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

Modelo oculto de Markov para imputação de genótipos de marcadores moleculares: Uma aplicação no mapeamento de QTL utilizando a abordagem bayesiana

Elias Silva de Medeiros

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

Piracicaba
2014

Elias Silva de Medeiros
Bacharel em Estatística

Modelo oculto de Markov para imputação de genótipos de marcadores moleculares: Uma aplicação no mapeamento de QTL utilizando a abordagem bayesiana

versão revisada de acordo com a resolução CoPGr 6018 de 2011

Orientadora:
Prof^{a} Dr^{a} **ROSELI APARECIDA LEANDRO**

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

Piracicaba
2014

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

Medeiros, Elias Silva de
Modelo oculto de Markov para imputação de genótipos de marcadores moleculares: Uma aplicação no mapeamento de QTL utilizando a abordagem bayesiana / Elias Silva de Medeiros.- - versão revisada de acordo com a resolução CoPGr 6018 de 2011. - - Piracicaba, 2014.
89 p: il.

Dissertação (Mestrado) - - Escola Superior de Agricultura “Luiz de Queiroz”, 2014.

1. Imputação de genótipos 2. Mapeamento de QTL 3. MCMC com Saltos Reversíveis I. Título

CDD 519.5
M488m

DEDICATÓRIA

Aos meus pais,

Amauri e Joziene,

por acreditarem no meu sonho e que sempre estiveram comigo.

Com amor, DEDICO.

AGRADECIMENTOS

A Deus, pois tudo que sou e tenho foi Ele que concedeu.

Aos meus pais Joziene e Amauri pelo amor, carinho e por toda educação que me foi dada. Aos meus irmãos Levi, Sara, Elizeu e Estefânia. A minha noiva Marina Maestre por me apoiar e estar comigo em todos os momentos. As provações foram muitas mas Deus sempre nos deu vitória.

A todos da minha cidade natal, São João do Cariri, que sempre me deram apoio e incentivaram nessa minha jornada. Em especial, aos amigos e colegas que fizeram parte da minha trajetória acadêmica, os quais durante cinco anos estiveram comigo viajando de segunda a sexta-feira da minha cidade natal a Campina Grande.

Aos professores da UEPB, em especial aos Professores do Departamento de Estatística. Ao casal de professores Tiago e Ana Patrícia (mainha) pelo acolhimento na minha chegada a Piracicaba.

Aos Professores e funcionários do Departamento de Ciências Exatas da ESALQ/USP pela amizade e formação.

A Professora Dr. Roseli Aparecida Leandro pela orientação e que sempre me deu todo apoio nas minhas decisões. Meu muito obrigado Roseli.

A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pela concessão da bolsa de estudos.

O meu muito obrigado!

"Desculpem se errei. Mas se errei foi tentando acertar."
Meu pai – Pr. José Amauri de Medeiros

"DEUS É BOM!!!"

SUMÁRIO

RESUMO

Modelo oculto de Markov para imputação de genótipos de marcadores moleculares: Uma aplicação no mapeamento de QTL utilizando a abordagem bayesiana

Muitas são as características quantitativas que são, significativamente, influenciadas por fatores genéticos, em geral, existem vários genes que colaboram para a variação de uma ou mais características quantitativas. As informações ausentes a respeito dos genótipos nos marcadores moleculares é um problema comum em estudo de mapeamento genético e, por conseguinte, no mapeamento dos locus que controlam estas características fenotípicas (QTL). Os dados que não foram observados ocorrem, principalmente, devido a erros de genotipagem e de marcadores não informativos. Para solucionar este problema foi utilizado o método do modelo oculto de Markov para inferir estes dados. Os métodos de acurácias evidenciaram o sucesso da aplicação desta técnica de imputação. Uma vez imputado, na inferência bayesiana estes dados não serão mais tratados como uma variável aleatória resultando assim, numa redução no espaço paramétrico do modelo. Outra grande dificuldade no mapeamento de QTL se deve ao fato de que não se conhece ao certo a quantidade destes que influenciam uma dada característica, fazendo com que surjam diversos problemas, um deles é a dimensão do espaço paramétrico e, consequentemente, a obtenção da amostra a posteriori. Assim, com o objetivo de contornar este problema foi proposta a utilização do método Monte Carlo via cadeia de Markov com Saltos Reversíveis, uma vez que este permite flutuar, entre cada iteração, modelos com diferentes quantidades de parâmetros. A utilização da abordagem bayesiana permitiu detectar cinco QTL para a característica estudada. Todas as análises foram implementadas no programa estatístico R.

Palavras-chave: Imputação de genótipos; Mapeamento de QTL; MCMC com Saltos Reversíveis

ABSTRACT

Hidden Markov model for imputation of genotypes of molecular markers: An application in QTL mapping using Bayesian approach

There are many quantitative characteristics which are significantly influenced by genetic factors, in general, there are several genes that contribute to the variation of one or more quantitative trait. The missing information about the genotypes in molecular markers is a common problem in studying genetic mapping and therefore the mapping of loci that control these phenotypic traits (QTL). The data were not observed occur mainly due to errors in genotyping and uninformative markers. To solve this problem the method of occult Markov model to infer this information was used. Techniques accuracies demonstrated the successful application of this technique of imputation. Once allocated, in the Bayesian inference this data will no longer be treated as a random variable thus resulting in a reduction in the parameter space of the model. Another great difficulty in mapping QTL is due to the fact that no one knows exactly the amount of these which influence a given characteristic, so that several problems arise, one of them is dimension of the parameter space and, consequently, obtaining the sample a posterior. Thus, in order to solve this problem using the method via Monte Carlo Markov chain Reversible Jump was proposed, since this allows fluctuate between each iteration, models with different numbers of parameters. The use of the Bayesian approach allowed five QTL detected for the studied trait. All analyzes were implemented in the statistical software R.

Keywords: Imputation of genotypes; QTL mapping; Reversible jump MCMC

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

O estudo detalhado dos locus que influenciam uma característica fenotípica, denominados de QTL (do inglês, *Quantitative trait loci*), é de fundamental importância em várias áreas da ciência, tais como, a agricultura, a medicina humana e a biologia evolutiva. Um mapeamento eficiente e robusto do genoma para posições desses genes é uma meta muito importante na genética quantitativa. A análise dos marcadores moleculares, em todo o genoma, fornece os meios para localizar e mapear os QTL de uma forma sistemática (E SILVA; ZENG, 2010).

Mas, sabe-se que, os grandes conjuntos de dados derivados desses marcadores contém uma quantidade significativa de genótipos ausentes. Os dados ausentes ocorrem, principalmente, devido a erros de genotipagem e de marcadores não informativos. De acordo com Roberts et al. (2007), na prática, existem algumas alternativas para lidar com este tipo de problema, tais como, repetir a genotipagem em regiões com genótipos ausentes (as vezes inviável, devido ao alto custo operacional); remover os marcadores que possuem genótipos ausentes (implicam perdas de informações); e o mais aconselhado, inferir os dados ausentes.

O intuito neste trabalho é inferir os genótipos dos marcadores não observados por meio de imputações. As informações ausentes a respeito dos genótipos nos marcadores moleculares é um problema comum em estudo de mapeamento genético e, por conseguinte, no mapeamento de QTL. Para solucionar este problema se faz necessária à utilização de técnicas de imputação para inferir os dados desses genótipos (HOWIE; MARCHINI; STEPHENS, 2011; LI et al., 2009). Existem diversos programas computacionais que são utilizados para imputação, como por exemplo, o IMPUTE (ZHAO, 2008) e o BEAGLE (BROWNING; BROWNING, 2009). Ambos os programas são baseados em modelos ocultos de Markov (HMM, do inglês, Hidden Markov model).

Os dados dos genótipos nos marcadores serão aqui utilizados para inferir as localizações de possíveis QTL, como também detecta-los no intervalo constituído entre dois marcadores. Assim, será realizada uma análise preliminar, no que diz respeito à imputação dos genótipos não observados nesses marcadores, para que, ao fazer inferência no intervalo entre dois marcadores, possam-se ter estimativas mais confiáveis e plausíveis. Com isso, a acurácia das técnicas para mapear QTL se torna maior. Tem-se também que, a imputação desses dados permite aos geneticistas avaliarem com precisão a evidência de

possíveis marcadores associados à QTL (BROWNING; BROWNING, 2009).

Neste trabalho o mapeamento de QTL será realizado por meio de métodos bayesianos, pois estes possibilitam tratar a quantidade de QTL como variável desconhecida, implicando em vantagens consideráveis para a modelagem. O grande problema quando se utiliza esta metodologia é o da obtenção da amostra aleatória da distribuição conjunta a posteriori, uma vez que, ao considerar a quantidade de QTL como uma incerteza, a dimensão do espaço paramétrico pode variar. Green (1995) propôs, como resolução deste problema, o algoritmo MCMC com Saltos Reversíveis, este algoritmo permite saltar entre modelos com dimensões diferentes por meio da especificação de distribuições propostas, ou seja, poderá ocorrer em cada nova iteração o nascimento ou morte de um QTL. Muitos trabalhos seguiram as ideias deste autor, tais como, (SATAGOPAN; YANDELL, 1996; STEPHENS; FISCH, 1998; YI, 2004; LEE; VAN DER WERF, 2006; YI et al., 2007), dentre outros.

2 REVISÃO BIBLIOGRÁFICA

2.1 Processo estocástico

Um processo estocástico é uma família de variáveis aleatórias $\{G(t), t \in T\}$ definidas em um espaço de probabilidade t pertencente ao conjunto T. O conjunto T é dito *espaço paramétrico*, os valores assumidos por $G(t)$ são denominados de estados e o conjunto de todos os possíveis estados é chamado de espaço de estados (KARLIN; TAYLOR, 1981)

Dado um valor fixo de t, $G(t)$ será uma variável aleatória que corresponde ao estado do processo no instante t. Para uma coleção finita $t_1, t_2, \ldots, t_n$, $G_{t_1}, G_{t_2}, \ldots, G_{t_n}$ será um conjunto de n variáveis aleatórias com distribuição conjunta. Conhecendo-se a distribuição conjunta ou a função de densidade para cada conjunto de variáveis aleatórias é possível determinar a estrutura de probabilidade do processo G_t. De acordo com Bruce e Disney (1970) análise do processo estocástico visa, principalmente, determinar estas distribuições conjuntas para prever o processo futuro, dado um determinado comportamento no passado.

Sabe-se que os valores que tomam as variáveis do processo serão chamados de estados e o conjunto C destes valores será o espaço de estados. Não necessariamente estes estados precisam ser de quantidade numérica, poder-se-á um conjunto de símbolos, por exemplo.

As seguintes possibilidades para os processos estocásticos podem ser classificadas como:

a) C enumerável e T enumerável: Processo a tempo discreto com espaço de estados discreto;

b) C enumerável e T intervalo: Processo a tempo contínuo com espaço de estados discreto;

c) C não enumerável e T enumerável: Processo a tempo discreto com espaço de estados contínuo;

d) C não enumerável e T intervalo: Processo a tempo contínuo com espaço de estados contínuo.

Em um processo estocástico o comportamento probabilístico não está apenas relacionado às distribuições marginais das variáveis, mas também pelas relações de dependência entres elas. Existem vários tipos de processos estocásticos, porém neste trabalho será discutido apenas um deles, o **processo de Markov** ou cadeia de Markov.

Segundo Bruce e Disney (1970) a estrutura de probabilidade de uma sequência aleatória, ou processo aleatório de parâmetro discreto, é determinada pelas probabilidades conjuntas que são expressas da forma,

$$p\left(j_0, j_1, \ldots, j_k\right) = P\left[G_0 = j_0, G_1 = j_1, \ldots, G_k = j_k\right]. \tag{1}$$

A expressão (1) será denominada de *processo de Markov* ou *cadeia de Markov* se, para cada k, a probabilidade condicional de que o sistema esteja em um dado estado após k, dependerá apenas do estado do passo imediatamente anterior $k-1$. Em outras palavras, para predizer o valor de G_k, todo o conhecimento de que se tem a respeito de $G_0, G_1, \ldots, G_{k-1}$ não será necessário, bastará apenas da informação de G_{k-1}.

Matematicamente, pode-se escrever uma expressão que represente tudo o que foi falado no parágrafo anterior:

$$p\left(j_0, j_1, \ldots, j_{k-1}\right) = P\left[G_k = j_k | G_{k-1} = j_{k-1}\right]. \tag{2}$$

Se a expressão (2) for verdadeira para todo k, então poder-se-á utilizar a identidade $P\left[A \bigcap B\right] = P\left[A\right] P\left[B|A\right]$ para alcançar o seguinte resultado,

$$p\left(j_0, j_1, \ldots, j_k\right) = p\left(j_0\right) p\left(j_1|j_0\right) \cdots p\left(j_k|j_{k-1}\right). \tag{3}$$

Para mais detalhes algébricos consultar as referências (BRUCE; DISNEY, 1970).

Na equação (3) as expressão $p(j_k|j_{k-1})$ e $p(j_0)$ são chamadas de probabilidades de transição e o conjunto de probabilidades iniciais, respectivamente. Assim, as probabilidades de transição de um passo são escritas da forma,

$$p_{ij} = P\left[G_k = j | G_{k-1} = i\right]. \tag{4}$$

E as probabilidades iniciais,

$$p_i^0 = P\left[G_0 = i\right]. \tag{5}$$

Pelo o que foi visto, tem-se que a expressão (3) pode ser reescrita da forma,

$$p\left(j_0, j_1, \ldots, j_k\right) = p_{j_0} p_{j_0 j_1} \cdots p_{j_{k-1} j_k}. \tag{6}$$

De forma análoga, essas probabilidades para os n passos são estabelecidas da forma,

$$p_{ij}^{(n)} = P\left[G_{k+n} = j | G_k = i\right]. \tag{7}$$

Na equação (7), $p_{ij}^{(n)}$ é a probabilidade de que o processo passe do estado i para o estado j em n passos.

Após esta breve revisão sobre processo estocástico, na próxima seção deste trabalho será abordado um caso especial do processo de Markov, os modelos ocultos de Markov.

2.2 Modelos ocultos de Markov: imputação dos genótipos dos marcadores

Por conveniência, algumas das notações utilizadas na seção Processo estocástico não serão mantidas, visando à analogia para os dados de marcadores moleculares que serão apresentados nesta e nas próximas seções.

Como visto anteriormente, uma cadeia de Markov, caso especial de um processo estocástico, é uma sequência de variáveis aleatórias $G_1, G_2, \ldots G_t, G_{t+1}, \ldots$, cuja distribuição de probabilidade de G_{t+1} está em função apenas de G_t, ou seja,

$$P(G_{t+1} = j | G_t = i, G_{t-1}, \ldots, G_1) = P(G_{t+1} = j | G_t = i). \tag{8}$$

Estas probabilidades podem ser representadas por meio de uma matriz de transição $\mathbf{A}$. A Figura 1 mostra um esboço de um diagrama para uma cadeia de Markov.

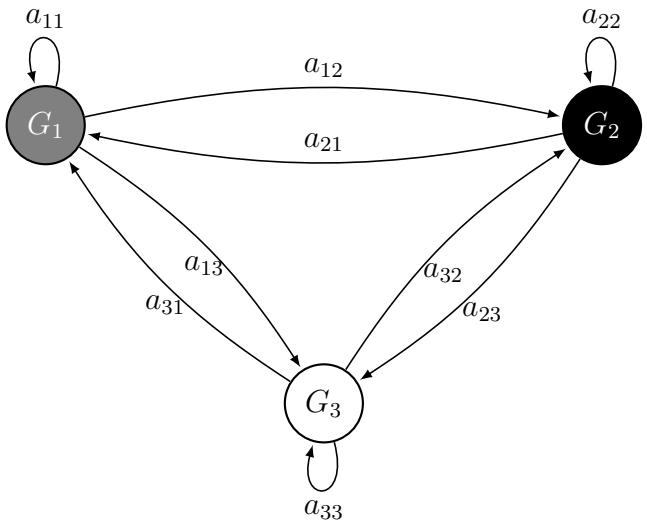


Figura 1 - Diagrama da matriz de transição de uma cadeia de Markov com três estados

Para esta ilustração a matriz de transição $\mathbf{A}$ é escrita da forma,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix},$$

em que, $\sum_{ij=1}^{N} a_{ij} = 1$, sendo neste caso, $N = 3$.

No entanto, em vários experimentos, os estados da cadeia de Markov não são diretamente observáveis, mas sim uma sequência de sinais resultantes de um conjunto de processos estocásticos que produzem uma sequência de observações, ou seja, a observação é uma função probabilística do estado, quando isto acontece tem-se um Modelo Oculto de Markov (HMM - do inglês *Hidden Markov Model*) (RABINER, 1989). Portanto, este tipo de modelo é caracterizado por conter uma sequência de estados que estão ocultos, mas podem ser previstos a partir de uma sequência dos estados observados (DUTHEIL et al., 2009).

Considere uma sequência de estados distintos $\mathcal{G} = \{S_1, S_2, \ldots, S_N\}$, em que o estado na posição ou no tempo t será representado por g_t, $t = 1, 2, \ldots, T$, e N é número total de estados distintos e uma sequência de símbolos de observações distintas $\mathcal{O} = \{V_1, V_2, \ldots, V_M\}$. Permita agora a seguinte ilustração.

Seja um experimento em que uma pessoa ficou confinada em sua própria

casa por alguns dias e em um dado momento lhe fora perguntado, como o "tempo" estava lá fora? Ensolarado, chuvoso ou nebuloso? Esta então seria sua sequência de estados distintos. Sendo que, neste mesmo experimento a única informação que a pessoa teria conhecimento era a forma de como seu zelador chegou a sua casa, com ou sem o guarda-chuva, ou seja, essa seria a sua sequência de símbolos observados.

Agora, dando continuidade a formulação do modelo.

A distribuição de probabilidade de transição estará representada na matriz $\mathbf{A}_{N\times N}$, e que cada elemento dessa matriz, a_{ij}, é calculado de acordo com a probabilidade, $a_{ij} = P\left(g_{t+1} = S_j | g_t = S_i\right), \quad ij = 1, 2, \ldots, N$. Todos os estados $S_i, S_j \in \mathcal{G}$. Ao elemento a_{ij} leia-se como a probabilidade de ocorrer o estado j no tempo $t+1$ dado o estado i no tempo t. No tempo $t = 0$, tem-se a definição da probabilidade do estado inicial $\pi_i = P\left(g_1 = S_i\right) \forall S_i \in \mathcal{G}$. Assim, a matriz $\mathbf{A}_{N\times N}$ é construída da seguinte forma,

$$\mathbf{A}_{N\times N} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{pmatrix}.$$

A distribuição de probabilidade dos símbolos observados nos estados $\mathcal{G}$ será apresentada na matriz $\mathbf{E}_{N\times M}$, sendo que cada elemento desta matriz é denotado por $e_{ik} = P\left(V_k | g_t = S_i\right)$, com $k = 1, 2, \ldots, M$. O elemento e_{ik} será denotado como a probabilidade de emissão. A matriz com estes elementos é escrita da forma,

$$\mathbf{E}_{N\times M} = \begin{pmatrix} e_{11} & e_{12} & \cdots & e_{1M} \\ e_{21} & e_{22} & \cdots & e_{2M} \\ \vdots & \vdots & & \vdots \\ e_{N1} & e_{N2} & \cdots & e_{NM} \end{pmatrix}.$$

As três medidas de probabilidades π_i, a_{ij} e e_{ik} especificam um HMM, por completo. Para cada uma destas medidas existirá um parâmetro e o conjunto de todos esses parâmetros será representado por $\boldsymbol{\theta}$.

A probabilidade de uma sequência de estados $\mathcal{G}$ proveniente de um HMM, composto pelo conjunto de parâmetros $\boldsymbol{\theta}$, é correspondente ao produto das probabilidades

de transição, que é escrito da forma,

$$\begin{aligned} p\left(\mathcal{G}|\boldsymbol{\theta}\right) &= \prod_{t=0}^{T} a_{g_t g_{t+1}} = a_{g_0 g_1} \prod_{t=1}^{T} a_{g_t g_{t+1}} \\ &= \pi_{g_1} \prod_{t=1}^{T} a_{g_t g_{t+1}}. \end{aligned} \tag{9}$$

A função de distribuição de uma sequência observável $\mathcal{O}$ dada uma sequência de estados $\mathcal{G}$ e um conjunto de parâmetros $\boldsymbol{\theta}$ é escrita da forma,

$$p\left(\mathcal{O}|\mathcal{G},\boldsymbol{\theta}\right) = \prod_{t=1}^{T} P\left(o_t|g_t,\boldsymbol{\theta}\right). \tag{10}$$

Assumindo que na equação (10) as observações são independentes, tem-se então,

$$p\left(\mathcal{O}|\mathcal{G},\boldsymbol{\theta}\right) = e_{g_1}\left(o_1\right) \times e_{g_2}\left(o_2\right) \times \ldots \times e_{g_T}\left(o_T\right), \tag{11}$$

sendo que, na equação (11), $e_{g_t}\left(o_t\right)$ é a probabilidade com que o estado g_t emite a observação o_t. Para uma sequência observável $\mathcal{O}$ ao longo de $\mathcal{G}$, a função de distribuição conjunta é composta como o produto de duas quantidades definidas nas equações (9) e (11), a qual é expressa da forma,

$$p\left(\mathcal{O},\mathcal{G}|\boldsymbol{\theta}\right) = p\left(\mathcal{G}|\boldsymbol{\theta}\right) \times p\left(\mathcal{O}|\mathcal{G},\boldsymbol{\theta}\right). \tag{12}$$

De acordo com Rabiner (1989) existem três problemas básicos que podem ocorrer diante de um HMM e que devem ser resolvidos para que o uso desse tipo de modelo seja útil, e que venha a ser aplicado nos experimentos. A seguir serão apresentados estes problemas.

1. Para uma sequência de observações $\mathcal{O}$ e o conjunto de parâmetros $\boldsymbol{\theta}$, como calcular a probabilidade, de maneira eficiente, $P\left(\mathcal{O}|\boldsymbol{\theta}\right)$.

Uma forma mais elegante de calcular $P\left(\mathcal{O}|\boldsymbol{\theta}\right)$, é determinando $P\left(\mathcal{O}|\mathcal{G},\boldsymbol{\theta}\right)$ para uma sequência de estados fixos $\mathcal{G}$, e em seguida, multiplicar por $P\left(\mathcal{G}|\boldsymbol{\theta}\right)$ e somar

sobre todos os possíveis elementos de $\mathcal{G}$. Ou seja,

$$\begin{aligned} P(\mathcal{O}|\mathcal{G},\boldsymbol{\theta}) &= \sum_{T} \left[P(\mathcal{O}|\mathcal{G},\boldsymbol{\theta}) \times P(\mathcal{G}|\boldsymbol{\theta})\right] \\ &= \sum_{T} \left[\pi_{g_1} \times e_{g_1}(o_1)\right] \times \left[a_{g_1,g_2} \times e_{g_2}(o_2)\right] \times \ldots \times \left[a_{g_{t-1},g_t} \times e_{g_t}(o_t)\right]. \end{aligned} \tag{13}$$

Na equação (13), tem-se que a soma envolverá $(2T-1)N^T$ multiplicações e N^T-1 adições, sendo necessário a utilização de um procedimento mais eficiente. Como solução são utilizados os algoritmos *forward* e *backward*, uma vez que estes reduzem significativamente o tempo computacional das análises (YU; KOBAYASHI, 2003). De acordo com Khreich et al. (2010) estes algoritmos são técnicas de programação dinâmica que constituem a base para determinar as estimativas dos parâmetros contido em um HMM.

Um dos algoritmos mais tradicionais em um HMM é o algoritmo *forward*. Este calcula a probabilidade de ocorrer toda a sequência de observações $\mathcal{O}$ dado o modelo, $P(\mathcal{O}|\boldsymbol{\theta})$ (NIELSEN; SAND, 2011). A seguir, um breve esboço de como este algoritmo é programado.

Considere então uma variável definida da forma,

$$\alpha_t(i) = P(o_1, o_2, \ldots, o_t, g_t = S_i|\boldsymbol{\theta}). \tag{14}$$

A equação (14) é entendida como a probabilidade de ter uma sequência parcial de observações até o instante t e neste momento o modelo se encontra no estado S_i, condicionado ao conjunto de parâmetros $\boldsymbol{\theta}$ (RABINER, 1989). A seguir serão apresentados os três passos para a execução do algoritmo *forward*.

a. Inicialização, $t = 1$:

$$\alpha_1(i) = P(o_1, g_1 = S_i|\boldsymbol{\theta}) = \pi_{g_i} \times e_{g_i}(o_1), \quad 1 \leq i \leq N.$$

b. Indução:

$$\alpha_{t+1}(j) = \left[\sum_{i=1}^{N} \alpha_t(i) \times a_{ij}\right] \times e_{g_j}(o_{t+1}), \qquad \begin{array}{l} t = T-1, T-2, \ldots, 1 \\ 1 \leq j \leq N. \end{array}$$

c. Finalização:

$$P\left(\mathcal{O}|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \alpha_T\left(i\right).$$

De maneira análoga, o algoritmo *backward* definido pela variável $\beta_t\left(i\right)$, será determinado a partir das probabilidades:

$$\beta_t\left(i\right) = P\left(o_{t+1}, o_{t+2}, \ldots o_T | g_t = S_i, \boldsymbol{\theta}\right),$$

em que $\beta_t\left(i\right)$ pode ser calculada por indução. Assim, sejam os três passos:

a. Inicialização:

$$\beta_t\left(i\right) = 1, \quad 1 \leq i \leq N.$$

b. Indução:

$$\beta_t\left(i\right) = \sum_{j=1}^{N} a_{ij} \times e_{g_j}\left(o_{t+1}\right) \times \beta_{t+1}\left(j\right), \qquad \begin{array}{l} 1 \leq t \leq T-1 \\ 1 \leq i \leq N. \end{array}$$

c. Finalização:

$$P\left(\mathcal{O}|\boldsymbol{\theta}\right) = \sum_{i=1}^{N} \pi_{g_i} \times \beta_1\left(i\right) \times e_{g_i}\left(o_1\right).$$

2. O segundo problema consiste em como definir uma sequência "ótima" de estados, dada uma sequência de observações $\mathcal{O}$ e um conjunto de parâmetros $\boldsymbol{\theta}$.

Seja então definida uma nova variável $\gamma_t\left(i\right)$, escrita da forma,

$$\gamma_t\left(i\right) = P\left(g_t = S_i|\mathcal{O}, \boldsymbol{\theta}\right), \tag{15}$$

ou seja, a probabilidade de iniciar o estado S_i no tempo ou na posição t, dada uma sequência de observações $\mathcal{O}$ e o conjunto $\boldsymbol{\theta}$. Conforme o teorema de Bayes, sabe-se que $P\left(g_t = S_i, \mathcal{O}|\boldsymbol{\theta}\right) = P\left(g_t = S_i|\mathcal{O}, \boldsymbol{\theta}\right) \times P\left(\mathcal{O}|\boldsymbol{\theta}\right)$. Sendo assim, a equação (15) pode ser

reescrita em função dos algoritmos *forward* e *backward*,

$$\gamma_t(i) = \frac{P(g_t = S_i, \mathcal{O}|\boldsymbol{\theta})}{P(\mathcal{O}|\boldsymbol{\theta})} = \frac{\alpha_t(i)\,\beta_t(i)}{\sum_{i=1}^{N} \alpha_t(i)\,\beta_t(i)}. \tag{16}$$

Na equação (16) verifica-se que $\sum_{i=1}^{N} \gamma_t(i) = 1$. Para determinar o estado mais provável de ocorrer no tempo t, basta fazer,

$$g_t = \underset{1 \leq i \leq N}{\arg\max} \left[\gamma_t(i)\right], \quad 1 \leq t \leq T. \tag{17}$$

Porém a equação (17) não assegura que a sequência de estados escolhida possa ser a "ideal". Por exemplo, quando uma dada probabilidade de transição é igual a zero, tem-se que o resultado para a sequência de estados poderá conter um estado inválido. Segundo Rabiner (1989) este problema acontece devido a equação (17) selecionar o estado mais provável para cada instante, sem levar em consideração a probabilidade de ocorrência de toda a sequência de estado. Com isso se faz necessário a aplicação de uma técnica para determinar uma sequência "ideal" plausível, baseada em métodos de programação dinâmica. Assim, para a resolução deste problema será utilizado o algoritmo de Viterbi (VITERBI, 1967). Segundo De Fonso, Aluffi-Pentini, Parisi (2007) e Viterbi (2006) o algoritmo de Viterbi foi projetado de modo a evitar uma enorme complexidade no que diz respeito a determinar o máximo de uma função. É um algoritmo computacionalmente eficiente para determinar a sequência mais provável de estados. Este faz uso de duas variáveis, $\delta_t(i)$ e $\psi_t(i)$, as quais serão definidas a seguir.

$$\delta_t(i) = \max_{g_1, g_2, \ldots, g_{t-1}} P\left[g_1, g_2, \ldots, g_t = S_i, o_1, o_2, \ldots, o_t|\boldsymbol{\theta}\right],$$

em que $\delta_t(i)$ representa a probabilidade máxima de uma única sequência de dentre todas as possíveis que terminam no estado S_i no tempo t. A segunda variável, $\psi_t(i)$, tem por finalidade permitir acompanhar a melhor sequência final no estado S_i no tempo t, a qual é definida da forma,

$$\psi_t(i) = \underset{g_1, g_2, \ldots, g_{t-1}}{\arg\max} P\left[g_1, g_2, \ldots, g_t = S_i, o_1, o_2, \ldots, o_t|\boldsymbol{\theta}\right].$$

Assim, a programação do algoritmo de Viterbi é composta dos seguintes passos:

a. Inicialização,

$$\begin{aligned}\delta_1(i) &= \pi_{g_i} \times e_{g_i}(o_i), \quad 1 \leq i \leq N \\ \psi_1(i) &= 0.\end{aligned}$$

b. Recursão,

$$\delta_t(j) = \max_{1 \leq i \leq N} (\delta_{t-1}(i) \times a_{ij}) \times e_{g_j}(o_t), \qquad 2 \leq t \leq T \;\; e \;\; 1 \leq j \leq N$$

$$\psi_t(j) = \arg\max_{1 \leq i \leq N} (\delta_{t-1}(i) \times a_{ij}).$$

c. Terminação,

$$\begin{aligned}P^*(\mathcal{O}|\boldsymbol{\theta}) &= \max_{1 \leq i \leq N} \delta_T(i) \\ g_T^* &= \arg\max_{1 \leq i \leq N} [\psi_T(i)].\end{aligned}$$

d. Retrocedendo,

$$G^* = \{g_1^*, g_2^*, \ldots, g_T^*\}, \text{ tal que, } g_t^* = \psi_{t+1}\left(g_{t+1}^*\right).$$

3. O terceiro problema esta relacionado a reestimação dos parâmetros do modelo, $\boldsymbol{\theta}$. Ou seja, dada uma sequência de observações $\mathcal{O}$ e $\boldsymbol{\theta}$, como ajustar os valores de $\mathbf{A}$, $\mathbf{E}$ e π de forma a maximar $P(\mathcal{O}|\boldsymbol{\theta})$.

O algoritmo EM é o principal instrumento para determinar as estimativas dos parâmetros no HMM. No entanto, este procedimento vem sendo substituído pelo algoritmo de Viterbi Training (VT), também conhecido na literatura como algoritmo K-médias (ou no inglês, segmental K-means), pois este é computacionalmente menos intenso e mais estável (HUMBURG; BULGER; STONE, 2008; LEMBER; KOLOYDENKO, 2008). Para um HMM e uma dada sequência de observações, o algoritmo VT realiza inferências sobre os parâmetros do HMM. Este algoritmo, em geral, converge mais rapidamente do que

outros algoritmos, como por exemplo o de Baum-Welch (BAUM; PETRIE, 1966), sendo que ambos podem convergir para um máximo local. Humburg, Bulger e Stone (2008) realizaram um estudo em que comparam estes dois algoritmos e verificaram que, para um grande número de iterações, acima de 60, o VT produz estimativas dos parâmetros similares ao algoritmo de Baum-Welch.

Em um HMM, o caminho mais provável para cada sequência do treinamento é obtido usando o algoritmo de decodificação Viterbi. Com base neste caminho, os estados de transição e os de emissão das observações são estimados e então utilizados para a reestimação dos parâmetros do HMM (AL-ANI, 2011; LAM; IRMTRAUD, 2010), ou seja, para cada iteração do algoritmo VT é gerado um novo conjunto para os parâmetros do modelo derivado a partir das probabilidades de transição e de emissão. Assim, o algoritmo VT é constituído dos seguintes passos:

a. Atribuir valores iniciais para os parâmetros do modelo.

b. Obter a sequência de estados mais provável $\mathcal{G}$ por meio do algoritmo de Viterbi.

c. Calcular, a_{ij} e $e_{g_i}(O_i)$ dado $\mathcal{G}$.

d. Estimar os parâmetros do novo modelo usando as ocorrências estimadas dos estados de transição e de emissão e retornar ao passo (b).

O algoritmo VT consiste em classificar os dados de acordo com as próprias informações contidas no experimento, por meio de comparações de distâncias. Para a implementação utiliza-se da distância euclidiana para realizações de tais comparações. Esta computa a semelhança por meio da distância entre duas distribuições vetoriais, quanto menor a distância entre as distribuições maior será a semelhança entre as mesmas (DU; CHANG, 2001). Para mais detalhes algébricos de como o algoritmo VT está relacionado a distância euclidiana consulte na referência (JUANG; RABINER, 1990).

Após esta revisão bibliográfica sobre HMM a próxima seção descreverá alguns conceitos genéticos que serão utilizados no decorrer deste trabalho.

2.3 Conceitos básicos

Neste trabalho serão abordados alguns conceitos básicos de genética com o intuito de facilitar a leitura e a compreensão do leitor.

Segundo Hardy[1] (1908) e Weinberg[2] (1908) apud Hallauer et al. (2010) em 1908, Hardy e Weinberg, independentemente demonstraram que numa grande população de acasalamento aleatório, as frequências genotípicas permaneceram constantes de geração a geração, e as proporções genotípicas atingiram um equilíbrio estável. Portanto, tal população é dita estar em equilíbrio de Hardy-Weinberg e permanece assim a menos que qualquer força perturbadora mude seu gene ou frequência genotípica.

Sabe-se ainda que, um gene que segregar numa população pode afetar o fenótipo de uma característica. Para uma característica complexa ou quantitativamente herdada, os genes que a determinam podem ser numerosos e suas relações com o meio ambiente podem ser complicadas. Considere uma característica quantitativa, com valor fenotípico F, o qual é determinado pelo valor genotípico G e o desvio ambiental E,

$$F = G + E.$$

Considere-se um gene com dois alelos, A_1 e A_2, com respectivas frequências, p_1 e p_0 em uma população F_2. Sejam P_2, P_1 e P_0 populações de frequências nos três genótipos A_1A_1, A_1A_2 e A_2A_2 cujos valores e as frequências dos genótipos na população em equilíbrio de Hardy-Weinberg é expressa da seguinte forma:

Genótipo	Valor genotípico	Frequência
A_1A_1	$\mu_2 = \mu + a$	$P_2 = p_1^2$
A_1A_2	$\mu_1 = \mu + d$	$P_1 = 2p_1p_0$
A_2A_2	$\mu_0 = \mu - a$	$P_0 = p_0^2$

Sabe-se que a soma das frequências será igual ao valor 1. O ponto médio μ entre os genótipos homozigotos, medirá o afastamento, $+a$ ou $-a$, de cada genótipo homozigoto em relação à média e, d mede o afastamento de cada genótipo heterozigoto em relação μ. Se $d = 0$, não existirá nenhuma dominância e a interação alélica é denominada aditiva; se $d = a$, indicará interação alélica de dominância completa; se $0 < d < a$, então a interação é de dominância parcial; e se $d > a$ conclui-se que a interação é de sobredominância.

O grau de dominância (GD) que descreve o tipo de interação alélica é des-

[1]HARDY, G.H. Mendelian proportions in a mixed population. **Science**, Cambridge - England, v.78, p.49-50, 1908.

[2]WEINBERG, W. Über den Nachweis der Vererbung beim Menschen. **Jahreshefte Verein f. vaterl.** Naturk, Wurtemberg, v.64, p.368-382, 1908.

crito pela relação $\left|\widehat{d}/\widehat{a}\right|$. Quando o valor desta expressão for menor do que $0,2$, a interação será do tipo aditiva, se o valor pertencer ao intervalo $0,2$ a $0,8$ haverá dominância parcial, caso esteja entre os dois valores, $0,8$ e $1,2$, tem-se dominância completa, caso contrário haverá sobredominância. O quadro apresentado a seguir exemplifica a interpretação sobre o GD.

$GD < 0,2$	$0,2 < GD < 0,8$	$0,8 < GD < 1,2$	$GD > 1,2$
Aditiva	Dominância Parcial	Dominância Completa	Sobredominância

Os locus podem interagir em pares ou em números mais elevados, e como mencionado anteriormente, as interações podem ser de vários tipos diferentes, mas no valor genotípico agregado, interações de todos os tipos são tratadas em conjunto, como um único desvio de interação.

A média do desvio da interação de todos os genótipos em uma população é zero quando os valores são expressos como desvios da média da população. O desvio de interação não é apenas uma propriedade dos genótipos de interação, mas depende também das frequências dos genótipos na população, e sucessivamente das frequências gênicas.

2.4 Populações utilizadas no mapeamento genético

Para obtenção das populações utilizadas no mapeamento deve-se partir de linhagens que sejam altamente contrastantes nas características fenotípicas, para tanto, na maioria das espécies cultivadas, por exemplo, as populações F_2 ou de retrocruzamento são as mais utilizadas. A seguir uma representação gráfica de como são obtidas essas populações.

Na Figura 2 têm-se que, P_1 e P_2 são dois parentais genitores de linhagens puras. A combinação dos dois gametas, P_1 e P_2, dão origem a geração F_1, que é heterozigota. O cruzamento de F_1 com um dos genitores formam a população de retrocruzamento, ou seja, $F_1 \times P_1$ formam a RC_1 e $F_1 \times P_2$ formam a RC_2. A população F_2 é obtida por autofecundação da geração F_1.

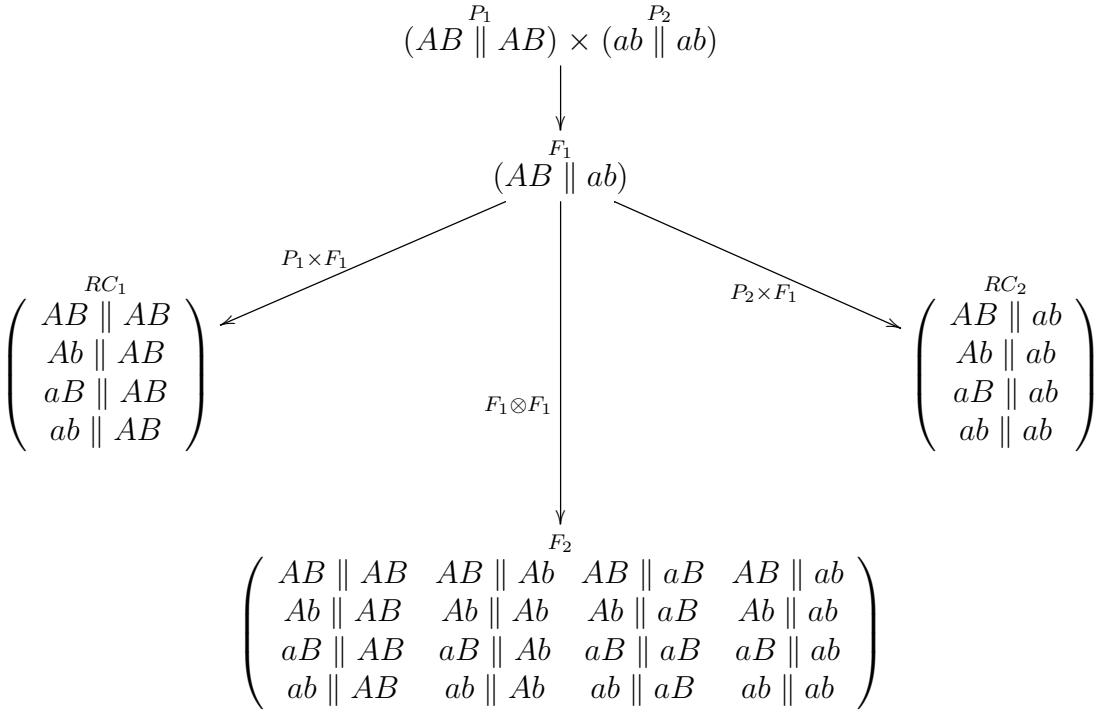


Figura 2 - Delineamentos experimentais utilizados nas análises de ligação entre marcadores

2.5 Fração de recombinação e funções de mapeamento

A razão entre o número de gametas recombinantes e o número total de gametas produzidos é definida com a fração de recombinação ou frequência de recombinação entre dois locus. O espaço paramétrico da fração de recombinação r é $0 \leq r \leq 0,5$. Quando r for igual a 0 tem-se que existe uma perfeita ligação entre os locus, já quando r for igual a $0,5$, indicará independência entre eles.

A função de mapeamento é uma função matemática que converte a fração de recombinação entre dois locus em uma distância genética d que os separam. Por exemplo, dois locus que apresentam uma fração de recombinação igual a 1% estão separados a 1 centimorgan (cM) no mapa genético. O parâmetro r pode ser representado pela fórmula de Mather (LANGE, 2002), a qual é escrita da forma,

$$r = \frac{1}{2} Pr\left(N_{[A,B]} > 0\right) = \frac{1}{2}\left[1 - Pr\left(N_{[A,B]} = 0\right)\right], \tag{18}$$

em que, na equação (18), $N_{[AB]}$ é o número de eventos de recombinação entre os locus

A e B no mesmo cromossomo. Sendo que $Pr\left(N_{[A,B]}=0\right)$ é a probabilidade de não ocorrer um quiasma (encontro) entre dois locus. A distância de mapa d é definida como $d=\frac{1}{2}E\left[N_{[A,B]}\right]$, representando a metade do número de quiasmas no intervalo $[A,B]$, e d é medida em unidades de cM.

Assim, a função de mapeamento de Morgan baseia-se no fato de que a probabilidade de ocorrer um quiasma numa distância de mapa d, é igual ao número esperado de permutações gênicas por gameta nesta distância, sendo assim, $E\left[N_{[A,B]}\right]=2d$. Assim, a função de mapeamento de Morgan é expressa da forma,

$$\begin{aligned} r &= \tfrac{1}{2}\left[1-Pr\left(N_{[A,B]}=0\right)\right] \\ &= \tfrac{1}{2}2d \\ &= d. \end{aligned} \tag{19}$$

Assumindo que cada permutação gênica ocorre de forma aleatória e independente, logo a ocorrência desta permutação entre dois locus num determinado cromossomo é modelada por uma distribuição de Poisson, em que $E\left[N_{[A,B]}\right]=2d$ (WU; CASELLA; MA, 2007). Assim a função de mapeamento de Haldane (1919) é escrita da forma,

$$\begin{aligned} r &= \tfrac{1}{2}\left[1-Pr\left(N_{[A,B]}=0\right)\right] \\ &= \tfrac{1}{2}\left[1-\tfrac{e^{-2d}(2d)^0}{0!}\right] \\ &= \tfrac{1}{2}\left[1-e^{-2d}\right]. \end{aligned} \tag{20}$$

Reescrevendo a equação (20) em função da distância de mapa tem-se,

$$d=-\frac{1}{2}\ln\left(1-2r\right). \tag{21}$$

Kosambi (1944) mostrou que a relação entre a distância de mapa d e a fração de recombinação r é estabelecida da forma,

$$2r=\tanh\left(2d\right) \tag{22}$$

Escrevendo a equação (22) em função da distância de mapa tem-se,

$$d=\frac{1}{4}\ln\left(\frac{1+2r}{1-2r}\right) \tag{23}$$

A Figura 3 mostra a representação gráfica das três funções de mapeamento que foram especificadas nas equações (19), (21) e (23).

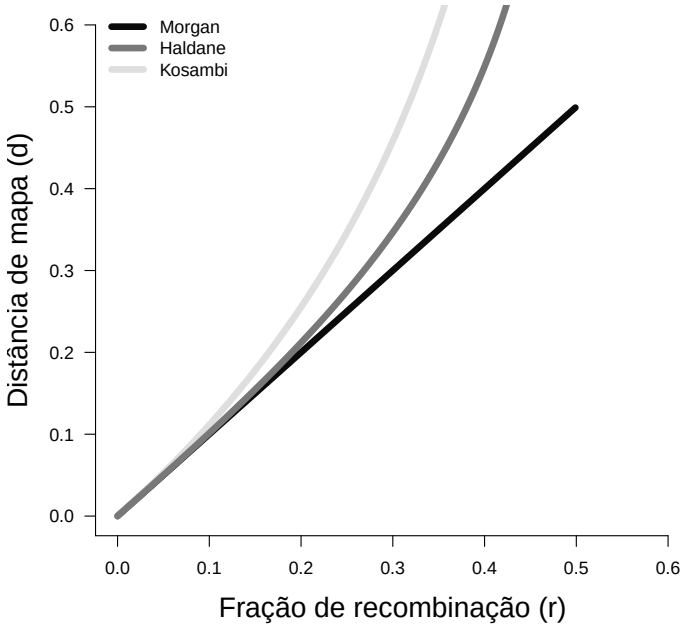


Figura 3 - Representação gráfica das três funções de mapeamento: Morgan, Haldane e Kosambi

2.6 Mapeamento de QTL

Doerge (2002) definiu um QTL como sendo uma determinada região do genoma que é responsável pela variação da característica quantitativa de interesse. Porém a identificação dessas regiões não é uma tarefa simples, devido ao grande número de QTL que pode conter em todo o genoma.

Edwards, Stuber e Wendel (1987) realizaram um trabalho em que utilizaram as informações dos marcadores moleculares para localizar QTL em milho. Neste mesmo trabalho os autores verificaram que as regiões ligadas aos marcadores explicaram entre 8% e 40% da variação fenotípica em um conjunto de 25 características avaliadas.

Lander e Botstein (1989) utilizaram o Mapeamento por Intervalo (IM) para estudar os efeitos dos QTL. Este tipo de mapeamento consiste em localizar QTL por meio da análise de marcadores flanqueadores. Esta técnica de mapear QTL foi de fundamental

importância, pois outras técnicas surgiram a partir desta, como o Mapeamento por Intervalo Composto e o Mapeamento de Múltiplos Intervalos (ZENG, 1993; KAO; ZENG; TEASDALE, 1999).

Modelo QTL

Seja uma população de mapeamento derivada de cruzamentos controlados, e que a característica quantitativa de interesse seja afetada por L QTL, o vetor dos valores fenotípicos observados $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ (n é a quantidade de indivíduos), pode ser descrito pelo seguinte modelo de regressão linear,

$$\mathbf{y} = \boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{24}$$

No modelo (24), $\boldsymbol{\mu}$ é a constante, $\boldsymbol{\beta}$ é o vetor dos efeitos genéticos, $\mathbf{Z}$ é a matriz do delineamento e $\boldsymbol{\varepsilon}$ é o vetor de erros aleatórios modelado por uma distribuição normal, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$. A matriz do delineamento e o vetor de efeitos genéticos dependerão, exclusivamente, da população utilizada no mapeamento. A seguir, serão detalhados os modelos para uma população de retrocruzamento, como também para uma população $F2$.

Em uma população de retrocruzamento o modelo linear utilizado é dado por:

$$y_i = \mu + z_i a + \varepsilon_i, \tag{25}$$

em que y_i é o valor fenotípico do indivíduo i, μ é a média geral, z_i é a variável indicadora que representa o genótipo do QTL do indivíduo i e é definido da forma

$$z_i = \begin{cases} 1, & \text{se o genótipo do QTL é } Qq, \\ 0, & \text{se o genótipo do QTL é } qq, \end{cases}$$

a é o efeito aditivo do QTL e ε_i é o erro aleatório.

A extensão do modelo (25) se faz necessária para estimar e testar os efeitos genéticos dos QTL para uma população $F2$ cujos genótipos são: QQ, Qq, e qq. O modelo

é então expresso da forma:

$$y_i = \mu + z_{1i}a + z_{2i}d + \varepsilon_i, \tag{26}$$

em que os parâmetros a e d representam os efeitos genéticos aditivo e dominante do QTL, respectivamente. Ainda no modelo (26) as variáveis indicadoras z_{1i} e z_{2i} são definidas como,

$$z_{1i} = \begin{cases} 1, & \text{se o genótipo do QTL é } QQ, \\ -1, & \text{se o genótipo do QTL é } qq, \end{cases}$$

$$z_{2i} = \begin{cases} 0, & \text{se o genótipo do QTL é } QQ \text{ ou } qq, \\ 1, & \text{se o genótipo do QTL é } Qq. \end{cases}$$

Na prática, numa população de mapeamento, não é possível a observação dos genótipos dos QTL, embora a suposição seja necessária quando se trata de um modelo estatístico. Na verdade, são usados os marcadores, uma vez que estes são observados para prever esses QTL por meio do estudo da ligação entre os marcadores e QTL. O procedimento se dá na determinação dos genótipos dos marcadores associados a uma ou mais características quantitativas e a partir desta determinação realiza-se inferência para o efeito de um QTL putativo na variação fenotípica. Vale ressaltar que a utilização de um único marcador não é suficiente para a análise, haja vista que se pretende saber em qual lado do marcador (direito ou esquerdo) o QTL está localizado (WU; CASELLA; MA, 2007).

Probabilidades condicionais

Os elementos da matriz $\mathbf{Z}$ serão compostos dos genótipos do QTL Q_{ij}, que por sua vez, não são observáveis. Assim,

$$y_{ij}|Q_{ij} \sim N\left(\mu + z_{ij}\beta_j; \sigma^2\right), \quad \begin{matrix} i = & 1, \ldots, n. \\ j = & 1, \ldots, L. \end{matrix}$$

A seguir, será detalhada a obtenção da distribuição de probabilidade $y_{ij}|Q_{ij}$ obtida por meio da fração de recombinação.

Considere a função de distribuição marginal de $\mathbf{y}$, $f(\mathbf{y})$, expressa da forma,

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

Para obter a solução da função de distribuição conjunta será necessária a construção da distribuição de probabilidade de cada Q_{ij}. De acordo com Satogopan et al. (1996) a distribuição de probabilidade dos genótipos dos L QTL, dado os genótipos dos marcadores moleculares M_{ik} e a distância entre eles, podem ser obtidas por meio da fração de recombinação entre estas marcas. Supondo que o j-ésimo QTL está na posição λ_j entre os marcadores k_j e k_{j+1}, e que a posição λ_j pertence ao intervalo D_{k_j} e $D_{k_{j+1}}$, $D_{k_j} \leq \lambda_j \leq D_{k_{j+1}}$, a função de distribuição de probabilidade para os genótipos dos QTL será escrita da forma,

$$f(Q_i|\lambda, M_i, D) = \prod_{j=1}^{L} f(Q_{ij}|\lambda_j, M_i, D). \tag{27}$$

Assumindo que os locus segregam de forma independente, a equação (27), será reescrita da seguinte maneira,

$$f(Q_i|\lambda, M_i, D) = \prod_{j=1}^{L} f\left(Q_{ij}|\lambda_j, M_{ik_j}, D_{k_j}, D_{k_{j+1}}\right). \tag{28}$$

A equação (28) é calculada utilizando as informações das Tabelas 1 e 2. Nestas tabelas r, r_1 e r_2 são as frações de recombinação obtidas a partir do mapa genético, sendo r_1 a fração de recombinação entre o marcador k e o QTL, r_2 a fração entre o QTL e o marcador $k+1$ e $r = r_1 + r_2 + 2r_1r_2$ a fração de recombinação entre os marcadores k e $k+1$.

A função de verossimilhança para o parâmetro λ e o conjunto de parâmetros $\boldsymbol{\theta} = (\mu, \beta, \sigma^2)^T$ é expressa da forma,

$$L(\lambda, \theta|\mathbf{y}) = \prod_{i=1}^{n} \sum_{j=1}^{L} f(y_i|Q_i = q_i, \theta) \times f(Q_i = q_i|\lambda)$$

Utilizando-se das probabilidades condicionais, a distribuição genotípica do

QTL, embora não seja observada, pode ser mensurada a partir dos genótipos dos marcadores flanqueadores (KAO; HO, 2012),

$$p\left(Q_{k'}|M_k, M_{k+1}\right) = \frac{p\left(M_k, Q_{k'}, M_{k+1}\right)}{p\left(M_k, M_{k+1}\right)}, \tag{29}$$

em que $Q_{k'}$ é o genótipo do suposto QTL que está localizado entre os marcadores flanqueadores M_k e M_{k+1}. As probabilidades condicionais resultantes da equação (29), de maneira simplificada, serão iguais as frequências conjuntas divididas pelas frequências marginais correspondentes aos genótipos dos marcadores.

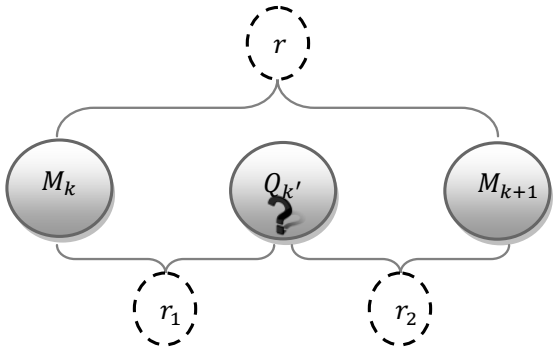


Figura 4 - Esquema de um QTL flanqueado entre dois marcadores

Na Figura 4 é apresentada uma esquematização de um QTL flanqueado entre dois marcadores e, como mencionado anteriormente, o genótipo deste QTL será estimado a partir dos genótipos destes marcadores, que são observáveis, e para isto será utilizada a fração de recombinação que será convertida em distância de mapa d para realização dos cálculos. As frequências e as probabilidades condicionais dos genótipos do QTL dado os marcadores flanqueados para as populações de retrocruzamento e F_2 podem ser observadas nas Tabelas 1 e 2, respectivamente.

Tabela 1 - Frequências dos genótipos dos marcadores MM, Mm em uma população de Retrocruzamento

Marcador		**Genótipo do QTL**	
Genótipo	$p\left(M_k, M_{k+1}\right)$	$p\left(M_k, Q_{k'}, M_{k+1}\right)$ MM	Mm
MM MM	$1/2\left(1-r\right)$	$1/2\left(1-r_1\right)\left(1-r_2\right)$	$1/2r_1r_2$
MM Mm	$1/r$	$1/2\left(1-r_1\right)r_2$	$1/2\left(1-r_2\right)r_1$
Mm MM	$1/r$	$1/2\left(1-r_2\right)r_1$	$1/2\left(1-r_1\right)r_2$
Mm Mm	$1/2\left(1-r\right)$	$1/2r_1r_2$	$1/2\left(1-r_1\right)\left(1-r_1\right)$

Tabela 2 - Frequências dos genótipos dos marcadores MM, Mm e mm em uma população F_2

Marcador		**Genótipo do QTL**		
Genótipo	$p(M_k, M_{k+1})$	$p(M_k, Q_{k'}, M_{k+1})$		
		MM	Mm	mm
MM MM	$1/4(1-r)^2$	$1/4(1-r_1)^2(1-r_2)^2$	$1/2r_1r_2(1-r_1)(1-r_2)$	$1/4r_1^2r_2^2$
MM Mm	$1/2(1-r)r$	$1/4(1-r_1)^2(1-r_2)r_2$	$1/2(1-r_1)\left(1-2r_2+2r_2^2\right)r_1$	$1/4(1-r_2)r_1^2r_2$
MM mm	$1/4r^2$	$1/4(1-r_1)^2r_2^2$	$1/2(1-r_1)(1-r_2)r_1r_2$	$1/4(1-r_2)^2r_1^2$
Mm MM	$1/2(1-r)r$	$1/4(1-r_1)(1-r_2)^2r_1$	$1/2(1-r_2)\left(1-2r_1+2r_1^2\right)r_2$	$1/4(1-r_1)r_2^2r_1$
Mm Mm	$1/2\left(1-2r+2r^2\right)$	$1/4(1-r_1)(1-r_2)r_1r_2$	$1/2\left(1-2r_1+2r_1^2\right)\left(1-2r_2+2r_2^2\right)$	$1/4(1-r_1)(1-r_2)r_1r_2$
Mm mm	$1/2(1-r)r$	$1/4(1-r_1)r_1r_2^2$	$1/2\left(1-2r_1+2r_1^2\right)(1-2r_2)r_2$	$1/4(1-r_1)(1-r_2)^2r_1$
mm MM	$1/4r^2$	$1/4(1-r_2)^2r_1^2$	$1/2(1-r_1)(1-r_2)r_1r_2$	$1/4(1-r_1)^2r_2^2$
mm Mm	$1/2(1-r)r$	$1/4(1-r_2)r_1^2r_2^2$	$1/2(1-r_1)\left(1-2r_2+2r_2^2\right)r_1$	$1/4(1-r_1)^2(1-r_2)r_2$
mm mm	$1/4(1-r)^2$	$1/4r_1^2r_2^2$	$1/2(1-r_1)(1-r_2)r_1r_2$	$1/4(1-r_1)^2(1-r_2)^2$

Por meio da Tabela 2 é possível estimar os coeficientes associados com os efeitos aditivo e dominante para o genótipo do QTL (Tabela 3).

Tabela 3 - Genótipos dos marcadores MM, Mm e mm e os efeitos aditivo (a) e dominante (d) dos genótipos dos QTL em uma F_2

Genótipo	Efeitos	
	a	d
MM MM	$\frac{(1-r_1)^2(1-r_2)^2-r_1^2r_2^2}{(1-r)^2}$	$\frac{2(1-r_1)(1-r_2)r_1}{(1-r)^2}$
MM Mm	$\frac{(1-r_1)^2(1-r_2)r_2-r_1^2r_2(1-r_2)}{r(1-r)}$	$\frac{r_1(1-r_1)(1-r_2)^2+(1-r_1)r_1r_2^2}{r(1-r)}$
MM mm	$\frac{(1-r_1)^2r_2^2-(1-r_2)^2r_1^2}{r^2}$	$\frac{2(1-r_1)(1-r_2)r_1r_2}{r^2}$
Mm MM	$\frac{r_1(1-r_1)(1-r_2)^2-(1-r_1)r_1r_2^2}{r(1-r)}$	$\frac{(1-r_1)^2(1-r_2)r_2+(1-r_2)r_1^2r_2}{r(1-r)}$
Mm Mm	0	$\frac{\left(1-2r_1+r_1^2\right)\left(1-2r_2+r_2^2\right)}{r^2+(1-r)^2}$
Mm mm	$\frac{(1-r_1)r_1r_2^2-(1-r_1)(1-r_2)^2r_1}{(1-r)r}$	$\frac{(1-r_1)^2(1-r_2)r_2+(1-r_2)r_1^2r_2}{(1-r)r}$
mm MM	$\frac{(1-r_2)^2r_1^2-(1-r_1)^2r_2^2}{r^2}$	$\frac{2(1-r_1)(1-r_2)r_1r_2}{r^2}$
mm Mm	$\frac{(1-r_2)r_1^2r_2-(1-r_1)^2(1-r_2)r_2}{(1-r)r}$	$\frac{(1-r_1)(1-r_2)^2r_1+(1-r_1)r_1r_2^2}{(1-r)r}$
mm mm	$\frac{r_1^2r_2^2-(1-r_1)^2(1-r_2)^2}{(1-r)^2}$	$\frac{2(1-r_1)(1-r_2)r_1r_2}{(1-r)^2}$

As próximas seções desta revisão bibliográfica apresentará a inferência bayesiana no mapeamento de QTL.

2.7 Inferência bayesiana

Na estatística é de fundamental importância o conhecimento sobre a quantidade de interesse θ, sendo este tratado como uma quantidade desconhecida. A análise bayesiana destina-se a obtenção da densidade a posteriori a cerca dos parâmetros de interesse, para isto, é combinada a informação prévia a respeito dos parâmetros (distribuição a priori), e o conhecimento que se tem sobre o parâmetro contido na amostra (função de verossimilhança).

Teorema de Bayes

Suponha que $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ é um vetor de n observações cuja distribuição de probabilidade é $p(\mathbf{y}|\boldsymbol{\theta})$, em que $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)^T$ é um vetor paramétrico de dimensão d. Suponha que a incerteza sobre $\boldsymbol{\theta}$ seja modelada por uma distribuição de probabilidade, $p(\boldsymbol{\theta})$. Pelo teorema de Bayes, tem-se que,

$$p(\mathbf{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta}) = p(\mathbf{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y}) \times p(\mathbf{y}). \tag{30}$$

A equação (30) pode ser reescrita da forma,

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) = \frac{p(\mathbf{y}|\boldsymbol{\theta}) \times p(\boldsymbol{\theta})}{p(\mathbf{y})}. \tag{31}$$

A equação (31) é a fórmula usual do teorema de Bayes. Na inferência bayesiana, $p(\mathbf{y}|\boldsymbol{\theta})$ será denotado por $L(\boldsymbol{\theta}|\mathbf{y})$ (função de verossimilhança), $p(\boldsymbol{\theta})$ é a distribuição a priori e, $p(\mathbf{y})$ é a integral definida num intervalo de valores possíveis de $\boldsymbol{\theta}$. A função $p(\mathbf{y})$ neste caso é expressa por

$$p(\mathbf{y}) = \int_{\boldsymbol{\theta}} L(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta}) \partial \boldsymbol{\theta}. \tag{32}$$

Observe que a equação (32) não dependerá de $\boldsymbol{\theta}$ e, portanto, esta quantidade representará apenas uma constante. Logo, a forma usual do teorema de Bayes em (31) é

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) = c \times L(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta}) \tag{33}$$

De acordo com Box e Tiao (1992) na equação (33), $c^{-1} = p(\mathbf{y})$ é uma constante normalizada necessária para assegurar que a distribuição a posteriori $p\left(\boldsymbol{\theta}|\mathbf{y}\right)$ após integrada resulte no valor um.

Assim, da equação (31), observa-se que $p\left(\boldsymbol{\theta}|\mathbf{y}\right)$ é proporcional à função de verossimilhança multiplicada pela priori:

$$p\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto L(\boldsymbol{\theta}|\mathbf{y}) \times p(\boldsymbol{\theta}) \tag{34}$$

A função de verossimilhança desempenha um papel muito importante na fórmula de Bayes. É por meio dela que os dados podem modificar o conhecimento a priori

sobre $\boldsymbol{\theta}$. Essa função pode ser vista como a representação do que os dados têm a informar a respeito de $\boldsymbol{\theta}$.

2.8 Distribuições a priori

De acordo com Ehlers (2011) a utilização de informação a priori em inferência bayesiana requer a especificação de uma distribuição a priori para o(s) parâmetro(s) de interesse, $\boldsymbol{\theta}$. Esta distribuição deve descrever o conhecimento que se tem sobre $\boldsymbol{\theta}$ antes mesmo da realização do experimento. Existem algumas formas de especificação das distribuições a priori, como por exemplo, as distribuições a priori conjugadas e as não informativas. A designação de prioris conjugadas é devido ao fato de que as distribuições a priori e a posteriori pertençam a uma mesma classe de distribuições e assim a atualização do conhecimento que se tem sobre $\boldsymbol{\theta}$ envolve apenas uma mudança nos hiperparâmetros. O uso de prioris conjugadas é muito importante na estatística bayesiana, pois o aspecto sequencial do método bayesiano pode ser explorado definindo-se apenas uma regra de atualização dos hiperparâmetros já que as distribuições permanecem as mesmas. Porém, a utilização destas prioris, às vezes, está mais relacionada à facilidade de implementação computacional do que a modelagem adequada do parâmetro de interesse (TOLEDO, 2006). A priori não-informativa refere-se ao caso em que pouca ou nenhuma informação é disponível antes de realizar o experimento. O termo "não informativa" é usado para descrever a falta de crenças subjetivas utilizadas na formulação de tal priori (ENO, 1999). De acordo com Meyer (2009) uma forma de atribuir distribuição a priori não-informativa é designar distribuições de forma que, todos os possíveis valores para um dado parâmetro, tenham a mesma chance de ocorrer. A ideia inicial é utilizar a distribuição uniforme para representar esta situação, e assim, $p(\boldsymbol{\theta}) \propto$ constante. Jeffreys (1961), baseando-se na informação de Fisher, propôs uma classe de prioris não informativas invariantes, contudo, possivelmente impróprias.

A seguir, serão apresentadas algumas distribuições a priori que são utilizadas para o mapeamento de QTL.

a) Número de QTL com efeitos detectáveis

   O número esperado de QTL a ser considerado no modelo, l_0, pode ser determinado utilizando métodos clássicos, como por exemplo, o mapeamento por intervalo composto, para que, em seguida, seja determinado um valor plausível para L (número

de QTL com efeitos detectáveis). O valor para L terá grande influência sobre as estimativas dos parâmetros a posteriori (GREEN, 1995).

Para determinar um valor de L, a distribuição de Poisson será atribuída como uma priori, com média l_0. Para que o valor de L seja suficientemente grande, tem-se que a probabilidade $Pr(l > L)$ será pequena. Pelos princípios de aproximações de distribuições, aproximando a distribuição de Poisson a uma distribuição normal, tem-se que L será $l_0 + 3\sqrt{l_0}$ (YI et al., 2005). Quando não há efeito da interação entre pares de QTL, o valor de L se reduz a $3\sqrt{l_0}$.

b) Número de QTL incluídos e os seus efeitos genéticos associados

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{L} \omega_j^{\gamma_j} (1 - \omega_j)^{1-\gamma_j} . \tag{35}$$

Na equação (35), $\omega_j = p(\gamma_j = 1)$ é a probabilidade a priori referente ao j-ésimo efeito. Cada QTL entra no modelo, independentemente de quaisquer outro QTL, com uma probabilidade $p(\gamma_j = 1) = 1 - p(\gamma_j = 0) = \omega_j$. No mapeamento de QTL, por simplicidade, $\omega_j = \omega$. Como caso particular, quando $\omega = 1/2$, tem-se que, $p(\boldsymbol{\gamma}) = \frac{1}{2}L$. Esta aproximação é muito utilizada, como uma priori pouco informativa, para soluções de problemas envolvendo seleção de variáveis (YI, 2004).

c) Posição do QTL

Em geral, para o parâmetro que representa a posição do j-ésimo QTL, λ_j, assumi-se uma distribuição a priori Uniforme no intervalo $[a, b]$. Dado que o comprimento do genoma é K, $\lambda \sim U[0, K]$ (SATAGOPAN et al., 1996). Isso ocorre devido ao fato de não se ter nenhum conhecimento prévio a respeito das posições dos QTL. Assim, duas restrições podem ser adotadas para reduzir o espaço paramétrico do modelo, sobre a distribuição a priori para as posições dos QTL. A primeira diz respeito à distância entre múltiplos QLT ligados. Já a segunda restringe o número de QTL detectáveis em cada cromossomo (BANERJEE; YANDELL; YI, 2008).

d) Efeitos genéticos

Yi et al. (2005) propuseram prioris hierárquicas para os efeitos genéticos,

$$\boldsymbol{\beta}_j \sim N\left(\mathbf{0}, \gamma_j c\sigma^2 \left(\mathbf{x}'_{.j}\mathbf{x}_{.j}\right)^{-1}\right). \tag{36}$$

Na equação (36), $\mathbf{x}_{.j} = \{x_{1j}, x_{2j}, \ldots, x_{nj}\}$ é um vetor contendo os coeficientes de $\boldsymbol{\beta}_j$ e c é um fator de escala positiva. Estes autores afirmam que muitas sugestões foram propostas para a escolha de c no que diz respeito a problemas de seleção de variáveis, eles comentam que tomaram $c = n$, em que n é o tamanho amostral, justificando que esta é uma escolha trivial.

e) Média geral e a variância residual

Para a média μ, em geral, é atribuída uma distribuição a priori Normal, $\mu \sim N(\eta_0, \tau_0^2)$, em que η_0 e τ_0^2 serão a média e a variância amostral, respectivamente. Já para a variância residual σ^2, quase sempre é estabelecida uma priori "não informativa", $p(\sigma^2) = \frac{1}{\sigma^2}$.

Baseados no trabalho de Chipman (2004), Yi et al. (2007) atribuíram distribuições a priori hierárquicas para os parâmetros presentes nos efeitos genéticos. Estes mesmos autores propuseram a utilização de modelos bayesianos hierárquicos (RUIZ et al., 2003; GELMAN, 2006).

2.9 Monte Carlo com Cadeia de Markov

O grande desafio, na maioria dos experimentos em que é aplicada a inferência bayesiana, é a obtenção da distribuição conjunta a posteriori. Pois, uma vez obtida uma amostra desta distribuição é possível calcular estatísticas relacionadas aos parâmetros de interesse. Mas, na maioria dos casos, não há uma solução analítica para este tipo de distribuição. Para contornar esse problema gera-se uma amostra das distribuições marginais a posteriori por meio dos Métodos de Monte Carlo com Cadeias de Markov - MCMC (TIERNEY, 1994; GAMERMAN; LOPES, 2006).

A seguir, é apresentado o algoritmo mais utilizado nos métodos MCMC, o Metropolis-Hastings (M-H), e que tem como caso particular o amostrador de Gibbs. De acordo com Gamerman e Lopes (2006) o algoritmo consiste em simular um passeio aleatório no espaço de $\boldsymbol{\theta}$ que convirja para uma distribuição estacionária, a qual é a de interesse no problema.

Algoritmo de Metropolis-Hastings

Segundo Chib e Greenberg (1995) o algoritmo M-H é um método utilizado para obtenção de amostras aleatórias relacionadas a uma distribuição de probabilidade, como por exemplo, a distribuição a posteriori. De acordo com Meyer (2009) este algoritmo é indicado em casos, em que, a distribuição condicional completa a posteriori não possui uma forma fechada.

Considere que a cadeia de Markov esteja no estado θ. O algoritmo M-H gera um valor candidato θ' de uma distribuição proposta, $q(\cdot, \theta)$. Vale ressaltar que, poderá depender, ou não, do estado atual da cadeia. Assim, o valor candidato, θ', é aceito com probabilidade:

$$\alpha(\theta, \theta') = min\left(1, \frac{\pi(\theta')\, q(\theta, \theta')}{\pi(\theta)\, q(\theta', \theta)}\right). \tag{37}$$

Na equação (37), π representa a distribuição de interesse. Para simplificar, o algoritmo M-H pode ser estabelecido de acordo com os seguintes passos:

1. Atribua um valor inicial, θ^0 na iteração $t = 0$;
2. Gere um valor candidato, θ', ao próximo estado, da distribuição $q(\cdot, \theta)$;
3. Calcule a probabilidade de aceitação, $\alpha(\theta, \theta')$;
4. Gere u de uma distribuição $U(0, 1)$;
5. Se $u < \alpha$, aceite o novo valor, caso contrário, rejeite-o e faça $\theta^{(t+1)} = \theta$;
6. Incremente o contador de t para $t + 1$ e retorne ao passo 2.

2.10 Monte Carlo com Cadeia de Markov e Saltos Reversíveis

Muitas são as características quantitativas que são extremamente influenciadas por fatores genéticos, no que diz respeito a sua variabilidade fenotípica, em geral, existem vários genes que colaboram para isto.

A utilização de metodologias baseadas em inferência bayesiana usando métodos MCMC vem sendo utilizada para mapear QTL (SATAGOPAN et al., 1996; SILLANPÄÄ; ARJAS, 1998; YI; XU, 2000; GAFFNEY, 2001; YI, 2004; BANERJEE; YANDELL; YI, 2008; MANICHAIKUL et al., 2009; LI; SILLANPÄÄ, 2012).

Para Satagopan et al. (1996) o número de QTL é uma quantidade conhecida. Assim, as estimativas a posteriori para os parâmetros podem ser inferidas utilizando o método MCMC tradicional. Ou seja, para a construção das cadeias de Markov são utilizados os algoritmos M-H e o Amostrador de Gibbs.

De acordo com Silva e Leandro (2009) o mapeamento de QTL por meio de métodos bayesianos possibilita tratar o número de QTL como uma quantidade desconhecida, implicando em vantagens consideráveis para a modelagem. O grande problema quando se utiliza esta metodologia, é o da obtenção da amostra aleatória da distribuição conjunta a posteriori, uma vez que, ao considerar o número de QTL como uma incerteza, a dimensão do espaço do modelo (número de parâmetros) pode variar. Green (1995) propôs, como resolução deste problema, o algoritmo MCMC com Saltos Reversíveis, este algoritmo permite saltar entre modelos com dimensões diferentes por meio da especificação de distribuições propostas, ou seja, poderá ocorrer em cada nova iteração o nascimento ou morte de um QTL. Muitos trabalhos seguiram as ideias deste autor, tais como, (STEPHENS; FISCH, 1998; YI, 2004; LEE; VAN DER WERF, 2006; YI et al., 2007), dentre outros.

De acordo com Ehlers (2011) o algoritmo MCMC com Saltos Reversíveis é executado da seguinte forma:

a) Considere que o estado atual da cadeia é $(\mathcal{C}, \boldsymbol{\theta})$. Ou seja, neste momento tem-se o modelo $\mathcal{C}$ composto pelo conjunto de parâmetros $\boldsymbol{\theta}$;

b) Seja agora que, um novo modelo $\mathcal{C}'$ com $\boldsymbol{\theta}'$ parâmetros é proposto com probabilidade $p_{\mathcal{C},\mathcal{C}'}$;

c) Por simplicidade, o novo modelo, $\mathcal{C}'$, tem um maior número de parâmetros, $n_{\mathcal{C}'} > n_{\mathcal{C}}$

A partir desta estrutura o seguinte algoritmo é utilizado:

1. Proponha a mudança$(\mathcal{C}, \boldsymbol{\theta}) \rightarrow (\mathcal{C}', \boldsymbol{\theta})$ com probabilidade $p_{\mathcal{C},\mathcal{C}'}$;

2. Gere um vetor aleatório, $\boldsymbol{\kappa} \sim q(\boldsymbol{\kappa})$, com dimensão $n_{\mathcal{C}'} > n_{\mathcal{C}}$;

3. Faça $\boldsymbol{\theta} = g(\boldsymbol{\theta}, \boldsymbol{\kappa})$, para uma função determinística g;

4. Aceite $(\mathcal{C}', \boldsymbol{\theta}')$ com probabilidade $\min(1, A)$. Em que,

$$A = \frac{\pi(\mathcal{C}', \boldsymbol{\theta}')}{\pi(\mathcal{C}, \boldsymbol{\theta})} \times \frac{p_{\mathcal{C}',\mathcal{C}}}{p_{\mathcal{C},\mathcal{C}'} q(\boldsymbol{\kappa})} \left| \frac{\partial g(\boldsymbol{\theta}, \boldsymbol{\kappa})}{\partial(\boldsymbol{\theta}, \boldsymbol{\kappa})} \right|.$$

Modelo de espaço composto

Segundo Meyer (2009) quando existem incertezas em relação a dimensão do espaço paramétrico, uma alternativa que surge é a utilização do modelo de espaço composto. A abordagem do espaço composto, que é uma modificação direta do conceito do produto dos espaços (CARLIN; CHIB, 1995), proporciona soluções no que diz respeito a uma grande variedade de problemas que envolvem a seleção de modelos (GODSILL, 2001). Este tipo de modelo considera um espaço que contempla todos os possíveis parâmetros envolvidos e que também inclui uma variável aleatória que informa quais parâmetros estão presentes no modelo. Os parâmetros por sua vez, incluídos neste modelo, determinam a função de verossimilhança, já os parâmetros não utilizados, estarão presentes na distribuição conjunta a posteriori em forma de pseudo-prioris.

2.11 Comparação de modelos

O fator de Bayes (FB), introduzido por Jeffreys (1961), é uma alternativa bayesiana para testar hipóteses relacionadas à comparação de dois modelos (M_1 e M_2), ou seja, esta estatística é utilizada para determinar qual, dentre dois quaisquer modelos, melhor descreve os dados, $\mathbf{y}$.

O FB além de incluir a incerteza do modelo, permite também que os modelos não encaixados sejam comparados. Pelo FB, para comparação de dois modelos, será utilizada a razão das verossimilhanças marginais dos dados. Assim, esta razão pode ser escrita da forma,

$$B_{12} = \frac{f(\mathbf{y}|M_1)}{f(\mathbf{y}|M_2)}. \tag{38}$$

Jeffreys (1961) sugeriu a seguinte interpretação para a estatística B_{12} (Tabela 4).

Tabela 4 - Classificação do Fator de Bayes

$\log_{10}(B_{12})$	B_{12}	Classificação
$< 0,50$	$1,00$ a $3,20$	Evidência a favor de M_2
$0,50$ a $1,00$	$3,20$ a $10,00$	Evidência positiva a favor de M_2
$1,00$ a $2,00$	$10,00$ a $100,00$	Forte evidência a favor de M_2
$> 2,00$	$> 100,00$	Evidência decisiva a favor de M_2

3 MATERIAL E MÉTODOS

3.1 Material

O conjunto de dados utilizado neste trabalho foram apresentados por Sibov et al. 2003a, 2003b apud Meyer (2009) e Pereira (2012). A seguir serão apresentados detalhes do delineamento utilizado para mapear QTL no trabalho destes autores.

No cruzamento entre duas linhagens endogâmicas $L-08-05F$ e $L-14-4B$ foi obtida uma população que possuía características contrastantes para a produção de grãos de milho. Deste cruzamento obteve-se a progênie F_1, sendo que, quatro plantas desta geração foram autofecundadas, dando origem a 400 plantas da população F_2 das quais foram obtidas 400 progênies $F_{2:3}$, que foram cruzadas entre si e semeadas em linhas com vinte plantas para aumentar a quantidade de sementes necessarias para análise do experimento.

A partir das 400 progênies foram criados quatro grupos com 100 progênies cada. Em cada grupo foi realizado um delineamento látice 10×10, com duas repetições cada um. Neste experimento foram avaliados vários caracteres, entretanto neste trabalho restrigiremos a analisar um deles, a produção de grãos de milho.

O mapa de ligação utilizado para detecção de QTL foi composto por 117 locus de marcadores microssatélites, os quais foram distribuídos em dez grupos de ligação. O mapa genético ficou com comprimento de 1634,20 cM e distância média entre as marcas de 14 cM.

3.2 Métodos

3.2.1 HMM para imputação dos genótipos dos marcadores moleculares

As análises de marcadores moleculares que contém informação genotípica do indivíduo são importantes para identificar associações de genes. Os grandes conjuntos de dados derivados desses marcadores contêm uma quantidade significativa de genótipos ausentes. De acordo com Roberts et al. (2007), na prática, existem algumas alternativas para lidar com este problema, tais como, repetir a genotipagem em regiões com genótipos ausentes, remover os marcadores com dados em falta e inferir os dados ausentes. Neste trabalho, o objetivo é inferir os genótipos ausentes por meio de imputações.

As informações ausentes a respeito dos genótipos nos marcadores molecu-

lares é um problema comum em estudo de mapeamento genético e, por conseguinte, no mapeamento de QTL. Os dados ausentes ocorrem devido a erros de genotipagem, marcadores não informativos, dentre outros motivos. Para solucionar este problema se faz necessária à utilização de técnicas de imputação para inferir os dados desses genótipos (HOWIE; MARCHINI; STEPHENS, 2011; LI et al., 2009). Existem diversos programas computacionais que são utilizados para imputação, como por exemplo, o IMPUTE (ZHAO, 2008) e o BEAGLE (BROWNING; BROWNING, 2009). Ambos os programas são baseados em HMM.

Os dados dos genótipos dos marcadores serão aqui utilizados para inferir as localizações de possíveis QTL, como também detectar QTL no intervalo constituído entre dois marcadores. Assim, será realizada uma análise preliminar, no que diz respeito à imputação dos genótipos não observados nos marcadores, para que, ao fazer inferência no intervalo entre dois marcadores, possam-se ter estimativas mais confiáveis e plausíveis. Com isso, a acurácia das técnicas para mapear QTL se torna maior. A imputação desses dados permite aos geneticistas avaliarem com precisão a evidência de possíveis marcadores associados à QTL (BROWNING; BROWNING, 2009).

Sabe-se que no mapeamento de QTL, utilizando uma abordagem bayesiana, os genótipos ausentes são tratados como uma variável aleatória fazendo com que a quantidade de parâmetros a serem estimados no modelo cresça. Logo, ao se fazer imputação destes genótipos a dimensão do espaço paramétrico diminuirá significativamente, aumentando assim a eficiência das estimativas e reduzindo o tempo computacional das análises.

O HMM, esquematizado na Figura 5, será estruturado neste trabalho da seguinte forma: g_i representa um estado não observado da cadeia de Markov, o_i é uma variável aleatória observável, sendo que o_i depende apenas de g_i. Os elementos a_{ij} e e_{ik} representam as probabilidades de transição e de emissão, respectivamente.

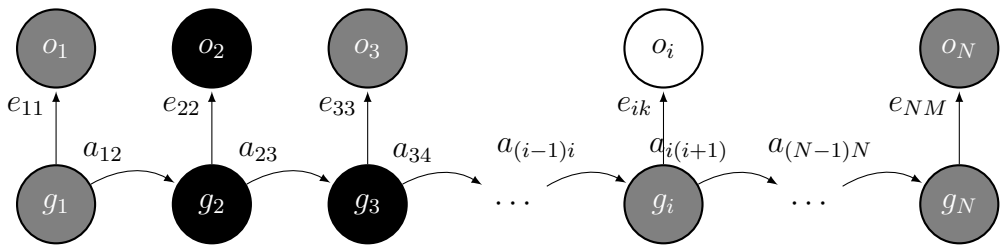


Figura 5 - Ilustração de uma cadeia de Markov oculta

A utilização do HMM para imputação de genótipos dos marcadores ocorre da seguinte maneira.

Considere um indivíduo originário de um retrocruzamento de duas linhagens puras, A e B, em que o pai F_1 foi cruzado novamente com A. Sendo assim, os possíveis valores genotípicos serão, $\mathcal{G} = \{\mathcal{AA}, \mathcal{AB}\}$.

O conjunto dos símbolos que serão emitidos, sequência observável, será expresso por, $\mathcal{O} = \{\mathcal{A}, \mathcal{H}, \texttt{NA}\}$. Ou seja, $\mathcal{AA}$ emitirá o símbolo $\mathcal{A}$, $\mathcal{AB}$ emitirá H. Já NA representará um valor não observado.

As probabilidades iniciais assumindo as regras de Mendel serão $\pi(AA) = \pi(AB) = 1/2$. As probabilidades de transição estarão em função da fração de recombinação r, $a_{ij} = r$, para $i \neq j$, em que r denota a fração de recombinação. Naturalmente, $a_{ij} = 1 - r$, para $i = j$. Para determinar as expressões para as probabilidades de emissão, assume-se uma taxa de erro constante na genotipagem, ϵ, então $e_{g_i}(AA, A) = e_{g_i}(AB, H) = 1 - \epsilon$, e $e_{g_i}(AA, H) = e_{g_i}(AB, A) = \epsilon$. Tem-se ainda que, $e_{g_i}(AA, \texttt{NA}) = e_{g_i}(AB, \texttt{NA}) = 1$, pois $\texttt{NA} = \{A \text{ ou } H\}$ de modo que $e_{g_i}(AA, \texttt{NA}) = e_{g_i}(AA, A) + e_{g_i}(AA, H) = 1$.

Para uma população F_2, as expressões dessas probabilidades são determinadas de maneira análoga (Tabela 5).

Tabela 5 - Probabilidades de transição em uma população F_2

g	g'		
	AA	AB	BB
AA	$(1-r)^2$	$2r(1-r)$	r^2
AB	$r(1-r)$	$(1-r)^2 + r^2$	$r(1-r)$
BB	r^2	$2r(1-r)$	$(1-r)^2$

Em uma população F_2 os possíveis símbolos observados serão $\mathcal{O}_d = \{A, H, B, \texttt{NA}\}$, com $A, B,$ e H correspondentes a dois homozigotos e um heterozigoto, respectivamente, NA corresponde a um valor completamente ausente, $\mathcal{G}_d = \{AA, AB, BB\}$, os possíveis valores genotípicos (Tabela 6).

Tabela 6 - As probabilidades de emissão em uma população F_2

g	$\mathcal{O}_d$			
	A	H	B	NA
AA	$1-\epsilon$	$\epsilon/2$	$\epsilon/2$	1
AB	$\epsilon/2$	$1-\epsilon$	$\epsilon/2$	1
BB	$\epsilon/2$	$\epsilon/2$	$1-\epsilon$	1

De acordo com as regras de Mendel, as probabilidades iniciais serão $\pi(AA) = \pi(BB) = \frac{1}{4}, \pi(AB) = \frac{1}{2}$.

3.2.2 Métodos para avaliar a acurácia

Após realizada a imputação, foi feito um estudo no qual foram retiradas certas quantidades de observações da própria matriz com todos os genótipos dos marcadores presentes. A porcentagem de valores retirados em cada indivíduo no decorrer dos 117 marcadores variou de 1% a 40%.

Para validação do método empregado para imputação dos genótipos dos marcadores moleculares foram utilizadas duas técnicas descritas a seguir.

A raiz quadrada do erro quadrático médio normalizado - NRMSE (do inglês, *normalized root mean squared error*) foi calculada para determinar a acurácia da imputação (KIM et al., 2004; HU et al., 2006; XIANG et al., 2008). A NRMSE é obtida de acordo com a seguinte expressão,

$$NRMSE = \sqrt{\frac{\frac{1}{Q}\sum_{q=1}^{Q}\left(\widehat{g}_q - g_q\right)^2}{\frac{1}{Q}\sum_{q=1}^{Q} g_q^2}} \tag{39}$$

Na equação (39) $q = 1, 2, \ldots, Q$ representa a quantidade de valores a serem imputados, g_q representa o valor real que foi ocultado da matriz completa dos genótipos dos marcadores e o seu respectivo valor imputado $\widehat{g}_q$. Quanto menor for a $NRMSE$, melhor será para a validação na acurácia da imputação.

Para uma avaliação cuidadosa da eficiência do algoritmo de imputação, além da NRMSE, foram calculados os coeficientes de correlação de Pearson (R) nos cenários estudados, baseando-se em toda informação dos marcadores. Quanto maior R, melhor será a acurácia da imputação.

Os valores médios obtidos a partir de 1000 iterações destas medidas foram utilizados para avaliação.

Após a especificação do método de imputação dos genótipos nos marcadores moleculares o próximo passo agora é detalhar os métodos bayesianos que foram utilizados para o mapeamento de QTL neste trabalho.

3.2.3 Modelo de Múltiplos QTL

De acordo com Yi et al. (2005) em uma população de mapeamento tem-se que, os valores de uma determinada característica, $\mathbf{y}$, e os genótipos dos marcadores, $\mathbf{g}$, para cada indivíduo da população, são observáveis. Considere que o genoma está particionado em H locus, $\mathbf{P} = \{P_1, P_2, \ldots, P_H\}$ e que o QTL ocorre nestas posições. Sabe-se que os genótipos dos QTL, $\mathbf{q}$, nas posições $\mathbf{P}$, não são observados. Mas, a partir dos marcadores observados é possível com o uso da distribuição de probabilidades condicionais, $p(\mathbf{q}|\mathbf{P}, \mathbf{g})$, inferir estes valores. Esta distribuição de probabilidade será utilizada no contexto bayesiano, como uma distribuição a priori para os genótipos dos QTL. Ainda de acordo com estes autores, o problema que surge no momento de inferir o número e as posições dos múltiplos QTL é equivalente ao problema da seleção de um subconjunto de $\mathbf{P}$ que explique completamente a variação fenotípica.

Assim, seja o seguinte modelo linear,

$$y_i = \boldsymbol{\mu} + \sum_{j=1}^{H} \mathbf{x}_{ij}\boldsymbol{\beta}_j + \varepsilon_i \quad i = 1, 2, \ldots, n \tag{40}$$

em que, no modelo (40), $\boldsymbol{\mu}$ representa a média geral, $\mathbf{x}_{ij}$ denota o genótipo do j-ésimo QTL do i-ésimo indivíduo, $\boldsymbol{\beta}_j$ é um vetor contendo os efeitos genéticos associados ao j-ésimo QTL e ε_i é o erro residual que é modelado por uma distribuição normal com média zero e variância constante (σ^2). Para determinação desses efeitos será utilizada a parametrização de Cockerham (KAO; ZENG, 2002). Em uma população F_2, para o modelo (40), os elementos da matriz $\mathbf{X}$ são estabelecidos da seguinte forma,

$$\begin{aligned} x_{ij1} &= z_{ij} - 1 \\ x_{ij2} &= (1 + x_{ij1}) \times (1 - x_{ij1}) - 0,5. \end{aligned} \tag{41}$$

Na equação (41) z_{ij} é a quantidade de alelos dominantes do genótipo do j-ésimo QTL para o i-ésimo indivíduo.

Considerando o vetor $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$, de dimensão L que contém as posições dos L QTL. Cada QTL pode afetar a variação fenotípica por meio de seus efeitos principais.

Seja um vetor $\boldsymbol{\gamma}$, de variáveis aleatórias binárias, γ_j, para indicar a inclusão

de possíveis locus ($\gamma_j = 1$) ou exclusão ($\gamma_j = 0$). O vetor $\boldsymbol{\gamma}$ denominará a quantidade de QTL que foi incluída e os seus efeitos genéticos associados. Com isso, as posições dos QTL incluídos no modelo serão denotadas por $\boldsymbol{\lambda\gamma}$. Logo, o conjunto $(\boldsymbol{\gamma}, \boldsymbol{\lambda\gamma})$ representará a arquitetura genética, o quantidade e as posições dos QTL, como também a sua ação gênica (YI et al., 2007). Com base nessas considerações o modelo (40) pode ser reescrito da forma,

$$y_i = \boldsymbol{\mu} + \sum_{j=1}^{L} \gamma_j \mathbf{x}_{ij} \boldsymbol{\beta}_j + \varepsilon_i, \quad i = 1, 2, \ldots, n. \tag{42}$$

3.2.4 MCMC com Saltos Reversíveis para o mapeamento de QTL

Sabe-se que no algoritmo MCMC tradicional existe uma incapacidade no que diz respeito a cadeia se mover de um modelo H_l para outro modelo $H_{l'}$, sendo l e l' as dimensões do espaço paramétrico dos dois modelos, ou seja, para o mapeamento de QTL utilizando abordagem bayesiana, seria como a cadeia mover-se de um modelo contendo l QTL para um modelo com l' QTL, sendo l' igual a $l + 1$ (nascimento de um QTL) ou igual a $l - 1$ (morte de um QTL).

Como solução para o problema de dimensão do espaço paramétrico, Green em 1995 propôs que os saltos entre os modelos H_l e $H_{l'}$ pudessem ser decompostos entre movimentos. E que, se a dimensão de H_l for maior do que a dimensão de $H_{l'}$ e se o movimento de H_l para $H_{l'}$ puder ser representado por uma transformação determinística de θ_l então, $\theta_{l'} = T(\theta_l)$.

No MCMC com Saltos Reversíveis cada estado da cadeia de Markov U_i contempla dois componentes, o indicador da quantidade de QTL, $L_i \in \{1, 2, \ldots, l\}$ e o vetor estocástico das possíveis dimensões dos parâmetros desconhecidos, $\mathbf{z}_i$. O vetor $\mathbf{z}$ toma valores num conjunto C, definido como a união de espaços $c = \mathbb{R}^{n_l}$, $n_l \geq 1$. Dado $L = l$, $\mathbf{z}$ pode tomar valores somente em C_l. Supondo que (l, z) é o estado atual da cadeia de Markov denotado por $U^{(t)}$ e que uma proposta $U^{(t+1)} = \left(L^{(t+1)}, Z^{(t+1)}\right)$ é gerada para um novo estado da cadeia. Com probabilidade $b_{ll'}$ a proposta $L^{(t+1)}$ é igual a l' QTL. Então, dado $L^{(t+1)} = l'$, a proposta $Z^{(t+1)}$ é gerada em $C_{l'}$. Considerando $\mathbf{u}$ um vetor aleatório em $\mathbb{R}^{n_{ll'}}$ com $n_{ll'} \geq 1$, o qual tem densidade de proposta $q_{ll'}(z, \mu)$ e $\mathbb{R}^{n_s + n_{ss'}} \longrightarrow \mathbb{R}^{n_{s'}}$, levando assim, a um mapeamento determinístico.

Segundo Silva (2006) quando considerado um movimento de estado (l, z)

para $(l', z') = \left(l', g_{ll'(z,\mu)}\right)$ e o movimento reverso de (l', z') para (l, z) para $(l', z') = \left(l', g_{l'l(z',\mu')}\right)$ os vetores de estado da cadeia de Markov e as variáveis aleatórias propostas (z, μ) e (z', μ') possuírão a mesma dimensão. A proposta $U^{(t+1)}$ é então aceita com probabilidade de aceitação:

$$a_{ll'}(z, z') = min\left\{1, \frac{\pi(l', z', |\mathbf{y})\, q_{l'l}(z', \mu')\, b_{l'l}}{\pi(l, z, |\mathbf{y})\, q_{ll'}(z, \mu)\, b_{ll'}} \times \left|\frac{\partial\left(g_{ll'}\right)(z, \mu)}{\partial z \partial \mu}\right|\right\}$$

De maneira didática esta probabilidade pode ser expressa da forma,

$$min\left\{\begin{array}{l} 1, (\text{razão a posteriori}) \quad \times (\text{razão proposta para o Salto}) \times (\text{probabilidade de Saltar}) \\ \qquad\qquad\qquad\qquad \times (\text{Jacobiano da transformação}). \end{array}\right\}$$

Green (1995) definiu os três possíveis movimentos entre os modelos:

1. **Nascimento de um QTL**

   Neste caso é proposto saltar de um modelo com l QTL para um modelo com $l+1$ QTL com probabilidade b_n. Assim, a probabilidade de aceitação é expressa da forma,

   $$a_{l(l+1)}(z, z+1) = min\left\{1, \frac{\pi(l+1, z+1, |\mathbf{y})\, q_{(l+1)l}(l+1, z+1|l, z)\, b_m}{\pi(l, z, |\mathbf{y})\, q_{l(l+1)}(l, z|l+1, z+1)\, b_n}\right\}.$$

2. **Morte de um QTL**

   Aqui, o modelo contendo l QTL saltará para o modelo com $l-1$ QTL com probabilidade b_m. A probabilidade de aceitação é dada da forma,

   $$a_{l(l+1)}(z, z+1) = min\left\{1, \frac{\pi(l, z, |\mathbf{y})\, q_{l(l+1)}(l, z|l+1, z+1)\, b_n}{\pi(l+1, z+1, |\mathbf{y})\, q_{(l+1)l}(l+1, z+1|l, z)\, b_m}\right\}.$$

3. **Permanência do QTL**

   Neste caso a quantidade de QTL entre os modelos permanecerá inalterada com probabilidade $b_p = 1 - (b_n + b_m)$.

3.2.5 Espaço composto

No modelo descrito na Equação (42), o vetor de variáveis indicadoras, $\boldsymbol{\gamma} = \{\gamma_j\}_{j=1}^{L}$, denota a quantidade de QTL. Seja $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_j\right\}_{j=1}^{L}$, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\mu}, \sigma^2)^T$,

$\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^{L}$, $\mathbf{x} = \{x_{ij}\}_{(i,j)=(1,1)}^{(n,L)}$ e a partição, $\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right)$, quando for inclusão, $\gamma_j = 1$, e $\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{x}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}}\right)$, quando for exclusão, $\gamma_j = 0$. Em que, $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = \left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \sigma^2_{\boldsymbol{\gamma}}\right)$ e $\boldsymbol{\theta}_{-\boldsymbol{\gamma}} = \boldsymbol{\beta}_{-\boldsymbol{\gamma}}$.

Para este mesmo modelo, a função de verossimilhança em um determinado $\boldsymbol{\gamma}$, dependerá de $\mathbf{x}_{\boldsymbol{\gamma}}$ e de $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$, e será escrita da forma,

$$L\left(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{x}, \boldsymbol{\theta}\right) = L\left(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right). \tag{43}$$

As distribuições a priori de $(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{x}, \boldsymbol{\theta})$ são então fatoradas em três componentes,

$$\begin{aligned} p\left(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{x}, \boldsymbol{\theta}\right) &= p\left(\boldsymbol{\lambda}\right) \times p\left(\boldsymbol{\theta}, \boldsymbol{\lambda}, \mathbf{x}|\boldsymbol{\gamma}\right) \\ &= p\left(\boldsymbol{\gamma}\right) \times p\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) \times p\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{x}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right). \end{aligned} \tag{44}$$

O produto entre a função de verossimilhança (43) e as distribuições a priori (44) resulta na distribuição completa a posteriori para o modelo de espaço composto,

$$\begin{aligned} p\left(\boldsymbol{\gamma}, \boldsymbol{\lambda}, \mathbf{x}, \boldsymbol{\theta}|\mathbf{y}\right) \propto \; & L\left(\mathbf{y}|\boldsymbol{\gamma}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right) \times \\ & p\left(\boldsymbol{\gamma}\right) \times p\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) \times \\ & p\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{x}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right). \end{aligned} \tag{45}$$

No modelo (45) a função conjunta a priori, $p\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right)$, será reescrita em função do produto de outras três prioris, como segue,

$$p\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) = p\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) \times p\left(\mathbf{x}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) \times p\left(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \mathbf{x}_{\boldsymbol{\gamma}}\right). \tag{46}$$

Ainda no modelo (45), para a distribuição conjunta a priori, $p\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{x}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right)$, será fatorada em três componentes,

$$p\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}, \mathbf{x}_{-\boldsymbol{\gamma}}, \boldsymbol{\theta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}, \boldsymbol{\lambda}_{\boldsymbol{\gamma}}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right) = p\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right) \times p\left(\mathbf{x}_{-\boldsymbol{\gamma}}|\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}\right) \times p\left(\boldsymbol{\theta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right). \tag{47}$$

Nos modelos (46) e (47), as distribuições $p\left(\boldsymbol{\lambda}_{\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right)$ e $p\left(\boldsymbol{\lambda}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma}\right)$ representam as prioris para as posições dos QTL. Já $p\left(\mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\lambda}_{\boldsymbol{\gamma}}\right)$ e $p\left(\mathbf{x}_{-\boldsymbol{\gamma}}, \boldsymbol{\lambda}_{-\boldsymbol{\gamma}}\right)$ são as distribuições de probabilidades para os genótipos dos QTL, que por sua vez são calculadas utilizando

análise multiponto (JIANG; ZENG, 1997).

De acordo com Yi (2004) a grande vantagem que o modelo de espaço composto oferece é a de que a dimensão dos parâmetros mantém-se inalterável.

3.2.6 Especificando as prioris

A partir dos modelos especificados anteriormente, dois grandes problemas são inevitáveis, a especificação de distribuições a priori para cada parâmetro do modelo e o cálculo da distribuição a posteriori. A seguir será descrita as prioris utilizadas neste trabalho.

Para o número esperado de QTL a ser considerado no modelo l_0, será aqui utilizado métodos clássicos, tal como, o mapeamento por intervalo composto, para que, em seguida, seja determinado um valor plausível para L (quantidade de QTL com efeitos detectáveis), uma vez que o valor para L terá grande influência sobre as estimativas dos parâmetros a posteriori (GREEN, 1995).

A seguir, uma breve descrição da análise clássica que foi utilizada neste trabalho.

Sabe-se que, para o modelo bayesiano, o número esperado de QTL terá que ser inferido para realização dos cálculos a posteriori. Assim, este número esperado será estimado por meio de uma análise utilizando Mapeamento por Intervalo Composto (CIM). Nesta abordagem, para a identificação de QTL nos cromossomos será utilizada a estatística de LOD score. De acordo com Bromam e Sen (2009) o LOD score indica evidência à presença de QTL. Segundo estes autores para a construção desta estatística, considera-se a hipótese nula H_0 sobre ausência de QTL. Esta hipótese é construída da seguinte forma. Seja $\mathbf{y} = (y_1, y_2, \ldots, y_i, \ldots, y_n)$, em que y_i denota o fenótipo do i-ésimo indivíduo. Sob a suposição de que a distribuição normal modela bem os valores fenotípicos independentemente dos genótipos, tem-se, $\mathbf{y} \sim N(\mu, \sigma^2)$. Seja a função de verossimilhança $L_0(\mu, \sigma^2) = P(\mathbf{y}|H_0) = \prod_{i=1}^{n} \phi\left(y_i; \mu, \sigma^2\right)$, em que ϕ é a densidade de uma distribuição normal. Os estimadores pelo Método da Máxima Verossimilhança (EMV) para μ e σ^2 são $\bar{y}$ e $\sum_{i=1}^{n} (y_i - \bar{y})^2 / n = SQRes_0/n$, respectivamente, sendo $SQRes_0$ a soma de quadrados residual sob a hipótese nula. Sob a hipótese alternativa H_1, de que existe ao menos um QTL no marcador, assume-se que $y_i|g_i \sim N(\mu_{g_i}, \sigma^2)$, em que g_i é o genótipo referente ao i-ésimo indivíduo, μ_{AA} e μ_{AB} são as médias fenotípicas para os dois grupos de genótipos e

σ^2 é a variância residual, a mesma para ambos os grupos. A função de verossimilhança é $L_1(\mu_{AA}, \mu_{BB}, \sigma^2) = \prod_{i=1}^{n} \phi\left(y_i; \mu_{g_i}, \sigma^2\right)$ e os EMVs para μ_i são as médias fenotípicas dentro dos dois grupos de genótipos. O EMV para σ^2 é uma estimativa agrupada, $SQRes_{_1}/n$, em que $SQRes_{_1} = \sum_{i=1}^{n} (y_i - \widehat{\mu}_{g_i})^2$ é a soma de quadrados residual sob H_1. Portanto, o LOD score é definido da seguinte maneira,

$$LOD = \frac{n}{2} \log_{10} \left(\frac{SQRes_{_0}}{SQRes_{_1}} \right). \tag{48}$$

Segundo Grier et al. (1998), se o LOD score for maior do que um limite superior definido, A, então a hipótese alternativa H_1 não é rejeitada. Mas se o LOD score for inferior a um valor mínimo definido, B, a hipótese nula H_0 não será rejeitada. Porém, se o LOD score estiver entre os limitantes superior e inferior, A e B, respectivamente, então conclui-se que não há dados suficientes que indique a não-rejeição de qualquer uma das hipóteses, sendo assim, os autores recomendam que mais observações sejam coletadas.

Para a quantidade total de QTL L, será atribuída uma distribuição de Poisson com média $3\sqrt{l_0}$. Para o número de QTL incluídos no modelo e aos seus efeitos genéticos associados será atribuída uma priori descrita pela expressão,

$$p(\boldsymbol{\gamma}) = \prod_{j=1}^{L} \omega_j^{\gamma_j} (1 - \omega_j)^{1-\gamma_j}. \tag{49}$$

No modelo (49) $\omega_j = p(\gamma_j = 1)$ é a probabilidade a priori referente ao j-ésimo efeito. Cada QTL entra no modelo, independentemente de qualquer outro QTL, com uma probabilidade $p(\gamma_j = 1) = 1 - p(\gamma_j = 0) = \omega_j$.

Para a posição do j-ésimo QTL, λ_j, assumi-se uma distribuição a priori Uniforme no intervalo $[a, b]$. Mas, dado que o comprimento do genoma é K, tem-se que $\lambda \sim U[0, K]$. Neste trabalho, para os efeitos genéticos, serão atribuídas prioris hierárquicas,

$$\beta_j \sim N\left(0, \gamma_j n \sigma^2 \left(\mathbf{x}_{.j}^{'} \mathbf{x}_{.j}\right)^{-1}\right). \tag{50}$$

Na equação (50) $\mathbf{x}_{.j} = \{x_{1j}, x_{2j}, \ldots, x_{nj}\}^T$ é um vetor contendo os coeficientes de β_j e um fator de escala positiva e n é a quantidade de indivíduos. Como os efeitos genéticos

são particionados em grupos, aditivo e dominante, para os efeitos genéticos do mesmo grupo k, foram estabelecidas as mesmas prioris, $\beta_{k,j} \sim N(0, \sigma_k^2)$. Para a variância, σ_k^2, foi atribuída uma hiperpriori χ^2-Inversa, $\sigma_k^2 \sim \text{Inv-}\chi^2(v_k, s_k^2)$, em que, v_k são os graus de liberdade, recomenda-se, $v_k = 6$ (CHIPMAN, 2004). De acordo com Gaffney (2001), s_k^2 é um parâmetro de escala que tem por objetivo controlar a região de confiança para a variância explicada pelo β_{kj}.

Para a média geral, μ, será atribuída uma distribuição a priori Normal, $\mu \sim N(\eta_0, \tau_0^2)$, em que η_0 e τ_0^2 serão a média e a variância amostral, respectivamente. Já para a variância residual, σ^2, será estabelecida uma priori "pouca informativa", $p(\sigma^2) = \frac{1}{\sigma^2}$.

3.2.7 Cálculos a posteriori

Considere a função de verossimilhança escrita na equação (43). A distribuição conjunta a posteriori será proporcional ao produto da função de verossimilhança, com as distribuições a priori dos parâmetros especificados na seção anterior. O algoritmo MCMC com Saltos Reversíveis será utilizado para efetuar os cálculos. Assim, considere a posteriori escrita da forma,

$$\begin{aligned} p(\boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{q}, \boldsymbol{\beta}, \boldsymbol{\Omega}, \mu, \sigma^2 | \mathbf{y}) \propto \; & L\left(\mathbf{y} | \boldsymbol{\gamma}, \mathbf{x}_{\boldsymbol{\gamma}}, \boldsymbol{\theta}_{\boldsymbol{\gamma}}\right) \times \\ & p(\boldsymbol{\lambda}) \times p(\boldsymbol{\gamma}) \times p(\mathbf{q} | \boldsymbol{\lambda}) \times p(\boldsymbol{\beta} | \boldsymbol{\Omega}) \times p(\boldsymbol{\Omega}) \times p(\mu) \times p(\sigma^2). \end{aligned} \tag{51}$$

Em (51), $\boldsymbol{\Omega}$ foi inserido para representar todas as variâncias de $\boldsymbol{\beta}$. Yi et al. (2007) escreveram um algoritmo que de forma aleatória atualiza os parâmetros $\boldsymbol{\lambda}, \boldsymbol{\gamma}, \mathbf{q}, \boldsymbol{\beta}, \boldsymbol{\Omega}, \mu, \sigma^2$. De acordo com estes autores, os parâmetros μ e $\boldsymbol{\beta}$ podem ser atualizados dado (Ω, σ), a partir de uma distribuição normal e todos os elementos de (Ω, σ), a partir de independentes, χ^2-Inversa, distribuições condicionais a posteriori dado $(\mu, \boldsymbol{\beta})$. A distribuição condicional a posteriori de cada elemento de $\boldsymbol{\lambda}$ não possui fórmula explícita sendo necessária a implementação do algoritmo Metropolis–Hastings (Yi et al., 2005). Já a distribuição condicional a posteriori para cada um dos elementos de $\mathbf{q}$ será uma multinomial.

A partir da distribuição condicional a posteriori, por meio do amostrador de Gibbs, é possível gerar todas as variáveis indicadoras, λ_j, da seguinte maneira,

$$p\left(\gamma_j = 1 | \boldsymbol{\gamma}_{-j}, \mathbf{X}, \boldsymbol{\beta}_{-j}, \boldsymbol{\Omega}, \mathbf{y}\right) = \frac{\omega L_1}{(1-\omega) L_0 + \omega L_1}, \tag{52}$$

na equação (52), "$_{-j}$" significa que o j-ésimo elemento não será incluso, ω é a probabilidade a priori de inclusão do j-ésimo elemento, $p\left(\gamma_j = 1|\boldsymbol{\gamma}_{-j}\right)$, e $L_m = p\left(\mathbf{y}|\gamma_j = m, \boldsymbol{\gamma}_{-j}, \mathbf{X}, \boldsymbol{\beta}_{-j}, \boldsymbol{\Omega}\right)$, $m = \{0, 1\}$.

Para realização das análises, será utilizado neste trabalho o programa R/qtlbim (YANDELL et al., 2007). Neste programa está implementado o algoritmo de Metropolis-Hastings modificado. De acordo com Yi et al. (2007), este novo algoritmo faz atualizações de $\boldsymbol{\gamma}$, o qual é computacionalmente mais eficiente em relação ao amostrador de Gibbs, uma vez que o número dos possíveis efeitos genéticos é grande. Este algoritmo, proposto por estes autores, procede da seguinte maneira,

a) Seja C igual ao valor zero ou um, o valor atual para γ_j;

b) Agora, considere um novo valor para γ_j, P = (0 ou 1), a partir da distribuição condicional a priori, $p\left(\gamma_j = C|\boldsymbol{\gamma}_{-j}\right)$;

c) Se $P = C$ a probabilidade de aceitação será 1, fazendo com que γ_j permaneça em C, não havendo necessidade de calcular quaisquer valores;

d) Se $P \neq C$, γ_j será atualizado a partir do valor de C cuja proposta é $1 - C$.

Assim, a probabilidade de aceitação é expressa da forma,

$$\alpha = \min\left(1, \frac{L_{1-C}}{L_C}\right). \tag{53}$$

Na equação (53) os valores L_0 e L_1 podem ser calculados a partir da coluna da matriz $\mathbf{X}$ condicionada a priori da variância relacionada ao parâmetro β_j.

3.2.8 Componentes de variância

Uma maneira comum de calcular a estimativa da variância ambiental é calculando a expressão, $\widehat{\sigma}^2 = RSS\left(\widehat{\boldsymbol{\theta}}\right)/gl$. Nesta expressão $RSS\left(\widehat{\boldsymbol{\theta}}\right) = \sum\left(\mathbf{y} - \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\Gamma}\boldsymbol{\beta}\right)^2$ e $gl = n - 1 - \sum\gamma$. A estimativa bayesiana para σ^2 é a sua própria média a posteriori. A herdabilidade, que expressa a relação entre as variâncias fenotípica e genotípica, ou seja, que mede o quanto da variação da característica fenotípica é explicada por fatores genéticos e também ambientais, é calculada da forma,

$$h^2 = \frac{TSS - RSS\left(\widehat{\boldsymbol{\theta}}\right)}{TSS} \times 100\%. \tag{54}$$

Na equação (54) $TSS = \sum (\mathbf{y} - \bar{\mathbf{y}})^2$, que é compreendida como a soma de quadrado total.

3.2.9 Fator de Bayes

Para comparar duas arquiteturas genéticas diferentes (modelos com diferentes quantidades de QTL) será aqui calculado o Fator de Bayes (FB) da seguinte forma,

$$FB = \frac{p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{m})}{p(\boldsymbol{\gamma})} \times \frac{p(0)}{p(0|\mathbf{y})}. \tag{55}$$

Na equação (55), $p(\boldsymbol{\gamma}|\mathbf{y}, \mathbf{m})$ e $p(0|\mathbf{y})$ são obtidas por meio da amostra da distribuição a posteriori para a quantidade de QTL, $p(0|\mathbf{y})$ é a posteriori para o modelo com a menor quantidade de QTL. Já $p(0)$ e $p(\boldsymbol{\gamma})$ são as prioris, que por sua vez são obtidas por meio da distribuição de Poisson com média 3. Supondo que na análise foram considerados modelo com diferentes quantidade de QTL $(l, l+1, l+2, \ldots)$. Logo, as prioris para cada quantidade serão calculadas da forma,

$$P(X = l) = \frac{3^l e^{-3}}{l!}$$

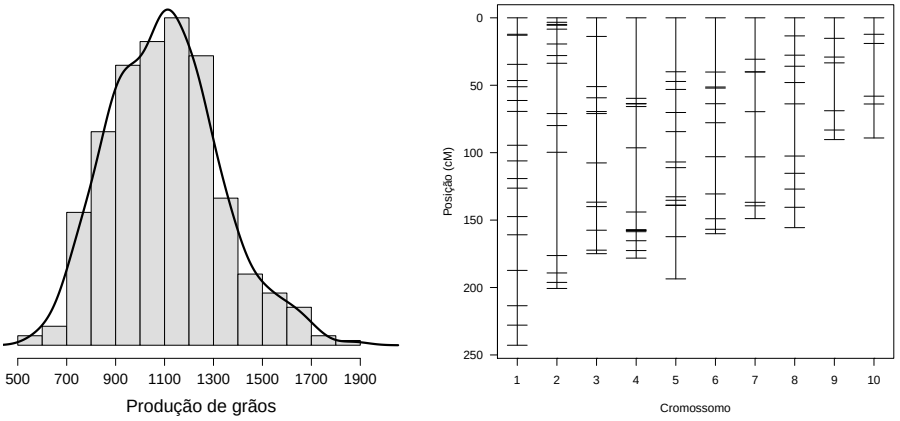
Após esta descrição da material e dos métodos que foram utilizados neste trabalho a próxima etapa agora é descrever os resultados e discutí-los.

4 RESULTADOS E DISCUSSÃO

4.1 Análise exploratória

Na Figura 6(a) encontra-se o histograma da característica fenotípica, produção de grãos. A linha tracejada na cor preta representa a densidade desta variável. Nota-se que há uma assimetria positiva de $0,38$. Para verificar se esta variável segue uma distribuição normal foi feito o teste de Shapiro-Wilk, que apresentou um valor p de $0,005$. Como este valor calculado é menor do que o nível de significância de $0,05$, logo existem fortes evidências para rejeição da hipótese de que a distribuição normal modela adequadamente à produção de grãos. Assim, com o intuito de sanar este problema de não normalidade desta característica foi proposta a transformação log.

O mapa genético, Figura 6(b), é composto de 117 marcadores microssatélites alocados em 10 cromossomos. Este mapa tem comprimento total de $1634,20$ cM e distância média entre as marcas de 14 cM. O comprimento dos cromossomos variou de $89,10$ cM (cromossomo 10) a $242,80$ cM (cromossomo 1) e o número de marcas em cada cromossomo variou de 6 (cromossomo 10) a 18 (cromossomo 1). Nota-se ainda que, as marcas encontram-se distribuídas de forma aleatória por todo o genoma.

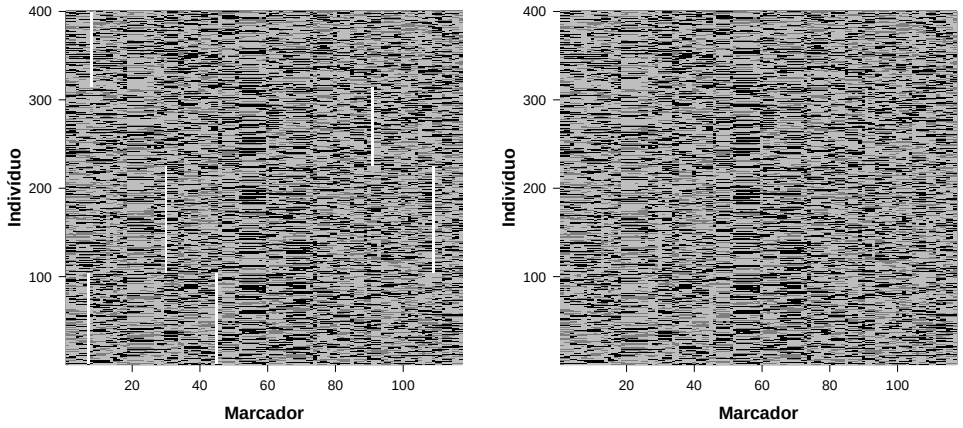


(a) (b)

Figura 6 - Histograma da característica fenotípica produção de grãos (a) e o mapa genético (b)

4.2 Imputação dos genótipos

Na Figura 7(a) tem-se uma representação gráfica da matriz dos genótipos dos marcadores observados. O conjunto de dados é constituído de 400 indivíduos e 117 marcadores, ou seja, cada indivíduo foi observado 117 vezes, resultando numa matriz de ordem 400×117. Retornando a Figura 7(a), os tons em branco, aproximadamente, 2% dos dados, representam a ausência de genótipos nos marcadores. Como na inferência bayesiana os valores perdidos são considerados como quantidades desconhecidas, ocasionando assim mais incertezas para o modelo, logo, se faz necessária, antes de realizar a inferência bayesiana para o mapeamento de QTL, a imputação destes genótipos.



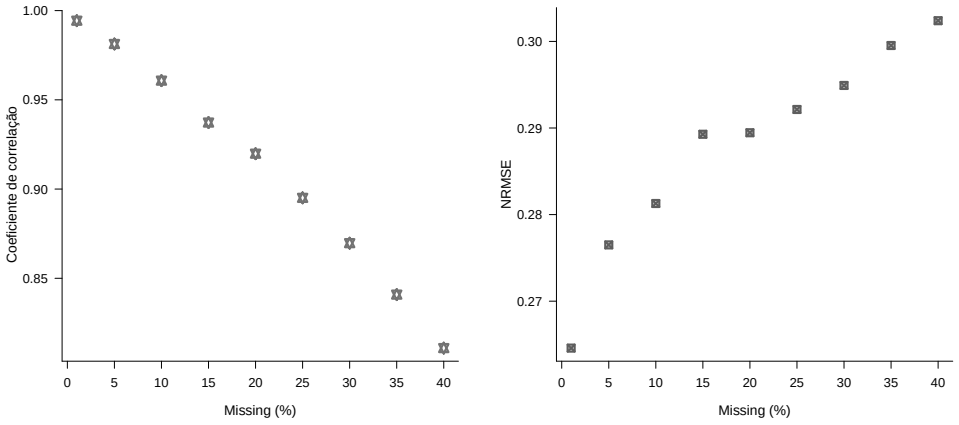
(a) (b)

Figura 7 - Representações gráficas das matrizes dos marcadores observados (a) e desses marcadores após a imputação (b)

Conforme descrito na metodologia deste trabalho será utilizado o algoritmo VT para imputação das quantidades não observadas. A seguir, uma breve descrição da programação deste algoritmo no trabalho.

Para o parâmetro relacionado as probabilidades de transição foi utilizado um valor inicial de $0,10$. Já para o parâmetro que compõe as probabilidades de emissão foi inserido um valor inicial de $0,01$. Para as probabilidades iniciais, foram utilizados valores inicias conforme regra de Mendel. A quantidade máxima de iterações para convergência deste algoritmo foi de tamanho 100. Todas as análises foram realizadas no programa estatístico R (R CORE TEAM, 2013). Os algoritmos necessários para realizar estas

análises, encontram-se implementados no pacote HMM do programa estatístico R. Na Figura 7(b), tem-se a representação gráfica da matriz dos marcadores observados após a imputação.



(a) (b)

Figura 8 - Coeficiente de correlação de Pearson (a) e raiz quadrada do erro quadrático médio normalizado (NRMSE) (b)

Para verificação da acurácia na imputação, na Figura 8(a) estão os valores médios do coeficiente de correlação de Pearson obtidos a partir de 1000 iterações para os conjuntos de dados imputados e dos valores reais. Observa-se que, a medida que a porcentagem de valores ausentes aumenta há um decréscimo na correlação entre os valores imputados e os valores reais. Na Figura 8(b) está a representação gráfica da NRMSE. Verificou-se que, em média, as estimativas da NRMSE, encontram-se dentro de um patamar considerável para a validação da acurácia.

4.3 Análise bayesiana: MCMC com Saltos Reversíveis

A implementação computacional foi realizada utilizando o pacote *qtlbim* do programa R (R CORE TEAM, 2013). A construção da amostra a posteriori foi realizada utilizando os métodos MCMC com Saltos Reversíveis. Foram realizadas 120 mil iterações, havendo um aquecimento da cadeia de tamanho 5 mil e utilizado um espaçamento de 40 iterações. Assim, após a execução para obtenção da amostra foi possível verificar a convergência da cadeia para os parâmetros do modelo: média geral, variâncias ambiental

e genética (ANEXO A - Figura 13).

Com base na amostra a posteriori é possível identificar regiões do genoma que influenciam na variação da característica fenotípica e, em seguida, estimar os efeitos genéticos destes QTL. A seguir, será detalhada como foi feita a construção do modelo para obtenção desta amostra.

Para o cálculo das probabilidades condicionais foi utilizada a função de mapeamento de Haldade e uma distância de $0,20$ cM entre os marcadores flanqueadores. Para a quantidade esperada de QTL no modelo foi realizada a seguinte análise.

Por meio do teste de permutação, 1000 iterações foram executadas e, ao nível de $0,10$ de significância obteve um limiar crítico de $3,65$ cM. Em seguida, foi realizado o mapeamento por intervalo composto, utilizando 10 marcadores como covariáveis para o modelo. Após este procedimento, foram detectados três possíveis QTL. Na Figura 9 tem-se um esboço gráfico desta análise, os picos que se destacam indicam possíveis presenças de QTL nessas regiões.

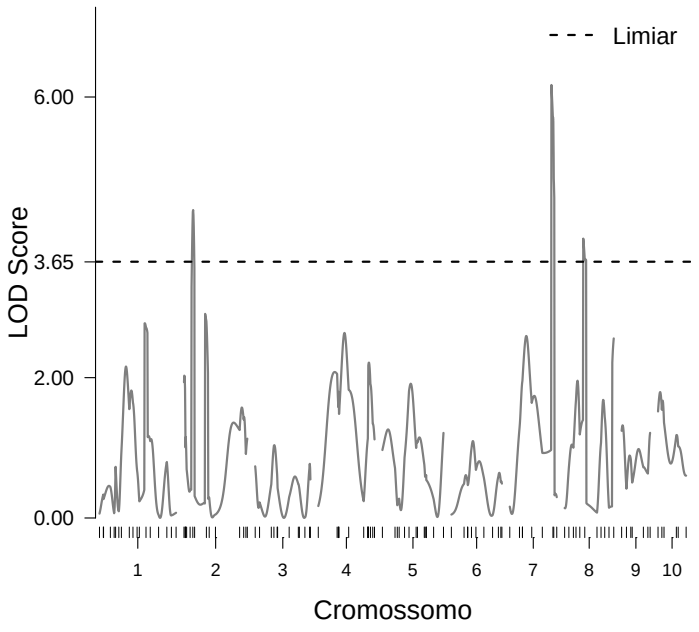
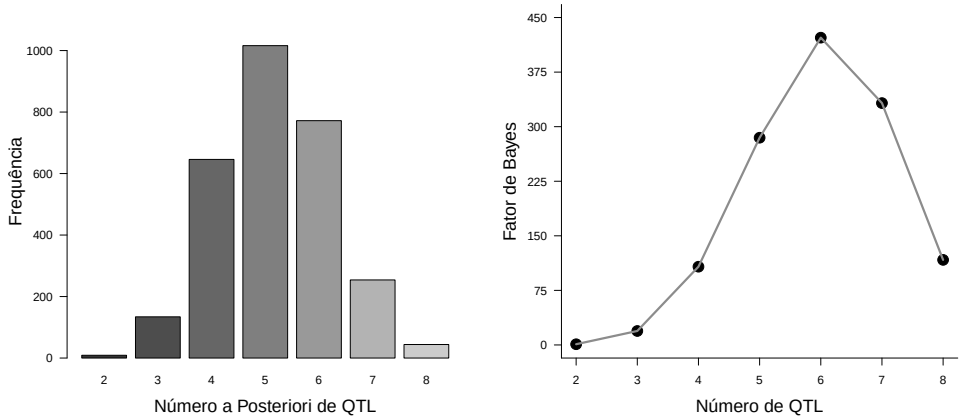


Figura 9 - Mapeamento por Intervalo Composto

Assim, no mapeamento bayesiano será considerada a quantidade de três para o número esperado de QTL no modelo. Essa mesma quantidade foi considerada

para os efeitos aditivo e dominante. Por simplicidade, as prioris foram consideradas independentes.

Após a execução da cadeia, as maiores frequências a posteriori encontram-se entre os modelos com quatro, cinco e seis QTL (Figura 10(a)) .



(a) (b)

Figura 10 - Frequência a posteriori para o número de QTL (a) e o Fator de Bayes para cada quantidade de QTL (b)

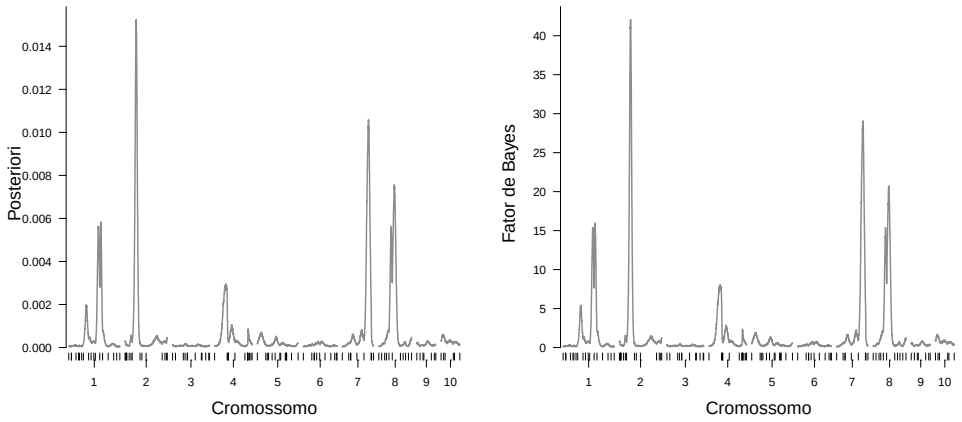
Pelo Fator de Bayes, Figura 10(b), os modelos com cinco, seis e sete QTL foram os que apresentaram os maiores índices. Para uma análise mais criteriosa é apresentada na Tabela 7 um resumo numérico dos valores mostrados na Figura 10.

Tabela 7 - Fator de Bayes para determinação do número de QTL presentes no modelo

Número de QTL	Posteriori	Priori	Fator de Bayes
2	$2,78 \times 10^{-2}$	$22,40 \times 10^{-2}$	$1,00$
3	$5,33 \times 10^{-2}$	$22,40 \times 10^{-2}$	$19,20$
4	$22,38 \times 10^{-2}$	$16,80 \times 10^{-2}$	$107,56$
5	$35,57 \times 10^{-2}$	$10,08 \times 10^{-2}$	$284,85$
6	$26,37 \times 10^{-2}$	$5,04 \times 10^{-2}$	$422,34$
7	$8,90 \times 10^{-2}$	$2,16 \times 10^{-2}$	$332,44$
8	$1,17 \times 10^{-2}$	$0,81 \times 10^{-2}$	$116,93$

Na Figura 11 é esboçada uma análise gráfica para a amostra a posteriori, Figura 11(a), como também calculado o Fator de Bayes, Figura 11(b), para cada região do genoma. Nos eixos horizontais, na parte externa estão os cromossomos e na parte interna os marcadores e nos eixos verticais estão as frequências a posteriori e o Fator de Bayes.

Em ambas as figuras observa-se uma grande concentração de picos nos cromossomos 1, 2, 4, 7 e 8, ou seja, há fortes evidência da presença de QTL nesses locais.



(a) (b)

Figura 11 - Análise unidimensional dos efeitos principais em cada marca do mapa genético para a posteriori (a) e para o Fator de Bayes (b)

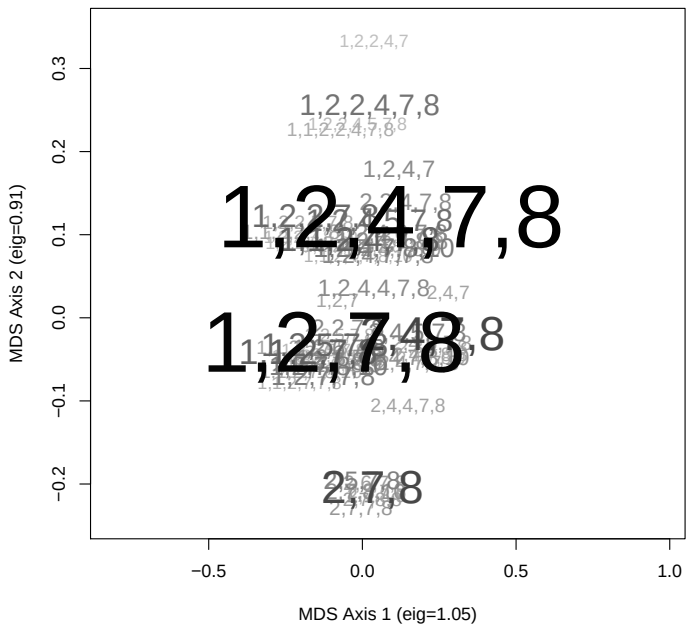


Figura 12 - Arquitetura genética de acordo com as estimativas da variância de cada QTL

Observando os resultados obtidos nas Figuras 10 e 11 e na Tabela 7 ficou evidenciado que o modelo com cinco QTL se ajustou melhor aos dados. Esse resultado é reforçado de acordo com a melhor arquitetura genética, Figura 12. Nesta figura o tamanho da fonte de um padrão é determinado pela probabilidade a posteriori. Observa que o modelo que se sobrepõe é formado pelos cromossomos $1, 2, 4, 7$ e 8.

Sabendo das regiões dos QTL que influenciam, significativamente, a variabilidade da característica fenotípica o último passo agora é determinar os efeitos genéticos de cada um destes QTL. A seguir um resumo destas estimativas.

A Tabela 8 apresenta as estimativas para a localização, para os efeitos genéticos (aditivo e dominante), tipo de interação alélica e a proporção da variância fenotípica explicada pelos QTL. Os efeitos genéticos aditivos do QTL para a característica fenotípica variaram de $-5,72\times10^{-2}$ a $+5,92\times10^{-2}$. Já os de dominância variaram de $-3,11\times10^{-2}$ a $+1,31\times10^{-2}$. A proporção da variância fenotípica explicada pelos QTL variam de $2,01\%$ para o QTL localizado no cromossomo 1 na posição $142,80$ cM a $4,10\%$ para o QTL localizado no cromossomo 8 na posição $68,20$ cM. Os cinco QTL detectados explicam $15,73\%$ da variação da característica fenotípica.

Tabela 8 - Estimativas da localização, dos efeitos aditivos ($\widehat{a}$) e dominantes ($\widehat{d}$), do grau de dominância (GD) e da herdabilidade ($\widehat{h}^2$) para cada QTL

Cromossomo	Posição cM	$\widehat{a}\times10^{-2}$	$\widehat{d}\times10^{-2}$	GD	$\widehat{h}^2$
1	$142,80$	$-3,57$	$-1,62$	$0,45$	$2,01\%$
2	$53,20$	$-5,72$	$+1,31$	$0,23$	$4,08\%$
4	$54,20$	$-4,40$	$-0,44$	$0,01$	$2,09\%$
7	$121,80$	$-5,44$	$-3,11$	$0,57$	$3,45\%$
8	$68,20$	$+5,92$	$-1,83$	$0,31$	$4,10\%$

As estimativas dos efeitos genéticos como também os valores de $\left|\widehat{d}/\widehat{a}\right|$ para cada QTL mostraram ocorrência de diferentes interações alélicas, sendo a de maior presença a de dominância parcial (4 QTL) e apenas um QTL apresentou interação do tipo aditiva.

Neste trabalho constatou-se que a utilização de modelos ocultos de Markov para imputação dos genótipos nos marcadores moleculares foi eficiente, fato este comprovado por meio de simulações realizadas com os valores observados. A partir do momento que o conjunto de dados não apresentava mais valores faltantes a construção do modelo se tornou mais eficaz e precisa, pois a dimensão do espaço paramétrico do modelo se reduziu

consideravelmente.

Para detecção de QTL utilizando uma abordagem bayesiana foi visto que não é preciso determinar um limiar crítico, uma vez que a quantidade de QTL é estabelecida pelo Fator de Bayes. A identificação de QTL se fez importante, pois sabendo a região do genoma que influenciam a variabilidade da característica fenotípica, os geneticistas poder-se-ão concentrar seus experimentos e análises em locais mais precisos do genoma e que contribuem significativamente para variações associadas a características fenotípicas, tal como, a produção de grãos de milho. Acrescenta-se ainda que, a grande dificuldade no mapeamento genético diz respeito ao fato de que não se conhece ao certo a quantidade de QTL significativos, ocasionando assim diversos problemas, um deles é a dimensão do espaço paramétrico. Como não se sabe ao certo esta quantidade, o desafio consiste em obter uma distribuição conjunta a posteriori para os parâmetros, uma vez que esta quantidade pode ser considerada como uma variável aleatória. Assim, com o objetivo de contornar este problema foi proposto a utilização dos métodos MCMC com Saltos Reversíveis e o espaço composto. Porém a complexidade de implementação e o entendimento da metodologia é um fator a ser descrito, existem poucos programas que implementaram estas técnicas e cujo códigos não encontram-se detalhados.

Em razão da limitação do tempo não foram apresentados neste trabalho detalhes teóricos no que diz respeito aos cálculos envolvendo as amostras a posteriori. Para trabalhos futuros tem-se a necessidade de compreensão da implementação computacional do mapeamento de QTL utilizando abordagem bayesiana com métodos MCMC com Saltos Reversíveis.

5 CONCLUSÃO

Os resultados obtidos neste trabalho possibilitaram as seguintes conclusões:

i) A utilização dos modelos ocultos de Markov possibilitou a imputação dos valores ausentes dos genótipos dos marcadores moleculares.

ii) Por meio de simulações verificou-se que a metodologia utilizada para imputação foi eficiente e eficaz, fato este comprovado por meio de técnicas de acurácia.

iii) Por meio da abordagem bayesiana utilizando o método MCMC com Saltos Reversíveis foram detectados cinco QTL. Os efeitos destes QTL mostraram diferentes tipos de interações alélicas, sendo a principal delas a de dominância parcial.

REFERÊNCIAS

AL-ANI, T. Hidden Markov models in dynamic system modelling and diagnosis. In: DYMARSKI, P. (Ed.). **Hidden Markov models, theory and applications**. Croatia: InTech, 2011. chap. 2, p. 27-50.

BANERJEE, S.; YANDELL, B.S.; YI, N. Bayesian quantitative trait loci mapping for multiple traits. **Genetics**, Austin, v. 179, p. 2275-2289, 2008.

BAUM, L.E.; PETRIE, T. Statistical inference for probabilistic functions of finite state Markov chains. **The annals of mathematical statistics**, Austin, p. 1554-1563, 1966.

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

BROMAN, K.W.; SEN, S. **A guide to QTL mapping with R/qtl**. New York: Springer, 2009. 396 p.

BROWNING, B.L.; BROWNING, S.R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. **The American Journal of Human Genetics**, Auckland, v. 84, p. 210-223, Feb. 2009.

BRUCE, C.A.; DISNEY, R.L. **Probability and Random Process for engineers and scientist**. New York: John Wiley & Sons, 1970. 338 p.

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

CHIB, S.; GREENBERG, E. Understanding the metropolis-hastings algorithm. **The American Statistician**, Washington, v. 49, n. 4, p. 327-335, 1995.

CHIPMAN, H. Prior distributions for Bayesian analysis of screening experiments. **Springer**, New York, p. 235-267, 2004.

DE FONZO, V.; ALUFFI-PENTINI, F.; PARISI, V. Hidden Markov models in bioinformatics. **Current Bioinformatics**, Oak Park, v. 2, p. 46-61, Jan. 2007.

DOERGE, R.W. Mapping and analysis of quantitative trait loci in experimental populations. **Nature Reviews Genetics**, London, v. 3, p. 43-52, 2002.

DU, Q.; CHANG, C.I. Hidden Markov model approach to spectral analysis for hyperspectral imagery. **Optical Engineering**, Chrnivtsi, v. 40, p. 2277-2284, Oct. 2001.

DUTHEIL, J.Y.; GANAPATHY, G.; HOBOLTH, A.; MAILUND, T.; UYENOYAMA, M.K.; SCHIERUP, M.H. Ancestral population genomics: the coalescent hidden Markov model approach. **Genetics**, Bethesda, v. 183, p. 259-274, Sept. 2009.

E SILVA, L.D.C.; ZENG, Z.B. Current progress on statistical methods for mapping quantitative trait loci from inbred line crosses. **Journal of Biopharmaceutical Statistics**, London , v. 20, p. 454-481, 2010.

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

EHLERS, R.S. **Inferência Bayesiana**. Disponível em: <http://www2.icmc.usp.br/ ehlers/bayes/bayes.pdf>. Acesso em: 4 set. 2011.

ENO, D.R. **Noninformative prior bayesian analysis for statistical calibration problems**. 1999. 124p. Thesis (Doctor) - Faculty of the Virginia Polytechnic Institute and State University, Virginia, 1999.

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. Thesis (Philosophy in Statistics Doctor) - University of Wisconsin, Madison, 2001.

GAMERMAN, D.; LOPES, H.F. **Markov Chain Monte Carlo: stochastic simulation for Bayes inference**. London: Chapman & Hall, 2006. 333 p.

GELMAN, A. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). **Bayesian analysis**, Pittsburgh, v. 1, p. 515-534, 2006.

GODSILL, S.J. On the relationship between Markov chain Monte Carlo methods for model uncertainty. **Journal of Computational and Graphical Statistics**, Alexandria, v. 10, p. 230-248, 2001.

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

GRIER, P.P.; CHRISTOPHER I.A.; ERIC B. The quantitative LOD score: test statistic and sample size for exclusion and linkage of quantitative traits in human sibships. **The American Journal of Human Genetics**, London, v. 62, p. 962-968, 1998.

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

HALLAUER, A.R.; CARENA, M.J.; MIRANDA FILHO, J.B. **Quantitative genetics in maize breeding.** 3. ed., New York: Springer, 2011. 663 p.

HOWIE, B.; MARCHINI, J.; STEPHENS, M. Genotype imputation with thousands of genomes. **G3: Genes, Genomes, Genetics**, Oxford, v. 1, p. 457-470, Nov. 2011.

HU, J.; LI, H.; WATERMAN, M.S.; ZHOU, X.J. Integrative missing value estimation for microarray data. **BMC bioinformatics**, London v. 7, p. 449, Oct. 2006.

HUMBURG, P.; BULGER, D.; STONE, G. Parameter estimation for robust HMM analysis of ChIP-chip data, **Bmc Bioinformatics**, London, v. 9, p. 343, Aug. 2008.

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

JIANG, C.; ZENG, Z.B. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. **Genetica**, Netherlands, v. 101, p. 47-58, 1997.

JUANG, B.H.; RABINER, L.R. The segmental K-Means algorithm for estimating parameters of hidden Markov models. **Acoustics, Speech and Signal Processing, IEEE Transactions on**, San Mateo, v. 9, p. 1639-1641. Jan. 1990.

KAO, C.H.; HO, H.A.A score-statistic approach for determining threshold values in QTL mapping. **Frontiers in Bioscience**, Taiwan, p. 2670-2682, Jun. 2012.

KAO, C.H.; ZENG, Z.B. Modeling epistasis of quantitative trait loci using Cockerham's model. **Genetics**, Austin, v. 160, p. 1243-1261, 2002.

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

KARLIN, S.; TAYLOR, H.M. **A second course in stochastic processes**. Gulf Professional Publishing, 1981. 557 p.

KHREICH, W.; GRANGER, E.; MIRI, A.; SABOURIN, R. On the memory complexity of the forward-backward algorithm. **Pattern Recognition Letters**, North Holland, v. 31, p. 91-99, Sept. 2010.

KIM, K.Y.; KIM, B.J.; YI, G.S. Reuse of imputed data in microarray analysis increases imputation efficiency. **BMC bioinformatics**, London, v. 5, n. 1, p. 160, Oct. 2004.

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

LAM, T.; IRMTRAUD, M.M. Efficient algorithms for training the parameters of hidden Markov model using stochastic expectation maximization (EM) training and Viterbi Training. **Algorithms for molecular biology: AMB**, Vancouver, v. 5, p. 1-16, Dec. 2010.

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

LANGE, K. **Mathematical and statistical methods for genetic analysis.** New York: Springer, 2002. 361 p.

LEE, S.H.; VAN DER WERF, J.H. Simultaneous fine mapping of multiple closely linked quantitative trait loci using combined linkage disequilibrium and linkage with a general pedigree. **Genetics**, Austin, v. 173, p. 2329-2337, 2006.

LEMBER, J.; KOLOYDENKO, A. The adjusted Viterbi training for hidden Markov models. **Bernoulli**, The Hague, v. 14, p. 180-206, Mar. 2008.

LI, Z.; SILLANPÄÄ, M.J. Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. **Genetics**, Austin, v. 190, p. 231-249, 2012.

LI, Y.; WILLER, C.; SANNA, M.; ABECASIS, G. Genotype imputation. **Genomics Hum Genet**, Michigan, v. 10, p. 387-406, Sept. 2009.

MANICHAIKUL, A.; MOON, J.Y.; SEN, S.; YANDELL, B.S.; BROMAN, K.W. A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. **Genetics**, Austin, v. 181, p. 1077-1086, 2009.

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

NIELSEN, J.; SAND, A. Algorithms for a parallel implementation of Hidden Markov Models with a small state space. In: IEEE INTERNATIONAL SYMPOSIUM, 2011, Anchorage. **abstracts...** Anchorage: IEEE, p. 452-459.

PEREIRA, R.N. **Modelo hierárquico bayesiano na determinação de associação entre marcadores e QTL em uma população F2**. 2012. 126 p. Tese (Doutorado em Agronomia) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2012.

R: A Language and Environment for Statistical Computing, R Core Team, R Foundation for Statistical Computing, Vienna, 2013. Disponível em: <http://www.R-project.org>. Aceso em: 20 jan. 2014

RABINER, L. R.A tutorial on hidden Markov models and selected applications in speech recognition. **Proceedings of the IEEE**, San Mateo, v. 77, p. 257–286, Feb. 1989.

ROBERTS, A.; MCMILLAN, L.; WANG, W.; PARKER, J.; RUSYN, I.; THREADGILL, D. Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. **Bioinformatics**, Oak Park, v. 23, p. 401-407, 2007.

RUIZ, R.; DEMÉTRIO, C.G.; ASSUNÇÃO, R.M.; LEANDRO, R.A. Modelos hierárquicos bayesianos para estudar a distribuição espacial da infestação da broca do café em nível local. **Revista Colombiana de Estadística**, Bogota, v. 26, p. 1-24, 2003.

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

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

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

SILLANPÄÄ, M.J.; ARJAS, E. Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. **Genetics**, Austin, v. 151, p. 1605-1619, 1998.

SILVA, J.P. **Uma abordagem bayesiana para mapeamento de QTLs utilizando métodos MCMC com saltos reversíveis**. 2006. 80 p. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2006.

SILVA, J.P.; LEANDRO, R.A.A bayesian approach to map QTLs using reversible jump MCMC. **Ciência e Agrotecnologia**, Lavras, v. 33, p. 1061-1070, 2009.

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

TIERNEY, L. Markov chains for exploring posterior distributions. **The Annals of Statistics**, Minnesota, v. 1, p. 1701-1728, 1994.

TOLEDO, E.R. **Mapeamento de QTLs utilizando as abordagens Clássica e Bayesiana**. 2006. 99p. Tese (Mestrado) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, 2006.

VITERBI, A.J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm.. **Information Theory, IEEE Transactions on**, Piscataway, v. 13, p. 260-229, 1967.

VITERBI, A.A personal history of the Viterbi algorithm. **Signal Processing Magazine, IEEE**, Piscataway, v. 23, p. 120-142, Jul. 2006.

WU, R.; CASELLA, G; MA, C. **Statistical genetics of quantitative traits: linkage, maps, and QTL.** New York: Springer-Verlag, 2007. 365 p.

XIANG, Q.; DAI, X.; DENG, Y.; He, C.; WANG, J.; FENG, J.; DAI, Z. Missing value imputation for microarray gene expression data using histone acetylation information. **BMC bioinformatics**, London, v. 9, p. 252, May. 2008

YANDELL, B.S.; MEHTA, T.; BANERJEE, S.; SHRINER, D.; VENKATARAMAN, R.; MOON, J.Y.; NEELY, W.W.; WU, H.; SMITH, R.; YI, N. R/qtlbim: QTL with Bayesian interval mapping in experimental crosses. **Bioinformatics**, London, v. 23, p. 641-643, 2007.

YI, N; XU, S. Bayesian mapping of quantitative trait loci for complex binary traits. **Genetics**, Austin, v. 155, p. 1391-1403, 2000.

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.

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

YU, S.Z.; KOBAYASHI, H. An efficient forward-backward algorithm for an explicit-duration hidden Markov model. **Signal Processing Letters, IEEE**, Piscataway, v. 10, p. 11-14, Jan. 2003.

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

ZHAO, Z.; TIMOFEEV, N.; HARTLEY, S.; CHUI, D.; FUCHAROEN, S.; PERLS, T.; STEINBERG, M.H.; BALDWIN, C.T.; SEBASTIANI, P. Imputation of missing genotypes: an empirical evaluation of IMPUTE. **BMC genetics**, London, v. 9, p. 85, Dec. 2008.

APÊNDICE

APÊNDICE A - Programação no R para imputação dos genótipos dos marcadores

```
# Função para imputação
require(HMM)
mx.imp<-function(qimp)           {
mark.ret<-markers.imput
for(i in 1:400){
mark.ret[i, sample(seq(1:ncol(mark.ret)), size=qimp)]<-"NA"}
mark.ret[mark.ret==1]<-"A"; mark.ret[mark.ret==2]<-"H"; mark.ret[mark.ret==3]<-"B"
return(mark.ret)
                                  }

imputed <- function(qimp, mr, r,e){
hmm1 = initHMM(c("AA","AB","BB"), c("A","H","B","NA"), c(1/4,1/2,1/4),
matrix(c((1-r)^2,r*(1-r),r^2,  2*r*(1-r),r^(2)+(1-r)^2,2*r*(1-r),
r^2,r*(1-r),(1-r)^2),ncol=3, nrow=3),
matrix(c((1-e)^2,e*(1-e),e^2,  2*e*(1-e),e^(2)+(1-e)^2,2*e*(1-e),
e^2,e*(1-e),(1-e)^2, 1,1,1),ncol=4,nrow=3))

# Sequência de observações
probImp1<- vt1<- observation1 <- vector("list", nrow(mr))
for(i in 1:400){
observation1[[i]] = mr[i,]

#Algoritmo VT
vt1[[i]] = viterbiTraining(hmm1,observation1[[i]], 60, pseudoCount=0.0001)

#Probabilidades a posteriori
probImp1[[i]]<- posterior(vt1[[i]]$hmm, observation1[[i]])[,which(mr[i,]=="NA")]
}

#Possiveis genótipos em uma população F2
gen1=c("A", "H", "B")
mmm<- mr

#Calculando as probabilidade para os missing
g<-vector("list", 400)
for(j in 1:400){
for(i in 1:ncol(probImp1[[j]])){
g[[j]][[i]]<- sample(gen1, size=1, prob= probImp1[[j]][,i])
               }
#Substituindo na matrix de missing
mmm[j,][which(mmm[j,]=="NA")]<-g[[j]]
mmm[mmm=="A"]<-1; mmm[mmm=="H"]<-2; mmm[mmm=="B"]<-3
               }
return(mmm)
}

#Proporção de missing
qimp=round(ncol(markers.imput)*c(0.013, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30,
```

```
0.35, 0.40))

nint = 1000 #Quantidade de iterações

mtx.imputed <- mx.imps<- vector("list",nint)
for(j in 1:nint){
for(i in 1:length(qimp)){
mx.imps[[j]][[i]] <- mx.imp(qimp[i])
mtx.imputed[[j]][[i]]<- imputed(qimp[i], mr=mx.imps[[j]][[i]], r=0.10,e=0.01)
} }

#Acurácia da imputação
RMSE<- matrix(c(NA), nrow=nint, ncol=length(qimp),
dimnames =  list(paste("Iteração", 1:nint),
paste("Missing", c("1%","5%","10%","15%","20%","25%","30%","35%","40%"))))

#Coeficiente de correlação de Pearson
R<- matrix(c(NA), nrow=nint, ncol=length(qimp),
dimnames =  list(paste("Pearson", 1:nint),
paste("Missing", c("1%","5%","10%","15%","20%","25%","30%","35%","40%"))))

for(j in 1:nint){
for(i in 1:length(qimp)){

RMSE[j,i]<- sqrt(mean((markers.imput[which(mx.imps[[j]][[i]]=="NA", arr.ind=T)]-
as.numeric(mtx.imputed[[j]][[i]][which(mx.imps[[j]][[i]]=="NA",
arr.ind=T)]))^2) / mean(markers.imput[which(mx.imps[[j]][[i]]=="NA", arr.ind=T)]^2))

R[j,i] <- cor(as.numeric(markers.imput), as.numeric(mtx.imputed[[j]][[i]]))
          }}

plot(c(1,(1:8)*5), apply(2,2,mean), xlab="Missing (%)", axes=F, lwd=3,
col="grey45", ylab="Coeficiente de correlação", cex=1.3, cex.lab=1.2, pch=11)
axis(1, seq(-5,45,5), cex.axis=1.1)
axis(2, seq(0.6,1,0.05), cex.axis=1.1, las=1)

plot(c(1,(1:8)*5), apply(RMSE,2,mean), xlab="Missing (%)", axes=F, lwd=3,
col="gray37", ylab="NRMSE", cex=1.3, cex.lab=1.2, pch=7)
axis(1, seq(-5,45,5), cex.axis=1.1)
axis(2, seq(0.3,0.5,0.01), cex.axis=1.1, las=1)
```

APÊNDICE B - Programação no *software* R para mapeamento de QTL utilizando o pacote *qtlbim*

```
#Entrada dos dados
prodG <- read.cross("gary",  genfile="foo.dat",
                    mapfile="markerpos.txt", phefile="phenotrans.dat",
                    chridfile="chrid.dat", mnamesfile="mnames.txt",
                    pnamesfile=NULL)
prodG <- jittermap(prodG, amount=1e-6)
```

- `gary`: É o formato do arquivo, que poderia ser, por exemplo, csv;
- `genfile=foo.dat`: É a matriz dos genótipos dos marcadores;
- `mapfile=markerpos.txt`: É um vetor com as posições destes marcadores;
- `phefile=pheno.dat`: É um vetor com a(s) característica(s) fenotípica(s);
- `chridfile=chrid.dat`: Aqui será escrito a quantidade que cada marcador se repete (M1, M1,...,M2,...,M117, M117,...);
- `mnamesfile=mnames.txt`: É um vetor contendo os nomes destes marcadores.

As probabilidades condicionais

Sabe-se que no mapeamento de QTL apenas são conhecidas a(s) característica(s) fenotípica(s) e os genótipos dos marcadores. Mas para realizar a análise de QTL, é necessária a informação dos genótipos dos QTL, os quais são obtidos por meio das probabilidades condicionais que são extraídas utilizando os marcadores flanqueadores. Para isso, será executada a função `qb.genoprob`. A `qb.genoprob` a ser executada no programa R será escrita da forma,

```
1 prodG.prob <- qb.genoprob(prodG, map.function="haldane",
2                           step=0.2)
```

- `map.function`=haldane: A função de mapeamento será a de Haldane;
- `step=0.2`: O espaçamento entre dois QTL consecutivos será de 0,2 cM.

Na função `qb.genoprob` são calculadas as localizações dos futuros QTL (`create.map`) e em seguida as frações de recombinação.

O modelo

O modelo considerado neste trabalho contemplará os efeitos aditivos e de dominância.

```
1 qbModel <- qb.model(prodG.prob, epistasis=F, main.nqtl=3, pheno.col=1,
2                     chr.nqtl = rep(2,nchr(prodG)))
```

- `prodG.prob`: Objeto que contém as probabilidades condicionais;
- `epistasis=F` : Indicando que não haverá efeito de interação entre pares de QTL;
- `pheno.col=1`: Na entrada dos dados, havia mais de uma característica fenotípica, pois isso, é necessário indicar em qual coluna se encontra a característica que será analisada. Neste caso, a variável, produção de grãos, está localizada na primeira coluna;
- `main.nqtl=3`: Número de QTL com efeitos principais, neste caso, três QTL. Como não há epistasia, o número máximo de QTL que será aceito pelo modelo será, $\texttt{main.nqtl} + 3 \times \sqrt{\texttt{main.nqtl}}$;
- `chr.nqtl` = rep(2,nchr(prodG)): No máximo dois QTL serão considerados em cada cromossomo.

Preparando os dados

A entrada dos dados se dá pela função `qb.data`. Nela, será especificadas as características que estarão no modelo. No R, ela pode ser escrita da forma,

```
qbData <- qb.data(prodG.prob, pheno.col=1, trait='normal', boxcox = F,
                  fixcov = 0, rancov = 0)
```

- `prodG.prob` : Aqui é especificado o objeto que contém as probabilidades condicionais;
- `trait='normal'`: A Característica fenotípica será modelada por uma distribuição Normal;
- `boxcox = F`: Esta função realizada uma transformação Box Cox na variável resposta.
- `fixcov = 0, rancov = 0`: Como não há covariáveis nem de efeito fixo e nem de efeito aleatório, em ambos os argumentos deverá ser atribuído o valor zero.

Construindo amostra a posteriori

Para executar o algoritmo MCMC com saltos reversíveis, a função responsável por isto será a `qb.mcmc`. Os comandos utilizados neste trabalhos são os seguintes.

```
qb.f2 <- qb.mcmc(prodG.prob, data=qbData, model = qbModel, mydir = '.',
              n.iter=120000, n.thin=50, n.burnin=1000, genoupdate=T,
              seed=3013)
```

- `mydir = '.'`: O objeto `qb` após compilado irá salvar, automaticamente, as amostras a posteriori;
- `n.iter=120000`: Serão realizadas 120.000 iterações;
- `n.thin=40`: Será considerado um espaçamento entre as iterações de tamanho 50;
- `n.burnin=1000`: As 1000 primeiras iterações serão descartadas;
- `genoupdate=T`: Atualizará os genótipos dos QTL em cada iteração.
- `seed=3013`: Especificará a semente para o gerador de números aleatórios.

O objeto `qb.f2` armazenará as amostras a posteriori para os parâmetros do modelo.

Teste de convergência

```
coda.iterdiag = qb.coda(qb.f2, element="iterdiag", variables=c("mean","envvar", "var"))
summary(coda.iterdiag)
plot(coda.iterdiag)
```

Cálculo a posteriori para o número de QTL

```
iterdiag = qb.f2$mcmc.samples$pheno1$iterdiag
posterior = as.numeric(prop.table(table(iterdiag$nqtl)))

barplot(table(iterdiag$nqtl), xlab="Número a Posteriori de QTL", ylab="Frequência",
              ylim=c(0,1150), axes=T, font.lab=9, font=9, las=1, cex.lab=1.5,
              col=c("grey20","grey30","grey40","grey50","grey60","grey70","grey80"))
```

Cálculo do Fator de Bayes

```
nqtl.number = as.numeric(levels(factor(iterdiag$nqtl)))

#Prior
prior = dpois(nqtl.number, 3)

#Fator de Bayes para nqtl
posterior/prior
FB=NULL
for(i in 1:length(prior)) {
        FB[i]=(posterior[i]/prior[i])*(prior[1]/posterior[1])
                              }

plot(nqtl.number, FB, cex=2, pch=19, axes=F, xlab="", ylab="", xlim=c(2,8),
ylim=c(0,450))
lines(nqtl.number, FB, lwd=3, col="gray55")
axis(1, seq(-1,9), las=1, font=9)
axis(2, seq(-75,525,75), las=1, font=9)
title(xlab="Número de QTL", font.lab=9, cex.lab=1.5, ylab="Fator de Bayes")
```

Calculo a posteriori e Fator de Bayes para localização dos QTL

```
tempBF <- qb.scanone(qb.f2, type="BF", epistasis = FALSE)
tempPOST <- qb.scanone(qb.f2, type="posterior", epistasis = FALSE)
plot(tempBF, main="", xlab="", axes=F, ylab="", scan="main", col="gray55")
axis(2, seq(0,50,5), las=1, font=9)
title(xlab="Cromossomo", font.lab=9, cex.lab=1.5, ylab="Fator de Bayes")

plot(tempPOST, main="", xlab="", axes=F, ylab="", scan="main", col="gray55")
axis(2, seq(0,0.016,0.002), las=1, font=9)
title(xlab="Cromossomo", font.lab=9, cex.lab=1.5, ylab="Posteriori")
```

Arquitetura genética

```
best = qb.best(qb.f2)
summary(best)
plot(best,main="")
```

Calculando as estimativas dos efeitos genéticos

```
arch = qb.arch(best)
f2.sub = subset(dados.f2,chr=arch$qtl$chr)
f2.sub.prob = sim.geno(f2.sub, n.draws=16, step=0.2, error=0.01)
qtl = makeqtl(f2.sub.prob, chr=as.character(arch$qtl$chr), pos=arch$qtl$pos)
```

```
5
6  f2.step = step.fitqtl(f2.sub.prob, qtl, pheno.col=1, arch)
7  summary(f2.step$fit)
8
9  mod.int <-
10 fitqtl(f2.sub.prob,qtl=qtl,get.ests=T,formula=y~(Q1+Q2+Q3+Q4+Q5),
11 pheno.col=1, method='imp')
12 summary(mod.int)
```

ANEXO

ANEXO A - Figuras para diagnóstico da cadeia

Na Figura 13, de cima para baixo, estão as representações gráficas para validação da convergência da cadeia para os parâmetros do modelo: média geral, variância residual e variância genética, respectivamente. Nesta figura, no lado esquerdo, estão os traços para verificação de convergência da cadeia diante das 120 mil iterações e do lado direto as respectivas densidades.

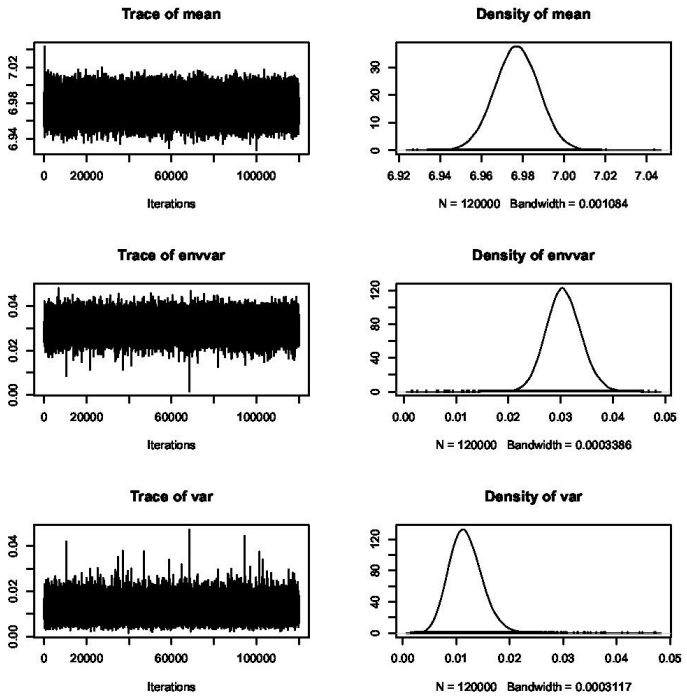


Figura 13 - Diagnóstico para convergência da cadeia