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

**Uma abordagem Bayesiana para o mapeamento de QTLs utilizando o método
MCMC com saltos reversíveis**

Joseane Padilha da Silva

Dissertação apresentada para obtenção de título
de Mestre em Agronomia. Área de concentração:
Estatística e Experimentação Agronômica.

Piracicaba
2006

Joseane Padilha da Silva
Estatístico

**Uma abordagem Bayesiana para o mapeamento de QTLs utilizando o método MCMC
com saltos reversíveis**

Orientador:
Profª Drª **ROSELI APARECIDA LEANDRO**

Dissertação apresentada, para obtenção de título
de Mestre em Agronomia. Área de concentração:
Estatística e Experimentação Agronômica

**Piracicaba
2006**

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Silva, Joseane Padilha da

Uma abordagem Bayesiana para o mapeamento de QTLs utilizando o método MCMC com saltos reversíveis / Joseane Padilha da Silva. - - Piracicaba, 2006.
80 p. : il.

Dissertação (Mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2007.
Bibliografia.

1. Inferência bayesiana 2. Genética estatística 3. Mapeamento genético 4. Marcador molecular 5. Métodos MCMC I. Título

CDD 519.542

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

DEDICATÓRIA

A Deus, pela iluminação e inspiração.

Aos meus pais,

Gilberto Gomes da Silva e Maria Aparecida Padilha da Silva, meus anjos terrenos; vocês representam a base desse trabalho.

Ao meu irmão,

Ricardo Padilha da Silva e esposa **Jaqueline Rodrigues**, pela amizade e grandes incentivos.

A minha sobrinha,

Milinha, por me proporcionar tantos momentos de alegria e descontração.

Aos meus avós,

Martiniano Gomes da Silva e Florinda Izabel de Souza, João Padilha e Júlia Elias Honorato Padilha, pelo incentivo e grande amor.

AGRADECIMENTOS

A professora Dr^a. **Dra. Roseli Aparecida Leandro**, pela enorme paciência, pela amizade e principalmente pela orientação à elaboração deste trabalho.

Ao CNPq pela concessão de bolsa de estudo.

Ao conselho do programa de Pós-graduação em Estatística e Experimentação Agronômica, os professores Dra. **Clarice Garcia Borges Demétrio**, Dr. **Décio Barbin** e Dr. **Edwin Moises Marcos Ortega** pelas sugestões.

Aos professores do Departamento de Ciências Exatas da ESALQ/USP, Dr. **César Gonçalves de Lima**, Dr. **Gerson Barretos Mourão**, e Dr. **Silvio Sandoval Zocchi**, pela amizade e formação.

Aos funcionários do Departamento de Ciências Exatas da ESALQ/USP, as secretárias **Solange de Assis Paes Sabadin** e **Luciane Brajão** e aos técnicos em informática **Jorge Alexandre Wiendl** e **Eduardo Bonilha**, pelos auxílios permanentes.

Aos colegas de estudo do mestrado e do doutorado, pela amizade, companheirismo e paciência.

Um agradecimento especial aos colegas, **Alexandre, Ana Paula, Cristiane, Fernanda, Juliana e Pâmela**, pelo incentivo, motivação, companheirismo, paciência e principalmente a grande amizade dedicada.

A **Cecilia**, pelo amor e paciência.

A minha amiga e irmã do coração **Andréia**, pela amizade incondicional.

Aos eternos amigos de graduação **Adriana, Ana Carolina, Felipe e Marcelo**, pelos momentos inesquecíveis; e as amigas de infância **Cris, Dariene, Sheila, Luciene e Rosana**.

A todas as pessoas que contribuíram direta ou indiretamente para a realização deste trabalho.

SUMÁRIO

	Página
RESUMO	8
ABSTRACT	9
1 INTRODUÇÃO	10
2 REVISÃO BIBLIOGRÁFICA	12
2.1 Conceitos Básicos de Genética	12
2.2 Marcadores Moleculares	14
2.3 Seleção de Mapeamento	16
2.3.1 Frequência de Recombinação e Distância no Mapa	18
2.4 Estabelecimento dos Grupos de Ligação	22
2.5 Determinação da Ordem dos Marcadores nos Grupos de Ligação	23
2.6 Abordagem Bayesiana	24
2.6.1 Visão Geral	24
2.6.2 Distribuições a Priori	25
2.6.2.1 Prioris Informativas	25
2.6.2.2 Prioris não Informativas	27
2.6.3 Métodos de Aproximação Baseados em Amostragem	28
2.6.3.1 Metodologia MCMC	29
2.6.3.2 Gibbs Sampling	30
2.6.3.3 Algoritmo de Metropolis-Hastings	32
2.6.4 Fator de Bayes	33
2.7 Mapeamento de QTL	34
2.7.1 Modelo QTL	35
2.8 Análise com número de QTLs conhecido	36

		6
2.9	Análise com número de QTLs desconhecido	38
3	MATERIAIS E MÉTODOS	40
3.1	Materiais	40
3.2	Métodos	40
3.2.1	MCMC com Saltos Reversíveis	42
3.2.2	Distribuições a Priori	45
3.2.3	Construção da amostra da distribuição conjunta a posteriori	47
3.2.3.1	Distribuições necessárias para a atualização dos parâmetros	47
4	RESULTADOS E DISCUSSÕES	50
4.1	Um QTL Afetando a Característica de Interesse	51
4.1.1	Distribuição a Priori Poisson	51
4.1.2	Distribuição a Priori Geométrica	53
4.1.3	Distribuição a Priori Uniforme	54
4.2	Dois QTLs Afetando a Característica de Interesse	56
4.2.1	Distribuição a Priori Poisson	56
4.2.2	Distribuição a Priori Geométrica	58
4.2.3	Distribuição Uniforme	59
4.3	Quatro QTLs Afetando a Característica de Interesse	61
4.3.1	Distribuição a Priori Poisson	61
4.3.2	Distribuição a Priori Geométrica	63
4.3.3	Distribuição a Priori Uniforme	65
4.4	Oito QTLs Afetando a Característica de Interesse	67
4.4.1	Distribuição a Priori Poisson	67
4.4.2	Distribuição a Priori Geométrica	68
4.4.3	Distribuição a Priori Uniforme	68

	7
4.5 Conclusões	69
5 CONSIDERAÇÕES FINAIS	71
REFERÊNCIAS	72
ANEXO	76

RESUMO

Uma abordagem Bayesiana para o mapeamento de QTLs utilizando o método MCMC com saltos reversíveis

A utilização de metodologias Bayesianas tem se tornado freqüente nas aplicações em Genética, em particular em mapeamento de QTLs usando marcadores moleculares. Mapear um QTL implica em identificar sua posição no genoma, bem como seus efeitos genéticos. A abordagem Bayesiana combina, através do Teorema de Bayes, a verossimilhança dos dados fenotípicos com distribuições a priori atribuídas a todos os parâmetros desconhecidos (número, localização e efeito do QTL) induzindo distribuições a posteriori a respeito dessas quantidades. Métodos de mapeamento Bayesiano podem tratar o número desconhecido de QTLs como uma variável aleatória, resultando em complicações na obtenção da amostra aleatória da distribuição conjunta a posteriori, uma vez que a dimensão do espaço do modelo pode variar. O Método MCMC com Saltos Reversíveis (MCMC-SR), proposto por Green(1995), é excelente para explorar distribuições a posteriori nesse contexto. O método proposto foi avaliado usando dados simulados no WinQTLCart, onde o maior objetivo foi avaliar diferentes prioris atribuídas para o número de QTLs.

Palavras-chave: mapeamento de QTLs, MCMC, MCMC com Saltos Reversíveis

ABSTRACT

A Bayesian approach to detect quantitative trait loci using reversible-jump MCMC

The use of Bayesian methodology in genetical applications has grown increasingly popular, in particular in the analysis of quantitative trait loci (QTL) for studies using molecular markers. In such analyses the aim is mapping QTLs, estimating their locations in the genome and their genotypic effects. The Bayesian approach proceeds by setting up a likelihood function for the phenotype and assigning prior distributions to all unknowns in the problem (number of QTL, chromosome, locus, genetics effects). These induce a posterior distribution on the unknown quantities that contains all of the available information for inference of the genetic architecture of the trait. Bayesian mapping methods can treat the unknown number of QTL as a random variable, which has several advantages but results in the complication of varying the dimension of the model space. The reversible jump MCMC algorithm offers a powerful and general approach to exploring posterior distributions in this setting. The method was evaluated by analyzing simulated data, where the major goal was evaluate if different priors distributions on the QTL numbers.

Keywords: mapping QTLs, MCMC, reversible jump MCMC

1 INTRODUÇÃO

A utilização de metodologias estatísticas tem se tornado um componente indispensável em análises de dados nas mais diversas áreas de conhecimento, sendo imprescindível recorrer ao auxílio delas para desenvolver pesquisas confiáveis e tomar decisões importantes. Dentre as diversas áreas que a estatística pode atuar, destaca-se o ramo da Biologia que estuda a herança e a variação de caracteres quantitativos: a Genética Quantitativa.

Caracteres quantitativos são aqueles cuja expressão fenotípica apresenta variações contínuas, atribuídas à segregação simultânea de muitos genes distribuídos pelo genoma, em regiões definidas como QTLs (“Quantitative Trait Loci”). Com o apoio de modelagens estatísticas pode-se realizar o mapeamento de QTLs. Mapear um QTL significa identificar quantos QTLs estão presentes no genoma, sua posição e estimar seus efeitos. É importante salientar, que um QTL pode apresentar efeitos de diferentes magnitudes.

Na Agronomia, é economicamente relevante localizar e estimar os efeitos genéticos de QTLs responsáveis pela variação genética de caracteres tais como: produção de grãos, altura da planta, teor de proteína, resistência a doenças, ganho de peso, consumo alimentar, produção de leite etc, que resultam da ação cumulativa de um conjunto de genes.

Conforme Paterson (1998) do ponto de vista prático, os resultados obtidos através do mapeamento de QTL podem auxiliar os pesquisadores no melhoramento genético de plantas ou animais e/ou responder questões básicas sobre os processos evolutivos.

Para que se possa realizar um mapeamento de QTLs, é fundamental que se disponha de um conjunto de marcadores genéticos, pois esses fornecem informações genéticas adicionais sobre os genótipos, já que, na prática, os genótipos dos QTLs são desconhecidos.

Neste trabalho, o mapeamento de QTLs foi realizado utilizando a inferência Bayesiana. A abordagem Bayesiana, através do Teorema de Bayes, combina a verossimilhança com a distribuição a priori atribuídas a todos os parâmetros desconhecidos (número, localização e efeitos genéticos dos QTLs) induzindo distribuições a posteriori a respeito destas

quantidades. Métodos Bayesianos de mapeamento podem tratar o número desconhecido de QTLs como uma variável aleatória, implicando em vantagens consideráveis para a modelagem. Entretanto, resulta em complicações na obtenção da amostra aleatória da distribuição conjunta a posteriori, uma vez que a dimensão do espaço do modelo pode variar. O algoritmo MCMC com Saltos Reversíveis é uma proposta para obtenção dessa amostra.

O Método MCMC com Saltos Reversíveis (MCMC-SR), proposto por Green(1995), permite saltar entre modelos de diferentes dimensões. No entanto, a complexidade dos passos do MCMC-SR aumenta a demanda computacional e podem impedir o bom desempenho do algoritmo.

O objetivo deste trabalho foi realizar uma revisão de literatura sobre o mapeamento de QTLs, considerando um modelo genético que inclui somente efeitos aditivos e de dominância, utilizando a abordagem Bayesiana, pesquisar pacotes estatísticos que realizam essa análise e analisar dados simulados.

2 REVISÃO BIBLIOGRÁFICA

2.1 Conceitos Básicos de Genética

A modelagem estatística empregada à procura de locos responsáveis por características fenotípicas de interesse constitui-se numa importante tarefa da pesquisa em melhoramento genético de plantas e animais. Dessa forma, a interação entre estatísticos e geneticistas, bem como a familiarização com termos básicos de genética é extremamente necessária para o desenvolvimento desse tipo de pesquisa.

Na genética, o gene é referido como a unidade básica da herança, e os alelos, são as formas alternativas da expressão de um gene. Define-se, ainda, o locus como o local, no cromossomo, onde reside o gene. Num mesmo par de cromossomos homólogos, os genes que estiverem localizados em posições correspondentes são denominados genes alelos. Quando um indivíduo possui dois genes alelos iguais, ele é chamado homozigoto dominante (ambos os alelos dominantes) ou homozigoto recessivo (ambos os alelos recessivos) (Cruz & Schuster, 2004). Por outro lado, se o indivíduo possui os genes alelos diferentes, o indivíduo é heterozigoto ou híbrido. Para facilitar a compreensão do genótipo de um indivíduo é usual o emprego de letras para representar os genes. Geralmente, emprega-se letra maiúscula (A) para indicar genes dominantes e letra minúscula (a) para indicar genes recessivos, como segue: AA - Homozigoto Dominante, aa - Homozigoto Recessivo e Aa - Heterozigoto.

Um indivíduo pode ter, ainda, alelos codominantes. Neste caso, permite-se identificar apenas duas classes fenotípicas (A_1 e a_1).

O **valor fenotípico** é uma expressão do genótipo (constituição genética) do indivíduo para a característica de interesse mais o componente de ambiente (clima, alimentação, manejo, saúde etc), ou seja,

$$\text{Fenótipo} = \text{Genótipo} + \text{Ambiente.} \quad (1)$$

O efeito genético de um gene (de um marcador, de um QTL) pode ser decomposto em dois componentes: efeito genético aditivo e efeito genético de dominância. Portanto,

a equação (1) pode ser reescrita como:

$$\text{Fenótipo} = (\text{Efeito Aditivo} + \text{Efeito de Dominância}) + \text{Efeito de Ambiente}$$

O efeito aditivo é a ação gênica em que os efeitos de uma característica genética sofrem alteração de cada alelo adicional introduzido e pode ser predito linearmente por meio desses alelos, enquanto que o efeito de dominância é a ação gênica ocorrida em razão dos desvios do efeito aditivo e não pode ser explicado linearmente.

Figura 1 - Decomposição dos efeitos genéticos. Da esquerda para a direita, os pontos da abscissa: 0, 1 e 2 representam os genótipos aa, Aa e AA, respectivamente. Os círculos pretos representam os verdadeiros valores genotípicos (G), enquanto que os círculos brancos são os valores esperados com base no efeito aditivo a , ($\hat{G}$). O desvio entre G e $\hat{G}$ para cada genótipo é chamado de desvio de dominância (d)

Com base na Figura 1, pode-se reescrever o valor genotípico como:

$$G = \hat{G} + d$$

e, então, o valor fenotípico pode ser representado da seguinte forma:

$$Y = G + E$$

em que, G é o valor genotípico; E é o efeito ambiental; $\hat{G}$ é o efeito aditivo do gene; d é o efeito de dominância do gene.

Analisando a Figura 1, observa-se que a contribuição dos homozigotos (AA, aa) é fornecida por a e do heterozigoto (Aa) por d , em relação à média. Portanto,

- (i) Se $d = 0$ então não há dominância entre os alelos;
- (ii) Se $d > 0$ então AA é dominante sobre aa;
- (iii) Se $d < 0$ então aa é dominante sobre AA.

Dessa forma, se um conjunto de indivíduos é genotipado para um certo loco cromossômico os efeitos aditivo e de dominância desse loco podem ser estimados. Contudo, no caso de mapeamento de QTL não é possível genotipá-lo, já que nem mesmo sua localização é conhecida. Sendo assim, Thoday (1961) propõe a utilização de genes marcadores para sistematicamente caracterizar e mapear poligenes individuais que controlam os caracteres quantitativos de interesse.

2.2 Marcadores Moleculares

Inicialmente, os estudos genéticos eram realizados utilizando-se marcadores morfológicos determinados por mutações simples em um gene particular, gerando alterações fenotípicas de fácil identificação no organismo. Os marcadores morfológicos contribuíram significativamente para o estabelecimento dos princípios teóricos do mapeamento genético e das análises de ligação gênica. No entanto, o número reduzido de marcadores fenotípicos disponíveis, a ausência de ligação desses com caracteres de importância econômica e os efeitos deletérios das mutações limitaram sua utilização.

Com o avanço das técnicas de biologia molecular, tornou-se possível a manipulação do DNA, que culminou com o surgimento de vários tipos de marcadores moleculares. O objetivo dessas novas técnicas é a criação de diagramas descritivos dos cromossomos com uma resolução cada vez mais apurada, descrevendo de maneira ordenada a constituição genética de espécies animais e vegetais, proporcionando um entendimento mais detalhado das funções biológicas dos organismos.

Atualmente, dispõe-se de um grande arsenal de métodos para a análise de marcadores moleculares, entretanto, Coelho (2000) ressalta a importância de se considerar o conteúdo informativo dos marcadores de natureza codominante (RFLPs e SSRs) em relação aos marcadores dominantes (RAPDs e AFLPs), além de aspectos como custo, tempo necessário para a realização das avaliações, e dificuldades práticas inerentes à execução de cada técnica. Marcadores do tipo RFLP (“restriction fragment length polymorphism”) são baseados em hibridação, em que uma pequena sequência é marcada radioativamente e ligada ao DNA dos indivíduos, usando princípios de pareamento das bases do DNA. Já os marcadores do tipo RAPD (“random amplified polymorphism DNA”) e o AFLP (“amplified fragment length polymorphism”) são baseados em amplificação, que têm como princípio o emprego de reações para aumento de sequências específicas.

Uma vez que os dados dos marcadores são avaliados, os cientistas constroem um mapa genético e designam os marcadores que segregam conjuntamente dentro do grupo de ligação (Lander *et al.*, 1987). A Figura 2 apresenta um exemplo de um mapa de ligação construído para uma espécie hipotética contendo três cromossomos os quais são indicados pela letra c.

Figura 2 - Mapa genético para uma população hipotética contendo três cromossomos

Para a construção do mapa genético é necessário que sejam adotados os seguintes procedimentos: i) seleção da população mais apropriada para mapear ii) agrupar os marcadores em diferentes grupos de ligação; iii) definir a ordem dos marcadores em cada grupo de ligação e iv) estimar a frequência de recombinação entre marcadores adjacentes.

2.3 Seleção de Mapeamento

A seleção da população de mapeamento envolve a escolha de genitores e a determinação do tipo de cruzamento, sendo considerada uma etapa crítica para o sucesso da construção do mapa. Além disso, tal seleção influencia diretamente no mapeamento QTL, já que esse processo depende dos marcadores genéticos espalhados por todo genoma.

Diversos tipos de populações podem ser utilizadas no mapeamento genético, sendo que a escolha deve estar de acordo com os objetivos, tempo e recursos disponíveis da pesquisa. Independentemente dessa escolha, duas condições básicas devem ser atendidas: as populações oriundas do cruzamento devem trazer o máximo de conteúdo de informação e possuir desequilíbrio de ligação (Paterson *et al.*, 1991; Tanksley, 1993). Essa idéia define o conceito de cruzamentos controlados, ou seja, permite-se fazer a escolha e acompanhamento das linhagens de forma a garantir a variabilidade genética.

As populações de mapeamento com maior conteúdo de informação são aquelas obtidas a partir do cruzamento entre dois indivíduos homozigotos contrastantes (puros). A população resultante desse tipo de cruzamento é denominada população F1, a qual tem desequilíbrio de ligação máximo. A Figura 3 ilustra o esquema de obtenção dessa população através do cruzamento entre dois pais contrastantes para um determinado caráter, mostrando que toda a descendência é Aa. Na Figura 4 são detalhadas duas das populações obtidas a partir de F1: população F2 e população de retrocruzamento (*Backcross*), que são muito utilizadas no mapeamento de QTLs, pois procuram explorar o desequilíbrio de ligação máximo existente na geração F1. Os indivíduos de uma população F2 são obtidos por autofecundação dos indivíduos oriundos da população F1. Nessas populações os marcadores codominantes segregam na proporção 1:2:1 (AA: Aa: aa). Considerando marcadores dominantes tem-se segregação na proporção 3:1 ($A_- : aa$).

Vale ressaltar, que nos programas de melhoramento genético de plantas é comum o uso de populações obtidas por autofecundação a partir da geração F2 (como F3, F4 etc) (Cruz & Schuster, 2004). Essas populações têm estrutura genética previsível e, portanto, podem ser utilizadas para mapeamento genético. Já os indivíduos oriundos de retrocruza-

Pode-se observar na Figura 5, que apenas dois genótipos são obtidos (heterozigotos: Aa; e homozigotos: aa). Nessas populações os marcadores codominantes segregam na proporção 1:1, quando envolver o genitor aa ou genitor AA. Caso haja dominância entre os alelos, a segregação fenotípica ocorrerá na proporção de 1:1 quando envolver o genitor aa, ou, não haverá segregação quando envolver o genitor AA. Segundo Staub *et al.*(1996) o máximo de informação genética é alcançado quando se usa uma população F2 e marcadores codominantes.

2.3.1 Freqüência de Recombinação e Distância no Mapa

Num mapa de ligação genética, a distância entre os genes é estimada por meio de recombinação entre os locos considerados, que por sua vez pode ser obtida a partir da freqüência dos gametas recombinantes, denotada por r . Portanto, é previamente importante entender o conceito de recombinação gênica ou *crossing-over*.

A recombinação gênica ou *crossing-over* é um fenômeno que ocorre durante a prófase I da meiose, em que as cromátides homólogas, mas não irmãs, se entrelaçam, sofrem quebras e fazem permuta de segmentos cromossômicos, havendo assim troca de genes, contribuindo para aumentar a variabilidade genética das células-filhas, uma vez que são produzidos tanto gametas parentais quanto recombinantes. Para esclarecer esse processo, considera-se o cruzamento entre dois indivíduos com genótipos AABB e aabb (Figura 6). Considerando-se, ainda que, A e B estão ligados e a fase um representa as quatro cromátides dos dois homólogos do cromossomo contendo os locos A e B, no indivíduo F1. A linha sólida representa o homólogo herdado do pai com genótipo AABB, e a linha pontilhada, o homólogo herdado do pai com genótipo aabb. Na fase dois, as cromátides se quebram, entrelaçam e fazem permuta de segmentos cromossômicos, evento denominado recombinação gênica ou *crossing-over*. Vale ressaltar que a permuta em uma região pode interferir na ocorrência de uma outra, nas suas proximidades e este fenômeno é denominado interferência e denotado por c . Na fase três, são produzidos dois novos tipos de gametas pelo indivíduo F1 (Ab e aB), além dos gametas parentais (AB e ab), os quais são apresentados na Tabela 1 com suas respectivas probabilidades.

Figura 6 - Esquema de uma recombinação gênica

Tabela 1 - Gametas produzidos pelo duplo-heterozigoto AB//ab

Gameta	Probabilidade	Tipo
AB	$\frac{1}{2}(1 - r)$	Parental
ab	$\frac{1}{2}(1 - r)$	Parental
Ab	$\frac{1}{2}r$	Recombinante
aB	$\frac{1}{2}r$	Recombinante

Diversos métodos estatísticos podem ser usados para estimar a frequência de recombinação, entre eles, o método dos momentos, o método dos mínimos quadrados e o da máxima verossimilhança, sendo o método de máxima verossimilhança o mais utilizado, devido a sua precisão.

Se todas as células germinativas sofrerem *crossing-over* entre os dois locos, serão produzidos quatro tipos de gametas com a mesma frequência de 25% para cada um e sendo assim, a soma da frequência dos gametas recombinantes será de 50%. Dessa forma, constata-se que a porcentagem máxima de recombinação é de 50%.

Quanto mais afastados estão dois genes, em um mesmo par de cromossomos homólogos, maior é a taxa de recombinação. Os mapas usam como unidade de distância entre genes a unidade de mapa que equivale a 1% de recombinação gênica e é representada por centimorgans (cM). Isso deve-se ao fato de que os trabalhos relativos ao estudo de ligação gênica foram desenvolvidos por Thomas Morgan.

Embora a frequência de recombinação indique a medida de distância entre dois locos ao longo de determinado cromossomo, elas são inadequadas como medida direta, uma vez que não apresentam aditividade. O ideal é que as distâncias usadas na construção de mapas genéticos sejam totalmente aditivas, de modo que, ao agregar novos locos (marcadores) ao mapa, não seja necessário ajustar as distâncias já estabelecidas. Para estabelecer a relação entre as distâncias genéticas entre dois locos no mapa e a frequência de gametas recombinantes criaram-se as funções de mapeamento (Lynch & Walsh, 1998). Um grande número de funções de mapeamento foram propostas, dentre elas destaca-se as funções de mapeamento de Haldane (Haldane, 1919) e Kosambi (Kosambi, 1944). Para um melhor entendimento, considere três locos (A, B e C), d_{AB} a distância entre A e B, d_{BC} a distância entre B e C e r_{AB} e r_{AC} como sendo suas respectivas porcentagens de recombinação, ou a distância em Morgan conforme mostra a Figura 7.

Figura 7 - Distância em porcentagem de recombinação entre os locos (A, B, C)

Nesse caso, tem-se que $r_{AC} \neq r_{AB} + r_{BC}$, ou seja não há aditividade entre as frequências de recombinação, portanto, deseja-se estabelecer a relação entre a porcentagem de recombinação entre dois locos r e sua distância d por meio de:

$$d = f(r).$$

Para isso, considera-se a seguinte relação:

$$r_{AC} = r_{AB} + r_{BC} - 2cr_{AB}r_{BC},$$

em que c indica a interferência entre as regiões cromossômicas. Para intervalos pequenos, a porcentagem de recombinação, r é estimada através da distância, d , entre os locos. Então, considerando que o valor de r_{BC} pequeno, tem-se que:

$$r_{BC} = d_{BC},$$

e, portanto,

$$r_{AC} = r_{AB} + d_{BC} - 2cr_{AB}d_{BC},$$

que pode ser reescrita como:

$$r_{AC} = r_{AB} + d_{BC}(1 - 2cr_{AB}).$$

Então,

$$\frac{d_{BC}}{r_{AC} - r_{AB}} = \frac{1}{1 - 2cr_{AB}}.$$

É razoável admitir também pequenos intervalos AC e que o produto $r_{AB}r_{BC}$ tende para zero, assim tem-se que:

$$r_{AC} \approx r_{AB} + r_{BC}.$$

Logo,

$$\frac{d_{BC}}{r_{BC}} = \frac{1}{1 - 2cr_{AB}},$$

que é válido somente para pequenos intervalos de distâncias ou pequenos valores de r. Utilizando-se a notação de derivada, tem-se:

$$f'(r) = \lim_{\Delta r \rightarrow 0} \frac{\Delta d}{\Delta r} = \frac{1}{1 - 2cr}$$

e, portanto,

$$F(r) - F(0) = \int_0^1 \frac{1}{1 - 2cr} dr = d.$$

a partir da qual obtêm-se todas as funções de mapeamento. A função de Haldane admite que eventos *crossing-over* ocorrem de modo independente (ausência de interferência ou $c = 1$), de forma que

$$d_H = \int_0^1 \frac{1}{1-2r} dr$$

em que d_H é a distância entre locos, estabelecida através da função de Haldane. Fazendo a mudança de variável, $u = 1 - 2r$, tem-se que:

$$d_H = -0,5 \ln((1 - 2r) - \ln(1)),$$

e, então,

$$d_H = 0,5 \ln(1 - 2r) \text{ e } r = \frac{1 - e^{2d_H}}{2}.$$

Dessa forma, se dois locos estão a uma distância de d cM, a probabilidade de recombinação é $r = 0,5(1 - \exp^{-2d})$. Quando $d \rightarrow +\infty$, $r \rightarrow \frac{1}{2}$ e vice-versa, tem-se que $r \in (0, \frac{1}{2})$. Dois locos não estão ligados se a razão de recombinação é 0,5. Isto também ocorre se eles estão em cromossomos diferentes.

2.4 Estabelecimento dos Grupos de Ligação

Uma vez estimadas as frequências de recombinação entre pares de locos, deve-se partir para a formação dos grupos de ligação. Os grupos de ligação são definidos como grupos de genes (marcadores) cujos locos estão localizados no mesmo cromossomo. Para demonstrar que dois locos estão ligados, utiliza-se a estatística LOD score.

Para definir a estatística LOD score, anteriormente é necessário definir o conceito de verossimilhança. Para isso, considere $Y_1, \dots, Y_n$ uma amostra aleatória de uma família de distribuição $f(y; \theta)$, $\theta \in \Theta$ e a função densidade de probabilidade conjunta dada por:

$$f(y; \theta) = \prod_{i=1}^n f(y_i; \theta).$$

Para um conjunto de dados fixos $(y_1, \dots, y_n)$, a função de verossimilhança é interpretada como função do vetor de parâmetros e denotada por $L(\theta|y)$ e dada por:

$$L(\theta|y) = \prod_{i=1}^n f(y_i; \theta). \quad (2)$$

A estatística LOD score é dada pelo logaritmo da razão de verossimilhança, isto é,

$$LOD = \log_{10} \frac{L(r|y)}{L(r = 0,5|y)}$$

em que, $L(r = 0,5|y)$ é a função de verossimilhança considerando $r = 0,5$ (ausência de ligação); $L(\hat{r}|y)$ é a função de verossimilhança avaliada em $r = \hat{r}$, em que $\hat{r}$ é a estimativa de máxima verossimilhança de r .

O logaritmo na base 10 é utilizado para facilitar a interpretação; um LOD igual a 3 indica que a ocorrência de ligação é mil vezes mais provável que a de segregação independente (Bearzoti, 2000).

Dois marcadores são considerados ligados quando a frequência de recombinação entre eles for inferior a um limite pré-definido (r_{max}) e/ou o LOD for superior a um limite também pré-definido (LOD_{min}). Uma vez definidos os critérios de (r_{max}) e (LOD_{min}), apenas um padrão de agrupamento é obtido, estando cada marcador presente em apenas um grupo (ou nenhum, no caso de não estar ligado a nenhum outro marcador).

2.5 Determinação da Ordem dos Marcadores nos Grupos de Ligação

Conhecendo-se as frequências de recombinação entre diversos locos do mesmo grupo de ligação, é possível estimar, com um certo grau de precisão, a melhor ordem para os marcadores dentro de cada grupo. Existem vários métodos para estimar a melhor ordem dos marcadores, por exemplo, para um grupo de ligação com apenas dois marcadores apenas uma ordem é possível. Para três marcadores são possíveis três ordens e para n marcadores, existem $n!/2$ ordens. Portanto, o estabelecimento da melhor ordem não é tarefa fácil, já que na prática dispõe-se de grande número de marcadores. Na literatura existem vários métodos

destinados a esse propósito: a) delineação rápida em cadeia (Doerge,1996); b) serialização; c) “simulated annealing” e d) ramos e conexões.

2.6 Abordagem Bayesiana

2.6.1 Visão Geral

Uma das principais finalidades da estatística é fazer inferências ou predições a cerca de parâmetros de interesse. Do ponto de vista da inferência clássica, os parâmetros desconhecidos do modelo são vistos como quantidades fixas e as inferências são baseadas nas informações contidas na amostra dos dados. Segundo Paulino (2003), a abordagem clássica foi adotada de forma quase unânime pelos estatísticos durante a primeira metade do século XX. No entanto, nos últimos quinze anos, devido aos avanços computacionais, renasce uma forte alternativa a esta abordagem: a inferência Bayesiana.

A inferência Bayesiana trata o vetor de parâmetros desconhecidos do problema como quantidades aleatórias e, portanto permite, com base em evidências a priori, atribuir distribuições de probabilidade a essas quantidades, conhecidas como distribuições a priori.

De uma maneira bem geral, a distribuição a priori é o único fator que diferencia a abordagem clássica da abordagem Bayesiana, sendo, portanto alvo de crítica para muitos dos estatísticos clássicos, que alegam que a seleção a priori é um processo arbitrário e subjetivo, já que dois pesquisadores podem ter diferentes graus de incertezas sobre uma quantidade desconhecida. No entanto, muitos autores argumentam que a abordagem Bayesiana é de fato subjetiva, mas por várias razões teóricas não é arbitrária.

As informações a priori capturam a opinião pessoal sobre a situação antes da observação dos dados. Através do Teorema Bayes, a informação contida nos dados amostrais é combinada com a informações a priori sobre os parâmetros desconhecidos induzindo a distribuições a posteriori sobre esses parâmetros. De acordo com Paulino (2003), o Teorema de Bayes é, para muitos, um dos poucos resultados da matemática que se propõe caracterizar a aprendizagem com a experiência, isto é, a modificação da atitude inicial em relação aos “an-

tecedentes”, “causas”, “hipóteses” ou “estados” depois de ter a informação adicional de certo acontecimento ou acontecimentos.

Considere uma quantidade desconhecida θ , a informação que se dispõe sobre esse parâmetro antes de observar os dados, denotada em termos de probabilidade por $\pi(\theta)$ e a função de verossimilhança definida na equação (2). O Teorema de Bayes estabelece que:

$$\pi(\theta|y) = \frac{L(\theta|y)\pi(\theta)}{\int L(\theta|y)\pi(\theta)d\theta}. \quad (3)$$

Visto que $\int L(\theta|y)\pi(\theta)d\theta$ independe do parâmetro θ , ela pode ser considerada como uma constante normalizadora. Então, a equação (3) pode ser:

$$\pi(\theta|y) = k^{-1}L(\theta|y)\pi(\theta)$$

em que, $L(\theta|y)$ fornece a plausibilidade ou verossimilhança para cada um dos possíveis valores de θ ; $\pi(\theta)$ é conhecida como a distribuição a priori de θ ; $\pi(\theta|y)$ é conhecida como a distribuição a posteriori de θ e $k^{-1} = \int L(\theta|y)\pi(\theta)d\theta = E_{\theta}[L(\theta|y)] = f(y)$.

Em problemas de estimação dos parâmetros é comum omitir a constante normalizadora k^{-1} , obtendo assim, a forma usual do Teorema de Bayes:

$$\pi(\theta|y) \propto L(\theta|y)\pi(\theta)$$

em outras palavras,

$$\text{Distribuição a Posteriori} \propto \text{Verossimilhança} \times \text{Distribuição a Priori}.$$

2.6.2 Distribuições a Priori

2.6.2.1 Prioris Informativas

A partir do conhecimento prévio sobre θ pode-se descrever, aproximadamente a distribuição que melhor o representa. Se num modelo observacional, as distribuições a priori e a posteriori pertencem a uma mesma classe de distribuições, a passagem da priori para

a posteriori envolve apenas uma simples mudança nos parâmetros da distribuição a priori (hiperparâmetros) sem a necessidade de cálculos adicionais, propriedade que define o conceito de conjugação. Portanto, sem dificuldades, é possível atualizar a distribuição de θ através da obtenção de novas observações.

Definição: Seja $F = \{f(y|\theta), \theta \in \Theta\}$ uma classe de distribuições amostrais, então uma classe de distribuições Ψ será conjugada a F se:

$$\forall f(y|\theta) \in F \text{ e } \pi(\theta) \in \Psi, \text{ então } \pi(\theta|y) \in \Psi.$$

É importante salientar que alguns cuidados com a utilização de análises com priors conjugadas devem ser considerados:

i) a classe Ψ pode ser muito ampla;

Exemplo: Seja a família $\Psi = \{\text{todas as distribuições}\}$ e F uma distribuição qualquer. Nesse caso, com certeza Ψ é conjugada a F , pois qualquer posteriori estará contida em Ψ e a definição deixa de ter valor prático.

ii) a classe Ψ pode ser muito restrita;

Exemplo: Seja a família $\Psi = \{P : P(\theta = \theta_0) = 1\}$. Isso significa que nenhuma amostra dará informação alguma, pois tem-se certeza a priori de que $\theta = \theta_0$, então esta certeza será mantida a posteriori.

Sendo assim, a classe Ψ deve ser ampla o suficiente para proporcionar a escolha de uma distribuição a priori conveniente e, ao mesmo tempo, deve ser restrita para dar sentido ao uso da definição.

Como visto, as distribuições a priori e a posteriori estão conectadas pela verossimilhança. Portanto, uma forma de construir a conjugada é se basear no núcleo da verossimilhança. Por exemplo, seja $Y_1, \dots, Y_n$ uma amostra de tamanho n da distribuição de Bernoulli com parâmetro θ , então a função de verossimilhança é:

$$L(\theta|y) = \theta^t(1 - \theta)^{n-t}, \quad 0 < \theta < 1 \text{ em que } t = \sum_{i=1}^n y_i$$

Do Teorema de Bayes, tem-se que a densidade a posteriori de θ será da forma:

$$\pi(\theta|y) \propto L(\theta|y)\pi(\theta) \propto \theta^t(1 - \theta)^{n-t}\pi(\theta)$$

Então, baseando-se no núcleo da verossimilhança que é da forma $\theta^a(1 - \theta)^b$, que caracteriza a família das distribuições Beta, especifica-se para θ uma priori Beta. Isto é,

$$\pi(\theta) = \frac{1}{B(a, \theta)} \theta^{a-1}(1 - \theta)^{\theta-1}, \quad \text{em que } a > 0 \text{ e } \theta > 0,$$

combinando-se a priori Beta com a verossimilhança obtém-se a densidade a posteriori:

$$\pi(\theta|y) \propto \theta^{a+t-1}(1 - \theta)^{\theta+n-t-1}.$$

Portanto, $\theta|y \sim \text{Beta}(a + \theta, \theta + n - t)$. Logo, a família de distribuições Beta é conjugada à Bernoulli.

Além da distribuição de Bernoulli, existem outras distribuições conjugadas, por exemplo: 1) Poisson conjugada à Gama com parâmetros a e θ ; 2) Exponencial conjugada à Gama com parâmetros $a + n$ e $\theta + \sum y_i$; 3) Normal com variância conhecida conjugada à Normal e 5) Normal com média conhecida e variância desconhecida conjugada à Gama com parâmetros $n_0/2$ e $n_0\sigma_0^2\phi$.

2.6.2.2 Prioris não Informativas

Conforme visto, a distribuição a priori é subjetiva e, portanto, um ponto preocupante entre os estatísticos de influência frequentista, é que possa alterar as conclusões. Nesse contexto, para conciliar esta crítica de influência frequentista, a análise Bayesiana conta com o auxílio do conceito de *priori não informativa*. Segundo Gamerman & Migon (1997), esse conceito foi formulado com o desejo de se realizar uma análise que fosse baseada no mínimo de informação subjetiva a priori, já que em algumas situações não existem informações

disponíveis para se determinar a distribuição a priori. Em outras palavras, permite-se que o experimento seja mais informativo do que a priori.

Para que nenhum valor particular de θ seja favorecido, inicialmente, utilizam-se distribuições uniformes para representar situações para as quais não se dispõe de informações ou não se deseja utilizar informações disponíveis, ou seja, $\pi(\theta) \propto \text{constante}$. No entanto, algumas dificuldades podem surgir com esta escolha: $\pi(\theta)$ é imprópria, isto é, $\int \pi(\theta) d\theta \rightarrow \infty$ ou $\sum_{\theta} \pi(\theta) \rightarrow \infty$.

Se $\phi = \phi(\theta)$ é uma transformação 1 a 1 de θ e se θ tem distribuição uniforme, então, pelo teorema de transformações de variáveis, a densidade de ϕ será:

$$\pi(\phi) = \pi(\theta(\phi)) \left| \frac{d\theta}{d\phi} \right| \propto \left| \frac{d\theta}{d\phi} \right|.$$

Esse raciocínio que leva à especificação $\pi(\theta) \propto \text{constante}$, deveria levar também a $\pi(\phi) \propto \text{constante}$, o que não acontece. Então, o ideal seria estabelecer uma regra que fosse invariante e que $\pi(\phi)$ não fosse imprópria.

No entanto, como o interesse está na obtenção da distribuição a posteriori, que é própria mesmo que a priori não o seja, não é dada importância à impropriedade das distribuições a priori. Jeffreys (1961), *apud* Gamerman & Migon (1997), propôs uma classe de prioris não informativas invariantes, porém, eventualmente imprópria, a qual se baseia na informação de Fisher.

Outra forma de permitir que os dados tenham maior peso é especificar prioris informativas com baixa precisão, essas prioris são ditas prioris vagas.

2.6.3 Métodos de Aproximação Baseados em Amostragem

Como visto anteriormente, quando as distribuições a priori e a distribuição a posteriori são conjugadas e pertencem a família de distribuições conhecidas como a normal, poisson, binomial etc a obtenção de resumos a posteriori de interesse é direta, ou seja, os

momentos de interesse são obtidos analiticamente. No entanto, na prática necessita-se trabalhar com modelos complexos que não permitem uma análise conjugada direta. Mesmo que a verossimilhança e a priori sejam simples, a combinação delas pode produzir uma distribuição a posteriori matematicamente intratável. Na verdade, o problema que se apresenta está no cálculo de integrais que aparecem no cálculo de medidas tais como:

$$E[h(\theta)|x] = \int h(\theta)\pi(\theta|x)d\theta = \frac{\int h(\theta)L(\theta|x)\pi(\theta)d\theta}{\int L(\theta|x)\pi(\theta)d\theta}.$$

Devido aos avanços computacionais, uma forma de resolver o problema é utilizar métodos baseados em amostragem que englobam, dentre outros, os métodos de Monte Carlo simples, os métodos de reamostragem por importância e o método de Monte Carlo com Cadeia de Markov (MCMC). Dentre esses, o método Monte Carlo com Cadeia de Markov é o mais utilizado.

2.6.3.1 Metodologia MCMC

O método MCMC permite obter uma amostra da distribuição $\pi(\theta|y)$ indiretamente. Através da simulação de uma amostra suficientemente grande podem-se calcular estimativas de quantidades de interesse dessa distribuição. Mais formalmente, para m iterações e p parâmetros desconhecidos, simulam-se amostras $\{\theta^1, \theta^2, \dots, \theta^m\}$ em que $\theta^i = (\theta_1^i, \theta_2^i, \dots, \theta_p^i)$, $i = 1, \dots, m$, cuja distribuição de equilíbrio seja $\pi(\theta|y)$. Então, para calcular a estimativa de $\theta^j, j = 1, \dots, p$, utiliza-se o fato de que:

$$\lim_{m \rightarrow \infty} \frac{1}{m} \sum_{i=1}^m \theta_j^i = \hat{\theta}^j.$$

Cabe lembrar que numa cadeia de Markov é um processo estocástico no qual o próximo estado da cadeia, θ^{t+1} , depende somente do estado atual, θ^t , e dos dados e não da história passada da cadeia e que existe uma dependência do estado inicial da cadeia, θ^0 , que é esquecido após um período de aquecimento denominado “burn-in”, o qual é geralmente descartado da análise. O período de “burn-in” não é claramente definido na literatura, por

exemplo, Health (1997) descarta as primeiras 200 amostras em suas análises; Geyer (1992) argumenta que descarta somente um por cento de sua amostra, existem autores que descartam 50% da amostra gerada.

Sob condições de regularidade, a cadeia irá convergir para uma única distribuição estacionária, denominada distribuição de equilíbrio. Além disso, o método exige que as probabilidades de transição de um estado para outro sejam invariantes; que cada estado da cadeia possa ser atingido a partir de qualquer outro em um número finito de iterações e que não haja estados absorventes.

A distribuição será uma distribuição de equilíbrio para a cadeia de Markov X , se for construído um Kernel de transição $P(\cdot, \cdot)$ que preserve π , ou seja, se $X_t \sim \pi$ implica $X_{t+1} \sim \pi$. Impor uma forte condição de reversibilidade com respeito a π , é suficiente para garantir que π seja invariante para a cadeia de Markov e assegura que (X_t, X_{t+1}) tem a mesma distribuição que no tempo reverso (X_{t+1}, X_t) quando $X_t \sim \pi$, isto é,

$$P((X_t, X_{t+1}) \in A \times B) = \int_A P(x, B) \pi(x) dx = \int_B P(x, A) \pi(x) dx = P((X_{t+1}, X_t) \in B \times A)$$

para subconjuntos $A, B \in \mathbb{R}^d$.

Os métodos de construção da cadeia de Markov cuja distribuição de equilíbrio seja $\pi(\theta|y)$, denominados métodos MCMC, mais utilizados são o Gibbs Sampling (Geman & Geman, 1984) e o algoritmo de Metropolis-Hasting (Metropolis et al., 1970).

2.6.3.2 Gibbs Sampling

Para simplificar o entendimento do método Gibbs Sampling, suponha uma distribuição conjunta a posteriori que dependa de dois parâmetros: θ_1 e θ_2 , denotada por $\pi(\theta_1, \theta_2|x)$, conhecida a menos de uma constante de proporcionalidade e que se tem interesse na obtenção de resumos estatísticos das densidades marginais a posteriori desses parâmetros, tais como, a média e a variância.

O método Gibbs Sampling permite construir uma amostra da distribuição a posteriori $\pi(\theta_1, \theta_2|y)$ supondo que as distribuições condicionais completas a posteriori de

$\pi(\theta_1|\theta_2, y)$ e $\pi(\theta_2|\theta_1, x)$ são conhecidas, ou seja, possuem forma fechada. Esse algoritmo gera uma sequência Gibbs, $(\theta_1^{(1)}, \theta_2^{(1)}), (\theta_1^{(2)}, \theta_2^{(2)}), \dots, (\theta_1^{(m)}, \theta_2^{(m)})$, partindo de um valor inicial arbitrário $(\theta_1^{(0)}, \theta_2^{(0)})$. A sequência Gibbs é obtida iterativamente através de gerações alternadas de valores, como descrito a seguir:

- (i) Na iteração j gere $\theta_1^{(j)}$ utilizando a distribuição condicional completa $\pi(\theta_1|\theta_2^{(j-1)}, y)$, isto é, $\theta_1^{(j)} \sim \theta_1|\theta_2^{(j-1)}, y$;
- (ii) Gere $\theta_2^{(j)}$ utilizando a distribuição condicional completa $\pi(\theta_2|\theta_1^{(j)}, y)$. Obtendo-se dessa maneira, $\theta^{(j)} = (\theta_1^{(j)}, \theta_2^{(j)})$;
- (iii) Execute os passos (i) e (ii) m vezes para a obtenção de uma sequência Gibbs de tamanho m .

Deve-se ressaltar que a ordem da geração do valor de cada componente é irrelevante.

Sob condições gerais de regularidade, a distribuição de $\theta_i^{(m)}, i = 1, 2$. converge para $\pi(\theta_i|y)$ (a distribuição a posteriori verdadeira) quando $m \rightarrow \infty$, ou seja, para m suficientemente grande a cadeia gerada pelo método Gibbs Sampling, na iteração m , converge para a distribuição de equilíbrio, na norma da variação total, a uma taxa geométrica em m , propriedade conhecida como ergodicidade (Casella & George, 1992).

O raciocínio desenvolvido pode ser generalizado para $\theta = (\theta_1, \theta_2, \theta_3, \dots, \theta_p)$, da seguinte forma:

- (i) Inicie o contador de iterações da cadeia, $j=1$, e escolha valores iniciais:

$$\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)}, \theta_3^{(0)}, \dots, \theta_{p-1}^{(0)}, \theta_p^{(0)}) ;$$

- (ii) Obtenha um novo vetor $\theta^{(j)} = (\theta_1^{(j)}, \dots, \theta_p^{(j)})$ a partir de $\theta^{(j-1)}$, através de sucessivas gerações de valores

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

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

...

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

(iii) repita ii) para $j = 2, 3, \dots, m$.

Se a indicação de convergência é atingida na iteração m , M iterações após as m iniciais representam uma amostra da cadeia em equilíbrio, ou seja, uma amostra de tamanho M da distribuição conjunta a posteriori $\pi(\boldsymbol{\theta} | y)$.

2.6.3.3 Algoritmo de Metropolis-Hastings

Desde que se conheçam as distribuições condicionais completas a posteriori dos parâmetros desconhecidos, o método Gibbs Sampling é um excelente algoritmo capaz de amostrar da distribuição a posteriori $\pi(\boldsymbol{\theta} | x)$. No entanto, na prática, nem sempre as distribuições condicionais completas dos parâmetros possuem forma fechada, surgindo a necessidade do uso do algoritmo de Metropolis-Hastings.

Seja X_t o estado atual da cadeia de Markov, o algoritmo de Metropolis-Hastings propõe um valor para o próximo estado da cadeia, X_{t+1}^* , utilizando uma distribuição proposta $q(X_{t+1}^* | X_t)$. Então o novo estado, X_{t+1} será igual a X_{t+1}^* com probabilidade $a(X_t, X_{t+1}^*)$ ou igual a X_t com probabilidade $1 - a(X_t, X_{t+1}^*)$, em que:

$$a(X_t, X_{t+1}^*) = \min \left\{ 1, \frac{\pi(X_{t+1}^* | y) q(X_t | X_{t+1}^*)}{\pi(X_t | y) q(X_{t+1}^* | X_t)} \right\}.$$

Mais formalmente, a implementação do algoritmo de Metropolis-Hastings é dada por:

1. Escolha um valor inicial X_0 ;
2. Gere $X_{t+1}^* \sim q(X_{t+1}^* | \cdot)$;
3. Calcule $a(X_t, X_{t+1}^*)$ e gere uma variável aleatória U de uma distribuição uniforme $[0,1]$, então, o novo estado da cadeia será:

$$X_{t+1} = \begin{cases} X_{t+1}^* & \text{se } U \leq a(X_t, X_{t+1}^*) ; \\ X_t & \text{caso contrário .} \end{cases}$$

4. Repita os passos 2 e 3, m vezes.

Vale ressaltar que a densidade proposta, $q(X_{t+1}^*|X_t)$, deve ser definida no domínio da densidade alvo, $\pi(\theta|\mathbf{x})$. Na prática, o algoritmo de Metropolis-Hastings é facilmente implementado quando a distribuição proposta é fácil de ser simulada. Geralmente, escolhem-se distribuições padrões como uniforme, beta, gama e normal.

Pode-se notar que o Gibbs Sampling é um caso especial do algoritmo de Metropolis-Hastings, no qual a distribuição proposta $q(X_{t+1}^*|X_t) = \pi(X_{t+1}^*|y)$. Então, $a(X_t, X_{t+1}^*) = 1$ e a proposta é sempre aceita.

2.6.4 Fator de Bayes

Em análise Bayesiana, o Fator de Bayes é utilizado para medir a evidência a favor de um modelo. Considere a razão a posteriori entre dois modelos M_1 e M_2 :

$$\frac{\pi(M_1|y)}{\pi(M_2|y)} = \frac{\pi(M_1)}{\pi(M_2)} B_{12}$$

em que,

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

A razão B_{12} conhecida como Fator de Bayes, compara o modelo M_1 com o modelo M_2 . Sendo assim, o Fator de Bayes é igual a razão a posteriori dividida pela razão a priori de dois modelos.

As quantidades $\pi(y|M_1)$ e $\pi(y|M_2)$ podem ser calculadas separadamente para cada modelo usando-se métodos MCMC. Uma calibração para o Fator de Bayes é dada por Jeffreys (1961) e é apresentada na Tabela 2.

Tabela 2 - Interpretação do fator de Bayes segundo Jeffreys (1961)

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

2.7 Mapeamento de QTL

Realizar um mapeamento de QTL significa identificar quantos QTLs estão presentes no genoma, sua posição e estimar os seus efeitos genéticos. Dentre os vários métodos de mapeamento de QTL existentes na literatura, destacam-se o método por marca simples ou também conhecido como análise pontual (Soller *et al.* 1976) e o método de mapeamento por intervalo (Lander & Botstein, 1989). No método de mapeamento por marca simples, a possibilidade de ocorrência do loco controlador da característica (QTL) é testada com base num único marcador, ou seja, utiliza-se um único marcador como unidade de análise, o que dificulta a estimação da posição exata do QTL, além de subestimar os efeitos dos genes quando o loco do QTL está distante do loco do marcador. No entanto, ressalta-se que esse método não deve ser desprezado, já que é útil para uma análise exploratória inicial.

No método de mapeamento por intervalo, testa-se a possibilidade de ocorrência do QTL em uma série de posições dentro de um intervalo flanqueado por dois marcadores (marcadores flanqueadores).

Para um melhor entendimento sobre o mapeamento por intervalo, considere o cruzamento entre duas linhagens homozigotas contrastantes que, ao serem cruzadas, proporcionam indivíduos da geração F_1 com genótipo M_1QM_2/m_1qm_2 . A Figura 8 ilustra o arranjo genotípico da geração F_1 produzido sendo que r_a é a frequência de recombinação entre o marcador M_1 e o QTL (Q) e r_b é a frequência de recombinação entre o QTL(Q) e marcador M_2 .

Para realização do mapeamento por intervalos é necessário conhecer as distâncias entre os marcadores M_1 e M_2 , ou seja, é preciso dispor do mapa de ligação.

Figura 8 - Esquema de dois marcadores flanqueadores, contendo um QTL (Q)

2.7.1 Modelo QTL

Considerando-se uma população originada de cruzamentos controlados, supondo-se que a característica de interesse é afetada por s QTLs e que não existe interação entre os QTLs, isto é, epistasia, o valor fenotípico observado para o i -ésimo indivíduo, y_i , pode ser descrito através do seguinte modelo de regressão linear (Satagopan *et al.*, 1996):

$$y_i = \mu + \sum_{j=1}^s a_j Q_{ij} + \sum_{j=1}^s d_j (1 - |Q_{ij}|) + \epsilon_i \quad (4)$$

em que μ denota a média geral, $a = [a_j]_{j=1}^s$ e $d = [d_j]_{j=1}^s$ denotam o efeito aditivo e o efeito de dominância relativos aos s QTLs e ϵ_i é o erro aleatório tal que $\epsilon_i \sim N(0, \sigma^2)$. Considerando-se uma população F_2 , Satogopan et al. (1996) assume a seguinte parametrização para Q_{ij} , se o genótipo do QTL for QQ , $Q_{ij} = 1$; Qq , $Q_{ij} = 0$; qq , $Q_{ij} = -1$.

A expressão (4) pode ser reescrita como

$$y_i = \mu + a \odot Q_i + d \odot (1 - |Q_i|) + \epsilon_i \quad (5)$$

em que $Q_i = [Q_{ij}]_{j=1}^s$ denota o genótipo dos s QTLs no i -ésimo indivíduo e $v_1 \odot v_2$ denota o produto interno entre os vetores v_1 e v_2 .

Ocorre que, na prática os genótipos dos QTLs, Q_{ij} , $j = 1, \dots, s$, $i = 1, \dots, n$ não são observáveis. No entanto, pode-se, utilizando informações de marcadores moleculares encontrar a sua distribuição de probabilidades. Observe que Q_{ij} é uma variável categórica.

Na prática, observamos o valor fenotípico $y = [y_i]_{i=1}^n$ e um conjunto de dados de marcadores genotípicos $M = [m_{ik}]_{i=1, k=1}^{n, K}$ em que n é o número de indivíduos e K é o número de marcadores. Admitindo-se

- (i) a existência de uma mapa de ligação. Sendo assim as localizações dos marcadores em cada cromossomo é conhecida, conforme apresentado na Figura 2;
- (ii) a existência de apenas um QTL entre marcadores adjacentes (essa é uma suposição razoável se o mapa for denso);
- (iii) a concatenação dos grupos de ligação feita pelos marcadores finais e iniciais de cada grupo (Figura 9)

Figura 9 - Concatenação dos grupos de ligação hipotéticos da Figura 2

Para a análise podem ser considerados dois casos:

- (i) o número de QTLs é conhecido (fixo);
- (ii) o número de QTLs é desconhecido (variável aleatória)

2.8 Análise com número de QTLs conhecido

Considerando-se o modelo dado pela expressão (5) e o número de QTLs, s , conhecido, o objetivo é fazer inferências sobre os parâmetros do modelo $\theta = (\mu, a, d, \sigma^2)$ e sobre

$\lambda = [\lambda_j]_{j=1}^s$ a localização dos s QTLs no genoma. Adotando-se a abordagem Bayesiana as inferências são baseadas na distribuição conjunta a posteriori das variáveis não observáveis dadas as observações by e os dados de marcadores $\mathbf{m}$, $\pi(\theta, \lambda \mid \mathbf{y}, M)$. A Figura 10 apresenta o esquema envolvido na modelagem envolvida no mapeamento Bayesiano de QTLs, considerando o número de QTLs conhecido.

Figura 10 - Esquema da modelagem de mapeamento Bayesiano com número de QTLs conhecido

A distribuição conjunta a posteriori, $\pi(\theta, \lambda \mid Y, M)$ é obtida combinando-se a verossimilhança, $L(\theta, \lambda)$, com a distribuição a priori para os parâmetros, $\pi(\theta, \lambda)$, isto é,

$$\pi(\theta, \lambda \mid y, M) \propto L(\theta, \lambda) \pi(\theta, \lambda)$$

ou equivalentemente,

$$\pi(\mu, a, d, \sigma^2, \lambda \mid Y, M) \propto L(\mu, a, d, \sigma^2, \lambda) \pi(\mu, a, d, \sigma^2, \lambda) \quad (6)$$

Assumindo-se independência entre os parâmetros do modelo a expressão (6) pode ser escrita como:

$$\pi(\mu, a, d, \sigma^2, \lambda \mid Y, M) \propto L(\mu, a, d, \sigma^2, \lambda) \pi(\mu) \pi(a) \pi(d) \pi(\sigma^2) \pi(\lambda)$$

ou ainda, assumindo-se independência entre os efeitos genéticos dos s QTLs,

$$\prod_{i=1}^n \pi(a_i) \prod_{i=1}^n \pi(d_i)$$

tem-se,

$$\pi(\mu, a, d, \sigma^2, \lambda \mid Y, M) \propto L(\mu, a, d, \sigma^2, \lambda) \pi(\mu) \prod_{i=1}^n \pi(a_i) \prod_{i=1}^n \pi(d_i) \pi(\sigma^2) \pi(\lambda). \quad (7)$$

Observa-se que essas suposições podem ser repensadas à luz dos dados e da experiência do pesquisador.

2.9 Análise com número de QTLs desconhecido

Considerando-se o modelo dado pela expressão (5) e o número de QTLs, s , como variável aleatória, o objetivo é fazer inferências sobre os parâmetros do modelo s , $\theta = (\mu, a, d, \sigma^2)$ e sobre $\lambda = [\lambda_j]_{j=1}^s$ a localização dos s QTLs no genoma. Adotando-se a abordagem Bayesiana as inferências são baseadas na distribuição conjunta a posteriori das variáveis não observáveis dadas as observações $\mathbf{y}$ e os dados de marcadores $\mathbf{m}$, $\pi(\theta, \lambda \mid \mathbf{y}, M)$. A Figura 11 apresenta o esquema envolvido na modelagem envolvida no mapeamento Bayesiano de QTLs, considerando-se o número de QTLs desconhecido.

A distribuição conjunta a posteriori, $\pi(s, \theta, \lambda \mid Y, M)$ é obtida combinando-se a verossimilhança, $L(\theta, \lambda)$, com a distribuição a priori para os parâmetros, $\pi(s, \theta, \lambda)$, isto é,

$$\pi(s, \theta, \lambda \mid y, M) \propto L(\theta, \lambda) \pi(s, \theta, \lambda)$$

ou equivalentemente,

Figura 11 - Esquema da modelagem de mapeamento Bayesiano com número de QTLs desconhecido

$$\pi(s, \mu, a, d, \sigma^2, \lambda \mid Y, M) \propto L(\mu, a, d, \sigma^2, \lambda) \pi(s, \mu, a, d, \sigma^2, \lambda) \quad (8)$$

Assumindo-se independência entre os parâmetros do modelo da expressão (4) e entre os efeitos genéticos dos s QTLs,

tem-se,

$$\pi(s, \mu, a, d, \sigma^2, \lambda \mid Y, M) \propto L(\mu, a, d, \sigma^2, \lambda) \pi(s) \pi(\mu) \prod_{i=1}^n \pi(a_i) \prod_{i=1}^n \pi(d_i) \pi(\sigma^2) \pi(\lambda). \quad (9)$$

Observa-se, novamente que, essas suposições podem ser repensadas à luz dos dados e da experiência do pesquisador.

3 MATERIAIS E MÉTODOS

3.1 Materiais

Foram simulados vários conjuntos de dados de experimentos QTL no *software* WinQTLCart. O WinQTLCart é um *software* inteiramente gratuito para universidades e pode ser obtido através da página <http://statgen.ncsu.edu/qtldcart/WQTLCart.htm>. Ele permite simular o mapa genético e a população de mapeamento para o experimentos QTL, assim como, analisá-los utilizando-se as abordagens clássica e Bayesiana. No presente trabalho, são analisados quatro desses conjuntos simulados.

Para simular estes conjuntos de dados foram construídos quatro mapas de ligação, denotados por Mapa I, Mapa II, Mapa III e Mapa IV, todos com dez cromossomos de comprimento 200cM contendo 20 marcadores moleculares dispostos em posições igualmente espaçadas. Ao longo de cada mapa simulado foram inseridos 1, 2, 4 e 8 QTLs, respectivamente. O tamanho amostral considerado foi de 300 indivíduos de uma população F2 e herdabilidade 80% para a característica fenotípica. A Tabela 3 apresenta os valores considerados para os parâmetros do modelo (5) na simulação dos dados.

3.2 Métodos

O objetivo principal em análise Bayesiana de QTLs é obter resumos a posteriori para o número, posição, genótipos e efeitos genéticos dos QTLs em um ou mais cromossomos. Como visto, a abordagem Bayesiana, através do Teorema de Bayes, combina a verossimilhança dos dados fenotípicos com as distribuições a priori atribuídas a todos os parâmetros de interesse, induzindo então, distribuições a posteriori a respeito das quantidades desconhecidas. Visto que as distribuições a posteriori (7) e (9) são analiticamente intratáveis, recorre-se a métodos de Monte Carlo com cadeia de Markov (MCMC) para gerar uma amostra da distribuição conjunta a posteriori. Entretanto, o método MCMC tradicional é útil somente num

Tabela 3 - Valores verdadeiros para simulação do modelo QTL

Localização			Efeitos	
Mapa	Crom	Posição	a	d
<i>I</i>	2	10,9	1,2060	-0,2925
<i>II</i>	1	27,6	-1,0733	-0,1169
	2	126,9	0,5502	0,2681
<i>III</i>	2	10,6	-0,9931	-0,1813
	7	66,4	0,6609	-0,1487
	1	138,4	0,0042	-0,0086
	2	160,8	-0,1680	-0,1747
<i>IV</i>	4	35,6	-0,3163	0,1693
	9	60,7	0,8878	-0,1337
	8	65,8	-0,3331	-0,1312
	1	70,3	0,1506	-0,0882
	2	82,6	0,868	-0,0914
	7	92,2	-0,3574	-0,0490
	1	95,2	-0,2867	0,0551
	2	143,9	0,3721	-0,0207

contexto em que o espaço paramétrico é fixo, pode-se observar que quando considera-se que o número de QTLs é desconhecido a dimensão do espaço paramétrico varia de acordo com o número s de QTLs presente no modelo, ou seja, considerando-se o modelo definido pela equação (9):

- (i) Se $s = 1$, ou seja, existe somente um QTL no modelo, os parâmetros a serem estimados a posteriori são:

$$(\mu, a_1, d_1, \sigma^2, \lambda_1).$$

- (ii) Se existirem 2 QTLs no modelo, deve-se estimar o seguinte vetor paramétrico:

$$(\mu, a_1, a_2, d_1, d_2, \sigma^2, \lambda_1, \lambda_2).$$

- (iii) Generalizando, se existirem s QTLs no modelo, deve-se estimar o vetor paramétrico:

$$(\mu, a_1, a_2, \dots, a_s, d_1, d_2, \dots, d_s, \sigma^2, \lambda_1, \lambda_2, \dots, \lambda_s)$$

Satagopan *et al.* (1996) considera que o número, s , de QTLs é fixo e com isso, as estimativas a posteriori para a média geral, os efeitos genéticos, a localização dos QTLs e a variância ambiental são baseadas no método MCMC tradicional e o Fator de Bayes é utilizado para estimar o número de QTLs presente no modelo. Para a construção das cadeias de Markov são utilizados os métodos Metropolis-Hasting (Metropolis *et al.*, 1953) e o Gibbs Sampling (Geman & Geman, 1984; Gelfand, 1990).

Quando o número de QTLs é considerado como uma variável aleatória não é mais possível utilizar o MCMC tradicional, visto que além da permanência do número de QTLs no modelo pode ocorrer em cada nova iteração ou o nascimento, ou a morte de um QTL. Ou seja, mudança para de dimensão do espaço paramétrico.

3.2.1 MCMC com Saltos Reversíveis

O problema da implementação do algoritmo de Monte Carlo com Cadeia de Markov tradicional é a incapacidade da cadeia mover-se de um modelo H_s para outro modelo $H_{s'}$, em que s e s' são as dimensões do espaço dos dois modelos. No contexto de mapeamento de QTLs este problema pode ser interpretado como a necessidade, por exemplo, da cadeia mover-se de um modelo contendo dois QTLs para um modelo contendo 3 QTLs e vice-versa. Uma solução proposta por Green (1995) sugere o seguinte: Se os saltos podem ser decompostos em movimentos entre dois modelos H_s e $H_{s'}$, a idéia é completar os espaços H_s e $H_{s'}$ de tal forma que eles possuam a mesma dimensão e criar uma bijeção entre ambas as competições. Se a dimensão de H_s for maior do que a dimensão de $H_{s'}$ e se o movimento de H_s para $H_{s'}$ puder ser representado através de uma transformação determinística de θ_s isto é, $\theta_{s'} = T(\theta_s)$.

No caso geral, em que θ_s é completado por (θ_s, u_1) e $\theta^{(s')}$ por $(\theta_{s'}, u_2)$ então o mapeamento determinístico entre $\theta^{(s')}$ e $\theta_{s'}$ é uma bijeção:

$$(\theta_{s'}, u_2) = T(\theta_s, u_1)$$

A idéia de Green(1995) é excelente, pois permite construir uma cadeia de

Markov reversível $(X_i)_{i \geq 1}$ com distribuição invariante π utilizando o algoritmo de Metropolis-Hastings modificado. A modificação consiste na inclusão do Jacobiano da transformação no cálculo da probabilidade de aceitação do algoritmo de Metropolis-Hasting de forma a contemplar a diferença existente na dimensão dos espaços paramétricos envolvidos no movimento. Em MCMC-SR, cada estado X_i contém dois componentes, isto é, $X_i = (S_i, Z_i)$, em que $S \in \{1, 2, 3, \dots, s\}$ é o indicador do número de QTLs presente no modelo e Z_i é o vetor estocástico das possíveis dimensões dos parâmetros desconhecidos. O vetor Z toma valores num conjunto C , definido como a união $C = \bigcup_{s=1}^p$ de espaços $c = R^{n_s}$, $n_s \geq 1$. Dado $S = s$, Z pode tomar valores somente em C_s . Mais especificamente, suponha que (s, z) é o estado atual da cadeia de Markov denotado por $X^{(t)}$ e que uma proposta $X^{(t+1)} = (S^{(t+1)}, Z^{(t+1)})$ é gerada para o novo estado da Cadeia de Markov. Com probabilidade $b_{ss'}$ a proposta $S^{(t+1)}$ é igual a s' QTLs. Então, dado $S^{(t+1)} = s'$, a proposta $Z^{(t+1)}$ é gerada em $C_{s'}$. Para realização deste processo deve-se expressar $Z^{(t+1)}$ como $Z^{(t+1)} = g_{ss'}(z, U)$, em que U é um vetor aleatório em $\mathbb{R}^{n_{ss'}}$, $n_{ss'} \geq 1$, o qual tem densidade proposta $q_{ss'}(z, u)$ e $\mathbb{R}^{n_s + n_{ss'}} \longrightarrow \mathbb{R}^{n_{s'}}$ é um mapeamento determinístico. Quando considera-se um movimento do estado (s, z) para $(s', z') = (s', g_{ss'}(z, u))$ e o movimento reverso de (s', z') para $(s, z) = (s, g_{s's}(z', u'))$, os vetores do estado da cadeia de Markov e as variáveis aleatórias propostas (z, u) e (z', u') devem ter mesma dimensão, ou seja:

$$n_s + n_{ss'} = n_{s'} + n_{s's}$$

A proposta, $X^{(t+1)}$ é finalmente aceita com probabilidade de aceitação dada por:

$$a_{ss'}(z, z') = \min \left\{ 1, \frac{\pi(s', z'|y) q_{s's}(z', u') b_{s's}}{\pi(s, z|y) q_{ss'}(z, u) b_{ss'}} \left| \frac{\partial(g_{ss'}(z, u))}{\partial z \partial u} \right| \right\} \quad (10)$$

Deve-se notar que o Jacobiano aparece devido a transformação determinística usada no mecanismo proposto por Green(2005). Intuitivamente, essa probabilidade pode ser vista como:

$$\min\{1, (\text{Razão a Posteriori}) \times (\text{Razão Proposta para o Salto} \times \text{Probabilidade de Saltar}) \times (\text{Jacobiano da Transformação})\}$$

Os movimentos entre modelos são vistos como: (1) passo de nascimento de QTL, neste caso propõe-se saltar de um modelo com s QTL para um modelo contendo s' QTLs, em que $s < s'$, com probabilidade b_n ; (2) passo morte de QTL, em que deseja-se saltar de um modelo contendo s QTLs para um modelo contendo s' QTLs, em que $s > s'$, com probabilidade b_m ; (3) passo de permanência, onde o número de QTLs permanece fixo para o próximo estado da cadeia com probabilidade $b_p = 1 - b_n - b_m$. Green(1995) propõe que $b_m + b_n \leq 0,9$.

De acordo com Gaffney (2001), não existe restrição quanto ao número de QTLs a serem adicionados ou removidos do modelo. No entanto, salienta que esta escolha pode afetar a razão de aceitação e portanto afetar o desempenho da cadeia gerada e consequentemente a estimação dos parâmetros. Sendo assim, restringe-se o passo de nascimento ou morte de apenas um QTL. Obviamente a proposta de nascimento ou morte de um QTL não garante sua aceitação e como em MCMC a razão de aceitação calculada, depende da contribuição da verossimilhança, da razão de prioris para os parâmetros e da densidade proposta. Além disso, é importante salientar, que em mapeamento de QTL o Jacobiano da equação (6) é igual a um.

A proposta do movimento para o nascimento de um QTL, ou seja, saltar de um modelo contendo s QTLs para um modelo contendo $s + 1$ QTLs é aceita de acordo com a seguinte probabilidade de aceitação:

$$a_{s(s+1)}(z, z + 1) = \min \left\{ 1, \frac{\pi(s + 1, z + 1|y)q_{(s+1)s}(s + 1, z + 1|s, z)b_m}{\pi(s, z|y)q_{s(s+1)}(s, z|s + 1, z + 1)b_n} \right\} \quad (11)$$

indicando essa probabilidade de aceitação para o nascimento de um QTL como $\min(1, A)$, então a probabilidade de aceitação para a morte de um QTL será $\min(1, \frac{1}{A})$.

Observe-se que quando o número de QTLs for considerado como uma variável aleatória não será mais necessária a utilização do Fator de Bayes para definição do número de QTLs visto que, nesse caso, tem-se a distribuição a posteriori para o número de QTLs.

No contexto Bayesiano são necessárias as especificações de prioris para os

parâmetros de interesse.

3.2.2 Distribuições a Priori

Uma exigência essencial para aplicação da teoria Bayesiana é a escolha cautelosa das distribuições a priori para os parâmetros de interesse. Esse cenário não é diferente no contexto de mapeamento de QTLs, sendo que, a distribuição a posteriori é extremamente sensível à escolha das distribuições a priori.

Nesse tipo de modelagem, assumem-se distribuições a priori para os seguintes parâmetros: média geral (μ); para cada um dos efeitos aditivo (a_j) e de dominância (d_j), $j = 1, \dots, s$, observe que em cada iteração o número de QTLs é fixo; para a variância ambiental (σ^2); para a localização de cada QTL (λ_j) e, finalmente, para o número de QTLs do modelo (s). Assumindo-se que os parâmetros θ , λ , e s são independentes, tem-se:

$$\pi(\mu, a, d, \sigma^2, \lambda, s) = \pi(\mu) \prod_{j=1}^s \pi(a_j) \prod_{j=1}^s \pi(d_j) \pi(\sigma^2) \prod_{j=1}^s \pi(\lambda_j) \pi(s)$$

Assume-se que, a priori, a média geral μ tem distribuição normal com média η e variância τ^2 . Satagopan *et al.* (1996) utiliza $\eta = 0$, enquanto que Raftery *et al.* (1997) sugerem escolher um valor para η que depende dos dados, por exemplo, $\eta = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Para a variância pode-se escolher $\tau^2 = s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$.

De acordo com Satagopan *et al.* (1996), existem outras possibilidades de escolha para os hiperparâmetros da distribuição a priori para μ que podem ser baseadas em informações do pesquisador ou através de experimentos similares realizados anteriormente.

Geralmente, para o j -ésimo efeito aditivo, a_j , e para o j -ésimo efeito de dominância, d_j assumem-se distribuições a priori normais com hiperparâmetros ν_a e ν_d para a média e σ_a^2 e σ_d^2 para a variância, respectivamente. Gaffney (2001) alerta que a escolha de σ_a^2 e σ_d^2 pode afetar a seleção do melhor modelo e, portanto, uma alternativa seria fixar valores para esses hiperparâmetros.

Para a variância ambiental, σ^2 , é comum escolher como priori uma distribuição gama-inversa com hiperparâmetros a e b . Gaffney(2001) sugere valores de $a = 3$ e $b = s_y^2$, enquanto que Satagopan et al. (1996) considera $a = b = 1$.

É também, comum atribuir para a posição do j -ésimo QTL, λ_j , distribuição a priori uniforme num intervalo $[a, b]$. Gaffney (2001) assume que os QTLs são distribuídos ao longo de todo o genoma. Por exemplo, se o genoma tem comprimento L , então, $\lambda_j \sim U[0, L]$.

Satagopan *et al.*(1996), considera apenas um grupo de ligação com os marcadores ordenados $\{1, 2, \dots, m\}$ e distâncias conhecidas $D = \{D_k\}_{k=1}^m$, em que D_k é a distância no mapa entre as marcas 1 e k (Figura 10). Neste caso, propõe-se uma distribuição uniforme para λ no intervalo $(\max(\lambda_{j-1}, \lambda_j - d), \min(\lambda_{j+1}, \lambda_j + d))$. Esta distribuição é denotada por $u(\lambda_j, \lambda_j^*)$ e mantém a ordem dos locos (nota-se que $\lambda_0 = 0$ e $\lambda_{s+1} = D_m$).

Para os genótipos dos QTLs, não é necessário atribuir distribuições a priori, visto que atribuiu-se distribuição a priori para a localização do QTL, λ_j .

Quanto ao número de QTLs total no modelo, existe na literatura três alternativas para a distribuição a priori para o número total de QTLs: Poisson, geométrica e uniforme, as quais serão exploradas neste trabalho.

No caso da distribuição Poisson com média κ , a probabilidade a priori de uma região conter s QTLs é:

$$\pi(s) = (\exp(-\kappa)\kappa^s)/s!, \quad s = 0, 1, \dots$$

No caso da distribuição Geométrica, a probabilidade a priori de uma região conter s QTLs é:

$$\pi(s) = (1 - r)^s r, \quad s = 0, 1, \dots$$

Dentre estas três possibilidades de prioris apresentadas, na prática, os pesquisadores têm optado por distribuições a priori Poisson ou geométrica.

3.2.3 Construção da amostra da distribuição conjunta a posteriori

A cadeia de Markov é inicializada em um ponto arbitrário $(\lambda^0, \theta^0, s^0)$, em que $\theta = (\mu, a, d, \sigma^2)$. Para a atualização de cada ponto da cadeia é necessário um ciclo de três passos, a saber: (1) atualização de s ; (2) atualização de λ , que implica na atualização de Q ; (3) a atualização de θ . A posição, λ do QTL é atualizada utilizando M-H; os demais parâmetros, θ , Q , possuem distribuições condicionais completas a posteriori conhecidas e portanto são atualizados utilizando o Gibbs Sampling.

3.2.3.1 Distribuições necessárias para a atualização dos parâmetros

Para a implementação dos métodos MCMC: Gibbs *sampling* ou Metropolis-Hastings, isto é, para a geração de uma amostra da distribuição (7) ou (9) é necessário obter todas as distribuições condicionais completas a posteriori, pois, os algoritmos Gibbs *sampling* e Metropolis-Hastings necessitam dessas distribuições para realizar as atualizações dos parâmetros.

(i) **Atualização de λ :** Em Satagopan(1996) os elementos de λ são modificados um a um usando o algoritmo de MH. Neste caso, os QTLs são considerados variáveis ordenadas ao longo do genoma. Para o j -ésimo locus, uma proposta λ_j^* é gerada de uma distribuição uniforme no intervalo $(\max(\lambda_{j-1}, \lambda_j - d), \min(\lambda_{j+1}, \lambda_j + d))$, denotada por $u(\lambda_j, \lambda_j^*)$, em que d é uma região em cada lado da posição atual de λ no cromossomo. A proposta é aceita com probabilidade:

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

em que $\lambda_{-j} = [\lambda_{j'} : 1 \leq j' \leq s, j \neq j']$. Se a proposta para o próximo estado da cadeia, t , não for aceita $\lambda^{(t)} = \lambda^{(t-1)}$, ou seja, a posição do QTL no estado da cadeia t

permanece igual a posição do estado anterior, $t - 1$ e o algoritmo segue para atualizar a próxima posição do QTL.

(ii) **Atualização de θ :** Visto que μ tem distribuição priori normal com hiperparâmetros η para média e τ^2 para a variância, a distribuição condicional completa para μ possui forma conhecida, ou seja:

$$\mu|\lambda, Q, a, d, \sigma^2, y \sim N\left(\frac{\sum_{i=1}^n (y_i - \sum_{j=1}^s a_j Q_{ij} - \sum_{j=1}^s d_j (1 - |Q_{ij}|)) + \eta \frac{\sigma^2}{\tau^2}}{n + \frac{\sigma^2}{\tau^2}}, \frac{\sigma^2}{n + \frac{\sigma^2}{\tau^2}}\right)$$

Assumindo que os efeitos genéticos a e d têm distribuição a priori normal com hiperparâmetros 0 para a média e σ_a^2 e σ_d^2 para a variância, respectivamente, tem-se que a distribuição condicional completa para a possui forma fechada, isto é:

$$a_{j*}|\lambda, Q, \mu, \{a_j\}_{j \neq j*}, \sigma^2, y \sim N\left(\frac{\sum_{i=1}^n (y_i - \mu - \sum_{j=1}^s a_j Q_{ij} - \sum_{j=1}^s d_j (1 - |Q_{ij}|)) Q_{ij}}{\sum_{i=1}^n Q_{ij}^2 + \frac{\sigma^2}{\sigma_a^2}}, \frac{\sigma^2}{\sum_{i=1}^n Q_{ij}^2 + \frac{\sigma^2}{\sigma_a^2}}\right)$$

Similarmente, para o efeito de dominância, a distribuição condicional completa possui a seguinte forma:

$$d_{j*}|\lambda, Q, \mu, \{d_j\}_{j \neq j*}, \sigma^2, y \sim N\left(\frac{\sum_{i=1}^n (y_i - \mu - \sum_{j=1}^s d_j (1 - |Q_{ij}|)(1 - |Q_{ij}|))}{\sum_{i=1}^n (1 - |Q_{ij}|)^2 + \frac{\sigma^2}{\sigma_d^2}}, \frac{\sigma^2}{\sum_{i=1}^n (1 - |Q_{ij}|)^2 + \frac{\sigma^2}{\sigma_d^2}}\right)$$

Dado que a variância residual, σ^2 , tem distribuição a priori gama-inversa, sua distribuição condicional completa a posteriori para σ^2 , $\pi(\sigma^2|\lambda, Q, \mu, a, d, y)$ é também gama-inversa, $GI(a + n/2, b + \sum_{i=1}^n \epsilon_i^2/2)$, em que b e c são escolhidos arbitrariamente.

Desta forma, para atualizar os parâmetros μ , a , d e σ^2 pode-se utilizar o Gibbs Sampling.

Para a construção da amostra da distribuição conjunta a posteriori dos parâmetros de interesse utilizou-se módulo *bim* do WinQTLCart. Este módulo permite gerar uma amostra aleatória da distribuição conjunta a posteriori para o número, localizações e efeitos genéticos dos QTLs utilizando o método MCMC tradicional ou MCMC-SR, sendo que, neste trabalho utilizou o método MCMC-SR.

- Priori para Variância: $\sigma^2 \sim IG(3, s_y^2)$;

- Priori para Média: $\mu \sim N(0, s_y^2)$, em que s_y^2 é a variância amostral dos dados fenotípicos;
- Priori para os efeitos: $a_j \sim N(0, betas_y^2)$ e $d_j \sim N(0, betas_y^2)$ em que $beta \sim Beta(2, 10)$;
- Priori para o número de QTLs: $Poisson(k)$, $Geométrica(k)$ e $Uniforme[0, L]$, em que $k = 1, 2, 3, 4$ e L é o comprimento do genoma.

Para cada um dos quatro conjuntos de dados simulados foram testadas todas as opções de distribuições a priori para o número de QTLs citadas acima. Para obter as amostras posteriori foram geradas cadeias com 100000 iterações, sendo que os resultados apresentados foram obtidos descartando as 1000 primeiras iterações e armazenando as 10000 posteriores. A indicação de de convergência foi verificada através da análise do comportamento da cadeia ao longo das iterações, embora métodos formais sobre a indicação de convergência devam ser utilizados.

4 RESULTADOS E DISCUSSÕES

Para a construção da amostra da distribuição conjunta a posteriori dos parâmetros de interesse utilizou-se módulo BIM do *software* WinQTLCart. Este módulo permite gerar uma amostra aleatória da distribuição conjunta a posteriori para o número, localizações e efeitos genéticos dos QTLs utilizando o método MCMC tradicional ou MCMC-SR.

Considerando-se que o número de QTLs, S , é uma variável aleatória e considerando-se as seguintes distribuições a priori:

- (i) Priori para Variância: $\sigma^2 \sim IG(3, s_y^2)$;
- (ii) Priori para Média: $\mu \sim N(0, s_y^2)$, em que s_y^2 é a variância amostral dos dados fenotípicos;
- (iii) Priori para os efeitos: $a_j \sim N(0, \beta s_y^2)$ e $d_j \sim N(0, \beta s_y^2)$ em que $\beta \sim Beta(2, 10)$;
- (iv) Priori para o número de QTLs: $Poisson(k)=P(k)$, $Geométrica(k)=Geom(k)$ e $Uniforme[0, L]=Unif[0, L]$, em que $k = 1, 2, 3, 4$ e L é o comprimento do genoma.

Neste trabalho utilizou-se o método MCMC-SR, ou seja, o número de QTLs foi considerado uma variável aleatória. Para gerar uma amostra da distribuição conjunta a posteriori dos parâmetros de interesse. A partir dessa amostra a distribuição marginal a posteriori para o número de QTLs, que afeta a característica de interesse, foi obtida e seu valor estimado.

Para cada um dos quatro conjuntos de dados simulados apresentados foram exploradas as diversas opções de prioris para o número de QTLs e foi verificada a sua influência na escolha do número de QTLs e a sua influência sobre a sensibilidade das estimativas a posteriori dos demais parâmetros de interesse. Para obter as amostras posteriori foram geradas cadeias com 100000 iterações, sendo que os resultados apresentados foram obtidos

descartando as 1000 primeiras iterações e armazenando-se as 10000 posteriores. A indicação de convergência foi verificada através da análise do comportamento da cadeia ao longo das iterações, embora métodos formais sobre a indicação de convergência devam, também, ser utilizados.

4.1 Um QTL Afetando a Característica de Interesse

Supondo-se que um único QTL afeta a característica de interesse e o considerando o conjunto de dados simulados I denotado por Mapa I. Apresenta-se, a seguir, a análise Bayesiana, realizada no *software* WinQTLCart, utilizando o método MCMC-SR, de acordo com a atribuição da distribuição a priori de Poisson, Geométrica e Uniforme para o número de QTLs.

4.1.1 Distribuição a Priori Poisson

A Figura 12 mostra a distribuição a posteriori para número de QTLs detectados, de acordo com a atribuição da distribuição a priori Poisson considerando-se os hiperparâmetros: um, dois, três e quatro e a Tabela 4 apresenta os respectivos resumos estatísticos a posteriori para o número de QTLs. A partir da Figura 12 e da Tabela 4 pode-se concluir que para este conjunto de dados as distribuições a priori Poisson com hiperparâmetro um e dois são as distribuições a priori que melhor se adequaram ao conjunto de dados simulados. No entanto, sabe-se que, na prática, nem sempre o pesquisador dispõe de informações a priori precisas sobre o número de QTLs que afetam a característica de interesse, então, se esse pesquisador atribuir distribuição a priori Poisson com hiperparâmetro três teria que decidir por um modelo contendo um ou dois QTLs. Essa poderia ser uma estimativa razoável, visto que, posteriormente, a análise de QTLs, técnicas mais refinadas poderiam ser empregadas para detectar os QTLs, por exemplo, técnicas de análise de experimentos de *microarrays*.

A Tabela 5 apresenta as estimativas a posteriori para a média, desvio padrão e intervalo de credibilidade 95% para a localização do QTL e na Tabela 6 são apresentadas as estimativas a posteriori para os efeitos genéticos e os intervalos de credibilidade de 95%,

pode-se notar que as diferentes distribuições a priori Poisson atribuídas para o número de QTLs não apresentaram influências significativas sobre as estimativas de localização e sobre o efeito genéticos do QTL simulado. As estimativas para a posição no genoma e efeito aditivo do QTL se aproximaram dos valores verdadeiros.

Figura 12 - Distribuição a posteriori para o número de QTLs - Mapa I (distribuição a priori Poisson)

Tabela 4 - Resumo estatístico a posteriori para o número de QTL - Mapa I (distribuição a priori Poisson)

Priori	Média	Desvio Padrão
P(1)	1,2671	0,5194
P(2)	1,6083	0,8095
P(3)	2,3046	0,9894
P(4)	2,3046	1,2033

Tabela 5 - Resumo estatístico a posteriori para a localização do QTL - Mapa I (distribuição a priori Poisson)

Priori	Crom	Posição	Desvio Padrão	IC95%
P(1)	2	10,7446	1,0578	(8,5977 ; 13,0676)
P(2)	2	10,5844	1,0778	(8,2527 ; 12,6036)
P(3)	2	10,7606	1,1178	(8,5541 ; 13,0629)
P(4)	2	10,8734	1,0614	(8,8250 ; 13,0432)

Tabela 6 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa I (distribuição a priori Poisson)

Priori	a	Dp	IC95%	d	Dp	IC95%
P(1)	1,1891	0,0803	(1,0336 ; 1,3470)	-0,0347	0,1071	(-0,2524 ; 0,1604)
P(2)	1,1860	0,0748	(1,0388 ; 1,3363)	-0,0234	0,1053	(-0,2255 ; 0,1867)
P(3)	1,1836	0,0762	(1,0387 ; 1,3215)	-0,0270	0,1047	(-0,2368 ; 0,1885)
P(4)	1,1963	0,0761	(1,0446 ; 1,3454)	-0,0251	0,1119	(-0,2271 ; 0,2133)

4.1.2 Distribuição a Priori Geométrica

Observa-se na Figura 13 que a probabilidade a posteriori para um QTL é superior às demais. Sendo assim, independentemente do valor do hiperparâmetro considerado, o número de QTLs estimado é igual a um. Na Tabela 7 observam-se que as prioris que melhor se adequaram a este conjunto de dados referem-se àquelas com hiperparâmetros igual a três e quatro. Nas Tabelas 8 e 9 nota-se que não houve diferenças significativas para as estimativas das localizações e efeitos dos QTLs. Além disso, deve-se notar que as estimativas dos efeitos de dominância permaneceram distantes dos valores verdadeiros.

Tabela 7 - Resumo estatístico a posteriori para o número de QTLs - Mapa I (distribuição a priori Geométrica)

Priori	Média	Desvio Padrão
Geom(1)	1,9364	1,3339
Geom(2)	1,4283	0,7122
Geom(3)	1,1883	0,4619
Geom(4)	1,1588	0,3878

Figura 13 - Distribuição a posteriori para o número de QTLs - Mapa I (distribuição a priori Geométrica)

Tabela 8 - Resumo estatístico a posteriori para a localização - Mapa I (distribuição a priori Geométrica)

Priori	Crom	Posição	Desvio Padrão	IC95%
Geom(1)	2	10,7796	1,1663	(8,3295 ; 13,0100)
Geom(2)	2	10,9429	1,1455	(8,5327 ; 13,1969)
Geom(3)	2	10,6268	1,1008	(8,2359 ; 12,7153)
Geom(4)	2	10,5711	1,1281	(8,1762 ; 12,8438)

Tabela 9 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa I (distribuição a priori Geométrica)

Priori	a	Dp	IC95%	d	Dp	IC95%
Geom(1)	1,1844	0,0752	(1,0442 ; 1,3320)	-0,0295	0,1094	(-0,2557 ; 0,1760)
Geom(2)	1,1953	0,0735	(1,0578 ; 1,3392)	-0,0259	0,1025	(-0,2253 ; 0,1728)
Geom(3)	1,1862	0,0760	(1,0305 ; 1,3364)	-0,0338	0,1067	(-0,2413 ; 0,1724)
Geom(4)	1,1841	0,0771	(1,0359 ; 1,3292)	-0,0371	0,1081	(-0,2428 ; 0,1702)

4.1.3 Distribuição a Priori Uniforme

A Figura 14 mostra a distribuição a posteriori para o número de QTLs detectados de acordo com atribuição da distribuição a priori uniforme e na Tabela 10 o resumo

estatístico a posteriori. Como nos casos anteriores, exceto a estimativa do efeito de dominância, as estimativas para a localização e efeitos aditivo se aproximaram dos valores verdadeiros, conforme mostram as Tabelas 11 e 12. Além disso, deve-se observar que as estimativas a posteriori para o número de QTL, efeitos genéticos e localizações são idênticas ao caso em que se atribuiu priori geométrica com hiperparâmetro um para o número de QTLs. Observa-se que L é o comprimento do genoma.

Figura 14 - Distribuição a posteriori para o número de QTLs - Mapa I (distribuição a priori $\text{Unifome}[0,L]$)

Tabela 10 - Resumo estatístico a posteriori para o número de QTLs - Mapa I (distribuição a priori $\text{Unifome}[0,L]$)

Priori	Média	Desvio Padrão
Unif[0,L]	1,9364	1,3339

Tabela 11 - Resumo estatístico a posteriori para a localização do QTL - Mapa I (distribuição priori Uniforme[0,L])

Priori	Crom	Posição	Desvio Padrão
Unif[0,L]	2	10,7796	1,1663

Tabela 12 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa I (distribuição a priori Uniforme[0,L])

Priori	a	Dp	IC95%	d	Dp	IC95%
Unif[0,L]	1,1844	0,0752	(1,0442 ; 1,3320)	-0,0295	0,1094	(-0,2557 ; 0,1760)

4.2 Dois QTLs Afetando a Característica de Interesse

Supondo-se que dois QTLs afetam a característica de interesse e considerando o conjunto de dados simulados II denotado por Mapa II. Apresenta-se, a seguir, a análise Bayesiana, realizada no *software* WinQTLCart, utilizando o método MCMC-SR, de acordo com a atribuição da distribuição a priori de Poisson; Geométrica e Uniforme para o número de QTLs.

4.2.1 Distribuição a Priori Poisson

Observando-se a Figura 15 e a Tabela 13 verifica-se que a atribuindo-se distribuições a priori Poisson com média um e dois para o número de QTLs obteve-se boas estimativas para esse conjunto de dados. Na prática, se o pesquisador atribuir priori Poisson com hiperparâmetro três obteria estimativas razoáveis, ou seja, ficaria em dúvida entre dois e três QTLs afetando a característica de interesse. Nas Tabelas 14 e 21 nota-se que não houve influência das diferentes distribuições a priori Poisson sobre as estimativas a posteriori para a localização e efeitos genéticos dos QTLs. Cabe observar, que diferente do caso em que existia somente um QTL afetando a característica, as estimativas a posteriori para os efeitos de dominância se aproximaram dos valores verdadeiros.

Figura 15 - Distribuição a posteriori para o número de QTLs - Mapa II (distribuição a priori Poisson)

Tabela 13 - Resumo estatístico a posteriori para o número de QTLs - Mapa II (distribuição a priori Poisson)

Priori	Média	Desvio Padrão
P(1)	2,2947	0,6074
P(2)	2,6325	0,8596
P(3)	3,5143	1,2046
P(4)	3,5143	1,3358

Tabela 14 - Resumo estatístico a posteriori para a localização do QTL - Mapa II (distribuição a priori Poisson)

Priori	Crom	Posição	Desvio Padrão	IC95%
P(1)	1	30,7418	1,6007	(27,2159 ; 33,6206)
	2	128,0388	5,0945	(117,3590 ; 135,9710)
P(2)	1	30,2326	1,8110	(26,8533 ; 33,4893)
	2	129,2226	5,6470	(121,8632 ; 136,6981)
P(3)	1	30,2355	1,7622	(26,8044 ; 33,2839)
	2	128,8551	7,1630	(122,6123 ; 137,1621)
P(4)	1	30,7929	1,6672	(26,9439 ; 33,6202)
	2	128,2905	8,0027	(122,1282 ; 134,7540)

Tabela 15 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa II (distribuição a priori Poisson)

Priori	a	Dp	IC95%	d	Dp	IC95%
P(1)	-1,0297	0,0819	(-1,1926 ; -0,8735)	-0,1112	0,1151	(-0,3415 ; 0,1093)
	0,3917	0,0929	(0,2074 ; 0,5672)	0,1951	0,1232	(-0,0451 ; 0,4270)
P(2)	-1,0282	0,0800	(-1,1703 ; -0,8697)	-0,0964	0,1135	(-0,3219 ; 0,1241)
	0,3876	0,0899	(0,2176 ; 0,5666)	0,2070	0,1266	(-0,0234 ; 0,4641)
P(3)	-1,0285	0,0794	(-1,1844 ; -0,8836)	-0,1035	0,1132	(-0,3572 ; 0,0959)
	0,3861	0,0959	(0,1944 ; 0,5677)	0,2027	0,1298	(-0,0268 ; 0,4509)
P(4)	-1,0292	0,0822	(-1,2003 ; -0,8644)	-0,1103	0,1114	(-0,3391 ; 0,1025)
	0,3835	0,0900	(0,2070 ; 0,5480)	0,1949	0,1241	(-0,0478 ; 0,4498)

4.2.2 Distribuição a Priori Geométrica

Similarmente ao caso em que existia somente um QTL afetando a característica de interesse, as distribuições a priori geométricas com parâmetros três e quatro apresentaram os melhores resultados para este conjunto de dados, conforme mostra a Figura 16 e Tabela 16. Nas Tabelas 17 e 18 pode-se observar que as estimativas para a localização e efeitos genéticos dos QTLs também se aproximaram dos valores verdadeiros.

Figura 16 - Distribuição a posteriori para o número de QTLs - Mapa II (distribuição a priori Geométrica)

Tabela 16 - Resumo estatístico a posteriori para o número de QTLs - Mapa II (distribuição a priori Geométrica)

Priori	Média	Desvio Padrão
Geom(1)	4,3458	2,9387
Geom(2)	2,5458	0,8766
Geom(3)	2,2286	0,6520
Geom(4)	2,2286	0,5576

Tabela 17 - Resumo estatístico a posteriori para a localização do QTL simulado (distribuição a priori Geométrica)

Priori	Crom	Posição	Desvio Padrão	IC95%
Geom(1)	1	30,6262	1,6473	(27,2410 ; 33,6104)
	2	128,3465	6,9052	(117,2982 ; 135,1382)
Geom(2)	1	30,4135	1,7169	(27,0227 ; 33,5445)
	2	128,8852	6,3374	(121,8816 ; 136,3200)
Geom(3)	1	30,7780	1,8194	(27,1381 ; 33,5876)
	2	127,6474	9,7346	(114,9255 ; 146,7650)
Geom(4)	1	30,4000	1,7685	(26,6813 ; 33,6429)
	2	129,3319	5,3259	(121,6807 ; 135,5054)

Tabela 18 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa II (distribuição a priori Geométrica)

Priori	a	Dp	IC95%	d	Dp	IC95%
Geom(1)	-1,0285	0,0860	(-1,1934 ; -0,8504)	-0,1059	0,1117	(-0,3214 ; 0,1115)
	0,3863	0,0940	(0,2093 ; 0,5545)	0,2024	0,1233	(-0,0464 ; 0,4355)
Geom(2)	-1,0291	0,0803	(-1,1898 ; -0,8673)	-0,1091	0,1163	(-0,3418 ; 0,1125)
	0,3848	0,0967	(0,1993 ; 0,5698)	0,1937	0,1222	(-0,0392 ; 0,4507)
Geom(3)	-1,0282	0,0838	(-1,1863 ; -0,8645)	-0,1193	0,1154	(-0,3541 ; 0,0921)
	0,3842	0,1106	(0,1870 ; 0,5656)	0,1984	0,1311	(-0,0635 ; 0,4568)
Geom(4)	-1,0210	0,0809	(-1,1898 ; -0,8801)	-0,1134	0,1156	(-0,3406 ; 0,1035)
	0,3903	0,0930	(0,1941 ; 0,5543)	0,2039	0,1198	(-0,0302 ; 0,4376)

4.2.3 Distribuição Uniforme

Conforme visto anteriormente, quando se atribui distribuição a priori Uniforme[0,L] para o número de QTLs, as estimativas a posteriori para os parâmetros dos modelos são idênticas ao caso em que se atribuiu distribuição geométrica com hiperparâmetro um.

Deve-se notar que estas prioris apresentam uma grande variabilidade para o número de QTLs detectados a posteriori.

Figura 17 - Distribuição a posteriori para o número de QTLs - Mapa II (distribuição a priori Uniforme[0,L])

Tabela 19 - Resumo estatístico a posteriori para o número de QTLs - Mapa II (distribuição a priori Uniforme[0,L])

Priori	Média	Desvio Padrão
Unif[0,L]	4,3458	2,9387

Tabela 20 - Resumo estatístico a posteriori para a localização do QTL - Mapa II (distribuição a priori Uniforme[0,L])

Priori	Crom	Média	Desvio Padrão	IC95%
Unif[0,L]	1	30,6262	1,6473	(27,2410 ; 33,6104)
	2	128,3465	6,9052	(117,2982 ; 135,1381)

Tabela 21 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa II (distribuição a priori Uniforme[0,L])

Priori	a	Dp	IC95%	d	Dp	IC95
Unif[0,1]	-1,0285	0,0860	(-1,193 ; -0,8504)	-0,1059	0,1117	(-0,3214 ; 0,1115)
	0,3863	0,0940	(0,2093 ; 0,5545)	0,2024	0,1233	(-0,0464 ; 0,4355)

4.3 Quatro QTLs Afetando a Característica de Interesse

Supondo-se que quatro QTLs afetam a característica de interesse e considerando o conjunto de dados simulados III denotado por Mapa III. Apresenta-se a seguir a análise Bayesiana, realizada no *software* WinQTLCart, utilizando o método MCMC-SR, de acordo com a atribuição da distribuição a priori de Poisson; Geométrica e Uniforme para o número de QTLs.

4.3.1 Distribuição a Priori Poisson

De acordo com a Figura 18 e Tabela 22, nota-se que há poucas evidências quanto a detecção de quatro QTLs afetando a característica. Nas Tabelas 23 e 24, nota-se que o modelo detectou corretamente apenas dois dos 4 QTLs simulados. As prioris que apresentaram o melhor desempenho para este conjunto de dados, foram as prioris Poisson com hiperparâmetro três e quatro. Sendo que dentre essas duas deve-se preferir a priori Poisson(3), devido a menor estimativa no desvio padrão.

Tabela 22 - Resumo estatístico a posteriori para o número de QTLs - Mapa III (distribuição a priori Poisson)

Priori	Média	Desvio Padrão
P(1)	2,4290	0,6604
P(2)	2,9148	0,9637
P(3)	3,2384	1,1118
P(4)	3,7807	1,3835

Figura 18 - Distribuição a posteriori para o número de QTLs - Mapa III (distribuição a priori Poisson)

Tabela 23 - Resumo estatístico a posteriori para a localização do QTL - Mapa III (distribuição a priori Poisson)

Priori	Crom	Posição	Desvio Padrão	IC95%
P(1)	2	11,2653	3,6754	(0,3586 ; 15,2375)
	7	60,2871	18,0840	(28,9083 ; 71,0425)
	-	-	-	-
	-	-	-	-
P(2)	2	11,4012	3,5361	(1,5725 ; 14,9090)
	7	37,6433	21,5805	(15,5007 68,4358)
	-	-	-	-
	-	-	-	-
P(3)	2	11,9328	2,7216	(2,7075 ; 15,0517)
	7	37,6312	21,8086	(13,6669 ; 69,1421)
	-	-	-	-
	-	-	-	-
P(4)	2	11,8707	2,7266	(2,6932 ; 14,7189)
	7	37,1431	21,4892	(9,5672 ; 70,0022)
	-	-	-	-
	-	-	-	-

Tabela 24 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa III (distribuição a priori Poisson)

Priori	a	Dp	IC95%	d	Dp	IC95%
P(1)	-0,9340	0,3800	(-1,1910 ; 0,1756)	-0,1668	0,1442	(-0,4851 ; 0,0452)
	0,5513	0,1512	(0,2904 ; 0,6564)	-0,0483	0,0306	(-0,0953 ; -0,0249)
	-	-	-	-	-	-
	-	-	-	-	-	-
P(2)	-0,9208	0,3716	(-1,2234 ; 0,0945)	-0,1791	0,1446	(-0,4522 ; 0,0767)
	0,3349	0,2396	(0,0419 ; 0,7958)	-0,0540	0,1359	(-0,2303 ; 0,2274)
	-	-	-	-	-	-
	-	-	-	-	-	-
P(3)	-0,9798	0,3289	(-1,2588 ; 0,0718)	-0,1895	0,1425	(-0,4912 ; 0,0474)
	0,3950	0,2348	(0,0912 ; 0,8079)	-0,1012	0,1252	(-0,3593 ; 0,1133)
	-	-	-	-	-	-
	-	-	-	-	-	-
P(4)	-1,0000	0,2924	(-1,2171 ; 0,0699)	-0,1804	0,1093	(-0,3939 ; 0,0094)
	0,3525	0,2430	(-0,0667 ; 0,7326)	-0,0929	0,1408	(-0,3790 ; 0,1384)
	-	-	-	-	-	-
	-	-	-	-	-	-

4.3.2 Distribuição a Priori Geométrica

Neste conjunto de dados as distribuição a prioris geométricas apresentaram resultados a posteriori para o número de QTLs (Figura 19) inferiores quando comparados com os resultados obtidos a partir da atribuição de prioris Poisson (Figura 18). Similarmente ocorre em relação as estimativas a posteriori para a localização e efeitos genéticos dos QTLs. Além disso, deve-se notar nas Tabelas 25 e 26 que somente dois dos quatro QTLs simulados foram detectados corretamente, no entanto, as estimativas distanciaram dos valores verdadeiros.

Tabela 25 - Resumo estatístico a posteriori para o número de QTLs - Mapa III (distribuição a priori Geométrica)

Priori	Média	Desvio Padrão
Geom(1)	4,6806	2,4717
Geom(2)	2,7483	0,9649
Geom(3)	2,3345	0,7859
Geom(4)	2,3345	0,6070

Figura 19 - Distribuição a posteriori para o número de QTLs - Mapa III (distribuição a priori Geométrica)

Tabela 26 - Resumo estatístico a posteriori para a localização do QTL - Mapa III (distribuição a priori Geométrica)

Priori	Crom	Posição	Desvio Padrão	IC95%
Geom(1)	2	11,7010	2,6817	(2,3628 ; 15,1460)
	7	43,9374	21,5233	(16,7541 ; 67,2841)
	-	-	-	-
	-	-	-	-
Geom(2)	2	11,6558	2,9097	(2,0533 ; 14,5816)
	7	38,2482	20,9122	(16,7349 ; 67,5511)
	-	-	-	-
	-	-	-	-
Geom(3)	2	12,1390	3,2330	(1,1225 ; 15,4131)
	7	40,2726	21,1795	(20,2133 ; 70,9809)
	-	-	-	-
	-	-	-	-
Geom(4)	2	12,5138	1,4459	(9,5807 ; 14,5730)
	7	40,6424	24,7152	(17,9277 ; 68,9980)
	-	-	-	-
	-	-	-	-

Tabela 27 - Resumo estatístico a posteriori para os efeitos genéticos do QTL - Mapa III (distribuições a priori Geométricas)

Priori	a	Dp	IC95%	d	Dp	IC95%
Geom(1)	-0,9663	0,3137	(-1,2284 ; 0,0000)	-0,1893	0,1329	(-0,5492 ; 0,0114)
	0,4192	0,2467	(0,0308 ; 0,7661)	-0,0885	0,1306	(-0,3472 ; 0,1606)
	-	-	-	-	-	-
	-	-	-	-	-	-
Geom(2)	-0,9730	0,3062	(-1,2230 ; 0,0873)	-0,1916	0,1318	(-0,5012 ; 0,0198)
	0,3513	0,2392	(0,0719 ; 0,7678)	-0,0967	0,1096	(-0,2780 ; 0,11585)
	-	-	-	-	-	-
	-	-	-	-	-	-
Geom(3)	-1,0324	0,3158	(-1,2695 ; 0,1180)	-0,1746	0,0926	(-0,3463 ; 0,0057)
	0,4070	0,2833	(0,1262 ; 0,9994)	-0,1234	0,1150	(-0,2945 ; 0,0538)
	-	-	-	-	-	-
	-	-	-	-	-	-
Geom(4)	-1,0501	0,1386	(-1,2344 ; -0,7392)	-0,1557	0,0736	(-0,2926 ; -0,0420)
	0,4169	0,2891	(0,1496 ; 0,8170)	-0,1338	0,0916	(-0,2689 ; -0,0400)
	-	-	-	-	-	-
	-	-	-	-	-	-

4.3.3 Distribuição a Priori Uniforme

Como pode-se observar na Figura 20, a distribuição a priori Uniforme[0,L] apresenta uma grande variabilidade no número de QTLs detectados. Além disso, nota-se na Tabelas 28 e 29 que as estimativas para a localização dos dois QTLs detectados corretamente estão distantes dos valores verdadeiros.

Tabela 28 - Resumo estatístico a posteriori para o número de QTLs - Mapa III (distribuição a priori Uniforme[0,L])

Priori	Média	Desvio Padrão
Unif[0,L]	4,6806	2,4717

Figura 20 - Distribuição a posteriori para o número de QTLs - Mapa III (distribuição a priori Uniforme[0,L])

Tabela 29 - Resumo estatístico a posteriori para a localização dos QTLs - Mapa III (distribuição a priori Uniforme[0,L])

Priori	Crom	Média	Desvio Padrão	IC95%
Unif[0,L]	2	11,7010	2,6819	(2,3628 ; 15,1460)
	7	43,9374	21,5233	(16,7541 ; 67,2841)
	-	-	-	-
	-	-	-	-

Tabela 30 - Resumo estatístico a posteriori para os efeitos genéticos - Mapa III (distribuição priori Uniforme[0,L])

Priori	a	Dp	IC95%	d	Dp	IC95%
Unif[0,L]	-0,9663	0,3137	(-1,2284 ; 0,0000)	-0,1893	0,1329	(-0,5492 ; 0,0114)
	0,4192	0,2467	(0,0308 ; 0,7661)	-0,0885	0,1306	(-0,3472 ; 0,1606)
	-	-	-	-	-	-
	-	-	-	-	-	-

4.4 Oito QTLs Afetando a Característica de Interesse

Supondo-se que oito QTLs afetam a característica de interesse e considerando o conjunto de dados simulados IV denotado por Mapa IV. Apresenta-se, a seguir, a análise Bayesiana, realizada no *software* WinQTLCart, utilizando o método MCMC-SR, de acordo com a atribuição da distribuição a priori de Poisson; Geométrica e Uniforme para o número de QTLs.

4.4.1 Distribuição a Priori Poisson

A partir da Figura 21 e Tabela 31, nota-se que a distribuição a priori Poisson que melhor se aproxima do número de QTLs simulado neste conjunto de dados é a distribuição a priori Poisson com média de quatro QTLs.

Figura 21 - Distribuição a posteriori para o número de QTLs - Mapa IV (distribuição a priori Poisson)

Tabela 31 - Resumo estatístico a posteriori para o número de QTLs - Mapa IV (distribuição a priori Poisson)

Priori	Média	Desvio Padrão
P(1)	4,00079	0,940107
P(2)	4,82595	1,199224
P(3)	5,5543	1,35714
P(4)	6,2101	1,5242

4.4.2 Distribuição a Priori Geométrica

Conforme mostra a Figura 22 e a Tabela 32, a distribuição a priori Geométrica não apresenta bons resultados para a detecção a posteriori para o número de QTLs neste conjunto de dados. Esse fato é confirmado por Stephens & Fisch (1998), os quais optaram por uma distribuição de Poisson com média de quatro QTLs quando analisaram seus dados simulados contendo oito QTLs.

Figura 22 - Distribuição a posteriori para o número de QTLs - Mapa IV (distribuição a priori Geométrica)

Tabela 32 - Resumo estatístico a posteriori para o número de QTLs - Mapa IV (distribuição a priori Geométrica)

Priori	Média	Desvio Padrão
Geom(1)	13,2559	6,1246
Geom(2)	5,7347	1,7535
Geom(3)	4,3061	1,3956
Geom(4)	4,3061	1,2424

4.4.3 Distribuição a Priori Uniforme

Nota-se a partir da Figura 23 e Tabela 33, que quanto maior o número de QTLs afetando a característica de interesse maior é a variabilidade no número de QTLs detectado a posteriori. Na prática não se recomenda a utilização da distribuição a priori Uniforme para o número de QTLs.

Figura 23 - Distribuição a posteriori para o número de QTLs - Mapa IV (distribuição a priori Uniforme[0,L])

Tabela 33 - Resumo estatístico a posteriori para o número de QTLs - Mapa IV (distribuição a priori Uniforme[0,L])

Priori	Média	Desvio Padrão
Unif[0,L]	13,2559	6,1246

4.5 Conclusões

Com base nos resultados apresentados, nota-se que as estimativas a posteriori para a média, efeitos genéticos, localização dos QTLs não são influenciadas pela escolha das prioris para o número de QTLs. No entanto, a estimativa para o número de QTLs foi extremamente sensível a escolha dessas prioris. Para um número pequeno de QTLs influenciando a característica de interesse as estimativas a posteriori para todos os parâmetros desconhecidos foram excelentes, no entanto, quando aumentou esse número surgiu dúvidas quanto a estimativa a posteriori sobre número de QTLs afetando a característica. Dentre as três possibilidades

de distribuições a priori para o número de QTLs testadas, a priori Uniforme[0,L] apresentou os piores resultados, o que é confirmado através da literatura, pois, a maioria dos autores sugerem usar priori Poisson ou Geométrica para o número de QTLs. Além disso, observou-se que para todos os conjuntos de dados a distribuição a priori Poisson com média de três QTLs apresentou resultados razoáveis. Deve-se ressaltar que o *default* do programa WinQTLCart para o número de QTLs é Poisson com média de três QTLs.

5 CONSIDERAÇÕES FINAIS

Neste trabalho, para obter as estimativas a posteriori dos parâmetros de interesse, foi utilizada uma proposta alternativa ao método proposto por Satagopan (1996): o algoritmo MCMC com Saltos Reversíveis (MCMC-SR) introduzido por Green(1995). Este método permitiu a simulação da distribuições a posteriori sobre espaços de dimensões diferentes.

Neste contexto, surge a necessidade de aprofundar as pesquisas em relação as prioris atribuídas para o número de QTLs, já que o modelo demonstrou ser extremamente sensível a sua escolha.

Visto que nesse trabalho utilizou-se de um programa já implementado, surgiram dificuldades para interpretar alguns resultados, como por exemplo, a estimativa a posteriori do efeito de dominância. Esse problema será solucionado através da programação do método MCMC-SR para o mapeamento de QTLs.

Dessa maneira, o desenvolvimento deste trabalho permite

- (i) indicar a necessidade de implementação de programas mais eficientes para o Mapeamento Bayesiano de QTLs;
- (ii) recomendar a utilização cautelosa do *software* WinQTLCart | módulo BIM para o mapeamento de QTLs;
- (iii) pesquisar metodologias alternativas para mapeamento de QTLs;
- (iv) considerar modelos com epistasia;
- (v) considerar o efeito de ambiente na modelagem.

REFERÊNCIAS

BEARZOTI, E. Mapeamento de QTL In: PINHEIRO, J.B. CARNEIRO, J. B. **Análise de QTL no melhoramento de plantas**: segunda jornada em genética e melhoramento de plantas. Goiânia: FUNAPE, 2000, p.63-209

CASELLA, G.; GEORGE, I. E. Explaining the Gibbs Sampler. **The American Statistician**, New York, v.46, p.167-169 , 1992.

COELHO, A. S. G. Considerações gerais sobre análise de QTL. In: PINHEIRO J.B,; CARNEIRO, I.F. **Análise de QTL no melhoramento de plantas**. Goiânia: FUNAPE, 2000, 232 p.

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

DOERGE, R. Constructing genetics maps by rapid chain delineation. **Journal of Quantitative Trait Loci**, New York, v.2, p.43-52, 1996.

GAFFNEY, P. J. **An efficient reversible jump markov chain monte carlo approach to detect multiple loci and their effects in inbred crosses**. 2001. 174 p. Tese (Doctor of Philosophy in Statistics) - University of Wisconsin, Madison, 2001.

GAMERMAN, D.; MIGON, H. S. **Inferência Bayesiana: uma abordagem integrada**, Rio de Janeiro, 1997. 27 p.

GEMAN, S.; GEMAN, D. Estochastic relaxation, gibbs distributions and the Bayesian restoration of images. **IEE Transactions on pattern analysis and machine intelligence**, New York, v.6 , p.169-193, 1984.

GELFAND, A. E.; SMITH, A.F.M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, Alexandria, v.85 , p.398-409, 1990.

GEYER, C. Pratical Markov chain Monte Carlo. **Statistical Science**, Baltimore, v.7, p.437-511, 1992.

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

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

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

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

LANDER, E.S.; GREEN, P.; ABRAHAMSON, J.; BARLOW, A.; DALY, M.J.; LINCOLN, S.E.; NEWBURG, L. Mapmaker: an interactive computer package for constructing genetic linkage maps of experimental and natural populations. **Genomics**, San Diego, v.1, p.174-181, 1987.

LANDER, E.S.; BOTSTEIN, D. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. **Genomics**, San Diego, v.121, p.185-199, 1989.

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

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **Journal of Chemical Physics**, v.21, , p. 1087-1092, 1953.

PAULINO, P. D. **Estatística Bayesiana**. Lisboa, 2003. 446 p.

PATERSON, A.H.; DAMON, S.; HEWITT, J.D.; ZAMIR, D.; RAQBINOWITCH, H.D. Mendelian factors underlying quantitative traits in tomato comparison across species, generations and environments. **Genetics**, Austin, v.127, p.181-197, 1991.

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

ROBERT, C. P.; CASELLA, G. **Monte Carlo Statistical Methods**. New York, p.46, 1999.

SATAGOPAN, J.M.,YANDELL, B.S.; NEWTON, M. A.; OSBORN, T.C. A bayesian approach to detece quantitative trait loci using Markov chain Monte Carlo . **Genetics** , Austin, p.805-816, 1996.

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, 1978.

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

STAUB, J.E, SERQUEN, F.C., GUPTA, M. Genetic markers, map construction, and their application in plant breeding. **HortScience**, New York, v.31, n.5, p.729-740, 1996.

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

TANKSLEY, S.D.; MILLER, J.; PATERSON, A.; BERNATSKY, R. **Molecular mapping of plant chromossomes**. New York, p.157-173, 1988.

TANKSLEY, S.D. Mapping polygenes. **Genetics**, Austin, v. 27, p.43-59, 1993.

THODAY, J.M. Location of polygenes. **Nature**, New York, v.191, n.47860, p.368-370, 1961.

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

ANEXO

Construção da distribuição marginal de Y

Visto que Q_i os genótipos dos QTLs do i -ésimo indivíduo, $i = 1, \dots, n$ são desconhecidos.

Tem-se que a distribuição conjunta de Y_i e Q_i é dada por:

$$f(y_i, Q_i) = f(y_i | Q_i)f(Q_i) \quad (12)$$

ou ainda,

$$f(y_i, Q_i) = f(y_i | Q_i)f(Q_i | \lambda, M) \quad (13)$$

Considerando-se que o genótipo do j -ésimo QTL do i -indivíduo é condicionalmente independente de outros genótipos do mesmo indivíduo tem-se:

$$f(y_i, Q_i) = f(y_i | Q_i) \prod_{i=1}^s f(Q_{ij} | \lambda, M) \quad (14)$$

Sendo assim, a distribuição marginal de Y_i será dada por:

$$f(y_i) = \sum_{q_{ij}} f(y_i | Q) \prod_{i=1}^s f(Q_{ij} = q_{ij} | \lambda, M) \quad (15)$$

ou seja, a distribuição marginal de Y_i é dada por uma mistura de distribuições, nesse caso, mistura de distribuições normais, conforme ilustra a Figura 10.

Figura 24: Exemplo ilustrativo da função de verossimilhança para o modelo QTL

A distribuição de probabilidade $f(Q_{ij} \mid \lambda, M)$ será obtida utilizando-se a informação de marcadores flanqueadores: M_k e M_{k+1} ($M_{k,k+1}$). as probabilidades condicionais para os genótipos do QTL, dados os genótipos dos marcadores são obtidas utilizando-se a expressão:

$$P(Q_{ij} \mid M_{k,k+1}) = \frac{P(Q_{ij} M_{k,k+1})}{P(M_{k,k+1})} \quad (16)$$

e são apresentadas na Tabela (34)

As 27 possibilidades para os genótipos dos QTLs dado os marcadores flanqueadores e suas probabilidades em função da frequência de recombinação r , r_1 e r_2 , em que r é a frequência de recombinação entre os marcadores M_k e $M_{k,k+1}$; r_1 , r_2 são as frequências de recombinação entre o QTL e as marcas M_k e M_{k+1} , respectivamente são apresentados na Tabela 34. Observamos que r é conhecida.

Tabela 34 - Probabilidade do Genótipo do QTL Dado os Marcadores Flanqueados de uma População F_2

m_k	m_{k+1}	$p(q_j = -1 m_k, m_{k+1})$	$p(q_j = 0 m_k, m_{k+1})$	$p(q_j = 1 m_k, m_{k+1})$
-1	-1	$\frac{n_1 n_2}{(1-R)^2}$	$\frac{2t_1 t_2}{(1-R)^2}$	$\frac{2d_1 d_2}{(1-R)^2}$
-1	0	$\frac{n_1 t_2}{R(1-R)}$	$\frac{t_1 (n_2 + d_2)}{R(1-R)}$	$\frac{d_1 t_2}{R(1-R)}$
-1	1	$\frac{n_1 d_2}{R^2}$	$\frac{2t_1 t_2}{R^2}$	$\frac{d_1 n_2}{R^2}$
0	-1	$\frac{t_1 n_2}{R(1-R)}$	$\frac{(n_1 + d_1) t_2}{R(1-R)}$	$\frac{t_1 d_2}{R(1-R)}$
0	0	$\frac{2t_1 t_2}{R^2 + (1-R)^2}$	$\frac{(n_1 + d_1)(n_2 + d_2)}{R^2 + (1-R)^2}$	$\frac{2t_1 t_2}{R^2 + (1-R)^2}$
0	1	$\frac{t_1 d_2}{R(1-R)}$	$\frac{(n_1 + d_1) t_2}{R(1-R)}$	$\frac{t_1 n_2}{R^2}$
1	-1	$\frac{d_1 n_2}{R^2}$	$\frac{2t_1 t_2}{R^2}$	$\frac{n_1 d_2}{R^2}$
1	0	$\frac{d_1 t_2}{R(R-1)}$	$\frac{t_1 (n_2 + d_2)}{R(R-1)}$	$\frac{n_1 t_2}{R(R-1)}$
1	1	$\frac{d_1 d_2}{(1-R)^2}$	$\frac{2t_1 t_2}{(1-R)^2}$	$\frac{n_1 n_2}{(1-R)^2}$

Ainda na Tabela 5, segue que $n_i = (1 - r_i)^2$, $t_i = (1 - r_i)$ e $d_i = r_i^2$, r_1 e r_2 são as razões de recombinação entre o QTL e as marcas M_k e M_{k+1} , respectivamente.

Sendo assim, a função densidade para um indivíduo i que tem genótipo dos marcadores flanqueadores $m_k = -1, m_{k+1} = -1$ deve ser ponderada pelas três possibilidades dos QTLs dados estes marcadores em cada posição λ no genoma, ou seja:

$$\begin{aligned}
 f(y_i | m_k = -1, m_{k+1} = -1, \lambda) = & \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{n_1 n_2}{(1-R)^2} \right] \exp \left(\frac{(y_i - \mu_{QQ})^2}{2\sigma^2} \right) + \\
 & + \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{2t_1 t_2}{(1-R)^2} \right] \exp \left(\frac{(y_i - \mu_{Qq})^2}{2\sigma^2} \right) + \\
 & + \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{2d_1 d_2}{(1-R)^2} \right] \exp \left(\frac{(y_i - \mu_{qq})^2}{2\sigma^2} \right).
 \end{aligned}$$

Por outro lado, se um indivíduo i tem genótipo $m_k = -1, m_{k+1} = 0$ a função densidade, será:

$$\begin{aligned}
f(y_i|m_k = -1, m_{k+1} = 0, \lambda) &= \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{n_1 t_2}{R(1-R)} \right] \exp \left(\frac{(y_i - \mu_{QQ})^2}{2\sigma^2} \right) + \\
&+ \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{t_1(n_2 + d_2)}{R(1-R)} \right] \exp \left(\frac{(y_i - \mu_{Qq})^2}{2\sigma^2} \right) + \\
&+ \frac{1}{\sqrt{2\pi\sigma^2}} \left[\frac{d_1 t_2}{R(1-R)} \right] \exp \left(\frac{(y_i - \mu_{qq})^2}{2\sigma^2} \right).
\end{aligned}$$