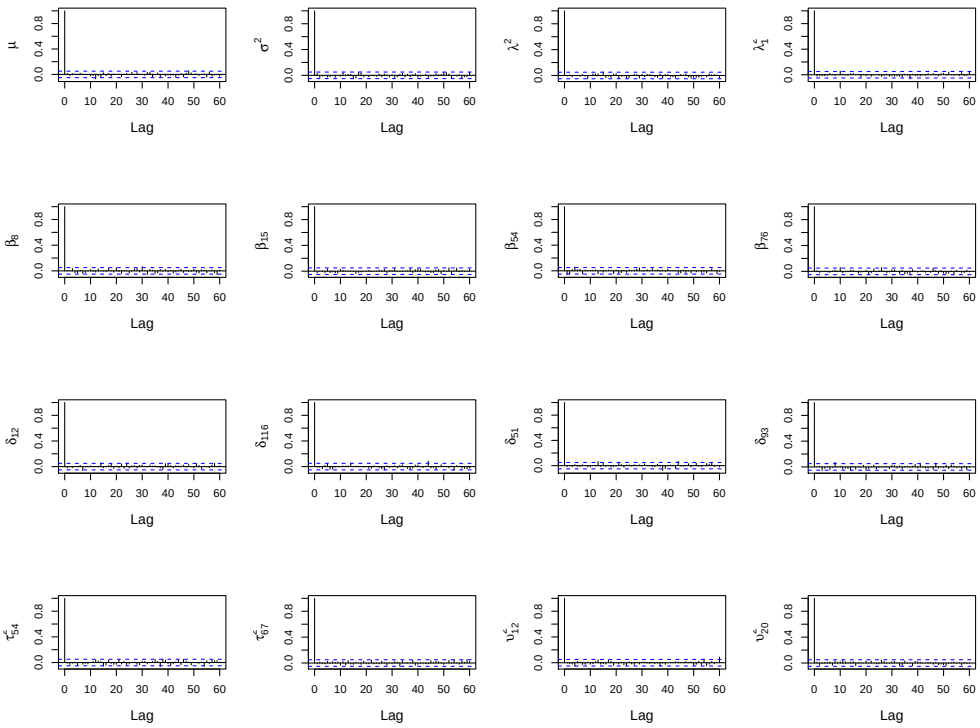

Figura 24 - Função de Autocorrelação das cadeias geradas utilizando o Modelo I

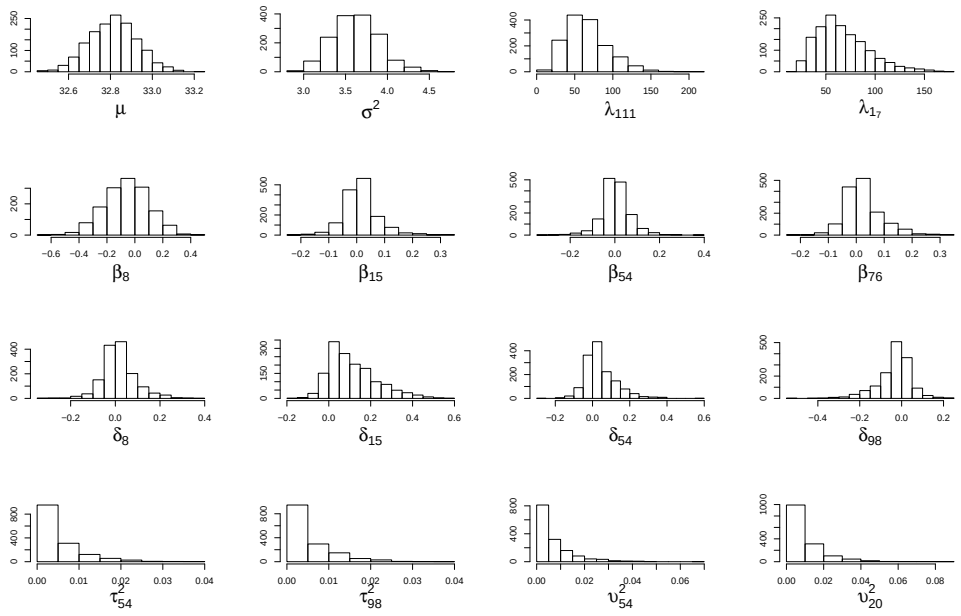

Figura 25 - Histograma dos valores gerados utilizando o Modelo II

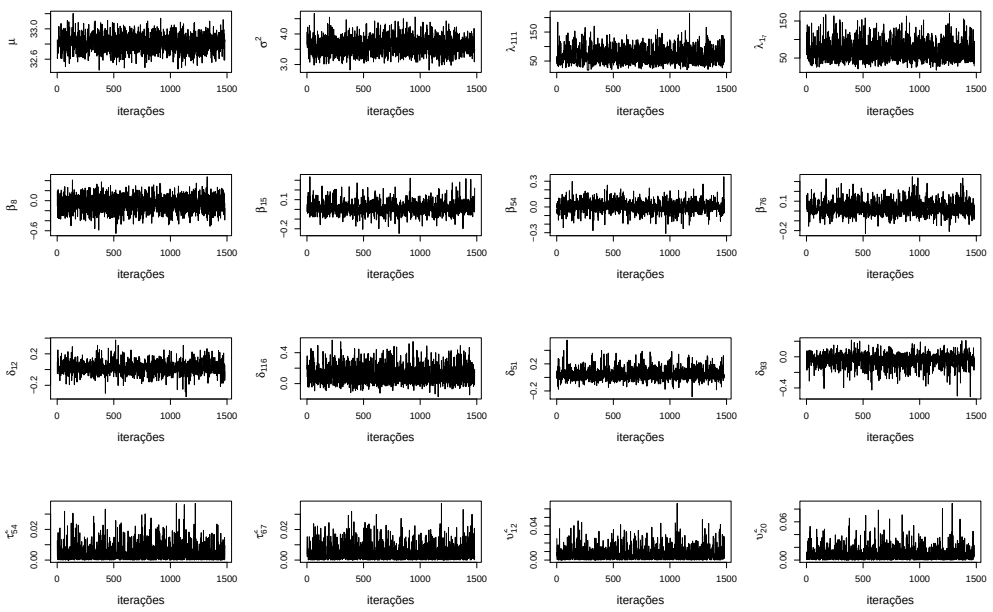

Figura 26 - Trajetória das cadeias geradas utilizando o Modelo II

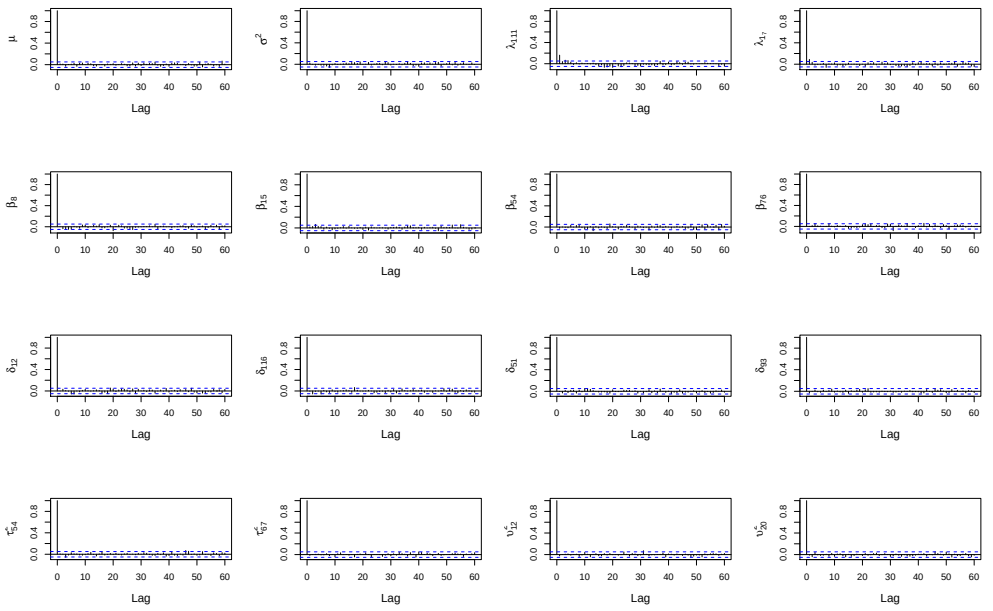

Figura 27 - Função de Autocorrelação das cadeias geradas utilizando o Modelo II

Observa-se nas Figuras 28 e 29 a mediana *a posteriori* dos efeitos aditivo e dominante estimados por meio dos Modelos I e II e à proporção da variância fenotípica explicada por cada efeito para os marcadores ao longo dos cromossomos, e nas Figuras de 30 e 31 os intervalos de credibilidade das estimativas dos parâmetros associados aos

marcadores com evidências de associação com QTL.

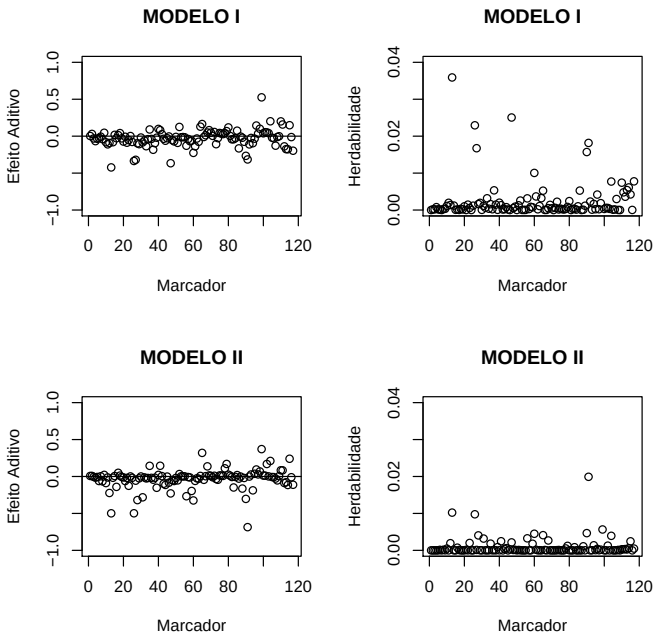

Figura 28 - Mediana *a posteriori* para o efeito aditivo de cada marcador e da herdabilidade

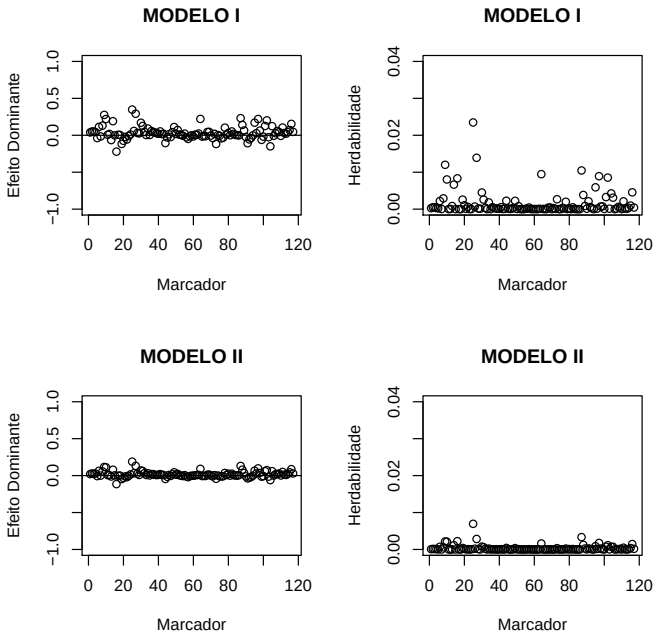

Figura 29 - Mediana *a posteriori* para o efeito dominante de cada marcador e da herdabilidade

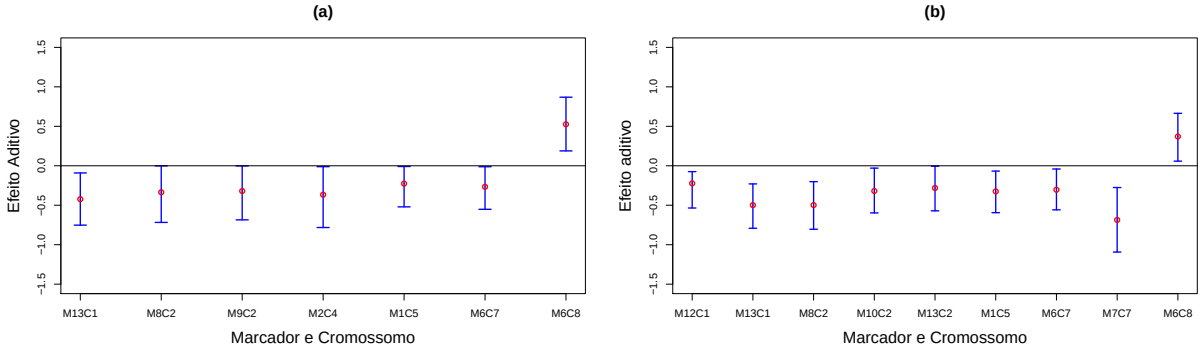

Figura 30 - Intervalo de Credibilidade 95% para o efeito aditivo dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

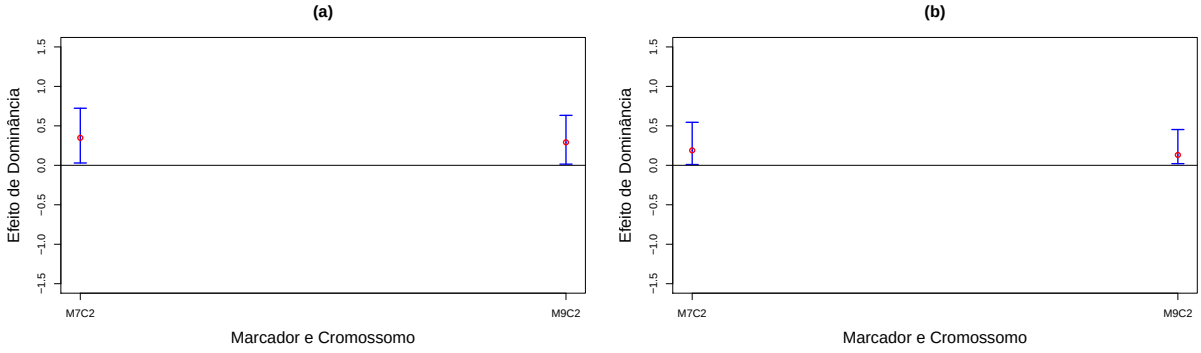

Figura 31 - Intervalo de Credibilidade 95% para o efeito dominante dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

Observa-se na Tabela 9 a lista completa dos marcadores selecionados por meio dos Modelos I e II e os cromossomos em que se encontram esses marcadores. Na Tabela 9 e nas Figuras 30 e 31 é possível observar que os resultados foram semelhantes para os Modelos I e II.

Tabela 9 - Marcadores com possíveis associações com QTL de acordo com os Modelos I e II

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

[1]Cromossomo onde está localizado o marcador.

*Não Significativo.

Combinando os efeitos aditivos e dominantes, foram identificados oito marcadores com evidências de associações a QTL utilizando o Modelo I. Esse número de marcadores foi igual a 10 pelo Modelo II. Nos dois modelos, esses marcadores estão localizados nos mesmos cromossomos (1,2,5,7,8), com exceção do marcador umc1652, que pode ser visto na Tabela 9, está localizado no cromossomo 4, identificado apenas pelo Modelo I. Observa-se uma diferença entre o número de marcadores identificados pelos modelos, mas a diferença ocorre porque para o Modelo II, ao invés de detectar apenas um marcador

associado a um possível QTL, foram identificados vários em sequência, formando uma região, aumentando as evidências de um QTL nesta região. Isso pode ser visto claramente com os marcadores 7, 8, 10 do cromossomo 2 na Tabela 9, considerando os efeitos aditivos e dominantes simultaneamente. Identificar marcadores em sequência, tornando-os como possíveis marcadores flanqueadores de um determinado QTL, não é exclusividade do Modelo II, embora, isso tenha ocorrido com maior frequência neste modelo.

Algumas análises com metodologias específicas foram realizadas anteriormente a este trabalho, utilizando-se o conjunto de dados milho tropical, fato que possibilita fazer comparações com os resultados da análise obtidos para o fenótipo produção de grãos. A comparação entre os diferentes métodos tem como objetivo principal analisar a concordância entre os resultados e não discriminar entre melhor ou pior método. Na Tabela 10 são apresentados os resultados para os cromossomos com evidências de QTL, obtidos:

1) pelo modelo CIM (Mapeamento por Intervalo Composto) apresentados em Sibov et al. (2003);

2) pela metodologia bayesiana para mapear QTL no caso em que efeitos de epistasia são incorporados no modelo, apresentado por Meyer et al. (2009);

3) e no presente trabalho.

Na Tabela 10 são apresentados apenas os marcadores dos Modelos I e II que coincidiram com os resultados destes dois trabalhos realizados anteriormente. Como se pode observar, uma parte dos marcadores selecionados neste trabalho está relacionada com QTL de acordo com as metodologias utilizadas por Sibov et al. (2003) e Meyer et al. (2009), como pode ser visto na Tabela 10. Comparando os resultados dos diferentes modelos, nota-se que há uma concordância na evidência de QTL nos cromossomos 2, 7 e 8, sendo que os dois modelos utilizados neste trabalho identificaram evidências de QTL em outros cromossomos (4,5).

Tabela 10 - Cromossomos com forte evidência da existência de QTL, considerando 4 modelos diferentes

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

[1] Cromossomo onde está localizado o QTL.

[2]Cromossomo onde está localizado o marcador.

4.3.2 Altura da espiga

Nas Figuras 32 a 37 estão as análises gráficas das cadeias de alguns parâmetros obtidas pelos Modelos I e II. Nota-se por meio das Figuras 32, 33, 35 e 36, que há indícios de convergência para os parâmetros. Além disso, nota-se por meio dos gráficos das Figuras 34 e 37 que o resultado da função de autocorrelação foi satisfatório para os parâmetros.

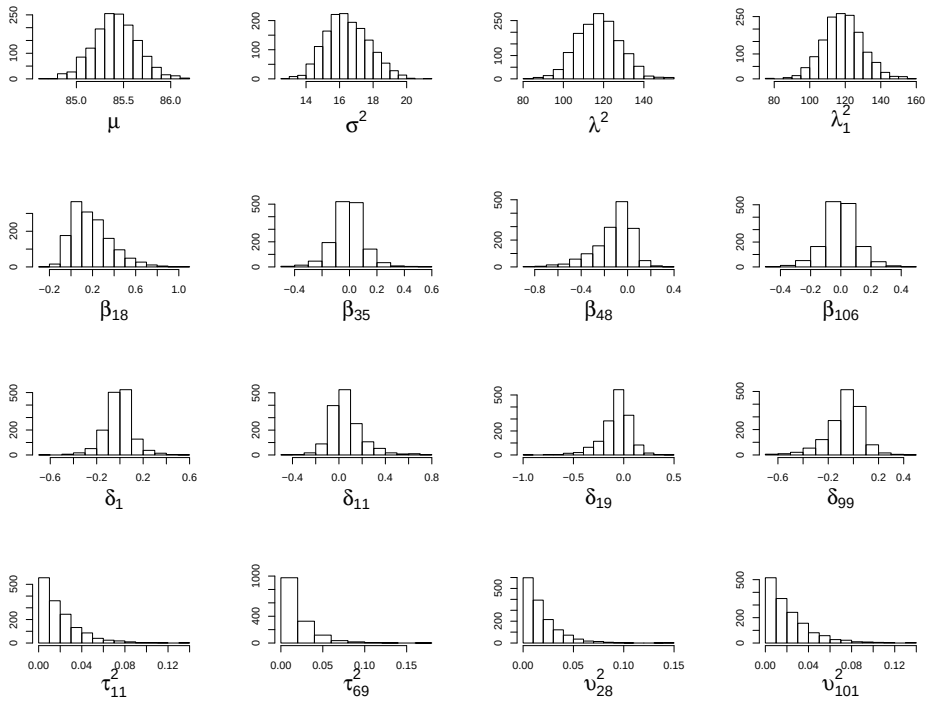

Figura 32 - Histograma dos valores gerados utilizando o Modelo I

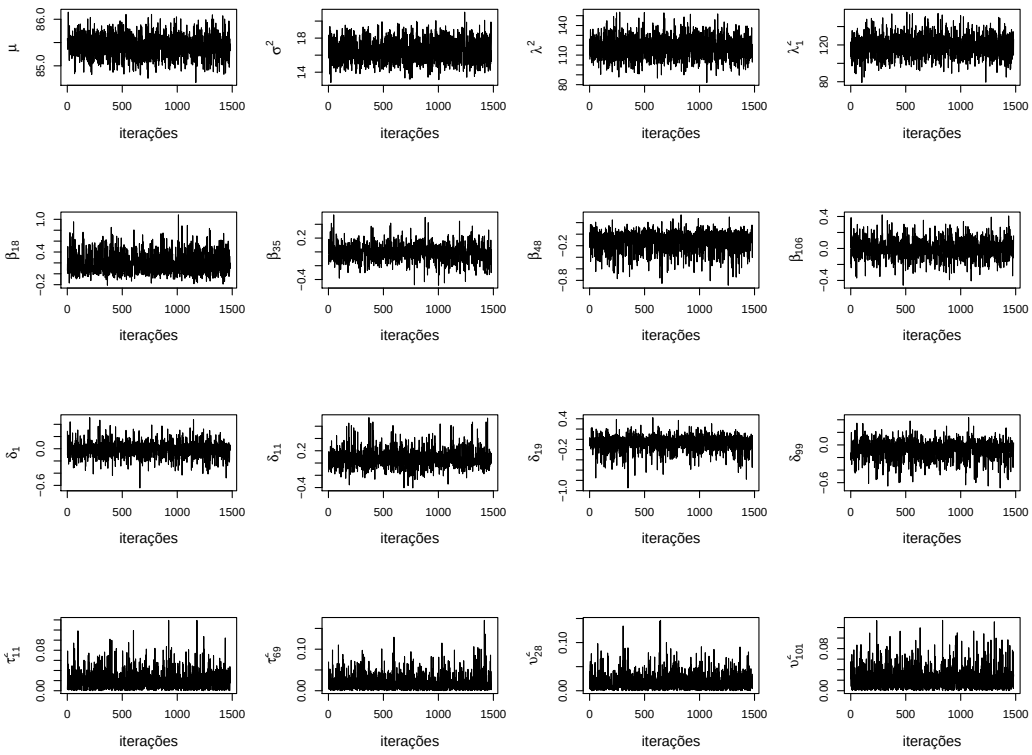

Figura 33 - Trajetória das cadeias geradas utilizando o Modelo I

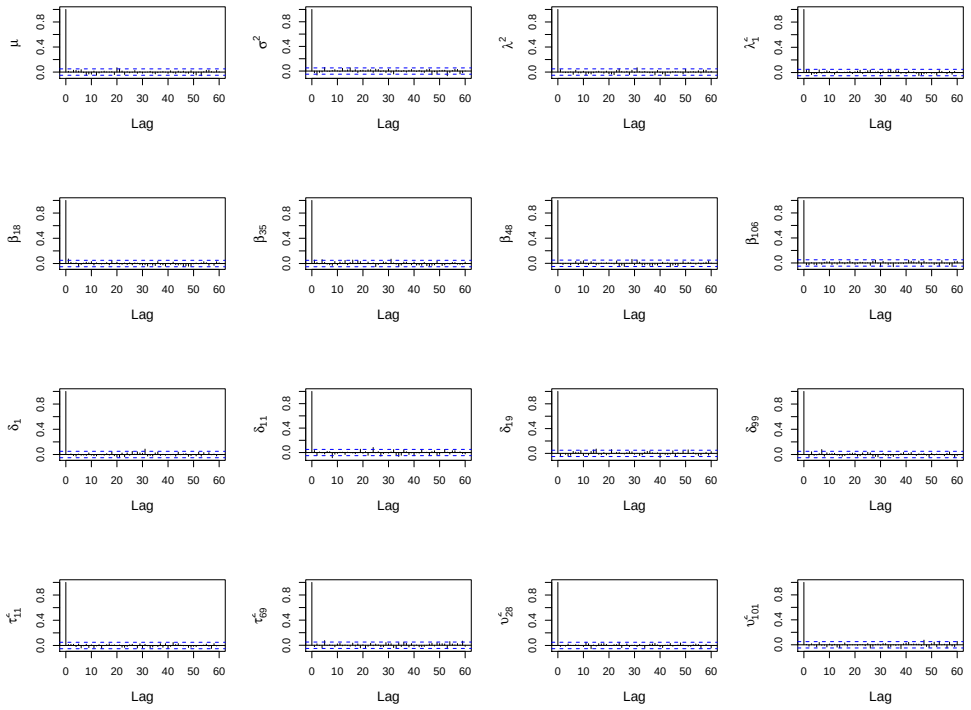

Figura 34 - Função de Autocorrelação das cadeias geradas utilizando o Modelo I

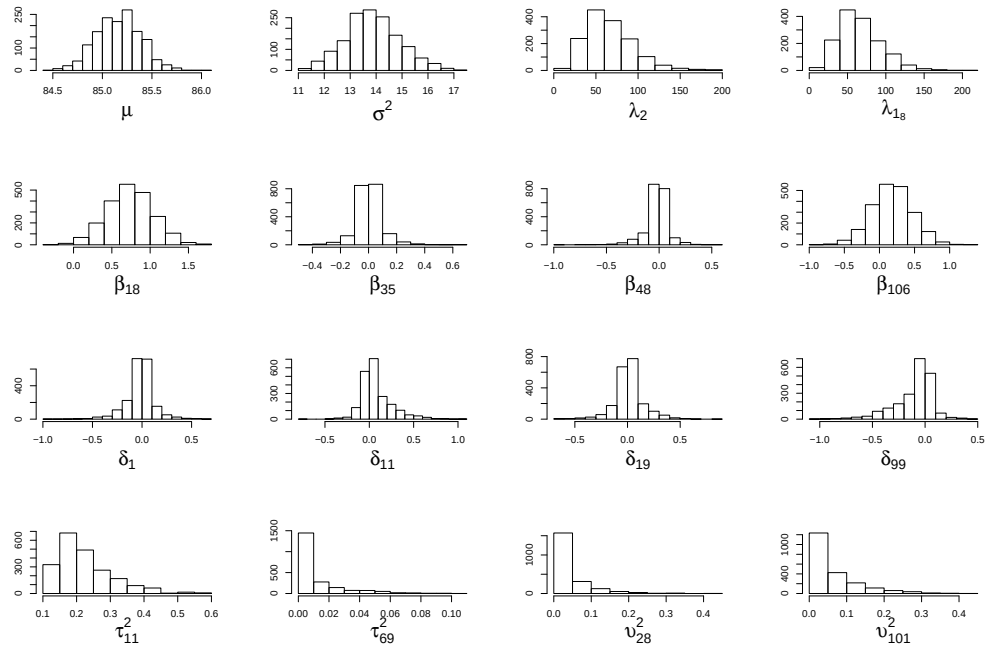

Figura 35 - Histograma dos valores gerados utilizando o Modelo II

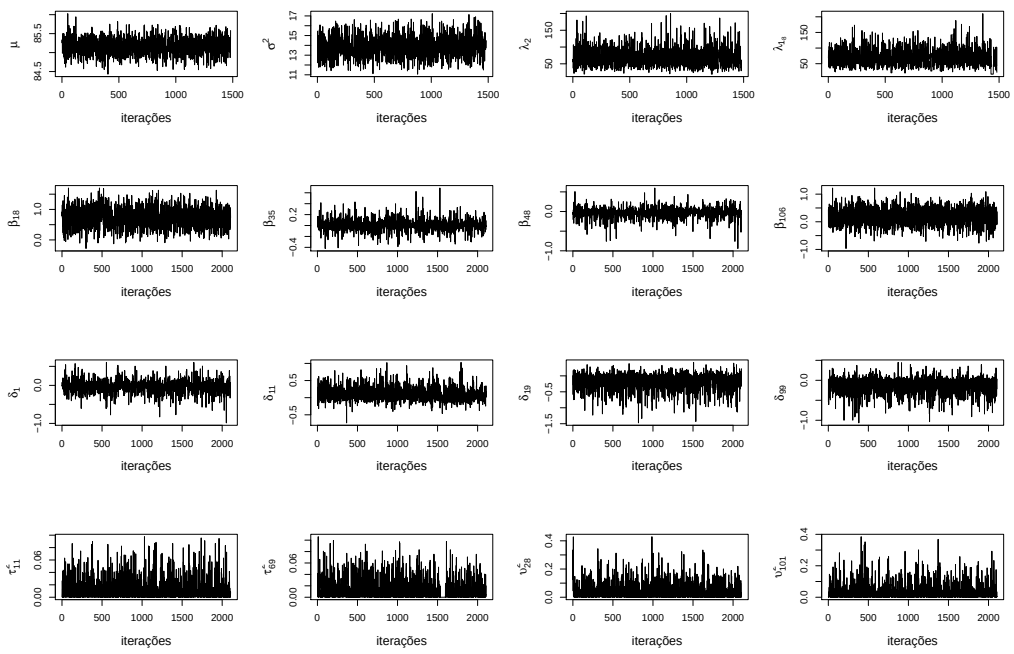

Figura 36 - Trajetória das cadeias e histograma dos valores gerados utilizando o Modelo II

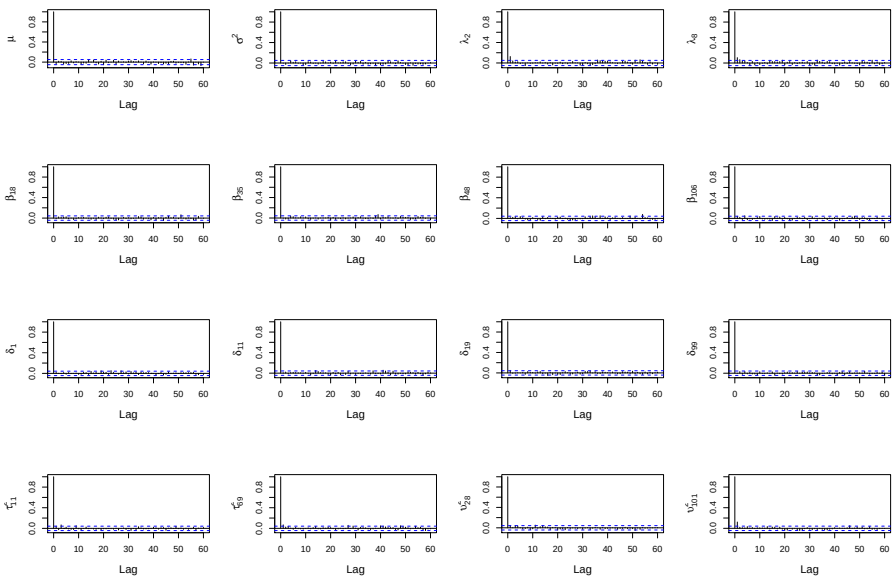

Figura 37 - Função de Autocorrelação das cadeias geradas utilizando o Modelo II

Observa-se por meio das Figuras 38 e 39 a mediana *a posteriori* dos efeitos aditivo e dominante estimados por meio dos Modelos I e II e à proporção da variância fenotípica explicada por cada efeito para os marcadores ao longo dos cromossomos, e nas

Figuras de 40 a 41 os intervalos de credibilidade das estimativas dos parâmetros associados aos marcadores com evidências de associação com QTL.

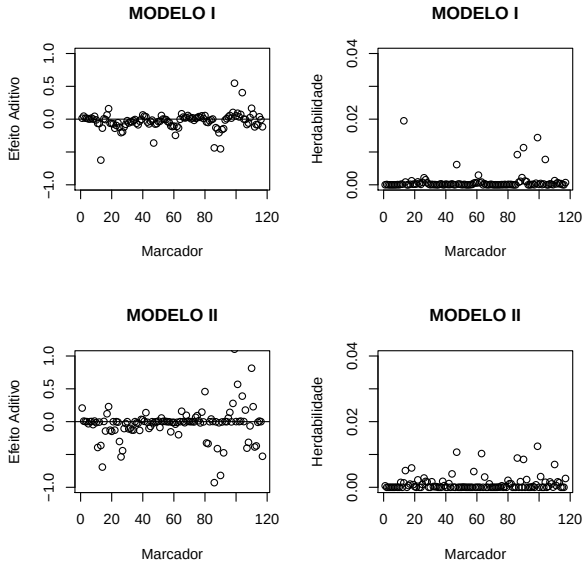

Figura 38 - Mediana *a posteriori* para o efeito aditivo de cada marcador e da herdabilidade

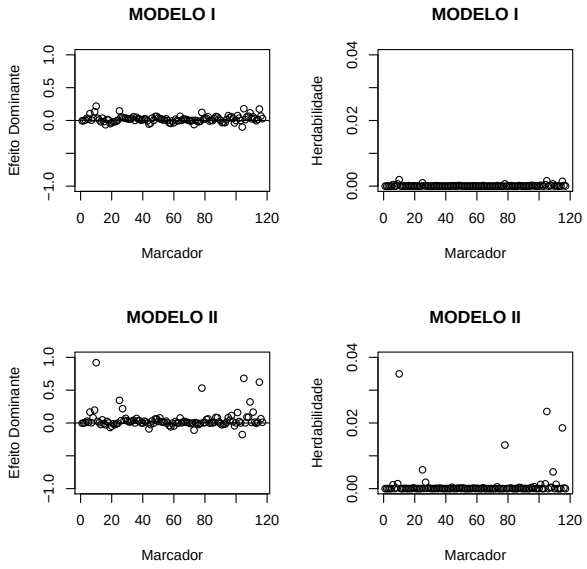

Figura 39 - Mediana *a posteriori* para o efeito dominante de cada marcador e da herdabilidade

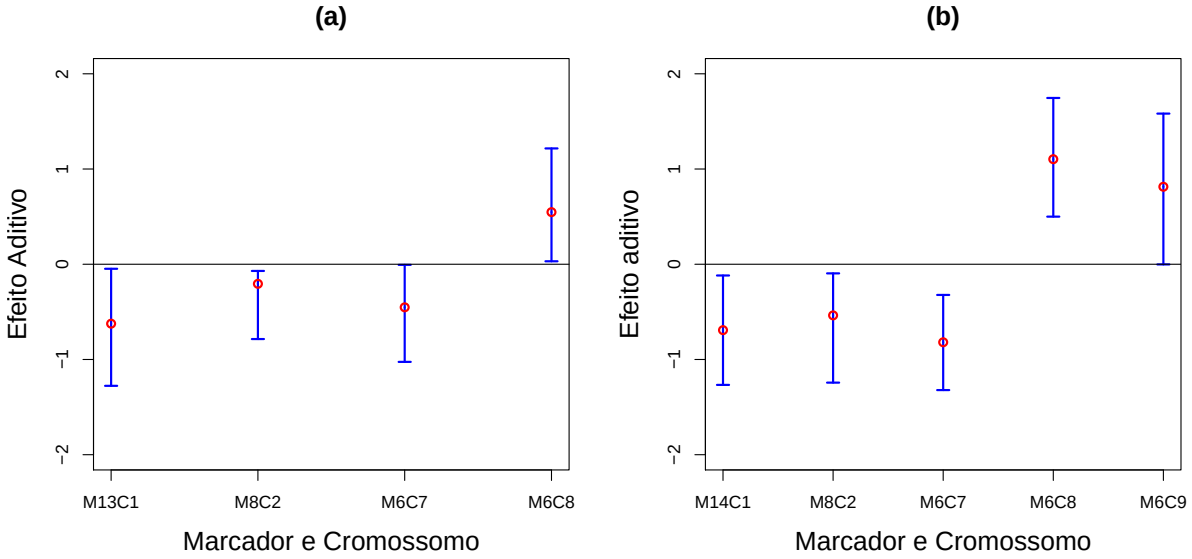

Figura 40 - Intervalo de Credibilidade 95% para o efeito aditivo dos marcadores considerados significativos de acordo com o Modelo I (a) e Modelo II (b)

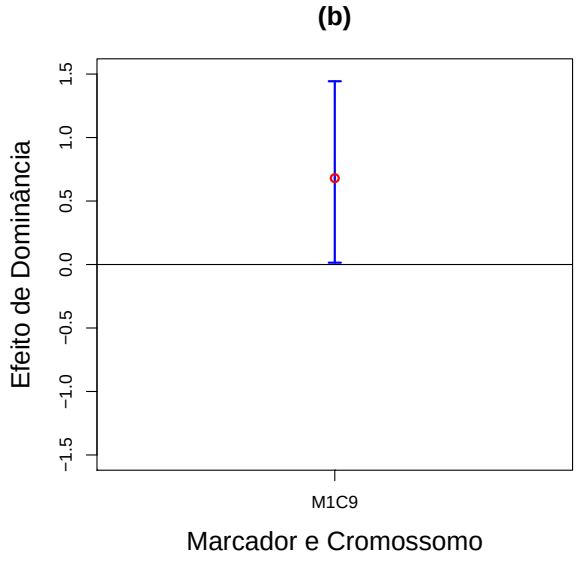

Figura 41 - Intervalo de Credibilidade 95% para o efeito dominante do marcador considerado significativo de acordo com o Modelo II (b)

Na Tabela 11, encontram-se os marcadores que podem estar associados a QTL e os cromossomos em que se localizam esses marcadores, de acordo com os dois modelos.

Tabela 11 - Marcadores com possíveis associações com QTLs de acordo com os Modelos I e II

			Efeito do marcador	
Modelo	Marcador/Crom,[1]	Marcador	Aditivo	Dominância
	13/1	bnlg0615	-0,62382	-0,01939*
	9/2	bnlg0166	-0,20535	0,05441*
	6/7	dupssr13	-0,45214	0,00110*
	6/8	bnlg1176	0,54746	-0,03706*
	14/1	phi0037	-0,69174	0,04778*
II	9/2	bnlg0166	-0,53673	0,03555*
	2/7	umc1632	-0,92901	0,01016*
	6/7	dupssr13	-0,81906	-0,00822*
	6/8	bnlg1176	1,10343	-0,04355*
	1/9	umc1893	0,00141*	0,68018
	6/9	bnlg0619	0,81395	0,00799*

[1]Cromossomo onde está localizado o marcador.

*Não Significativo.

É possível observar por meio da Tabela 11 e das Figuras 40 a 41 que os cromossomos em que se encontram localizados os marcadores selecionados são os mesmos, com exceção do cromossomo 9, cujo marcadores foram selecionados apenas pelo Modelo II. Nota-se também nesta tabela e figuras que o número de marcadores selecionados com associação ao efeito de dominância do QTL foi menor que os associados ao efeito aditivo, sendo que não houve marcadores selecionados para o efeito de dominância utilizando o Modelo I. Isso pode ter ocorrido porque o efeito de dominância do QTL talvez não seja muito expressivo, visto que no estudo de simulação, quando o efeito era considerado relativamente pequeno, não foi possível detectá-lo.

Assim como o fenótipo produção de grãos, este fenótipo foi analisado por Sibov et al. (2003) e Meyer et al. (2009), o que possibilita verificar a consistência dos resultados encontrados. Observa-se uma concordância nos resultados, porém o Modelo II detectou existência de um possível QTL no cromossomo 1, que, dentre os outros três modelos, havia sido detectado apenas pelo Modelo CIM de Sibov et al. (2003) e, a partir do

Modelo I não foi possível detectar associação a QTL no cromossomo 9, que foi identificado pelos outros modelos.

Tabela 12 - Cromossomos com forte evidência da existência de QTL, considerando 4 modelos diferentes

Modelo	QTL/Crom.[1]	Posição	Intervalo
CIM	1/1	150,38	bnlg0615-phi0037
	2/7	35,70	umc1632-umc1409
	3/7	91,70	bnlg0434-dupssr13
	5/9	1,01	umc1893-bnlg0430
Bayesiano	1/2	49,06	umc1845-bnlg0166
	2/7	19,46	umc1426-umc1632
	3/7	117,78	dupssr13-umc1154
	4/9	11,06	umc1893-bnlg0430
Modelo	Marcador/Crom.[2]	-	Marcador
I	13/1	-	bnlg0615
	9/2	-	bnlg0166
	6/7	-	dupssr13
II	14/1	-	phi0037
	9/2	-	bnlg0166
	2/7	-	umc1632
	6/7	-	dupssr13
	1/9	-	umc1893

[1] Cromossomo onde está localizado o QTL.

[2]Cromossomo onde está localizado o marcador.

4.3.3 Altura da Planta

Nas Figuras 42 a 47 estão as representações gráficas da trajetória das cadeias, função de autocorrelação e histogramas das densidades marginais *a posteriori* associadas aos efeitos aditivos, dominantes e demais parâmetros. Conforme discutido nos

resultados de herdabilidade baixa e alta, nota-se por meio das Figuras 42, 43, 45 e 46 que há indícios de convergência para os parâmetros. Observa-se por meio das Figuras 44 e 47 que o resultado da função de autocorrelação foi satisfatório para todos os parâmetros.

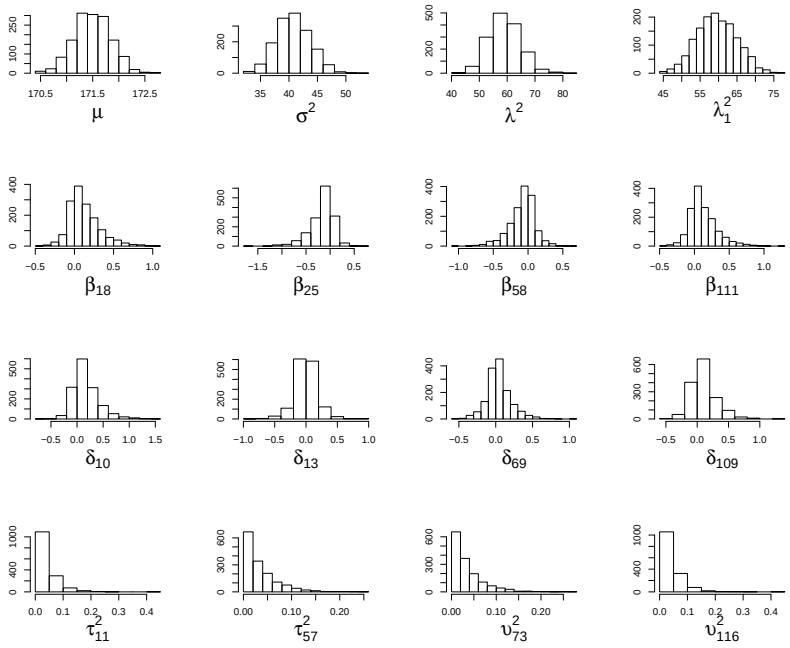

Figura 42 - Histograma dos valores gerados utilizando o Modelo I

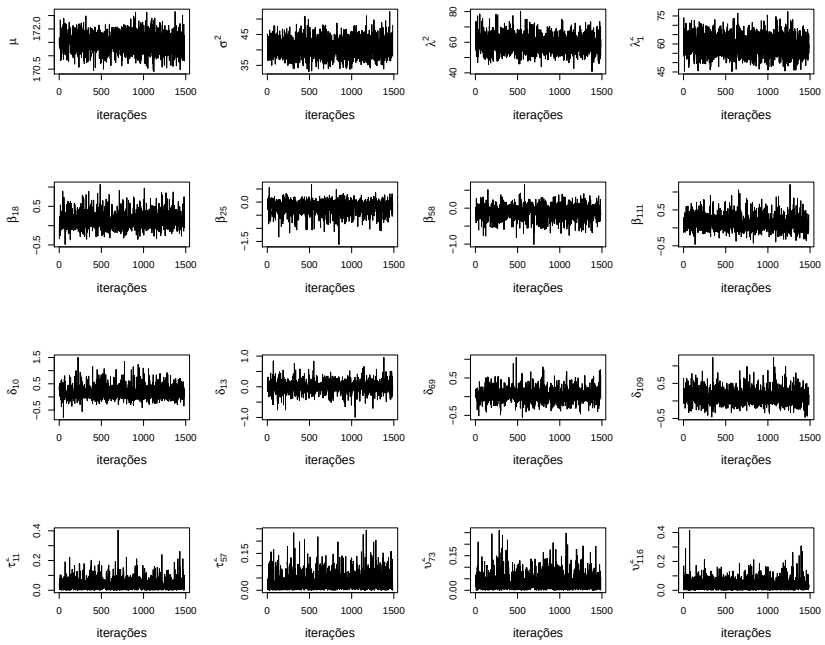

Figura 43 - Trajetória das cadeias geradas utilizando o Modelo I

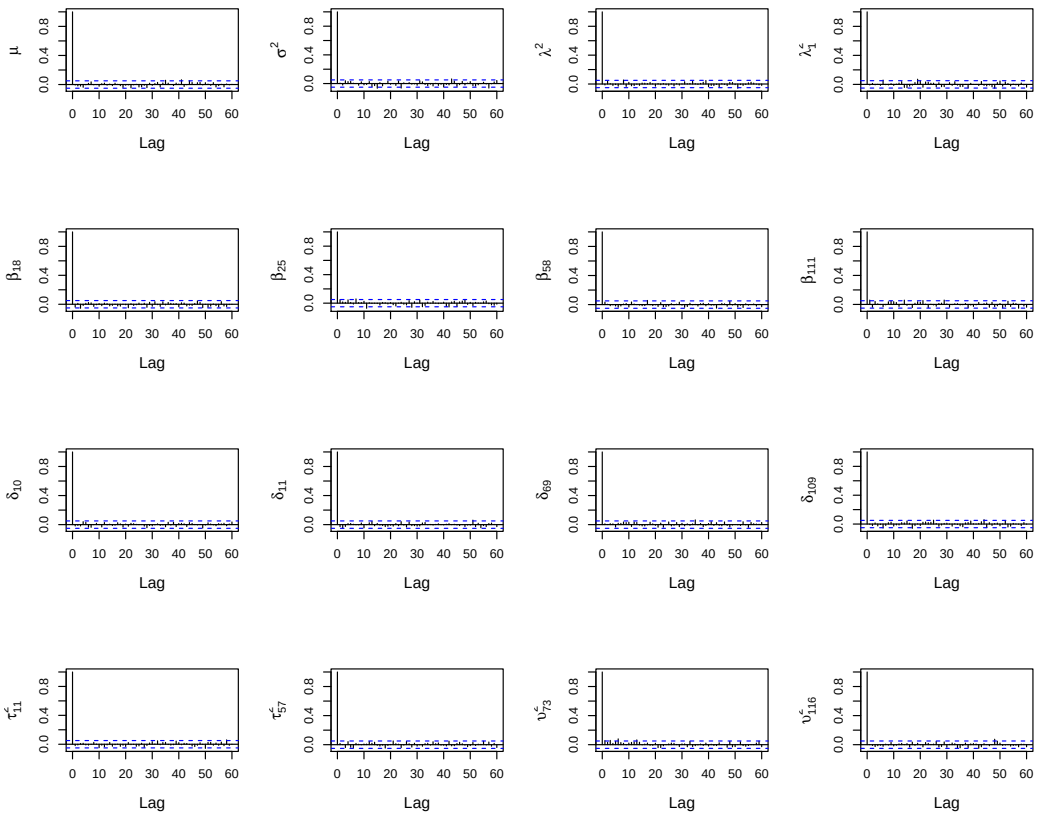

Figura 44 - Função de Autocorrelação das cadeias geradas utilizando o Modelo I

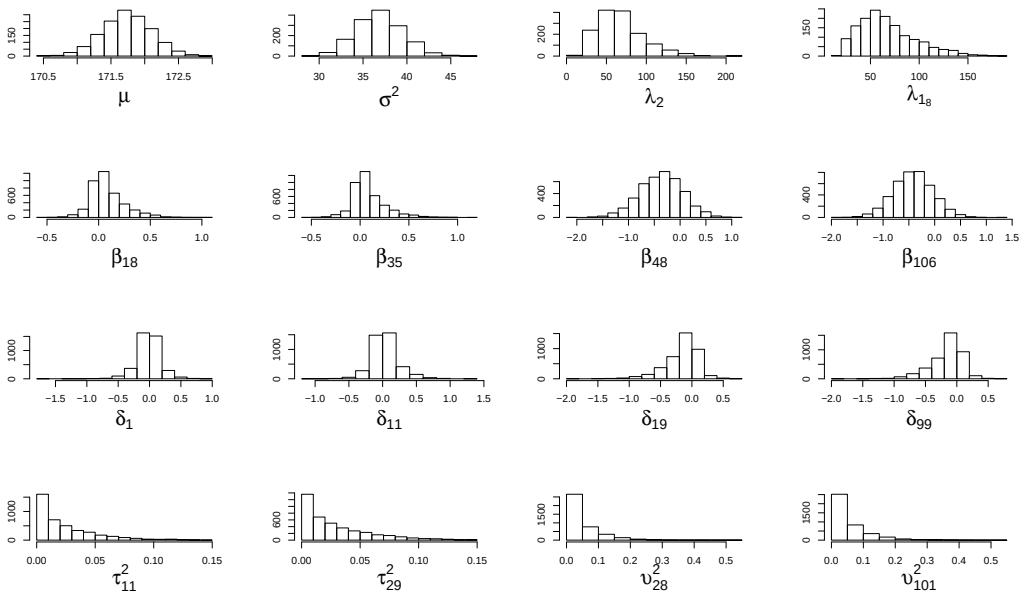

Figura 45 - Histograma dos valores gerados utilizando o Modelo II

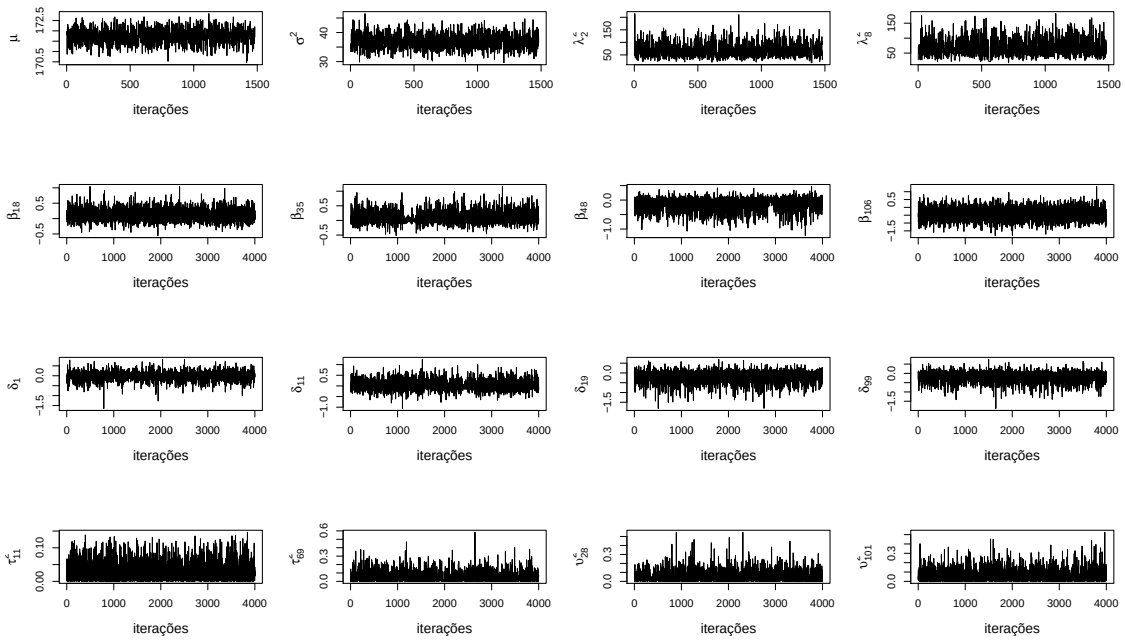

Figura 46 - Trajetória das cadeias geradas utilizando o Modelo II

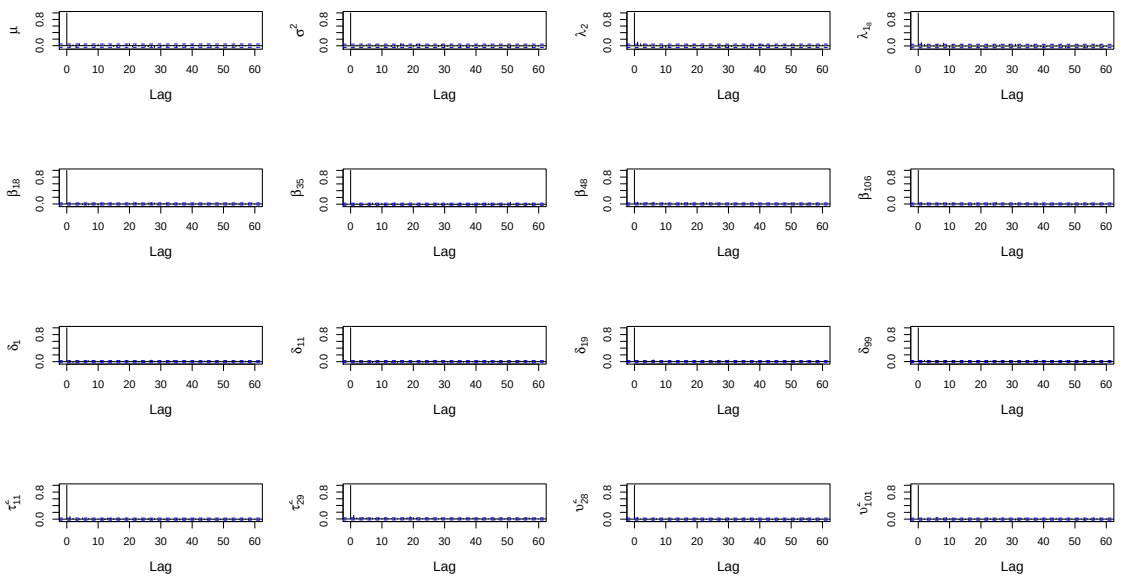

Figura 47 - Função de Autocorrelação das cadeias geradas utilizando o Modelo II

Observa-se nas Figuras 48 e 49 a mediana *a posteriori* dos efeitos aditivo e dominante estimados por meio dos Modelos I e II e à proporção da variância fenotípica explicada por cada efeito para os marcadores ao longo dos cromossomos, e nas Figuras de 30 e 31 os intervalos de credibilidade das estimativas dos parâmetros associados aos marcadores com evidências de associação com QTL.

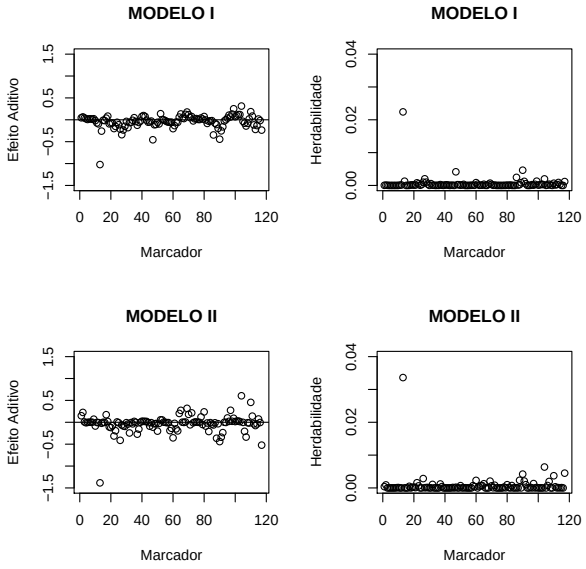

Figura 48 - Mediana *a posteriori* para o efeito aditivo de cada marcador e da herdabilidade

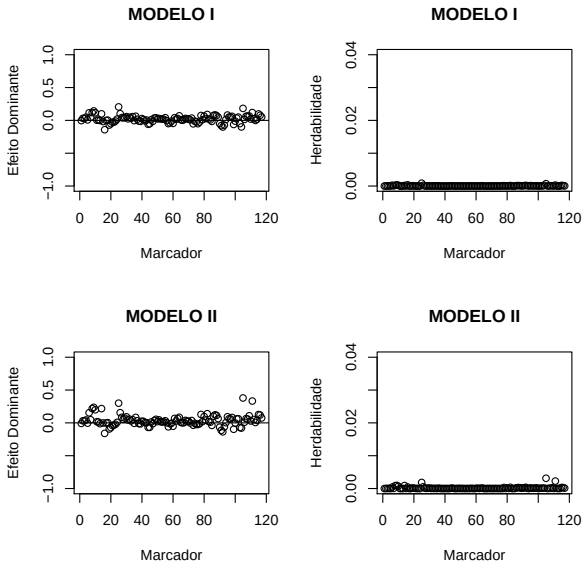

Figura 49 - Mediana *a posteriori* para o efeito dominante de cada marcador e da herdabilidade

Na análise do conjunto de dados referente à altura da planta foram ajustados os dois modelos, porém poucos marcadores foram selecionados com possíveis associações a QTL. Pelo Modelo I, o único marcador com possível associação utilizando o intervalo de credibilidade foi o marcador 13 do cromossomo 1, nomeado de bnlg0615. Já pelo Modelo II, foram selecionados 3 marcadores, os marcadores phi0037, umc1652 e dupssr13 que estão localizados nos cromossomos 1, 4 e 7 respectivamente. Observa-se na Tabela 13 os

marcadores que coincidiram com as análises do modelo CIM e o modelo bayesiano.

Tabela 13 - Cromossomos com forte evidência da existência de QTL, considerando 4 modelos diferentes

Modelo	QTL/Crom.[1]	Posição	Intervalo
CIM	1/1	79,38	bnlg2238-umc2025
	2/1	153,38	bnlg0615-phi0037
	3/2	53,88	umc1845-bnlg0166
	4/7	117,20	dupssr13-umc1154
Bayesiano	1/1	153,68	bnlg0615-phi0037
	2/2	53,29	umc1845-bnlg0166
	3/7	24,63	umc1428-umc1632
	4/7	121,77	dupssr13-umc1154
Modelo	Marcador/Crom.[2]	-	Marcador
I	13/1	-	bnlg0615
II	14/1	-	phi0037
	6/7	-	dupssr13

[1] Cromossomo onde está localizado o QTL.

[2]Cromossomo onde está localizado o marcador.

O marcador selecionado por meio do Modelo I e dois dos três selecionados por meio do Modelo II encontram-se na seleção do modelo CIM de Sibov et al (2003) e o modelo bayesiano de Meyer et al. (2009). Analisando os resultados do modelo bayesiano, percebe-se por meio deste modelo que há evidências da existência de QTL entre os marcadores umc1845-bnlg0166. Inspecionando os resultados das estimativas dos Modelos I e II para esses marcadores, observou-se que a estimativa foi expressiva e que LS do intervalo de credibilidade para o marcador bnlg0166 ultrapassou muito pouco o Zero.

Tabela 14 - Efeito e LI e LS dos intervalos para os marcadores bnlg0166 e umc1632

Modelo	Marcador	Efeito Dominante	LI	LS
I	bnlg0166	-0,3254	-1,1822	0,0636
I	umc1632	-0,3704	-1,1996	0,0633
II	bnlg0166	-0,6366	-1,7045	0,0390

4.3.4 Discussão

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

Os marcadores, em geral, que não estavam ligados a QTL no estudo de simulação, tiveram os efeitos estimados reduzidos para um valor o mais próximo possível de zero. Essa redução foi mais acentuada no Modelo II. Acredita-se que isso tenha ocorrido porque para este modelo considerou-se como distribuição a *priori* para as quantidades τ_j^2 a distribuição Half-Cauchy, que é extremamente concentrada no zero.

Observou-se na análise dos dados simulados que nem sempre os dois marcadores flanqueadores ao QTL são detectados, mas simplesmente um deles. Quando foi detectado apenas um marcador, foi exatamente o marcador em que a distância entre ele e o QTL era menor. Tomando como base as análises pelos modelos CIM e Bayesiano, isso se repetiu com os dados de milho tropical para a maioria dos casos. Observa-se na Tabela 10 que, de acordo com os modelos CIM e Bayesiano os marcadores bnlg1176 e bnlg1607 são marcadores flanqueadores de um QTL. Utilizando as posições do QTL informada na Tabela 10 e recorrendo ao mapa com a posição dos marcadores apresentado por Sibov et al. (2003) e que encontra-se no APÊNDICE B, percebe-se que o marcador que teria que

ser identificado, caso não identificasse os dois, teria que ser o bnlg1176. De fato, foi o que ocorreu com os Modelos I e II.

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

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

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

5 CONSIDERAÇÕES FINAIS

Os resultados obtidos nesta pesquisa levaram às seguintes conclusões:

1) Os Modelos I e II em geral apresentaram resultados semelhantes.

2) Os Modelos I e II quando comparados com o modelo Bayesiano proposto por Meyer et al.(2009) e o método CIM, apresentaram em comum alguns cromossomos com evidências de QTL.

3) Os dois modelos propostos neste trabalho apresentam algumas vantagens em relação a alguns métodos para mapeamento de QTL, como o CIM, IM e MIM, pois esses modelos são estatisticamente mais fáceis de entender e de implementar e o tempo computacional costuma ser menor.

Dessa forma, o desenvolvimento deste trabalho permite:

(i) recomendar a utilização de associação entre marcadores e QTL utilizando encolhimento bayesiano (*bayesian shrinkage*).

REFERÊNCIAS

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

BAE, K.; MALLICK, B. K. Gene selection using a two-level hierarchical bayesian model. **Bioinformatics**, Oxford, v.20, p.3423-3430, 2004.

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

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

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

CHE, X.; XU, S. Significance test and genome selection in bayesian shrinkage analysis. **International Journal of Plant Genomics**, New York, v.2010, p.1-11, 2010.

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

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

DUARTE, N. S. **Análise multivariada no mapeamento genético de traços quantitativos**. 2007. 125 p. (Mestrado em Estatística)- Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2007.

EFRON, B.; HASTIE, T.; JOHNSTONE.; TIBSHIRANI,R. Least angle regression. **Annals of Statistics**, Hayward, v.32, p.407-499, 2004.

EHLERS, R. S. **Introdução à Inferência Bayesiana**. Departamento de Estatística - UFPR, 2005. Disponível em: < http://leg.ufpr.br/ ehlers/bayes>. Acesso em: 15 dez. 2010.

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

FIGUEIREDO, M. A. T. Adaptive Sparseness for Supervised Learning. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, New York, v.25, p.1150-1159, 2003.

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

GELFAND, A.E. Gibbs Sampling. **Journal of the American Statistical Association**, Alexandria, v.95, n.452, p.1300-1304, 2000.

GELFAND, A.E.; SMITH , A. F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria , v.85, n.410, p.348-409, June 1990.

GELMAN, A. **Prior distributions for variance parameters in hierarquical models**. Bayesian Analysis, New York, v.1, n.3, p.515–533, 2006.

GELMAN, A.; CARLIN, J. B.; STER, H. S.; RUBIN, D.B. **Bayesian data analysis**. Boca Raton: Chapman & HAll/CRC, 2000. 526 p.

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

GELMAN, A.; RUBIN, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hoyward, v.7, p.457-511, 1992.

GEWEKE, J. Evaluating the accuracy of sampling-based approaches to calcualting posterior moments. **Bayesian Statistics**, New York: Oxford University, v.4, p.625-631, 1992.

GREEN, P. J. Reversible jump Markov Chain Monte Carlo Computation and Bayesian model determination. **Biometrika**, London, v.82, p.711-732, 1995.

HASTINGS,W. K. Monte Carlo Sampling methods using Markov chains and their application. **Biometrika**, London, v.57, n.1, p.97-109, 1970.

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

JEFREY, H. **Theory of probability**. Oxford: Clarendon Press, 1961. 447p.

JOHNSON, R.; KOTZ, S. **Continuos univariate distributions**. 2nd ed. Boston: Houghton Mifflin, 1995. v.2, 752 p.

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**, Alexandria, 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.

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

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

LANDER, E,S.; BOTSTEIN, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. **Genetics**, Austin, v.121, n.1, p.185-199, Jan. 1989.

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

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

METROPOLIS, N.; ROSEMBLUT, A. W.; ROSEMBLUT, M.N. ; TELLER, A. H.; TELLER , E. Equation of state calculations by fast computing machines. **Journal of Chemical Physics**, New York, v.21, p.1987-1092, 1953.

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

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

PAULINO, C.D.; TURKMAN, M. A.; MURTEIRA, B. **Estatística Bayesiana**. Lisboa: Fundação Calouste Gulbenkian, 2003. 446 p.

R DEVELOPMENT CORE TEAM. **R**: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Disponível em: <http://www.R-project.org>. Acesso em: 20 jan. 2011.

SABADIN, P. K.; de SOUZA JÚNIOR, C. L.; GARCIA, A. A. F. QTl mapping for yeield 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, Statistical Meeting, 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.

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

SIBOV, S. T.; de SOUZA JÚNIOR, C. L.; GARCIA, A. A. F.; GARCIA, A. F.; SILVA, A. R.; MANGOLIN, C.A.; BENCHIMOL, L.L.; SOUZA, A.P. Molecular mapping in tropical maize (*Zea mays L.*) using microsatellite markers. 2. Quantitative trait loci (QTL) for grain yield, plant height, ear height and grain moisture. **Hereditas**, Lund, v.139, 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, H. D. **Aspectos biométricos da detecção de QTLs (Quantitative Trait Loci) em espécies cultivadas.** 2001. 130 p. Tese (Doutorado em Estatística e Experimentação Agronômica) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2001.

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

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

SCHUSTER, I.; CRUZ, C. D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. Viçosa: UFV, 2004. 568 p.

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

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

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

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.; SHRINER,D. Advances in Bayesian multiple QTL mapping multiple in experimental designs. **Heredity**, New York, v.100, p.240-252, 2008.

Yi,N.; GEORGE,V.; and Allison,D.B. Stochastic search variable selection for identifying quantitative trait loci. **Genetics**, Austin, v.164, p.1129-1138, 2003.

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

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

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

YUAN, M.; LIN, Y. Model selection and estimation in regression with grouped variables. **Journal of the Royal Statistical Society**, London, Series B, v.68, p.49-67, 2005.

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

APÊNDICES

APÊNDICE A

A obtenção da distribuição de Laplace, especificada na expressão 15, por meio da equação 13 se dá da seguinte forma:

$$\begin{aligned}
\pi(\beta_j|\sigma^2) &= \int_0^\infty \pi(\beta_j|\sigma^2,\tau_j^2)\pi(\tau_j^2)d\tau_j \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2\tau_j^2}} e^{-\frac{1}{2\sigma^2\tau_j^2}\beta_j^2} \frac{\lambda^2}{2} e^{-\frac{\lambda^2\tau_j^2}{2}} d\tau_j^2 \\
&= \frac{\lambda}{2\sqrt{\sigma^2}} \int_0^\infty \frac{\lambda}{\sqrt{2\pi\tau_j^2}} e^{-\frac{\lambda^2}{2\tau_j^2}\left(\frac{\beta_j^2}{\lambda^2\sigma^2}+(\tau_j^2)^2\right)} d\tau_j^2 \\
&= \frac{\lambda}{2\sqrt{\sigma^2}} \int_0^\infty \frac{\lambda}{\sqrt{2\pi\tau_j^2}} e^{-\frac{\lambda^2}{2\tau_j^2}\left(\frac{\beta_j^2}{\lambda^2\sigma^2}+(\tau_j^2)^2\right)} \frac{e^{\frac{2\lambda^2}{2\tau_j^2}\tau_j^2\sqrt{\frac{\beta^2}{\lambda^2\sigma^2}}}}{e^{\frac{2\lambda^2}{2\tau_j^2}\tau_j^2\sqrt{\frac{\beta^2}{\lambda^2\sigma^2}}}} d\tau_j^2 \\
&= \frac{\lambda}{2\sqrt{\sigma^2}} \int_0^\infty \frac{\lambda}{\sqrt{2\pi\tau_j^2}} e^{-\frac{\lambda^2}{2\tau_j^2}\left[(\tau_j^2)^2-2\tau_j^2\sqrt{\frac{\beta^2}{\lambda^2\sigma^2}}+\left(\sqrt{\frac{\beta^2}{\lambda^2\sigma^2}}\right)^2\right]} e^{-\lambda\sqrt{\frac{\beta^2}{\sigma^2}}} d\tau_j^2 \\
&= \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda\frac{|\beta_j|}{\sqrt{\sigma^2}}} \int_0^\infty \frac{\lambda}{\sqrt{2\pi\tau_j^2}} e^{-\frac{\lambda^2}{2\tau_j^2}\left(\tau_j^2-\sqrt{\frac{\beta^2}{\lambda^2\sigma^2}}\right)^2} d\tau_j^2 \\
&= \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda\frac{|\beta_j|}{\sqrt{\sigma^2}}} \qquad (38)
\end{aligned}$$

Como os parâmetros $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ e $\boldsymbol{\tau^2} = (\tau_1^2, ..., \tau_p^2)$ são independentes, então

$$\begin{aligned}
\pi(\boldsymbol{\beta}|\sigma^2) &= \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda\frac{|\beta_1|}{\sqrt{\sigma^2}}} \times ... \times \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda\frac{|\beta_p|}{\sqrt{\sigma^2}}} \qquad (39) \\
&= \prod_{j=1}^{p} \frac{\lambda}{2\sqrt{\sigma^2}} e^{-\lambda|\beta_j|/\sqrt{\sigma^2}}
\end{aligned}$$

APÊNDICE B

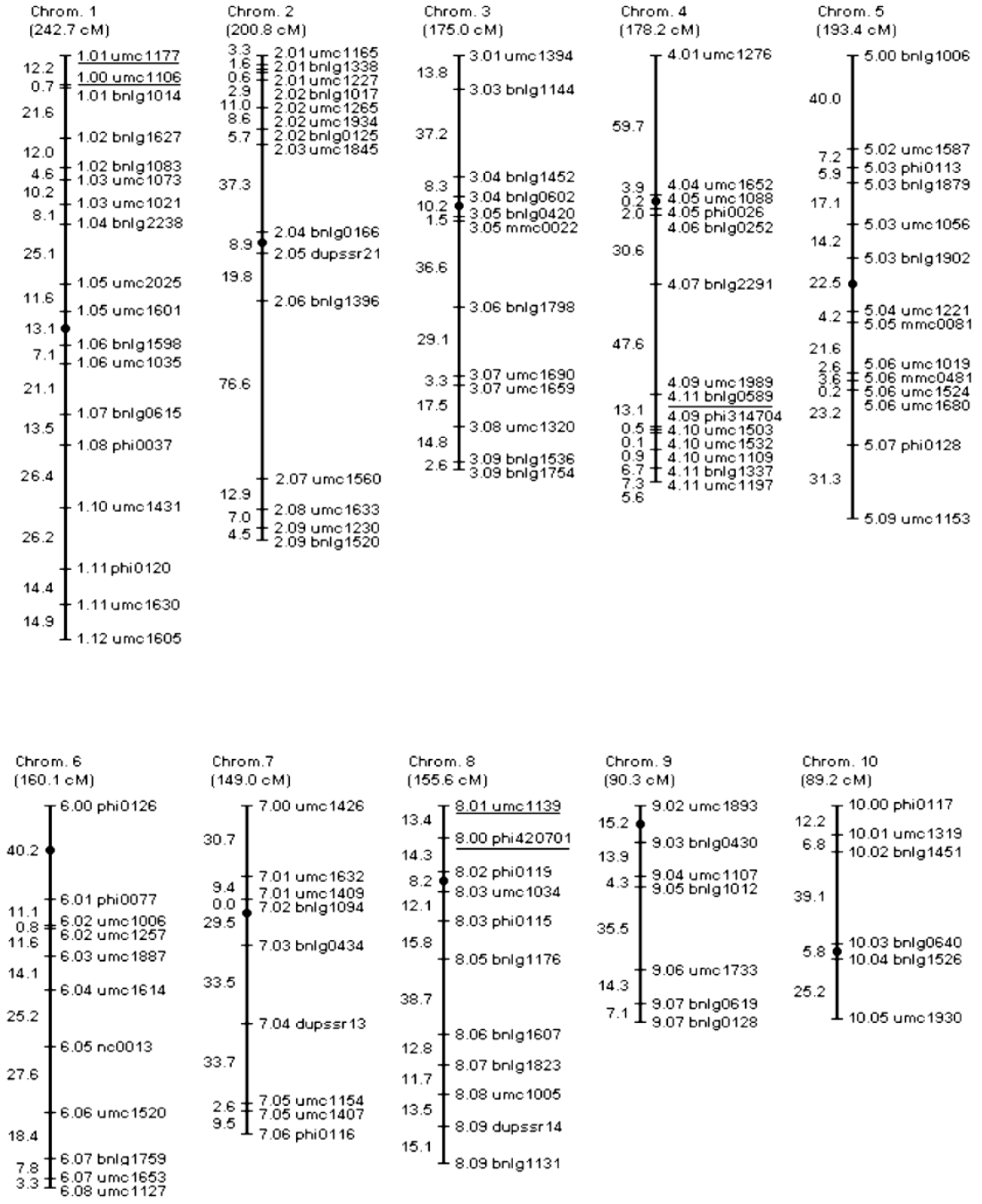

Figura 50 - Mapa de ligação com 117 locos de marcadores microsatélites distribuídos em dez grupos de ligação.

APÊNDICE C

Programa em C utilizado referente ao Modelo I.
Notação utilizada:
σ^2 = sig2; μ = mu; α = alpha; δ = delta; λ^2 = lambda2; λ_1^2 = lambda12; υ_j^2 = upsilonj; τ_j^2 = tauj.
Início do Programa

```
#include<R.h>
#include<Rmath.h>
#include<stdlib.h>

//Constantes Globais
#define n 400 // linhas de x (matriz com os genótipos dos marcadores)
#define coll 200 // colunas de x (números de marcadores)
#define iter 60000 // número de iterações

//Variáveis globais (Declaração)
//Ponteiros
float *Alocar_vetor_real (int na)
{
float *v; /* ponteiro para o vetor */
if (na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return (NULL);
}
/* aloca o vetor */
v = (float *) calloc (na, sizeof(float));
if (v == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
return (v); /* retorna o ponteiro para o vetor */
}
float *Liberar_vetor_real (float *v)
{
if (v == NULL) return (NULL);
free(v); /* libera o vetor */
return (NULL); /* retorna o ponteiro */
}

float **Alocar_matriz_real (int m, int na)
{
float **v; /* ponteiro para a matriz */
int i; /* variavel auxiliar */
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return (NULL);
}
/* aloca as linhas da matriz */
v = (float **) calloc (m, sizeof(float *)); /* Um vetor de m
ponteiros para float */
```

```
if (v == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
/* aloca as colunas da matriz */
for ( i = 0; i < m; i++ ) {
v[i] = (float*) calloc (na, sizeof(float)); /* m vetores de n floats */
if (v[i] == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
}
return (v); /* retorna o ponteiro para a matriz */
}

int **Alocar_matriz_inteira (int m, int na)
{
int **v2; /* ponteiro para a matriz */
int i; /* variavel auxiliar */
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return (NULL);
}
/* aloca as linhas da matriz */
v2 = (int **) calloc (m, sizeof(int *)); /* Um vetor de m
ponteiros para float */
if (v2 == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
/* aloca as colunas da matriz */
for ( i = 0; i < m; i++ ) {
v2[i] = (int*) calloc (na, sizeof(int)); /* m vetores de n floats */
if (v2[i] == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
}
return (v2); /* retorna o ponteiro para a matriz */
}


float **Liberar_matriz_real (int m, int na, float **v)
{
int i; /* variavel auxiliar */
if (v == NULL) return (NULL);
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return(v);
}
for (i=0; i<m; i++) free (v[i]); /* libera as linhas da matriz */
```

```
free(v); /* libera a matriz (vetor de ponteiros) */
return(NULL); /* retorna um ponteiro nulo */
}


int **Liberar_matriz_inteira (int m, int na, int **v2)
{
int i; /* variavel auxiliar */
if (v2 == NULL) return (NULL);
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return(v2);
}
for (i=0; i<m; i++) free (v2[i]); /* libera as linhas da matriz */
free(v2); /* libera a matriz (vetor de ponteiros) */
return(NULL); /* retorna um ponteiro nulo */
}

//Gerar valores da Inversa Gaussiana
double inversagauss(double mu, double lambda)
{

  double u, y, x1,mu2, l2;

  y = rnorm(0,1);
  y *= y;
  mu2 = mu * mu;
  l2 = 2.0*lambda;
  x1 = mu + mu2 *y/l2 - (mu/l2)* sqrt(4.0*mu*lambda*y + mu2 *y*y);

  u = runif(0,1);
  if(u <= mu/(mu + x1)) return(x1);
  else return(mu2/x1);

}

void programa(){

int i,j,l,c,h,k;
float m_alphaj[coll],var_alphaj[coll],alphak[coll];
float m_deltaj[coll],var_deltaj[coll];
float y[n],deltak[coll],prod1[n],prod2[n];
float alphaj[coll-1],deltaj[coll-1],aux[n],auxw[n];
float a,r,r1,b,a1,a2,a3,a4,b1,b2,b3,b4,sum,sig2mu,shape;
int iteracao,coluna,linha,colunaj;
//ponteiros
int **x,**xj;
float *mu,*sig2,*tau,*lambda2,**alpha,**delta,*lambda12;
float **upsilonj,**tauj,**m_upsilonj,**m_tauj,**w,**wj;
float aux1,aux2,aux6,cte1,var_mu,somay,m_mu;
float C,S,somay2,aux3,aux5,somax,somatau,somataud,somaw,calc;
```

```
iteracao=iter;
coluna=coll;
colunaj=coluna-1;
linha=n;

/* Importante: Inicializar ponteiros */
//vetores
mu = Alocar_vetor_real (iteracao);
sig2 = Alocar_vetor_real (iteracao);
tau = Alocar_vetor_real (iteracao);
lambda2 = Alocar_vetor_real (iteracao);
lambda12 = Alocar_vetor_real (iteracao);
//matrizes
alpha = Alocar_matriz_real (iteracao, coluna);
delta = Alocar_matriz_real (iteracao, coluna);
upsilonj = Alocar_matriz_real (iteracao, coluna);
tauj = Alocar_matriz_real (iteracao, coluna);
m_upsilonj = Alocar_matriz_real (iteracao, coluna);
m_tauj = Alocar_matriz_real (iteracao, coluna);
x = Alocar_matriz_inteira (linha, coluna);
xj = Alocar_matriz_inteira (linha, colunaj);
w = Alocar_matriz_real (linha, coluna);
wj = Alocar_matriz_real (linha, colunaj);
/* FIM */

//variáveis locais (se fizermos outras funções)
//--------------------------
a1=a2=3; b1=b2=2; a=5; b=5; a3=0.5; b3=0.5; sig2mu=1000;//alguns parâmetros.
a4=1; b4=0.5;

// Leitura dos arquivos
// Abre arquivo
FILE *ar;
ar=fopen("genotipo.txt","r");
FILE *ar2;
ar2=fopen("fenotipo.txt","r");
FILE *armu;
armu=fopen("mu.txt","w");
FILE *arsigma;
arsigma=fopen("sigma2.txt","w");
FILE *aralpha;
aralpha=fopen("alpha.txt","w");
FILE *ardelta;
ardelta=fopen("delta.txt","w");
FILE *arupsilonj;
arupsilonj=fopen("upsilonj.txt","w");
FILE *artauj;
artauj=fopen("tauj.txt","w");
FILE *arlambda2;
arlambda2=fopen("lambda2.txt","w");
FILE *arlambda12;
```

```
arlambda12=fopen("lambda12.txt","w");



if((ar == NULL)||(ar2==NULL)){
//Rprintf("Erro de leitura no txt, ar = %d ar2=d%, cuidado! \n", ar,ar2);
}
else{
// lê
for(i=0;i<n;i++){// linhas
fscanf(ar2,"%f",&y[i]);
for(j=0;j<coll;j++){
fscanf(ar,"%d",&x[i][j]);
x[i][j]--; // x=x-1
w[i][j]=(1+x[i][j])*(1-x[i][j])-0.5;
}
}
//FIM LEITURA


// Condições inicias para vetores e matrizes
for(i=0;i<iter;i++){
lambda2[i]=lambda12[i]=mu[i]=sig2[i]=0;
tau[i]=1;
}

for(j=0;j<coll;j++){
m_alphaj[j]=var_alphaj[j]=alphak[j]=m_deltaj[j]=var_deltaj[j]=deltak[j]=0;
}

for(i=0;i<iter;i++){
for(j=0;j<coll;j++){
alpha[i][j]=tauj[i][j]=delta[i][j]=upsilonj[i][j]=0;
}
}

//=======================================
// GERA A MÉDIA
// Calcular Média
sum=0;
for(i=0;i<n;i++){
sum=sum+y[i];
}
mu[0]=sum/n;

//=======================================
// GERA SIGMA2
//=======================================
GetRNGstate();
//srand(19821998);
```

```
sig2[0]=rgamma(a3,(1/b3)); // shape , scale

//=======================================
//%%%%%%%%%%%%%%LAMBDA%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lambda2[0]=rgamma(a3,(1/b3));
//#%%%%%%%%%%%%%%LAMBDA1%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lambda12[0]=rgamma(a3,(1/b3));
//#%%%%%%%%%%%alpha%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for(j=0;j<coll;j++){
alpha[0][j]=rnorm(0,1);
var_alphaj[j]=rgamma(0.5,0.5);
//##################Delta###########################################
delta[0][j]=rnorm(0,1);
var_deltaj[j]=rgamma(0.5,0.5);
//#%%%%%%%%%%TAU_J%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
upsilonj[0][j]=rgamma(a4,(1/b4));
tauj[0][j]=rgamma(a4,(1/b4));
}


// Atualização dos parâmetros
//=============================================================
for(i=1;i<iter;i++){
// produtos (l=linha,c=coluna)
//x%*%alpha[i-1,]  // prod1
//w%*%delta[i-1,] // prod2

for(l=0;l<n;l++){
aux1=0;
aux2=0;
for(c=0;c<coll;c++){
aux1=aux1+(x[l][c]*alpha[i-1][c]);
aux2=aux2+(w[l][c]*delta[i-1][c]);
}
prod1[l]=aux1;
prod2[l]=aux2;
}

//=====================================================================
// gerar mu
cte1= (n/sig2[i-1]) + (1/sig2mu);
var_mu=1/cte1;
// sum(y - (x%*%alpha[i-1,])-(w%*%delta[i-1,]))
somay=0;
for(l=0;l<n;l++){
somay=somay+y[l]-prod1[l]-prod2[l];
}
m_mu = somay/(sig2[i-1]*cte1);
mu[i]=rnorm(m_mu,sqrt(var_mu));
//=====================================================================
//gerar sig2
```

```
somay2=0;
for(l=0;l<n;l++){
somay2=somay2+pow((y[l]-mu[i]-prod1[l]-prod2[l]),2);
}
S = somay2/2. + b;
C = n/2. + a;
calc=(S/2. + 1./b);
tau[i] = rgamma(C,(1/calc));
sig2[i] = 1/tau[i];

//=========================================================================
// gerar alphaj
for(l=0;l<coll;l++){
// fazendo alphaj; xj;deltaj e wj
for(h=0;h<coll-1;h++){
if(h<l){
alphaj[h]=alphak[h];
for(k=0;k<n;k++){
xj[k][h]=x[k][h];
}
}
else if(h>l){
alphaj[h-1]=alphak[h];
for(k=0;k<n;k++){
xj[k][h-1]=x[k][h];
}
}
}//fim h

// fazendo xj%*%alphaj
for(k=0;k<n;k++){
aux[k]=0;
auxw[k]=0;
for(j=0;j<coll-1;j++){
aux[k]+=xj[k][j]*alphaj[j];
}
}

aux3=0; //sum(x[,l]*(y - mu[i,1]-(w%*%delta[i-1,]) - aux ))
somax=0;//sum(x[,l]^2
somatau=0;
for(k=0;k<n;k++){
somax+=pow(x[k][l],2);
aux3+=x[k][l]*(y[k]-mu[i]-prod2[k]-aux[k]);
}

m_alphaj[l]=aux3/(somax+ (sig2[i]/upsilonj[i-1][l])); //m.alphaj[1,l]
var_alphaj[l]=sig2[i]/(somax + (sig2[i]/upsilonj[i-1][l]));
alphak[l]= rnorm(m_alphaj[l],pow(var_alphaj[l],0.5));
alpha[i][l]=alphak[l];
//=========================================================================
```

```
//# gerar upsilonj
m_upsilonj[i][l]=sqrt(lambda2[i-1]/pow(alpha[i][l],2));
upsilonj[i][l]= 1/inversagauss(m_upsilonj[i][l],lambda2[i-1]);
somatau+=upsilonj[i][l]/2;
}// fim l
r=somatau + b1;


for(l=0;l<coll;l++){
//###obtendo delta
for(h=0;h<coll-1;h++){
if(h<l){
deltaj[h]=deltak[h];
for(k=0;k<n;k++){
wj[k][h]=w[k][h];
}
}
else if(h>l){
deltaj[h-1]=deltak[h];
for(k=0;k<n;k++){
wj[k][h-1]=w[k][h];
}
}
}//fim h

//=========================================================================
//gerar delta
// fazendo xj%*%alphaj e wj%*%deltaj
for(k=0;k<n;k++){
 auxw[k]=0;
for(j=0;j<coll-1;j++){
auxw[k]+=wj[k][j]*deltaj[j];// wj%*%deltaj
}
}

somaw=0; //sum(w[,l]^2)
aux5=0;//sum(w[,l]*(y - mu[i,1]-(x%*%alpha[i,]) - aux4 ))
somataud=0;
for(k=0;k<n;k++){
aux6=0;
for(c=0;c<coll;c++){
aux6=aux6+(x[k][c]*alpha[i][c]);
}
somaw+=pow(w[k][l],2);
aux5+=w[k][l]*(y[k]-mu[i]-aux6-auxw[k]);
}

m_deltaj[l]=aux5/(somaw+ (sig2[i]/tauj[i-1][l])); //m.deltaj[1,l]
var_deltaj[l]=sig2[i]/(somaw + (sig2[i]/tauj[i-1][l]));
deltak[l]= rnorm(m_deltaj[l],pow(var_deltaj[l],0.5));
```

```
delta[i][l]=deltak[l];
//=========================================================================
//# gerar tauj
m_tauj[i][l]=sqrt(lambda12[i-1]/pow(delta[i][l],2));
tauj[i][l]= 1/inversagauss(m_tauj[i][l],lambda12[i-1]);
somataud+=tauj[i][l]/2;
}// fim l2
r1=somataud + b2;
//=========================================================================
//#gerar lambda2
//#valor proposto
shape=coll + a1;
lambda2[i]=rgamma(shape,(1/r));
//=========================================================================
//#gerar lambda12
shape=coll + a2;
lambda12[i]=rgamma(shape,(1/r1));

}//Fim iter



for(k=0;k<iter;k++){
fprintf(armu,"%f \n",mu[k]);
fprintf(arsigma,"%f \n",sig2[k]);
fprintf(arlambda2,"%f \n",lambda2[k]);
fprintf(arlambda12,"%f \n",lambda12[k]);
for(h=0;h<coll;h++){
fprintf(aralpha,"%f ",alpha[k][h]);
fprintf(arupsilonj,"%f ",upsilonj[k][h]);
fprintf(ardelta,"%f ",delta[k][h]);
fprintf(artauj,"%f ",tauj[k][h]);
}
fprintf(aralpha,"\n");
fprintf(arupsilonj,"\n");
fprintf(ardelta,"\n");
fprintf(artauj,"\n");
}


/* Importante: Liberar ponteiros */
//vetores
mu = Liberar_vetor_real (mu);
sig2 = Liberar_vetor_real (sig2);
tau = Liberar_vetor_real (tau);
lambda2 = Liberar_vetor_real (lambda2);
lambda12 = Liberar_vetor_real (lambda12);
//matrizes
alpha = Liberar_matriz_real (iteracao, coluna, alpha);
delta = Liberar_matriz_real (iteracao, coluna, delta);
upsilonj = Liberar_matriz_real (iteracao, coluna, upsilonj);
```

```
tauj = Liberar_matriz_real (iteracao, coluna, tauj);
x = Liberar_matriz_inteira (linha, coluna, x);
m_upsilonj = Liberar_matriz_real (iteracao, coluna,m_upsilonj);
m_tauj = Liberar_matriz_real (iteracao, coluna,m_tauj);
xj = Liberar_matriz_inteira (linha, colunaj, xj);
w = Liberar_matriz_real (linha, coluna, w);
wj = Liberar_matriz_real (linha, colunaj, wj);

/* FIM */


PutRNGstate();

fclose(ar);
fclose(ar2);
fclose(armu);
fclose(arsigma);
fclose(aralpha);
fclose(ardelta);
fclose(arupsilonj);
fclose(artauj);
fclose(arlambda2);
fclose(arlambda12);

}//fim else leitura
}
```

Programa em C utilizado referente ao Modelo II.
Notação utilizada:
σ^2 = sig2; μ = mu; α = alpha; δ = delta; λ^2 = lambda2; λ_1^2 = lambda12; υ_j^2 = upsilonj; τ_j^2 = tauj.
Início do Programa

```
#include<R.h>
#include<Rmath.h>
#include<stdlib.h>

//Constantes Globais
#define n 400 // linhas de x
#define coll 117 // colunas de x
#define iter 11//8 // número de iterações
#define min(D,G)((D<G)?(D):(G))

//Variáveis globais (Declaração)
float *Alocar_vetor_real (int na)
{
float *v; /* ponteiro para o vetor */
if (na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return (NULL);
}
/* aloca o vetor */
v = (float *) calloc (na, sizeof(float));
if (v == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
return (v); /* retorna o ponteiro para o vetor */
}
float *Liberar_vetor_real (float *v)
{
if (v == NULL) return (NULL);
free(v); /* libera o vetor */
return (NULL); /* retorna o ponteiro */
}

float **Alocar_matriz_real (int m, int na)
{
float **v; /* ponteiro para a matriz */
int i; /* variavel auxiliar */
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return (NULL);
}
/* aloca as linhas da matriz */
v = (float **) calloc (m, sizeof(float *)); /* Um vetor de m ponteiros para float */
if (v == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
```

```
}
/* aloca as colunas da matriz */
for ( i = 0; i < m; i++ ) {
v[i] = (float*) calloc (na, sizeof(float)); /* m vetores de n floats */
if (v[i] == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
}
return (v); /* retorna o ponteiro para a matriz */
}

int **Alocar_matriz_inteira (int m, int na)
{
int **v2; /* ponteiro para a matriz */
int i; /* variavel auxiliar */
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return (NULL);
}
/* aloca as linhas da matriz */
v2 = (int **) calloc (m, sizeof(int *)); /* Um vetor de m ponteiros para float */
if (v2 == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
/* aloca as colunas da matriz */
for ( i = 0; i < m; i++ ) {
v2[i] = (int*) calloc (na, sizeof(int)); /* m vetores de n floats */
if (v2[i] == NULL) {
printf ("** Erro: Memoria Insuficiente **");
return (NULL);
}
}
return (v2); /* retorna o ponteiro para a matriz */
}


float **Liberar_matriz_real (int m, int na, float **v)
{
int i; /* variavel auxiliar */
if (v == NULL) return (NULL);
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return(v);
}
for (i=0; i<m; i++) free (v[i]); /* libera as linhas da matriz */
free(v); /* libera a matriz (vetor de ponteiros) */
return(NULL); /* retorna um ponteiro nulo */
}
```

```
int **Liberar_matriz_inteira (int m, int na, int **v2)
{
int i; /* variavel auxiliar */
if (v2 == NULL) return (NULL);
if (m < 1 || na < 1) { /* verifica parametros recebidos */
printf ("** Erro: Parametro invalido **\n");
return(v2);
}
for (i=0; i<m; i++) free (v2[i]); /* libera as linhas da matriz */
free(v2); /* libera a matriz (vetor de ponteiros) */
return(NULL); /* retorna um ponteiro nulo */
}

//Gerar valores da Inversa Gaussiana
double inversagauss(double mu, double lambda)
{

  double u, y, x1,mu2, l2;

  y = rnorm(0,1);
  y *= y;
  mu2 = mu * mu;
  l2 = 2.0*lambda;
  x1 = mu + mu2 *y/l2 - (mu/l2)* sqrt(4.0*mu*lambda*y + mu2 *y*y);

  u = runif(0,1);
  if(u <= mu/(mu + x1)) return(x1);
  else return(mu2/x1);

}



void programa(){
int i,j,l,c,h,k;
float m_alphaj[coll],var_alphaj[coll],alphak[coll],m_deltaj[coll];
float var_deltaj[coll],aceita1[coll],aceita2[coll];
float y[n],deltak[coll],prod1[n],prod2[n],aceita3[coll],aceita4[coll];
float alphaj[coll-1],deltaj[coll-1],aux[n],auxw[n];
float phi,phi1,a,b,a1,a2,b1,b2,a3,a4,a5,a6,sum,sig2mu,lambdastar1;
int iteracao,coluna,linha,colunaj;
//ponteiros
int **x,**xj;
float *mu,*sig2,*tau,**lambda2,**alpha,**delta,**lambda12;
float **upsilon,**tauj,**w,**wj;
float lamb1,lamb11,lamb22,v,lamb2,q,q1,v1;
double upsilon1,upsilon2,tauj1,tauj2,upsilonestcand;
double upsilonestpar,upsilonestcand1,upsilonestpar1;
double lambdastarcand,lambdastarpar,lambdastarcand1,lambdastarpar1;
float aux1,aux2,u22,aux6,cte1,var_mu,somay,m_mu,u1,u12,u2;
float C,S,somafenotipo,aux3,aux5,somax,somaw,lambdastar,calc,upsilonest,taujest;
```

```
iteracao=iter;
coluna=coll;
colunaj=coluna-1;
linha=n;

/* Importante: Inicializar ponteiros */
//vetores
mu = Alocar_vetor_real (iteracao);
sig2 = Alocar_vetor_real (iteracao);
tau = Alocar_vetor_real (iteracao);

//matrizes
alpha = Alocar_matriz_real (iteracao, coluna);
delta = Alocar_matriz_real (iteracao, coluna);
upsilon = Alocar_matriz_real (iteracao, coluna);
tauj = Alocar_matriz_real (iteracao, coluna);
lambda2 = Alocar_matriz_real (iteracao, coluna);
lambda12 = Alocar_matriz_real (iteracao, coluna);
x = Alocar_matriz_inteira (linha, coluna);
xj = Alocar_matriz_inteira (linha, colunaj);
w = Alocar_matriz_real (linha, coluna);
wj = Alocar_matriz_real (linha, colunaj);
/* FIM */


//variáveis locais
//=====================================================================
phi=phi1=2; a=5; b=5; a1=10; b1=0.1; sig2mu=1000;
a2=5; b2=5; a3=0.1; a4=10; a5=1; a6=0.5;

// Leitura dos arquivos w e y
// Abre arquivo
FILE *ar;
ar=fopen("genotipo.txt","r");
FILE *ar2;
ar2=fopen("fenotipo.txt","r");
FILE *araceita1;
araceita1=fopen("aceita1.txt","w");
FILE *araceita2;
araceita2=fopen("aceita2.txt","w");
FILE *araceita3;
araceita3=fopen("aceita3.txt","w");
FILE *araceita4;
araceita4=fopen("aceita4.txt","w");
FILE *armu;
armu=fopen("mu.txt","w");
FILE *arsigma;
arsigma=fopen("sigma2.txt","w");
FILE *aralpha;
aralpha=fopen("alpha.txt","w");
FILE *ardelta;
```

```
ardelta=fopen("delta.txt","w");
FILE *arupsilon;
arupsilon=fopen("upsilon.txt","w");
FILE *artauj;
artauj=fopen("tauj.txt","w");
FILE *arlambda2;
arlambda2=fopen("lambda2.txt","w");
FILE *arlambda12;
arlambda12=fopen("lambda12.txt","w");



if((ar == NULL)||(ar2==NULL)){
//Rprintf("Erro de leitura no txt, ar = %d ar2=d%, cuidado! \n", ar,ar2);
}
else{
// lê
for(i=0;i<n;i++){// linhas
fscanf(ar2,"%f",&y[i]);
for(j=0;j<coll;j++){
fscanf(ar,"%d",&x[i][j]);
x[i][j]--; // x=x-1
w[i][j]=(1+x[i][j])*(1-x[i][j])-0.5;
}
}

//FIM LEITURA

// Condições inicias para vetores e matrizes
for(i=0;i<iter;i++){
mu[i]=sig2[i]=0;
tau[i]=1;
}

for(j=0;j<coll;j++){
aceita1[j]=aceita2[j]=aceita3[j]=aceita4[j]=m_alphaj[j]=var_alphaj[j]=0;
alphak[j]=m_deltaj[j]=var_deltaj[j]=deltak[j]=0;
}

for(i=0;i<iter;i++){
for(j=0;j<coll;j++){
alpha[i][j]=tauj[i][j]=delta[i][j]=upsilon[i][j]=lambda2[i][j]=lambda12[i][j]=0;
}
}

//=====================================================================
// GERA A MÉDIA
// Calcular Média
sum=0;
for(i=0;i<n;i++){
sum=sum+y[i];
```

```
}
mu[0]=sum/n;
//Rprintf("media  = %lf \n \n",mu[0]); //teste
//=====================================================================
// GERA SIGMA2
//=====================================================================
GetRNGstate();
//srand(19821998);

sig2[0]=rgamma(a1,(1/b1)); // shape , scale
//Rprintf("sig2[0] = %lf \n \n",sig2[0]); //teste
/* teste
for(j=0;j<*iteste;j++){
teste[j]=rgamma(a1,(1/b1));
}
*/
//=====================================================================


//#%%%%%%%%%%%%%%%alpha%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
for(j=0;j<coll;j++){
alpha[0][j]=rnorm(0,1);
var_alphaj[j]=rgamma(0.5,0.5);
//%%%%%%%%%%%%%%%%Delta%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
delta[0][j]=rnorm(0,1);
var_deltaj[j]=rgamma(0.5,0.5);
//#%%%%%%%%%%upsilon,tauj,lambda2,lambda12%%%%%%%%%%%%%%%%%%%%%%%%%%%%
upsilon[0][j]=inversagauss(a3,(1/a4));    //rinvgauss(a3,a4)
tauj[0][j]=inversagauss(a3,(1/a4));   //rinvgauss(a3,a4)
lambda2[0][j]=rgamma(a1,(1/b1));
lambda12[0][j]=rgamma(a1,(1/b1));
}


// Atualização dos parâmetros
//=====================================================================
for(i=1;i<iter;i++){
// produtos (l=linha,c=coluna)
//x%*%alpha[i-1,]  // prod1
//w%*%delta[i-1,] // prod2

for(l=0;l<n;l++){
aux1=0;
aux2=0;
for(c=0;c<coll;c++){
aux1=aux1+(x[l][c]*alpha[i-1][c]);
aux2=aux2+(w[l][c]*delta[i-1][c]);
}
prod1[l]=aux1;
prod2[l]=aux2;
}
```

```
//=========================================================================
// gerar mu
cte1= (n/sig2[i-1]) + (1/sig2mu);
var_mu=1/cte1;
// sum(y - (x%*%alpha[i-1,])-(w%*%delta[i-1,]))
somay=0;
for(l=0;l<n;l++){
somay=somay+y[l]-prod1[l]-prod2[l];
}
m_mu = somay/(sig2[i-1]*cte1);
mu[i]=rnorm(m_mu,sqrt(var_mu));
somay2=0;
for(l=0;l<n;l++){
somafenotipo=somafenotipo+pow((y[l]-mu[i]-prod1[l]-prod2[l]),2);
}
S = somay2/2. + b;
C = n/2. + a;
calc=(S/2. + 1./b);
tau[i] = rgamma(C,(1/calc));
sig2[i] = 1/tau[i];

//=========================================================================
// gerar alphaj
for(l=0;l<coll;l++){
// fazendo alphaj; xj;deltaj e wj
for(h=0;h<coll-1;h++){
if(h<l){
alphaj[h]=alphak[h];
for(k=0;k<n;k++){
xj[k][h]=x[k][h];
}
}
else if(h>l){
alphaj[h-1]=alphak[h];
for(k=0;k<n;k++){
xj[k][h-1]=x[k][h];
}
}
}//fim h

// fazendo xj%*%alphaj e wj%*%deltaj
for(k=0;k<n;k++){
aux[k]=0;
auxw[k]=0;
for(j=0;j<coll-1;j++){
aux[k]+=xj[k][j]*alphaj[j];
}
}

aux3=0; //sum(x[,l]*(y - mu[i,1]-(w%*%delta[i-1,]) - aux ))
somax=0;//sum(x[,l]^2)
```

```
for(k=0;k<n;k++){
somax+=pow(x[k][l],2);
aux3+=x[k][l]*(y[k]-mu[i]-prod2[k]-aux[k]);
}

m_alphaj[l]=aux3/(somax+ (sig2[i]/upsilon[i-1][l])); //m.alphaj[1,l]
var_alphaj[l]=sig2[i]/(somax + (sig2[i]/upsilon[i-1][l]));
alphak[l]= rnorm(m_alphaj[l],pow(var_alphaj[l],0.5));

alpha[i][l]=alphak[l];
//=====================================================================
//# gerar upsilon
//# valor proposto
upsilonest=rgamma(a3,(1/a4));
//#upsilonest<-rinvgauss(coll,a3,a4)
//#Condicional Proposta
upsilon1=1/(sqrt(upsilonest)*(pow(upsilonest,2)+pow(lambda2[i-1][l],2)));
upsilon1*=exp(-1/(2*upsilonest)*pow(alpha[i][l],2));
//#Condicional Inicial
upsilon2=1/(pow(upsilon[i-1][l],0.5)*(pow(upsilon[i-1][l],2)+pow(lambda2[i-1][l],2)));
upsilon2*=exp(-1/(2*upsilon[i-1][l])*pow(alpha[i][l],2));
upsilonestcand=dgamma(upsilonest,a3,1/a4,1);
upsilonestpar=dgamma(upsilon[i-1][l],a3,1/a4,1);
q=(upsilon1*upsilonestcand)/(upsilon2*upsilonestpar);
u1=runif(0.,1.);
if(((u1<1)&&(q>=1))||((u1<q)&&(1>=q))){
upsilon[i][l]=upsilonest;
aceita1[l]=aceita1[l]+1;
}
else{
upsilon[i][l]=upsilon[i-1][l];
}
//=====================================================================
//#gerar lambda2
//#valor proposto
lambdastar=rgamma(a1,(1/b1));
//#Condicional Proposta
lamb1=(1/(pow(phi,2.)+pow(lambdastar,2.)));
lamb1*=(lambdastar/(pow(lambdastar,2.)+pow(upsilon[i][l],2.)));
//#Condicional Inicial
lamb2=(1/(pow(phi,2.)+pow(lambda2[i-1][l],2.)));
lamb2*=(lambda2[i-1][l]/(pow(lambda2[i-1][l],2.)+pow(upsilon[i][l],2.)));
lambdastarcand=dgamma(lambdastar,a1,1/b1,1);
lambdastarpar=dgamma(lambda2[i-1][l],a1,1/b1,1);
v=(lamb1*lambdastarcand)/(lamb2*lambdastarpar);
//Rprintf("v=%f\n",v);
u2=runif(0.,1.);
if(u2<min(1.,v)){
lambda2[i][l]=lambdastar;
aceita3[l]=aceita3[l]+1;
}
```

```
else{
lambda2[i][l]=lambda2[i-1][l];
}
}// fim l
//=========================================================================
//gerar delta
for(l=0;l<coll;l++){
for(h=0;h<coll-1;h++){
if(h<l){
deltaj[h]=deltak[h];
for(k=0;k<n;k++){
wj[k][h]=w[k][h];
}
}
else if(h>l){
deltaj[h-1]=deltak[h];
for(k=0;k<n;k++){
wj[k][h-1]=w[k][h];
}
}
}//fim h

// fazendo xj%*%alphaj e wj%*%deltaj
for(k=0;k<n;k++){
 auxw[k]=0;
for(j=0;j<coll-1;j++){
auxw[k]+=wj[k][j]*deltaj[j];// wj%*%deltaj
}
}

somaw=0; //sum(w[,l]^2)
aux5=0;//sum(w[,l]*(y - mu[i,1]-(x%*%alpha[i,]) - aux4 ))
for(k=0;k<n;k++){
aux6=0;
for(c=0;c<coll;c++){
aux6=aux6+(x[k][c]*alpha[i][c]);
}
somaw+=pow(w[k][l],2);
aux5+=w[k][l]*(y[k]-mu[i]-aux6-auxw[k]);
}

m_deltaj[l]=aux5/(somaw+ (sig2[i]/tauj[i-1][l])); //m.deltaj[1,l]
var_deltaj[l]=sig2[i]/(somaw + (sig2[i]/tauj[i-1][l]));
deltak[l]= rnorm(m_deltaj[l],pow(var_deltaj[l],0.5));

delta[i][l]=deltak[l];
//=========================================================================
//# gerar tauj
//# valor proposto
taujest=rgamma(a3,(1/a4));
//#Condicional Proposta
```

```
tauj1=1/(pow(taujest,0.5)*(pow(taujest,2)+pow(lambda12[i-1][l],2)));
tauj1*=exp(-1/(2*taujest)*pow(delta[i][l],2));
//#Condicional Inicial
tauj2=1/(pow(tauj[i-1][l],0.5)*(pow(tauj[i-1][l],2)+pow(lambda12[i-1][l],2)));
tauj2*=exp(-1/(2*tauj[i-1][l])*pow(delta[i][l],2));
upsilonestcand1=dgamma(taujest,a5,1/a6,1);
upsilonestpar1=dgamma(tauj[i-1][l],a5,1/a6,1);
q1=(tauj1*upsilonestcand1)/(tauj2*upsilonestpar1);
u12=runif(0.,1.);
//Rprintf("u12=%f\n",u12);
if(u12<min(1.,q1)){
tauj[i][l]=taujest;
aceita2[l]=aceita2[l]+1;
}
else{
tauj[i][l]=tauj[i-1][l];
}
//=====================================================================
//#gerar lambda2
//#valor proposto
lambdastar1=rgamma(a1,(1/b1));
//#Condicional Proposta
lamb11=(1/(pow(phi1,2.)+pow(lambdastar1,2.)));
lamb11*=(lambdastar1/(pow(lambdastar1,2.)+pow(tauj[i][l],2.)));
//#Condicional Inicial
lamb22=(1/(pow(phi1,2.)+pow(lambda12[i-1][l],2.)));
lamb22*=(lambda12[i-1][l]/(pow(lambda12[i-1][l],2.)+pow(tauj[i][l],2.)));
lambdastarcand1=dgamma(lambdastar1,a1,1/b1,1);
lambdastarpar1=dgamma(lambda12[i-1][l],a1,1/b1,1);
v1=(lamb11*lambdastarcand1)/(lamb22*lambdastarpar1);
u22=runif(0.,1.);
if(u2<min(1.,v1)){
lambda12[i][l]=lambdastar1;
aceita4[l]=aceita4[l] +1;
}
else{
lambda12[i][l]=lambda12[i-1][l];
}
}// fim l 2
}//Fim iter


for(h=0;h<coll;h++){
fprintf(araceita1,"%f \n",aceita1[h]);
fprintf(araceita2,"%f \n",aceita2[h]);
fprintf(araceita3,"%f \n",aceita3[h]);
fprintf(araceita4,"%f \n",aceita4[h]);
}

for(k=0;k<iter;k++){
fprintf(armu,"%f \n",mu[k]);
```

```
fprintf(arsigma,"%f \n",sig2[k]);
for(h=0;h<coll;h++){
fprintf(aralpha,"%f ",alpha[k][h]);
fprintf(arupsilon,"%f ",upsilon[k][h]);
fprintf(ardelta,"%f ",delta[k][h]);
fprintf(artauj,"%f ",tauj[k][h]);
fprintf(arlambda2,"%f ",lambda2[k][h]);
fprintf(arlambda12,"%f ",lambda12[k][h]);
}
fprintf(aralpha,"\n");
fprintf(arupsilon,"\n");
fprintf(ardelta,"\n");
fprintf(artauj,"\n");
fprintf(arlambda2,"\n");
fprintf(arlambda12,"\n");
}


/* Importante: Liberar ponteiros */
//vetores
mu = Liberar_vetor_real (mu);
sig2 = Liberar_vetor_real (sig2);
tau = Liberar_vetor_real (tau);
//matrizes
alpha = Liberar_matriz_real (iteracao, coluna, alpha);
delta = Liberar_matriz_real (iteracao, coluna, delta);
upsilon = Liberar_matriz_real (iteracao, coluna, upsilon);
tauj = Liberar_matriz_real (iteracao, coluna, tauj);
lambda2 = Liberar_matriz_real (iteracao, coluna, lambda2);
lambda12 = Liberar_matriz_real (iteracao, coluna, lambda12);
x = Liberar_matriz_inteira (linha, coluna, x);
xj = Liberar_matriz_inteira (linha, colunaj, xj);
w = Liberar_matriz_real (linha, coluna, w);
wj = Liberar_matriz_real (linha, colunaj, wj);

/* FIM */



PutRNGstate();

fclose(ar);
fclose(ar2);
fclose(armu);
fclose(arsigma);
fclose(aralpha);
fclose(ardelta);
fclose(arupsilon);
fclose(artauj);
fclose(arlambda2);
fclose(arlambda12);
```

```
fclose(araceita1);
fclose(araceita2);
fclose(araceita3);
fclose(araceita4);

}//fim else leitura
}
```