ERNANDES GUEDES MOURA

APLICAÇÃO DE MODELOS FUNCIONAIS NA SELEÇÃO GENÔMICA AMPLA

LAVRAS – MG

2017

ERNANDES GUEDES MOURA

APLICAÇÃO DE MODELOS FUNCIONAIS NA SELEÇÃO GENÔMICA AMPLA

Dissertação apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

Prof. Dr. Marcio Balestre
Orientador

LAVRAS – MG

2017

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Moura, Ernandes Guedes.
Aplicação de Modelos Funcionais na Seleção Genômica Ampla / Ernandes Guedes Moura. edicao – Lavras : UFLA, 2017.
54 p. : il.

Dissertação (mestrado acadêmico) - Universidade Federal de Lavras, 2017. Bibliografia.–Universidade Federal de Lavras, 2017.
Orientador: Prof. Dr. Marcio Balestre.
Bibliografia.

1. Modelos Bayesianos. 2. Seleção Genômica. 3. Regressão Bayesiana. I. Balestre, Marcio . . II. Título.

ERNANDES GUEDES MOURA

APLICAÇÃO DE MODELOS FUNCIONAIS NA SELEÇÃO GENÔMICA AMPLA

Dissertação apresentado à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para obtenção do título de Mestre.

APROVADA em 23 de janeiro de 2017.

Dr. Júlio Sílvio de Sousa Bueno Filho UFLA

Dr. Fabyano Fonseca e Silva UFV

Dr. Marcio Balestre

Orientador

LAVRAS - MG

2016

AGRADECIMENTOS

Primeiramente a Deus, por todas as bênçãos que tem me proporcionado e concedido mais essa dádiva.

Aos meus pais, Francisco de Assis Guedes Lima e Francisca Maria de Moura, e aos meus irmãos Francis James Guedes Moura e Juliana Guedes Moura, e o meu tio Pe. Cícero de Moura Filho, pelo apoio incondicional e compreensão da importância desse projeto.

Aos meus avós: Cícero Maroto de Moura, Maria José de Moura e Maria das Dores de Lima (in memoriam), pelo carinho dedicado.

Aos meus familiares, pelo carinho e preocupação e, principalmente, por compreenderem minha ausência em diversos momentos.

À Ana Teresa Pereira da Silva, pelo companheirismo e apoio em todos os momentos.

À Universidade Federal de Lavras, em especial ao Departamento de Estatística-DES, pela oportunidade.

Ao CNPq- Conselho Nacional de Desenvolvimento Científico e Tecnológico, pela concessão da bolsa de mestrado.

Ao Instituto Federal do Piauí, pela oportunidade proporcionada.

Ao professor Marcio Balestre, pela orientação, apoio e paciência.

Aos membros da banca, Júlio Silvio de Sousa Bueno Filho, José Airton Rodrigues Nunes e Fabyano Fonseca e Silva, pela disponibilidade e contribuição neste trabalho.

Aos diversos professores do Departamento de Estatística da UFLA que fizeram parte de minha formação, em especial os professores: Joel Augusto Muniz e Renato Ribeiro de Lima, pelo apoio e incentivo.

A todos os funcionários do DEX/UFLA.

Aos meus amigos e colegas de mestrado e doutorado, em especial Sérgio Domingos Simão, Carlos Pereira da silva, Michele Barbosa, Taís Alvarenga, Indalécio Cunha, Kelly Lima e Elias Medeiros, pela colaboração, e a Andrezza Kéllen Alves Pamplona, pela grande contribuição neste trabalho.

A todos os docentes do IFPI Campus Uruçuí, em especial os professores: Miguel Antônio Rodrigues e Gabriel dos Santos Pintos, pelo apoio e incentivo.

A todos os colegas técnicos administrativos do IFPI Campus Uruçuí, em especial Jéferson Peixoto e Lucivânia F. Miranda, pelo apoio e pela preocupação demonstrada.

Aos meus amigos de longe, em especial, Francisco José de Abreu, Cristiano Rodrigues, Dayse Batista dos Santos, Edylberto Lima, Mauro Sérgio Brasil e Almir Pereira da Silva Neto, pelo o incentivo e preocupação demonstrada.

Aos amigos do departamento DCS/UFLA: Jodean Alves da Silva, José Ferreira Lustosa Filho e Rodrigo Fonseca da Silva, pelo apoio sempre que necessário.

A todos que, direta ou indiretamente, me apoiaram nesta jornada.

RESUMO

Com o surgimento de marcadores de alta densidade SNPs, ao mesmo tempo em que surge um grande avanço no que diz respeito ao aumento da capacidade preditiva dos modelos, agravaram-se os problemas de multicolinearidade e alta dimensionalidade dos modelos na seleção genômica, gerando desafios estatísticos e computacionais. Objetivou-se neste trabalho propor um método e verificar sua eficiência na seleção genômica usando modelos funcionais. Dessa forma, propôs-se que os efeitos de um loco genético é função de sua respectiva localização no genoma. Para verificar a palpabilidade do modelo, simulou-se 300 indivíduos a três populações F2, conforme três herdabilidades (0,2; 0,5 e 0,8), em um total de 12150 marcadores SNPs, distribuídos em dez grupos de ligação. O modelo proposto no presente estudo obteve destaque nos cenários oligogênico e poligênico, e pode ser recomendado a estudos posteriores a dados reais e com diversas arquiteturas genéticas para conclusões mais consistentes.

Palavras-chave: Modelos Bayesianos. Seleção Genômica. Regressão Bayesiana

ABSTRACT

Upon the emergence of high-density SNP markers, along with great advancements related to the increase in the predictive ability of models, there have been problems of multicolinearity and high dimensionality of models in genomic selections, causing many statistic and computational challenges. This work aimed at proposing a method and checking its efficiency in genomic selections with functional models. Thus, we suggest that the effects of a genetic locum is a function of its respective genomic position. To verify the suitability of such models, we simulated 300 individuals in three populations F2, according to three heritabilities (0.2; 0.5 and 0.8) in a total of 12150 SNP markers distributed into ten bond groups. The model proposed in this study was successful with oligogenic and polygenic scenarios, therefore, further research with real data and several genetic frames is recommended for more consistent conclusions.

Keywords: Bayesian Models. Genomic Selection. Bayesian regression.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Há diversos mecanismos moleculares que causam a variabilidade e a diversidade genô mica das inúmeras características fenotípicas existentes na natureza. O maior interesse de melhoristas é garantir médias elevadas para as características em estudo sem exaurir a variabilidade genética (BERNARDO, 2010). Assim, esses profissionais realizam vários estudos a fim de entender as leis que regem os mecanismos genéticos responsáveis por tais características. No entanto, essa tarefa pode ser árdua, haja vista que a maioria desses caracteres quantitativos são governados por muitos genes de pequenos efeitos e poucos com grandes efeitos (HU; WANG; XU, 2012).

Na literatura há vários modelos estatísticos utilizados para estimar os Quantitative Trait Loci (QTLs) (covariáveis) em estudo no genoma, a partir de caracteres quantitativos (fenótipos – variável resposta), tais como RR-BLUP, Bayes A, Bayes B, etc., visando à seleção genômica. Esses modelos diferem em eficiência, capacidade preditiva, custo computacional e flexibilidade em lidar com diferentes arquiteturas genéticas. Assim, pela dinamicidade em que ocorre a ciência genômica, frequentemente são criados novos modelos, com a finalidade de melhorar os aspectos citados acima.

Com o uso de marcadores moleculares, a quantidade de informação disponível tem sido cada vez maior e, com isso, passaram a surgir problemas tais como a multicolinearidade e a alta dimensionalidade dos modelos (número de marcadores é maior que o número de indivíduos), o que gera alto custo computacional, tornando-se necessário estabelecer abordagens estatísticas para lidar com isto.

Assim, para lidar com tais problemas tem sido frequente o uso de modelos mistos baseados em regressão aleatória do tipo BLUP e, mais ainda, com abordagem bayesiana, por meio da Genome Wide Selection (GWS) proposta por Meuwissen, Haves e Goodard (2001), que são capazes de manusear situações em que o número de efeitos é maior que o número de observações (XU, 2003). E foi com essas preocupações que Hu, Wang e Xu (2012) desenvolveram o modelo denominado genoma contínuo, em que os autores dividiram o genoma em *bin* (pequenos intervalos do genoma) e analisaram seus efeitos em vez de efeitos das marcas, havendo com isso redução na dimensão do modelo, que por sua vez possibilita abordagem de técnicas mais simples já consagradas na literatura. A análise *bin* (HU; WANG; XU, 2012) mostrou melhoria significativa na acurácia preditiva do modelo em comparação aos métodos concorrentes: eBayes, G-Blup, Bayes B1, Bayes B2 e Lasso.

É evidente a necessidade de estratégias estatísticas e computacionais que auxiliem os profissionais que atuam com seleção genômica na identificação de alguma relação funcional entre os efeitos das marcas e da predição fenotípica. Nesse sentido, o presente estudo tem como objetivo propor outra abordagem ao método genoma contínuo, proposto por Hu, Wang e Xu (2012) e abordado por Xu (2013) com a finalidade de tratar os problemas de seleção genômica ampla sob a ótica de modelos funcionais, em que os autores supõem que o efeito genético de um loco individual é uma função da posição no genoma (quantidade contínua). Para isso, simularam-se três populações F2 com 300 indivíduos cada e de acordo com três níveis de herdabilidades (0,2; 0,5 e 0,8).

2 REFERENCIAL TEÓRICO

A crescente busca dos geneticistas e melhoristas da predição fenotípica por meio dos genótipos dos indivíduos (animais ou plantas) tem trazido vários avanços na ciência genômica. Nesse sentido, pode-se destacar o surgimento de marcadores moleculares que permitem quantificar a variabilidade genética existente no DNA.

2.1 Mapeamento de QTLs

Com o uso dos marcadores moleculares foi possível o mapeamento do genoma de animais e plantas e a sua associação com o valor fenotípico. Assim, saturando o genoma com marcadores, pode-se investigar, além dos efeitos dos principais genes, também suas respectivas localizações, sendo este processo denominado como mapeamento de QTLs (*Quantitative Trait Loci*) (LANDER; BOTSTEIN, 1989).

Existem, na literatura, várias abordagens para mapeamento de QTLs que vão desde a mais simples, em que a associação é realizada em um marcador por vez, sendo ligado ao caráter de interesse (EDWARDS; STUBER; WEDEL, 1987), até abordagens em que não se utilizam mapas de ligação e, pela saturação do genoma, cada marcador molecular é assumido como provável QTL (XU, 2003). Dessa forma, o principal interesse no estudo de QTLs é avaliar e identificar a posição do loco e estimar seus efeitos genéticos, em que esses podem ser aditivos, de dominância e outros efeitos presentes no modelo proposto (TOLEDO et al., 2008).

Nesse sentido, Singh e Singh (2015) ressaltam que, basicamente, os métodos de mapeamento de QTLs são utilizados para resolução de três grandes questões:

i. Os genótipos de QTLs de diferentes indivíduos não são observados e precisam ser estimados via marcadores moleculares que apresentam alto desequilíbrio de ligação.

ii. Uma vez que há, potencialmente, milhares de possíveis locos em todo o genoma, um modelo genético adequado deve ser selecionado para análise de QTLs, pois há um grande número de modelos possíveis.

iii. Os locos localizados no mesmo cromossomo são correlacionados e, como consequência, difíceis de separar.

De acordo com Singh e Singh (2015), a análise de QTLs vem sendo, e continua a ser, uma área de intensa investigação, uma vez que representa uma variedade de questões desafia-

doras que precisam ser resolvidas para a obtenção de resultados confiáveis e reprodutíveis por outros pesquisadores. Entretanto, essa metodologia só pode ser aplicada em populações estruturadas, com segregação conhecida ou com pedigrees complexos, pois quando se utilizam poucos marcadores pode ocorrer estimação errônea do efeito e da posição do QTL, podendo haver falsos positivos

2.2 Modelos de seleção genômica ampla (GWS)

Os primeiros estudos sobre GWS (*Genome-Wide Selection*) foram realizados por Meuwissen, Haves e Goodard (2001) e, essa metodologia tem sido muito utilizada no melhoramento genético. Essa técnica consiste em saturar o DNA com marcadores moleculares, predizer seus efeitos e, posteriormente, predizer adequadamente o valor genético genômico dos indivíduos da população em estudo, objetivando ganho na seleção para as características de interesse. Com isso, há explicação de grande parte da variação genética de um caráter quantitativo (MEUWISSEN; HAVES; GOODARD, 2001). Dessa forma, o propósito da GWS consiste em desenvolver modelos que obtenham melhor predição, para que o melhorista faça seleção com maior confiabilidade.

Nesse contexto, Meuwissen, Haves e Goodard (2001) desenvolveram o método RR-BLUP (*Random Regression Best Linear Unbiased Prediction*), que estima os efeitos de todas as marcas, simultaneamente, sendo considerados efeitos aleatórios com variâncias constantes (modelo homocedástico), isto é, cada SNP (Single Nucleotide Polymorphism) contribui igualmente para variância total dos SNPs . No entanto, a seguinte questão surge: será que, de fato, as marcas têm mesmas variâncias ? Provavelmente esta suposição seja muito forte, haja vista que pode haver marcas em regiões não associadas ao caráter de interesse e outras que se sobressaem. Entretanto, os autores supracitados propuseram também uma abordagem bayesiana, denominadas Bayes A e Bayes B.

Os métodos bayesianos supõem que alguns SNPs têm grande contribuição para variância e alguns SNPs têm uma contribuição pequena ou nula (VAN DEN BERG et al., 2015). Dessa maneira, automaticamente, assumem variâncias heterogêneas entre as marcas. Nesta metodologia, muitos efeitos de marcadores são assumidos como zero, *a priori*, reduzindo o número de efeitos a serem estimados, o que permite o enfoque em regiões do genoma em que realmente existem QTLs (MEUWISSEN; HAVES; GOODARD, 2001). Os autores supracitados detectaram superioridade teórica do método Bayes B em relação ao RR-BLUP. Entretanto, estudos

mais recentes, como o de Daetwyler et al. (2010) e de Van Den Berg et al. (2015) mostraram que essa superioridade é relativa, a depender da estrutura genética de cada situação em estudo. Esses autores verificaram que à medida que o número de marcadores moleculares "tendem ao infinito", isto é, no cenário infinitesimal, o modelo GBLUP (método RR-BLUP com efeito de marcas integrado para as componentes de variância) supera os modelos Bayesianos.

Xu (2013) usou alta densidade de marcadores SNP para inferir pontos de interrupção de recombinação baseados no desequilíbrio (LD) e convertê-los em dados *bin*. Neste caso, todos os marcadores dentro de um *bin* têm o mesmo padrão de segregação e cada *bin* é considerado um novo marcador e, com isso, o número *bins* pode ser substancialmente menor que o número original de marcadores. Nessa abordagem, o autor assumiu que os efeitos das marcas podem ser pensados como uma função contínua da posição delas.

Dessa forma, os dados *bin* são considerados novos dados genômicos. Hu, Wang e Xu (2012) e Xu (2013) utilizaram modelos funcionais para mapeamento genético e seleção genômica, baseando-se em dados *bin*. Eles verificaram que, com essa abordagem, houve menor erro quadrático médio (EQM) e maior coeficiente de determinação (R^2) em relação aos cinco métodos concorrentes (eBayes, G-Blup, Bayes B1, Bayes B2 e Lasso). O desenvolvimento de métodos estatísticos para análise de dados *bin* pode representar uma nova direção na GWS.

2.3 Modelos funcionais

O uso de regressão onde tanto a resposta como a covariável são funcionais tornou-se cada vez mais comum em muitos campos científicos (medicina, finanças, agricultura, entre outras). Esse tipo de regressão é chamado de regressão de "função na função" (modelos funcionais com resposta e covariáveis funcionais) e seu principal objetivo é investigar a associação entre resposta e covariáveis funcionais de previsão (KIM; MAITY; STAICU, 2015).

Outra aplicação de modelos funcionais é em problemas estatísticos onde se têm covariável funcional e variável resposta escalar. Essa metodologia permite que muitos problemas desafiadores sejam tratados. De acordo com Cardot e Sarda (2005) muitas áreas de investigação têm necessidade de lidar com dados funcionais.

Baseado no descrito acima, dado um vetor de observações $\mathbf{y} = y_1, y_2, \ldots, y_n$ conjunto de variáveis funcionais observadas $w_1(x), w_2(x), \ldots, w_n(x)$, o modelo funcional que descreve tal variável resposta é:

$$\boldsymbol{y}_i = \alpha + \int \boldsymbol{\beta}(x) w_i(x) dx + \boldsymbol{\varepsilon}_i \tag{2.1}$$

em que $\boldsymbol{\beta}(x)$ tem o papel de ponderar o quanto cada ponto $w_i(x)$ contribui para a integral. Da mesma forma como acontece em regressão linear tradicional, $\boldsymbol{\beta}(x)$ determina o efeito de $w_i(x)$ em $\boldsymbol{y}_i$; assim sendo, alterações em $w_i(x)$ não têm efeito sobre y_i quando $\boldsymbol{\beta}(x) = 0$ (JAMES; WANG; ZHU, 2009).

Utilizando o modelo 2.1, Hu, Wang e Xu (2012) desenvolveram um modelo infinitesimal baseado em marcadores para análise de QTL. Isto porque, ao contrário do modelo infinitesimal tradicional, os autores propõem que o efeito genético de um loco individual é uma função da posição no genoma (quantidade contínua). Este modelo está descrito na próxima seção.

2.4 Modelo genoma contínuo

Hu, Wang e Xu (2012) afirmam que, na era do genoma, o número de marcadores SNPs podem chegar a alguns milhões. Associar cada marcador a um QTL tem-se, praticamente, um modelo infinitesimal, o que se torna difícil de lidar. Diante disso, estes autores introduziram um novo modelo que denominaram genoma contínuo. Nele, substitui-se o somatório de termos infinitos por uma integral e, em seguida, utiliza-se uma abordagem de integração numérica para resolvê-la. A integração numérica pode ser resolvida através da divisão de todo o genoma em muitos pequenos intervalos (também chamados *bin*).

Assim, ao invés de estimar os efeitos de cada marcador, individualmente, estimam-se os efeitos dos *bin* (cada *bin* representa um novo marcador). Cada *bin* pode conter muitos marcadores e o efeito de um *bin* representa os efeitos totais de todos os marcadores dentro desse *bin* (HU; WANG; XU, 2012). Nesse sentido, Xu (2013) ressalta a possibilidade de mapeamento de QTLs utilizando dados *bin* e afirma, ainda, que tal mapeamento é muito mais fácil do que utilizar os marcadores originais, já que o modelo tem dimensão reduzida.

Antes de mostrar o modelo genoma contínuo, vamos entender o modelo infinitesimal. Seja y_j o valor fenotípico observado para o indivíduo $j$numa população de tamanho n. O modelo linear para a análise de regressão clássica é:

$$\boldsymbol{y}_j = \beta + \sum_{k=1}^{p} \boldsymbol{Z}_{jk}\gamma_k + \varepsilon_j, \qquad j = 1, \ldots, n \tag{2.2}$$

sendo β o intercepto, γ_k o efeito do loco k, $\boldsymbol{Z}_{jk}$ uma variável indicadora genotípica conhecida para indivíduo j no loco k, e é o erro j com uma variância desconhecida σ^2. A variável indica-

dora genotípica $\boldsymbol{Z}_{jk}$ para o loco é definida como:

$$\boldsymbol{Z}_{jk} = \begin{cases} +1, & \text{para} A_1A_1 \\ 0, & \text{para } A_1A_2 \\ -1, & \text{para } A_2A_2 \end{cases}$$

Em que A_1A_1, A_1A_2 e A_2A_2 são os três genótipos do loco k. Perceba que o símbolo p representa o número de loco incluídos no modelo e que quando $p \longrightarrow \infty$ o modelo se torna:

$$\boldsymbol{y}_j = \beta + \sum_{k=1}^{\infty} \boldsymbol{Z}_{jk}\gamma_k + \varepsilon_j, \qquad j = 1, \ldots, n \tag{2.3}$$

O modelo supracitado é o modelo infinitesimal de características quantitativas. O coeficiente de regressão γ_k, de acordo com Hu, Wang e Xu (2012), não pode ser estimado por que o modelo tem um tamanho infinito e é mal condicionado (alta multicolinearidade). Finalmente, substitui-se k pela localização correspondente do loco no genoma, denotado porλ, o qual é contínuo e varia de 0 a L, sendo L o tamanho do genoma. Assim, o modelo infinitesimal pode ser substituído por:

$$\boldsymbol{y}_j = \mu + \int_0^L \boldsymbol{Z}_j(\lambda)\gamma(\lambda)d\lambda + \varepsilon_j, \qquad j = 1, \ldots, n \tag{2.4}$$

em que $\boldsymbol{Z}_j(\lambda)$ é conhecido pelo genoma saturado com marcadores , μ representa o intercepto e $\gamma(\lambda)$é o efeito genético expresso como uma função desconhecida do loco no genoma. Dessa forma, o objetivo é obter $\gamma(\lambda)$ a partir dos dados.

O modelo citado em 2.4 foi proposto por Hu, Wang e Xu (2012) e também foi abordado por Xu (2013), como segue:

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \int_0^L \boldsymbol{Z}_j(\lambda)\gamma(\lambda)d\lambda + \varepsilon_j = \boldsymbol{X}_j\boldsymbol{\beta} + g_j(L) \tag{2.5}$$

em que λ é uma posição no genoma expressa como uma quantidade contínua; $\boldsymbol{Z}_j(\lambda)$ é uma variável indicadora; $\gamma(\lambda)$ é o efeito genético na posição λ expresso como uma função de λ; $\boldsymbol{X}_j$ e $\boldsymbol{\beta}$ representam algumas covariáveis e seus efeitos, que devem ser incluídos no modelo para reduzir o erro; e $\varepsilon_j \sim \mathbb{N}(0, \sigma^2)$ é o erro com uma variância desconhecida σ^2.

No modelo 2.5, $g_j(L)$ representa o valor genômico para o indivíduo j. Note, ainda, que esse modelo é um tipo de modelo linear funcional em que a variável resposta é um escalar e a covariável é uma função (CARDOT; SARDA, 2005; CARDOT; FERRATY; SARDA, 2003; MULLER; STADTMULLER, 2005).

Como a função $\gamma(\lambda)$ é desconhecida, a integral em 2.5 não é explícita, havendo necessidade de utilizar integração numérica. Assim, Xu (2013) utilizou a seguinte aproximação:

$$\boldsymbol{y}_j \approx \boldsymbol{X}_j\boldsymbol{\beta} + \sum_{k=1}^{m} \bar{Z}_j(\lambda_k)\bar{\gamma}(\lambda_k)\Delta_k + \boldsymbol{\varepsilon}_j \tag{2.6}$$

com

$$\bar{Z}_j(\lambda_k) = \Delta_k^{-1} \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} \boldsymbol{Z}_j(\lambda) d\lambda \tag{2.7}$$

$$\Delta_k = \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} d\lambda \tag{2.8}$$

$$\bar{\gamma}(\lambda_k)\delta_k = \Delta_k^{-1} \int_{\lambda_k - \frac{1}{2}\Delta_k}^{\lambda_k + \frac{1}{2}\Delta_k} \gamma(\lambda) d\lambda \tag{2.9}$$

em que m é número *bin*, $\bar{Z}_j(\lambda_k)$ a média de $\boldsymbol{Z}_j$ para todos os locos dentro do *bin* k, $\bar{\gamma}(\lambda_k)$ o efeito médio de todos os locos dentro do *bin* k e, Δ_k o tamanho do *bin* k. Assim, tomando$Z_{ij} = \bar{Z}_j(\lambda_k)$ e $\gamma_k = \bar{\gamma}(\lambda_k)\Delta_k$, o modelo 2.6 fica:

$$\boldsymbol{y}_j = \boldsymbol{X}_j\boldsymbol{\beta} + \sum_{k=1}^{m} \boldsymbol{Z}_{jk}\gamma_k + \boldsymbol{\varepsilon}_j \tag{2.10}$$

O efeito estimado $\bar{\gamma}_{\lambda_k}$ para o k-ésimo *bin* pode ser testado e, segundo Xu (2013) a estatística teste de Wald é o mais conveniente para testar a significância desses efeitos, que é equivalente ao teste F quando o grau de liberdade do numerador é um.

É importante ressaltar que o modelo supracitado proposto por Xu (2013) contorna um problema constante em genômica que é a alta dimensionalidade do modelo infinitesimal, transformando-o em um modelo de dimensão finita, podendo, assim, ser tratado com técnicas já consagradas na literatura.

Um questionamento que se faz em relação à forma como o modelo foi tratado pelos autores é: pelo fato de considerar a média do *bin*, assume-se pesos iguais para cada efeito dentro desse *bin*, com isso, houve necessidade de abordar métodos adaptativos (alternativos) para lidar com situações em que não se tem alto desequilíbrio de ligação, e ou não tenha efeito homogêneo de marcadores em cada janela (*bin*) do genoma. Outra possibilidade, e que se espera ser mais adequada, é atribuir pesos dados pela frequência de "visitas" de cada marcador dentro de um *bin*. Neste caso, o modelo genoma contínuo pode lidar também com populações com baixo desequilíbrio de ligação, sem precisar usar métodos alternativos para esse fim.

3 MATERIAL E MÉTODOS

3.1 Dados simulados

Utilizando o *software* QGenes (JOEHANES; NELSON, 2008), foram simulados dez grupos de ligação com tamanho de 120 cM cada e distância média de 0.001cM no genoma, totalizando 12150 marcadores SNPs,. Os QTLs (com efeitos simulados a partir de uma distribuição normal) foram escolhidos, aleatoriamente, de acordo com os seguintes cenários: para representar um modelo oligogênico, considerou-se 12 QTLs dentre os marcadores simulados; para representar um modelo poligênico, considerou-se 120 QTLs; e para um modelo 600 QTLs. Para cada configuração acima, três populações F_2 foram simuladas, com 300 indivíduos cada, de acordo com três níveis de herdabilidade: 0,2; 0,5 e 0,8.

Em todos os cenários, foram adotados as cinco configurações *bins* de acordo com o número de marcadores por *bin* (Tabela 3.1). Cada configuração foi assim calculada:

$$N \times q = m$$

em que N é número de SNPs, q uma proporção do número de SNPs e m total de *bins*. Contudo, o número de marcas por *bin* é:

$$T = \frac{N}{m} \longrightarrow \frac{N}{N \times q} = \frac{1}{q}$$

Dessa forma, para obter a proporção (q) citada acima, basta fixar o número de marcas por *bin* (T).

Tabela 3.1 – Número *bins* de acordo com o número de marcadores (SNPs) por *bin*.

Proporção *bins* (q)	N^oSNPs/*bin* (T)	Número *bins* (m)
0,1	10	1215
0,05	20	607
0,01	100	121
$0,006\bar{6}$	150	81
0,005	200	61

Fonte: Dados do autor (2016)

3.2 Implementação da análise

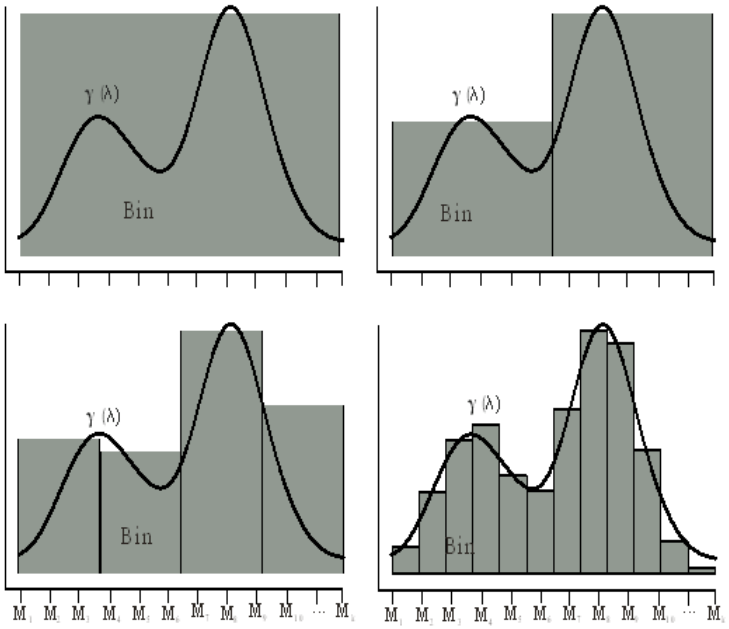
O algoritmo para o amostrador de Gibbs e Metrópolis-Hastings foi implementado utilizando-se o *software* R (R CORE TEAM, 2015). Considerou-se, nas análises dos dados simulados para todos os cenários, um número fixo de 5000 iterações. Para o ajuste do modelo aos dados simulados, foram descartadas as 1000 primeiras iterações para retirar o efeito dos valores iniciais.

3.3 Modelo

Neste trabalho, a proposta é utilizar o modelo proposto por Hu, Wang e Xu (2012), porém com uma abordagem diferente para a integração numérica, atribuindo pesos às marcas de forma individual dados pela probabilidade de ocorrência de cada efeito dentro de cada janela. É importante ressaltar que o modelo proposto nesse estudo, diferentemente do modelo proposto por Hu, Wang e Xu (2012), não é um modelo de análise *bin*; esses serão utilizados somente como estratégia de integração com o propósito de aumentar a chance de encontrar máximos locais e globais.

Em um genoma saturado, devido ao alto desequilíbrio de ligação entre os marcadores, não há necessidade de amostrar regiões próximas a um marcador amostrado M_m. Com isso, é possível utilizar a posição dos marcadores ao invés de gerar pseudo-marcadores. Dessa forma, não se observam todos os marcadores, mas sim, os marcadores (M) dados às posições (λ). O espaço das posições (λ) é contínuo, porém pelo fato de não serem conhecidas todas as posições λ e como existe grande quantidade de marcadores (com suas respectivas posições), amostram-se tais posições de forma discreta, em que cada posição amostrada ($\lambda_m = M_m$) terá um correspondente $\gamma(\lambda_m)$, conforme ilustrado na Figura 3.1.

Figura 3.1 – Cromossomo dividido em *bins* ($\lambda_m = M_m$).



A Figura 3.1 ilustra um cromossomo dividido em quatro configurações de *bins* e uma curva hipotética representando os efeitos dos marcadores. O modelo proposto nesse trabalho pode ser facilmente visualizado por essa ilustração: na primeira configuração, o cromossomo todo foi considerado um *bin* e, neste caso, espera-se que o processo MCMC convirja rápido através do algoritmo Metrópolis-Hastings (HASTINGS, 1970; METROPOLIS et al., 1953). Em que $\gamma(\lambda)$ é função contínua desconhecida em função de λ também desconhecido, porém para essa situação corre-se o risco de estacionariedade em um máximo local; na segunda configuração, o cromossomo foi dividido em dois *bin*, aumentando a chance de encontrar outros pontos de máximos e esse processo será tanto mais preciso quanto maior for a quantidade *bin* em que os cromossomos foram divididos. Todavia, segundo Hu, Wang e Xu (2012) um *bin* deve ser suficientemente grande para alcançar alta resolução.

Além disso, não se pode esquecer o aspecto computacional, sendo que o custo deste possui uma relação inversamente proporcional ao número *bin*, isto é, quanto menor for número *bin*, mais rápido será o processo. Diante disso, há necessidade de uma "conciliação" entre a acurácia preditiva do modelo e o tempo de análise; por isso, foram adotadas diferentes configurações *bin*, descritas em 3.1. Espera-se que mesmo quando adotados poucos *bin*, haja boa acurácia preditiva devido ao desequilíbrio de ligação (correlação entre as marcas).

Logo, seja $\boldsymbol{y}_i$ o valor fenotípico do indivíduo i, para $i = 1, \ldots, n$. O modelo adotado para y_i, em um cromossomo, é:

$$\boldsymbol{y}_i = \mu + \int_0^L \boldsymbol{Z}_i(\lambda)\gamma(\lambda)d\lambda + \varepsilon_i \tag{3.1}$$

sendo λ a posição no cromossomo expressa como uma quantidade contínua, L o tamanho do cromossomo, μ a média geral, $\gamma(\lambda)$ o efeito aditivo do marcador na posição λ (expresso como uma função desconhecida de λ), ε_i é o erro para o indivíduo i, assumindo ter distribuição $\mathbb{N}(0, \sigma^2)$ e $\boldsymbol{Z}_i(\lambda)$ o genótipo do marcador na posição λ para o indivíduo i definido como:

$$\boldsymbol{Z}_{ik} = \begin{cases} 2, & \text{para homozigoto dominante} \\ 1, & \text{para heterozigoto} \\ 0, & \text{para homozigoto recessivo} \end{cases}$$

Assim, se tivermos C cromossomos no genoma, o modelo 3.1 fica:

$$\boldsymbol{y}_i = \mu + \sum_{j=1}^{C} \int_0^L \boldsymbol{Z}_i(\lambda)\gamma(\lambda)d\lambda + \varepsilon_i \quad j = 1, \ldots, n \tag{3.2}$$

Pelo fato de a função $\gamma(\lambda)$ ser desconhecida, a integral em (3.2) não é explícita e, então, uma forma de integração numérica é necessária para calculá-la. Existem algumas possibilidades de integração numérica para encarar esse problema e, nesse estudo, foi adotada a estratégia de divisão dos cromossomos em *bins*. Por isso, daqui em diante, pensaremos no cromossomo particionado em *bins*. Para cada configuração de *bins* será realizada uma análise e, com isso, serão atualizadas as condicionais a posteriori para cada cenário descrito em 3.1.

Para a análise bayesiana, as variáveis observáveis são os valores fenotípicos ($\boldsymbol{y}$) e os genótipos dos marcadores ($\boldsymbol{Z}_i$). As variáveis não observáveis são os coeficientes de regressão (μ e γ), as variâncias (σ_e^2 e σ_λ^2) e as posições λ_{kj}, porém essas serão tomadas como as posições dos marcadores, dadas as posições λ_{kj}. As distribuições *a priori* escolhidas para a variância residual, a variância do marcador e para a média, neste trabalho, foram não informativas (*prioris de Jeffreys*):

$$p(\mu) \propto 1 \qquad p(\sigma_e^2) \propto \frac{1}{\sigma_e^2} \tag{3.3}$$

$$p(\lambda_{kj}) \propto \mathbb{N}(0, \sigma_{\lambda_{kj}}^2) \qquad p(\sigma_{\lambda_{kj}}^2) \propto \chi_{esc}^{-1}(\nu, S^2)$$

em que γ_{kj} é o efeito do marcador na posição λ_{kj}.

A posição λ_{kj} varia dentro dentro do k-ésimo *bin* no j-ésimo cromossomo e vamos assumir que ela coincide com a posição da marca M_m, ou seja, temos uma relação biunívoca $(\lambda_1 = M_1, \lambda_2 = M_2, \ldots, \lambda_m = M_m)$ e por isso, utilizamos a *priori* de que ela é uniformemente distribuída em cada *bin* no j-ésimo cromossomo (considerando Δ_{kj} como o tamanho do k-ésimo *bin* no j-ésimo cromossomo), temos:

Para simplificar a notação, tomando $\gamma = \{\gamma_{kj}\}$, $\sigma^2_\gamma = \{\sigma^2_{\lambda_{kj}}\}$ e $\lambda = \{\lambda_{kj}\}$ a priori conjunta para as variáveis não observáveis é dada por:

$$p(\mu, \sigma^2_e, \gamma, \sigma^2_\gamma) = p(\mu)p(\sigma^2_e)\prod_{k=1}^{m} p(\gamma)p(\sigma^2_\gamma) \tag{3.4}$$

em que m é o número de total de marcadores no cromossomo.

A verossimilhança foi descrita por:

$$p(\boldsymbol{y}|\mu, \sigma^2_e, \gamma, \sigma^2_\gamma) = \prod_{i=1}^{n} p(\boldsymbol{y}_i|\mu, \sigma^2_e, \gamma, \sigma^2_\gamma)$$

$$\propto (\sigma^2_e)^{\frac{n}{2}} exp\left\{ -\frac{1}{2\sigma^2_e}\sum_{i=1}^{n}\left(\boldsymbol{y}_i - \mu - \sum_{j=1}^{C}\int_0^L \boldsymbol{Z}_{ij}(\lambda)\gamma(\lambda)d\lambda\right)^2\right\}$$

em que $\boldsymbol{y} = \{y_1, y_2, \ldots, y_m\}$.

Por intermédio do teorema de Bayes, a distribuição a *posteriori* é dada por:

$$p(\mu, \sigma^2_e, \gamma, \sigma^2_\gamma|\boldsymbol{y}) \propto p(\boldsymbol{y}|\mu, \sigma^2_e, \gamma, \sigma^2_\gamma)p(\mu, \sigma^2_e, \gamma, \sigma^2_\gamma)$$

Problemas que apresentam componentes aleatórias podem ser descritos de forma aproximada baseados em informações probabilísticas. Os métodos que fazem uso de técnicas de simulação iterativas são os métodos de Monte Carlo via cadeia de Markov (Metrópolis-Hastings e Amostrador de Gibbs). Assim, utilizou-se o algoritmo Monte Carlo Cadeia de Markov (MCMC), via amostragem de Gibbs e Metrópolis-Hastings, para amostrar valores e obter aproximações das distribuições marginais dos parâmetros. Os passos MCMC são descritos a seguir.

1. Inicialização: Os parâmetros μ e σ^2_e são inicializados com a média e a variância dos dados fenotípicos, respectivamente; λ é inicializado com o valor zero correspondente ao efeito do marcador na posição do meio (λ_{kj}) do k-ésimo *bin* no j-ésimo cromossomo, σ^2_γ é inicializada com 0,5.

$$I^{(t)} = \left[\mu^{(t)}, \gamma_1^{(t)}, \ldots, \gamma_m^{(t)}, \sigma_e^{2(t)}, \sigma_{\lambda_1}^{(t)}, \ldots, \sigma_{\lambda_m}^{2(t)}\right] \tag{3.5}$$

em que t é o número da iteração atual, iniciando em zero.

2. Atualizar μ:

$$p(\mu|\ldots) \propto exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i-\mu-\sum_{j=1}^{m}Z_{ij}\gamma_k\right)^2\right\}\cdot 1$$

Fazendo algumas manipulações algébricas de completamento de quadrados, considerando que os termos que não dependem do parâmetro são constantes para este parâmetro, temos:

$$p(\mu|\ldots) \sim \mathbb{N}\left[\frac{\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m}Z_{ij}\gamma_k^t\right)}{n}, \frac{\sigma_e^{2(t)}}{n}\right] \tag{3.6}$$

Logo, a distribuição condicional a posteriori para μ representa o núcleo de uma distribuição normal com média e variância descritas em (3.6). A média μ amostrada é denotada por $\mu^{(t+1)}$ atualizada no lugar de $\mu^(t)$ em todos os processos de amostragem subsequentes.

3. Atualizar γ:

$$p(\mu|\ldots) \propto exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i-\mu-\sum_{j=1}^{m}Z_{ij}\gamma_k\right)^2\right\}\cdot exp\left\{-\frac{1}{2\sigma_\gamma^2}\sum_{i=1}^{n}\gamma^2(\lambda)\right\}$$

Fazendo algumas manipulações algébricas de completamento de quadrados, considerando que os termos que não dependem do parâmetro γ são constantes para este parâmetro, temos:

$$\begin{aligned}\bar{\gamma} &= \left(\sum_{i=1}^{n}Z_{ij}^2+\frac{\sigma_e^{2(t)}}{\sigma_\gamma^{2(t)}}\right)\sum_{i=1}^{n}Z_{ij}\left(y_i-\mu^{(t)}\sum_{j=1}^{m}Z_{ij}\gamma_k^t\right) \\ s_\gamma^2 &= \left(\sum_{i=1}^{n}Z_{ij}^2+\frac{\sigma_e^{2(t)}}{\sigma_\gamma^{2(t)}}\right)^{-1}\sigma_e^{2(t)}\end{aligned} \tag{3.7}$$

Logo, a distribuição condicional a *posteriori* para γ representa o núcleo de uma distribuição normal com média e variância descritas em (3.7). Os γ amostrados são denotados por $[\gamma]^{(t+1)}$ e atualizados no lugar de $[\gamma]^{(t)}$

4. Atualizar σ_e^2 :

$$p(\sigma_e^2|\cdots) \propto (\sigma_e^2)^{-\frac{n}{2}}exp\left\{\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i-\mu-\sum_{j=1}^{m}Z_{ij}\gamma_k\right)^2\right\}(\sigma_e^2)^{-1}$$

Fazendo algumas manipulações algébricas de completamento de quadrados, considerando que os termos que não dependem do parâmetro σ_e^2 são constantes para este parâmetro, temos:

$$p(\sigma_e^2|\cdots) \sim \chi_{esc}^{-2}\left(n, \sum_{i=1}^{n}\left(\boldsymbol{y}_i - \mu - \sum_{j=1}^{m}\boldsymbol{Z}_{ij}\gamma_k^t\right)^2\right) \tag{3.8}$$

Logo, a distribuição condicional a posteriori é "Qui-quadrada Invertida escalada", como descrita em (3.8). A variância amostrada $\sigma_e^{2(t+1)}$ é atualizada no lugar de $\sigma_e^{2(t)}$.

5. Atualizar σ_γ^2:

$$p(\sigma_\gamma^2|\cdots) \propto (\sigma_\gamma^2)^{-\frac{1}{2}} exp\left\{-\frac{1}{\sigma_\gamma^2}\sum_{i=1}^{n}\gamma^2\right\}(\sigma_\gamma^2)^{-(1+\frac{\nu}{2})} exp\left\{-\frac{1}{\sigma_\gamma^2}\nu S^2\right\}$$

Fazendo algumas manipulações algébricas, temos:

$$p(\sigma_\gamma^2|\cdots) \sim \chi_{esc}^{-2}\left(\nu+1, [\gamma^2(\lambda)]^{(t)} + S^2\right). \tag{3.9}$$

Logo, a distribuição condicional a posteriori é "Qui-quadrada Invertida escalada", como descrita em (3.9). A variância amostrada $\sigma_\gamma^{2(t+1)}$ é atualizada no lugar de $\sigma_\gamma^{2(t)}$.

O parâmetro λ_k não possui uma função conhecida, assim sendo, não se pode utilizar o amostrador de Gibbs. Diante das pressuposições, as marcas serão amostradas dadas as posições λ_k ao invés de amostrar diretamente as posições λ_k. Para isso, será utilizado o algoritmo Metropolis-Hastings (HASTINGS, 1970; METROPOLIS et al., 1953). Esse algoritmo não exige que o parâmetro tenha uma função de probabilidade conhecida, uma vez que faz uso de uma função auxiliar possível de ser amostrada, retirando valores candidatos que podem ser aceitos com α de probabilidade.

Neste caso, pode-se utilizar uma distribuição uniforme como auxiliar para λ_{kj} que será amostrada em cada *bin* em todos os cromossomos, sob um intervalo delimitado por $max(LI_{kj}, \lambda_{kj} - c)$ e $min(LS_{kj}, \lambda_{kj} + c)$, sendo c uma constante discreta que define o caminhamento (salto) dentro do k-ésimo *bin* no j-ésimo cromossomo, normalmente fixado um valor de 10 ou 20% do número de marcas deste *bin*. Esta função é denotada por $u(\lambda_{kj}^*, \lambda_{kj}^{(t)})$, para o k-ésimo *bin* no j-ésimo cromossomo, e a nova posição λ_{kj}^* será aceita na t-ésima iteração com $min(1, \alpha)$ de probabilidade. Assim, se α for aceito, uma nova posição é estabelecida e o marcador Z referente a essa posição é selecionado.

$$\alpha = \frac{p(\lambda_{kj}^{*}|\ldots)u(\lambda_{kj}^{*},\lambda_{kj}^{(t)})}{p(\lambda_{kj}^{(t)}|\ldots)u(\lambda_{kj}^{(t)},\lambda_{kj}^{*})} \tag{3.10}$$

sendo:

$$u(\lambda_{kj}^{(t)},\lambda_{kj}^{*}) = \begin{cases} \frac{1}{2c}, & \text{se } \lambda_{kj}^{(t)}+c \leq LS_{kj} \text{e } \lambda_{kj}^{(t)}-c \geq LI_{kj}. \\ \frac{1}{c+\lambda_{kj}^{(t)}-LI_{kj}}, & \text{se } \lambda_{kj}^{(t)}+c < LS_{kj} \text{e } \lambda_{kj}^{(t)}-c < LI_{kj}. \\ \frac{1}{c+LS_{kj}-\lambda_{kj}^{(t)}}, & \text{se } \lambda_{kj}^{(t)}+c > LS_{kj} \text{e } \lambda_{kj}^{(t)}-c > LI_{kj}. \\ \frac{1}{2c}, & \text{se } \lambda_{kj}^{*}+c \leq LS_{kj} \text{e } \lambda_{kj}^{*}-c \geq LI_{kj}. \\ \frac{1}{c+\lambda_{kj}^{(t)}-LI_{kj}}, & \text{se } \lambda_{kj}^{*}+c < LS_{kj} \text{e } \lambda_{kj}^{*}-c < LI_{kj}. \\ \frac{1}{c+LS_{kj}-\lambda_{kj}^{*}}, & \text{se } \lambda_{kj}^{*}+c > LS_{kj} \text{e } \lambda_{kj}^{*}-c > LI_{kj}. \end{cases}$$

Repete-se a sequência acima até a convergência da cadeia para uma distribuição estacionária. Na cadeia final, a média a posteriori dos parâmetros finais amostrados será utilizada para a estimação deles. Após a convergência, adota-se $f(\lambda)$ como a frequência de visitas realizadas na posição $M|\lambda$ dentro de um *bin* específico. Contudo, a integral $\int_0^L \boldsymbol{Z}_i \lambda \gamma(\lambda) d\lambda$ para recompor o valor genético genômico não é conhecida e $\gamma(\lambda)$ também não. Hu, Wang e Xu (2012) utilizaram o efeito médio dos bisn como frequência. Nesse estudo, relaxa-se a suposição do efeito médio e, como para cada λ pode-se atribuir a frequência do número de vezes que o modelo visitou o marcador correspondente, é possível substituir $\int_0^L \boldsymbol{Z}_i \lambda \gamma(\lambda) d\lambda$ por $f(\lambda)\boldsymbol{Z}_j\hat{\gamma}$.

Assim, tomando $g = \int_{k=1}^{m} \boldsymbol{Z}\lambda\gamma(\lambda)d\lambda$ pode ser recuperado por $P\boldsymbol{Z}_j\hat{\gamma}$, sendo P (frequência do número de vezes que o modelo visitou o marcador) o peso de cada marcador dado sua posição λ no genoma e, g é valor genético genômico para o indivíduo j.

3.4 Acurácia preditiva

Os dados obtidos nos cenários acima foram analisados com os modelos Bayes B e Lasso Bayesiano, utilizando a função BGLR contida no pacote BGLR (PÉREZ; DE LOS CAMPOS, 2014) e, o RR-BLUP utilizando a função mixed.solve (Endelman, 2011) no pacote rrBLUP, ambos são bibliotecas do *software* R (R CORE TEAM, 2015). Utilizou-se, também, o modelo descrito na seção 3.3 para analisar estes dados e comparar resultados.

Para avaliar a capacidade preditiva dos modelos, utilizou-se o erro quadrático médio (EQM), bem como seu "subproduto", o coeficiente de determinação R^2. O EQM é definido

como:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{y}_i)^2 \tag{3.11}$$

em que $\hat{\boldsymbol{y}}_j = \hat{\mu} + P\boldsymbol{Z}_j\hat{\gamma}$ o valor predito do j-ésimo indivíduo usando os parâmetros $\hat{\mu}$ e $\hat{\gamma}_j$, que são estimados das amostras excluindo o indivíduo j. Note que, quanto menor é o EQM, obtém-se uma melhor predição.

O coeficiente de determinação R^2 pode ser definido como:

$$R^2 = 1 - \frac{EQM}{SQT} = \frac{\sum_{i=1}^{n}(\boldsymbol{y}_i - \hat{y})^2}{\sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{y})^2} \tag{3.12}$$

A expressão (3.12) pode ser interpretada como a proporção da variância fenotípica explicada por todos os blocos gênicos (janelas), consequentemente, por todos os marcadores do genoma.

4 RESULTADOS

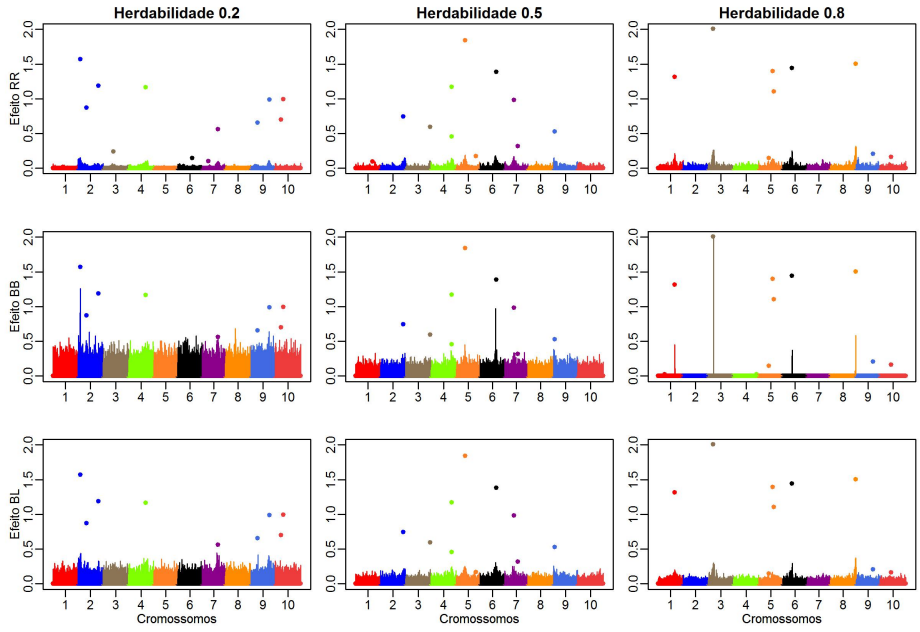
Para facilitar o entendimento, adotaram-se para essa seção, os termos Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005 nas diferentes herdabilidades (0,2, 0,5 e 0,8) e em cada cenário (oligogênico, poligênico e infinitesimal). Os tamanhos de cada janela (*bin*) são, respectivamente, 10%, 5%, 1%, 0,66̄ % e 0,5% do número de marcadores. Utilizou-se também, RR para representar o método RR-BLUP, BB para representar o método Bayes B e BL para o método Lasso Bayesiano.

4.1 Dados simulados

4.1.1 Cenário oligogênico

Para avaliar o presente cenário, realizou-se a simulação de 12 genes distribuídos em 10 grupos de ligação. Na Figura 4.1 ilustram-se os efeitos absolutos verdadeiros QTLs e os efeitos absolutos preditos com herdabilidades (0,2; 0,5 e 0,8) aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano.

Figura 4.1 – Efeitos verdadeiros e estimados no cenário oligogênico aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano. Pontos coloridos representam os 12 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.

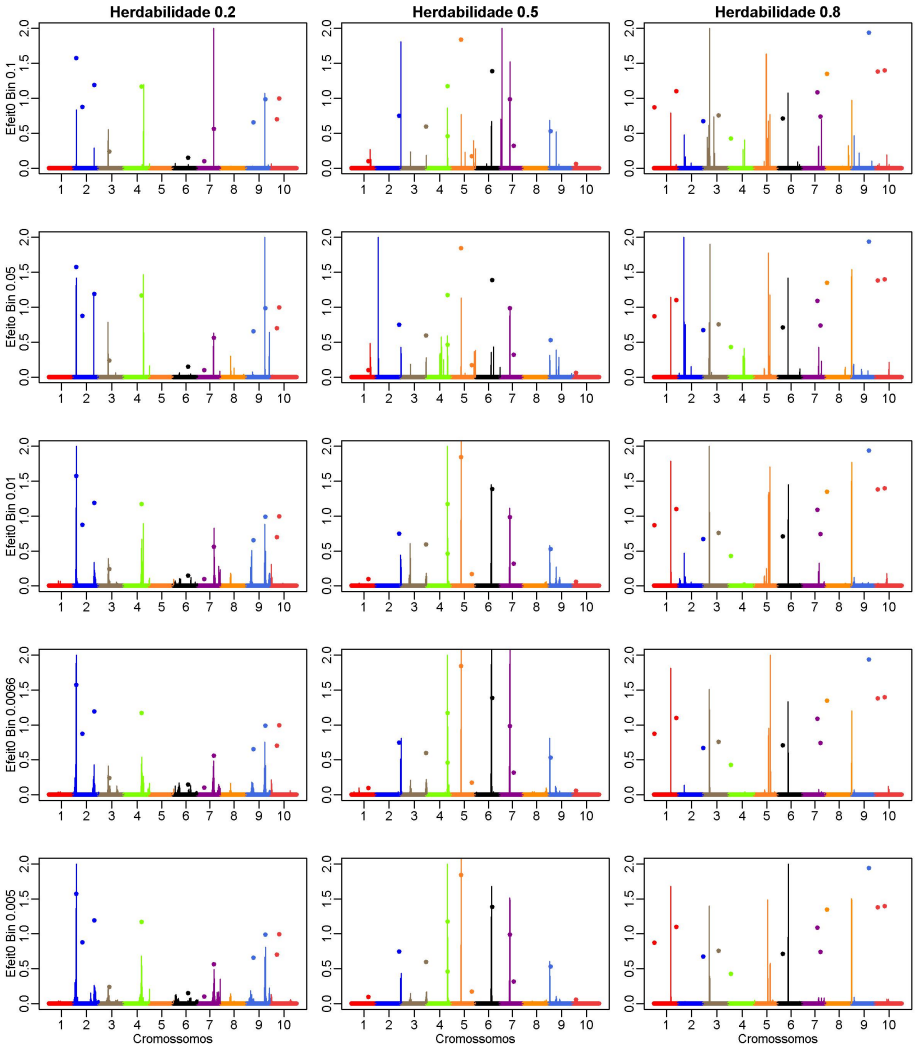


De acordo com a Figura 4.1, observou-se que grande parte dos QTLs de grande efeito simulados, foram mapeados pelos três métodos supracitados, isto é, em comparação com os

efeitos verdadeiros simulados, os efeitos preditos pelos modelos mostraram padrão semelhante (ver apêndice), mas com polarização para baixo.

Na Figura 4.2 são mostrados os efeitos absolutos verdadeiros dos QTLs e os efeitos absolutos preditos com as herdabilidades (0,2; 0,5 e 0,8) aos métodos, respectivamente, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0,0066 e Bin 0,005.

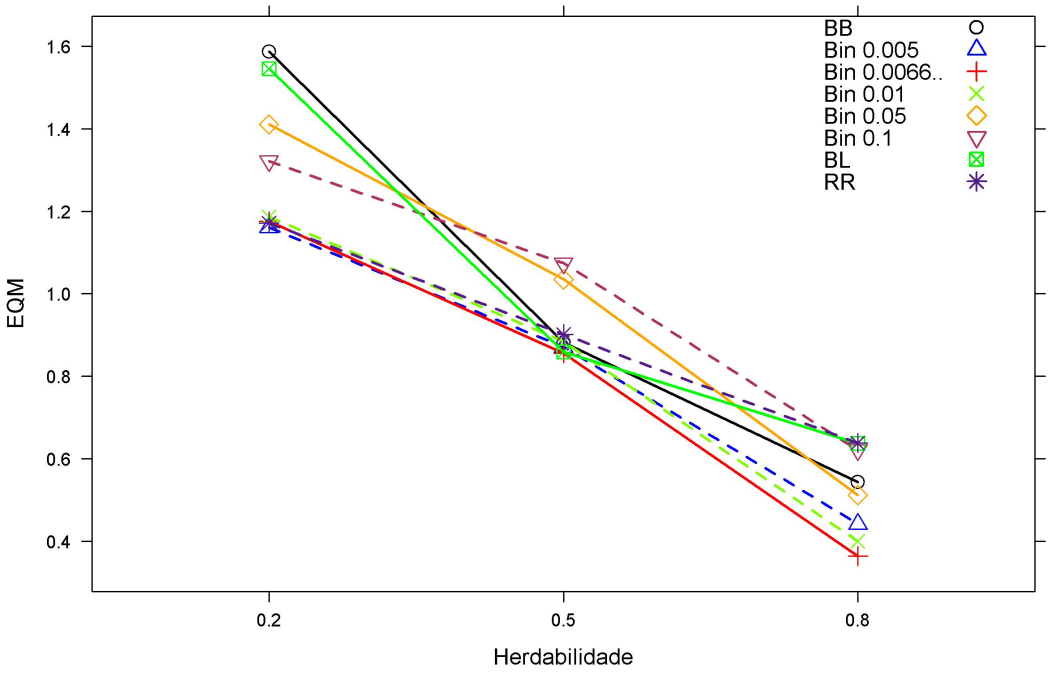
Figura 4.2 – Efeitos verdadeiros e estimados no cenário oligogênico aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005. Pontos coloridos representam os 12 QTLs distribuídos em 12150 SNPs em 10 grupos de ligação.



De acordo com a Figura 4.2, observou-se que a maioria dos QTLs simulados foram mapeados por todos os modelos *bins* (janelas), isto é, em comparação com os efeitos verdadeiros simulados, os efeitos preditos pelos modelos mostraram padrão semelhante (ver apêndice) e, em geral, obeteve estimativas mais próximas dos QTLs simulados que os modelos padrão (RR, BB e BL).

Na Figura 4.3 têm-se os valores de erro quadrático médio (EQM) dos modelos RR, BB, BL e as cinco configurações de janelas.

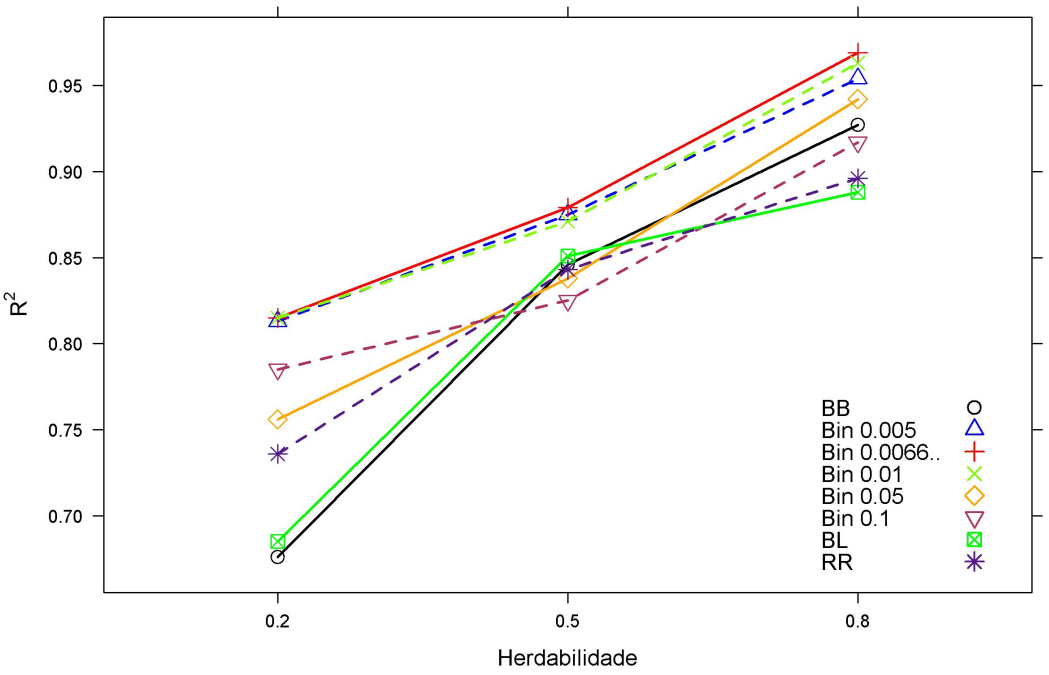
Figura 4.3 – Erro quadrático médio (EQM) no cenário oligogênico com herdabilidades (0,2; 0,5 e 0,8) aos métodos RR-BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.



Na Figura 4.3 observou-se que quando aumenta os níveis de herdabilidade os valores de EQM, em todos os modelos, decaem. Fixando-se a herdabilidade, para herdabilidade 0,2 o menor valor de EQM (1,046) é no modelo Bin 0.005 e o modelo que apresentou maior valor foi o BB, com EQM (1,587). Já na herdabilidade 0,5 os maiores valores de EQM foram Bin 0.1 (1,074) e Bin 0.05 (1,035) e o menor valor de EQM foi o Bin 0.0066 (0,856). Na herdabilidade 0,8 o menor valor de EQM foi o Bin 0.0066 (0, 364) e o maior valor de EQM foi o do modelo RR (0,638). Um fato curioso é que neste estudo verificou-se, no presente cenário, que o modelo RR, no que diz respeito ao EQM, seu valor decai de forma praticamente linear com aumento da herdabilidade. Isso não é sinônimo de vantagens, pois, neste estudo, para herdabilidade 0,8 o RR obteve menor e o Bin 0.0066 maior capacidade preditiva.

Na Figura 4.4 têm-se os valores do coeficiente de determinação (R^2) dos diferentes modelos RR, BB e BL com as cinco configurações de janelas para o cenário oligogênico.

Figura 4.4 – Coeficiente de determinação (R^2) no cenário oligogênico com herdabilidades (0,2; 0,5 e 0,8) nos métodos RR-BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.



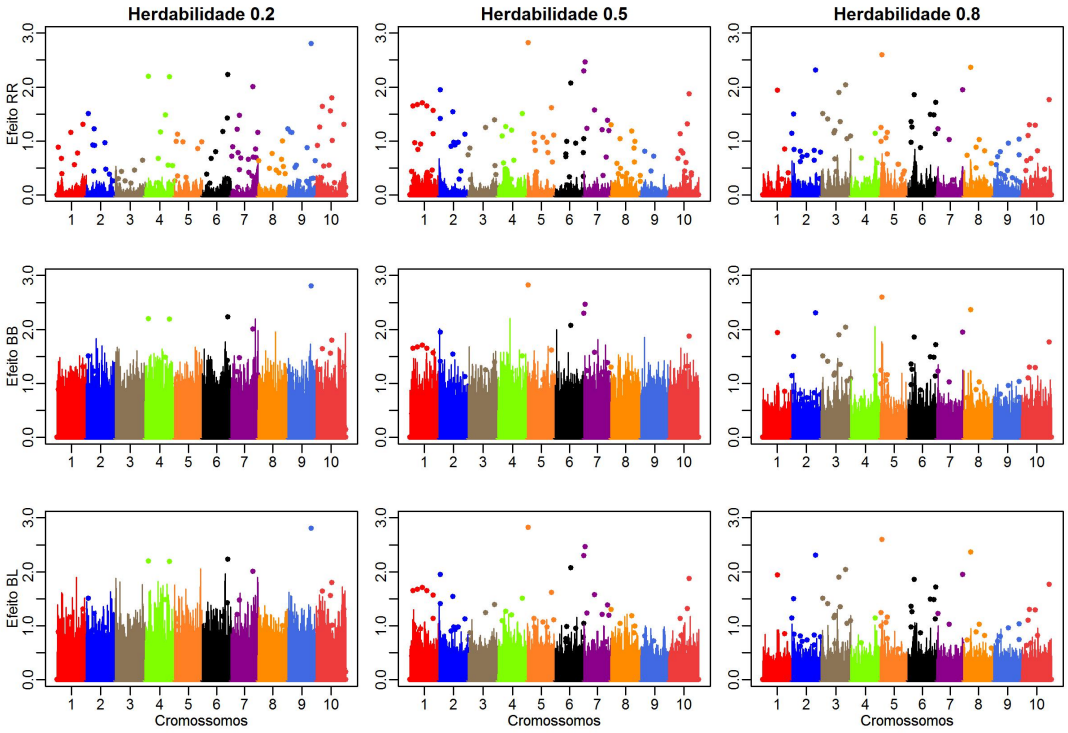
Na Figura 4.4 são mostrados os valores de R^2, em que para herdabilidade 0,2 os modelos *bins* (nas suas cinco configurações) obtiveram maiores coeficientes de determinação (R^2) e os modelos Bin 0.0066, Bin 0.05 e Bin 0.01 obtiveram maiores R^2 nos três níveis de h^2 (herdabilidade). Com aumento de h^2, todos os modelos aumentaram consideravelmente seus coeficientes de determinação. No entanto, notou-se que a taxa de crescimento de R^2 para os modelos padrão, quando passa de $h^2 = 0{,}5$ para $h^2 = 0{,}8$, é bem menor de que quando passa de $h^2 = 0{,}2$ pra $h^2 = 0{,}5$. Todavia, os modelos *bin* têm taxa de crescimento similar nesses dois intervalos. Culminando em vantagens de utilizar modelos funcionais (abordagens *bin*) na GWS, pois, para o presente cenário, apesar de configurações distintas de janelas causar diferentes R^2, a maioria das configurações propostas obtiveram coeficientes de determinação maiores que os outros modelos (RR, BB e BL).

4.1.2 Cenário poligênico

Para avaliar o presente cenário, realizou-se a simulação de 120 genes em 10 grupos de ligação. Na Figura 4.5 ilustram-se os efeitos absolutos verdadeiros dos QTLs e os efeitos

absolutos preditos com herdabilidades (0,2; 0,5 e 0,8) aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano.

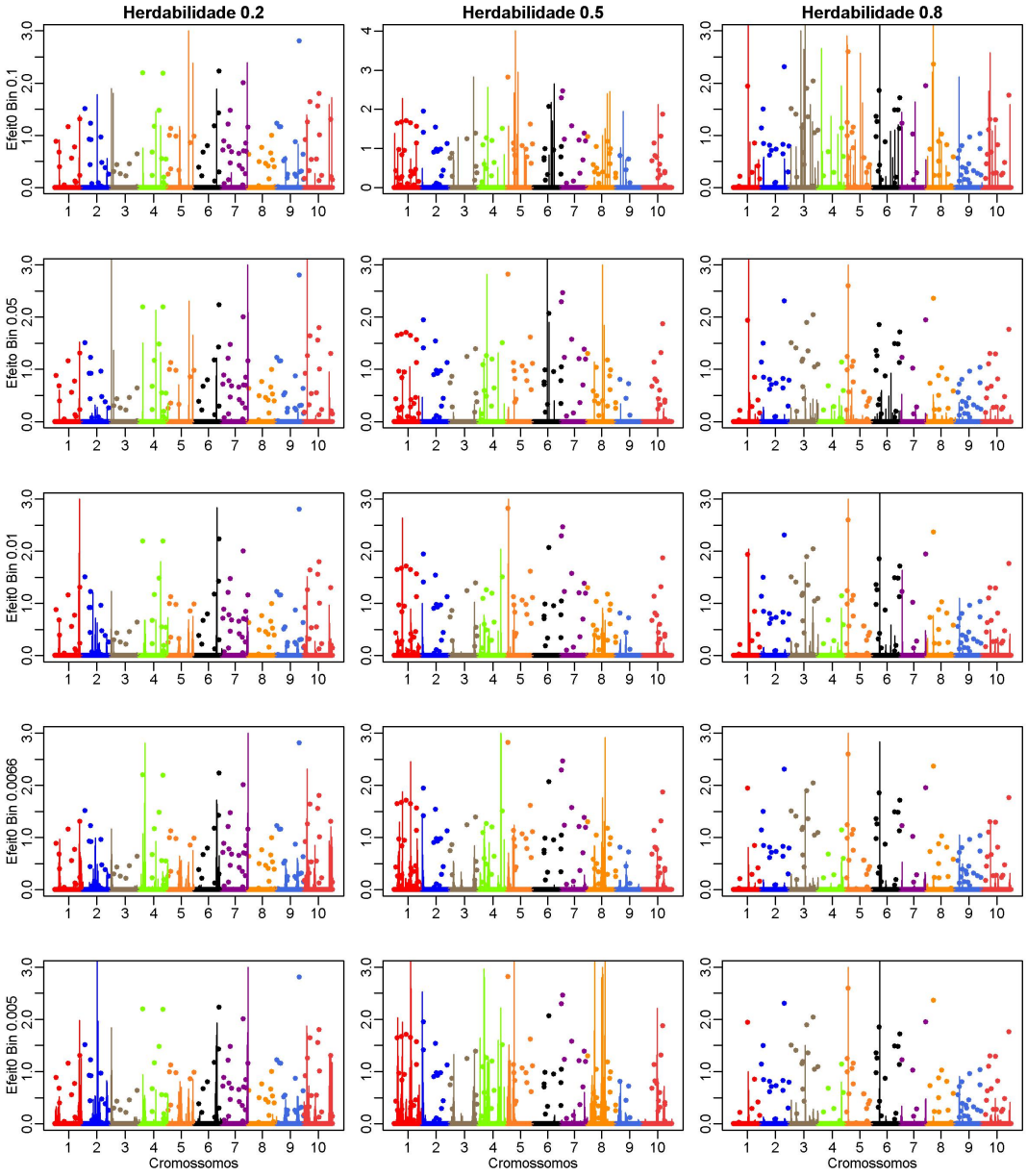
Figura 4.5 – Efeitos verdadeiros e estimados no cenário poligênico aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano. Pontos coloridos representam os 120 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.



De acordo com a Figura 4.5, observou-se que grande parte dos QTLs de grande efeito simulados, foram mapeados pelos três métodos citados acima, isto é, em comparação com os efeitos verdadeiros simulados, os efeitos preditos pelos modelos mostraram-se padrão semelhante (ver apêndice), mas com polarização considerável para baixo.

Na Figura 4.6 são apresentados os efeitos absolutos verdadeiros dos QTLs e os efeitos absolutos preditos com herdabilidades (0,2; 0,5 e 0,8) aos métodos, respectivamente, Bin 0.1, Bin 0.05 e Bin 0.01, Bin 0.0066 e Bin 0.005.

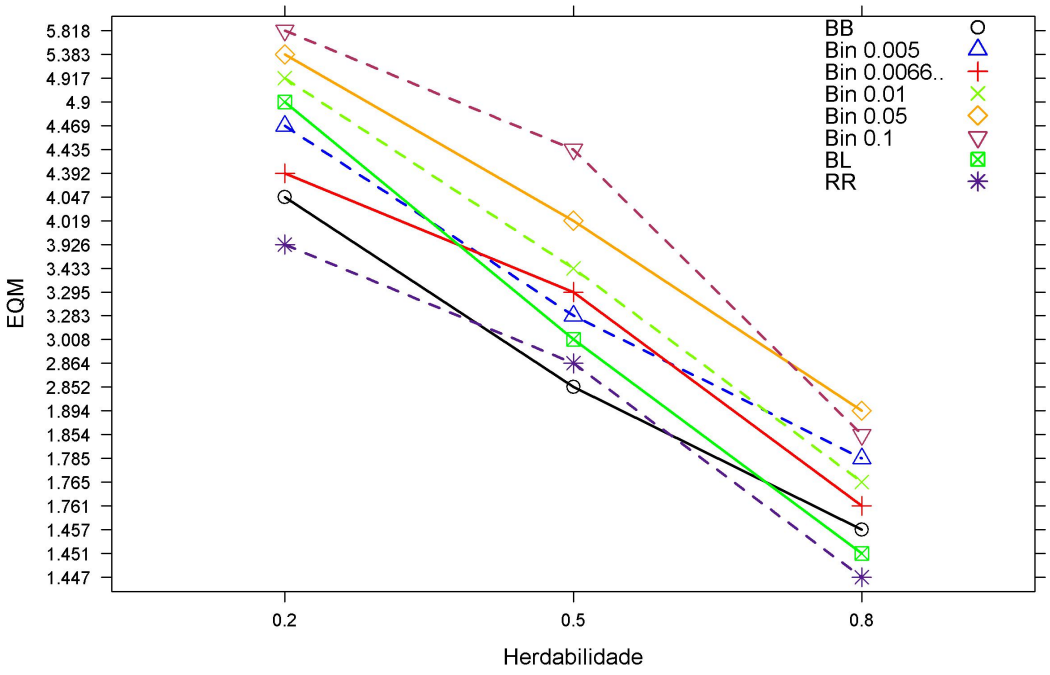
Figura 4.6 – Efeito no cenário poligênico aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin0.01, Bin 0.0066 e Bin 0.005. Pontos coloridos representam os 120 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.



De acordo com a Figura 4.6, observou-se que boa parte dos QTLs de grande e médio efeito simulados, foram mapeados pelos métodos citados acima, isto é, em comparação com os efeitos verdadeiros simulados, os efeitos preditos pelos modelos *bins* mostraram-se padrão semelhante (ver apêndice). Contudo, muitos QTLs de pequeno e médio efeito não foram mapeados. Mas, apesar de mapear menos QTLs que os modelos padrão, quando mapeia, suas estimativas são melhores que as dos modelos concorrentes.

Na Figura 4.7 têm-se os valores de EQM dos modelos RR, BB e BL com as cinco configurações de janelas (Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005).

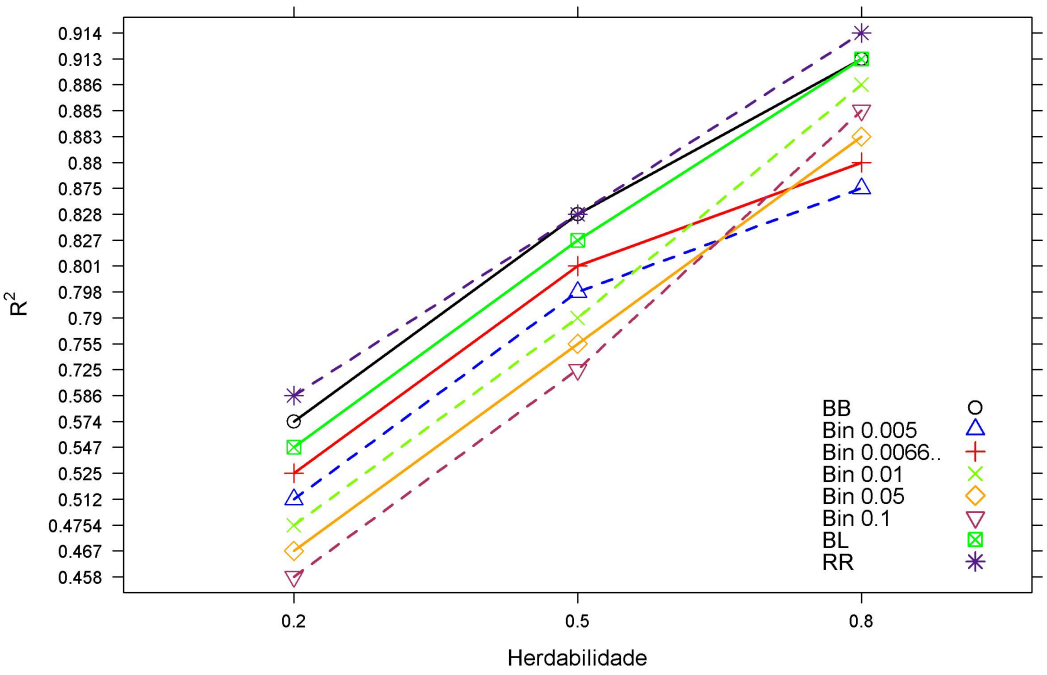
Figura 4.7 – Erro quadrático médio (EQM) no cenário poligênico com herdabilidades (0,2; 0,5 e 0,8) nos métodos RR-BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.



Os modelos RR e BB obtiveram menores valores de EQM nas herdabilidades 0,2 e 0,5. Notou-se que os modelos Bin 0.0066 e Bin 0.005 para $h^2 = 0.2$, superaram o método BL. Para herdabilidade alta, a diferença de EQM para os modelos padrão é pequena.

Na Figura 4.8 têm-se os valores do coeficiente de determinação dos diferentes modelos RR, BB e BL com as cinco configurações de janelas para o cenário poligênico.

Figura 4.8 – Coeficiente de determinação (R^2) no cenário poligênico com herdabilidades (0,2; 0,5 e 0,8) nos métodos RR-BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.



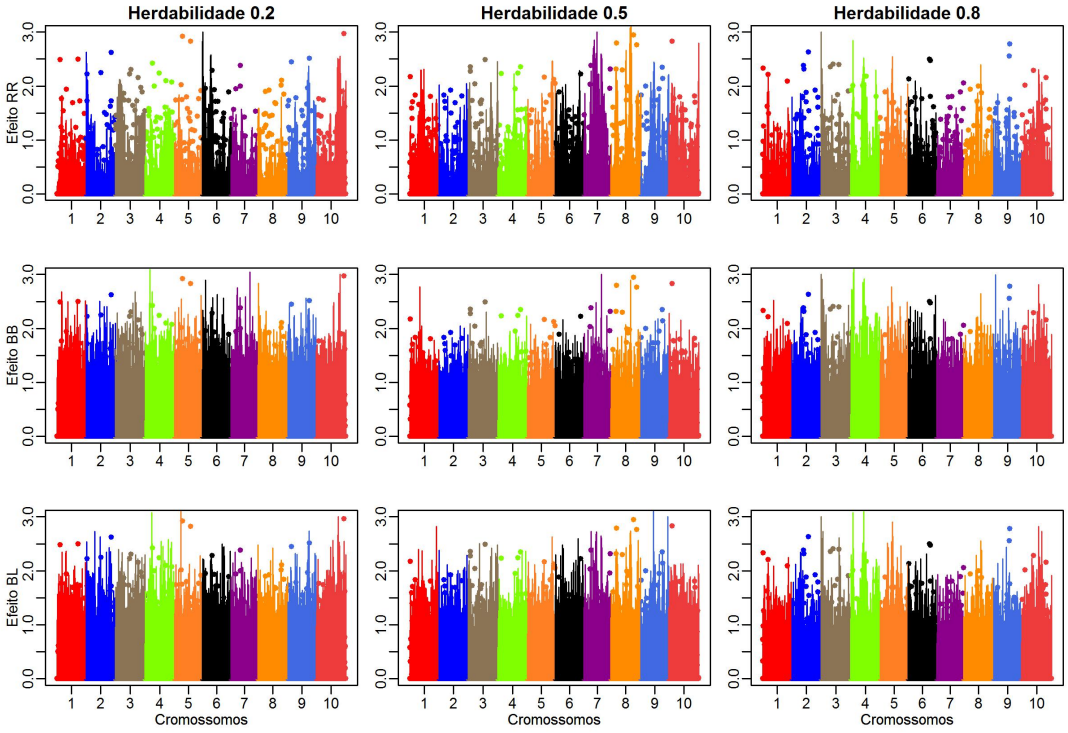
Os valores do coeficiente de determinção (R^2) dos modelos RR, BB e BL foram maiores em relação aos modelos com diferentes tamanhos de janelas para as diferentes herdabilidades (Figura 4.8), sendo que o modelo RR tem maior valor de R^2 que os demais. Note que o coeficiente de determinação para todos os modelos crescem à medida que aumenta a herdabilidade. Para herdabilidade alta, os modelos padrão convergem para o mesmo coeficiente de determinação e os modelos Bin 0.1 e Bin 0.01 parecem seguir um padrão com o aumento da herdabilidade.

Os valores de R^2 possui relação inversa aos valores de EQM, pois os modelos que têm maior capacidade preditiva (menor EQM) têm maiores valores de R^2.

4.1.3 Cenário infinitesimal

Para avaliar o presente cenário, realizou-se a simulação de 600 genes em 10 grupos de ligação. Na Figura 4.9 ilustram-se os efeitos absolutos verdadeiros dos QTLs e os efeitos absolutos preditos com herdabilidades (0,2; 0,5 e 0,8) aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano.

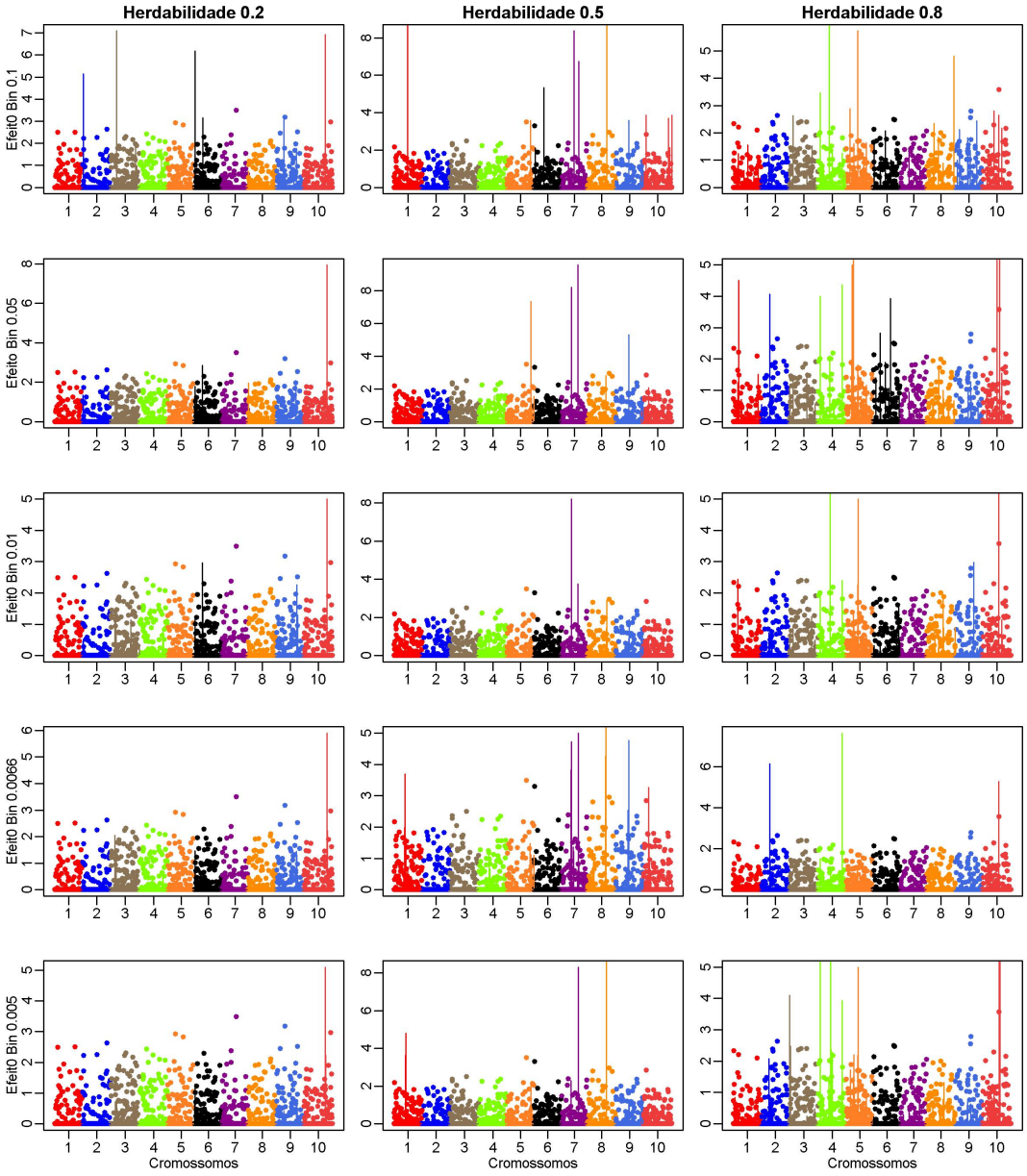
Figura 4.9 – Efeitos verdadeiros e estimados no cenário infinitesimal aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano. Pontos coloridos representam os 600 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.



De acordo com a Figura 4.9, observou-se que grande parte dos QTLs de grande e médio efeito simulados foram mapeados pelos três métodos supracitados, isto é, em comparação com os efeitos verdadeiros simulados, os efeitos preditos pelos modelos mostraram-se padrão semelhante (ver apêndice), mas com ligeira polarização para baixo. Contudo, o modelo RR apesar de ter estimativas menores que os outros dois modelos pelo fato do efeito *shrinkage*, mesmo assim parece apontar picos onde realmente têm QTLs.

Na Figura 4.10 são mostrados os efeitos absolutos verdadeiros dos QTLs e os efeitos absolutos preditos com herdabilidades (0,2; 0,5 e 0,8) aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin 0.01 Bin 0.0066 e Bin 0.005.

Figura 4.10 – Efeito no cenário infinitesimal aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin 0.01 Bin 0.0066 e Bin 0.005. Pontos coloridos representam os 600 QTLs distribuídos em 12150 SNPs em 10 grupos de ligação.

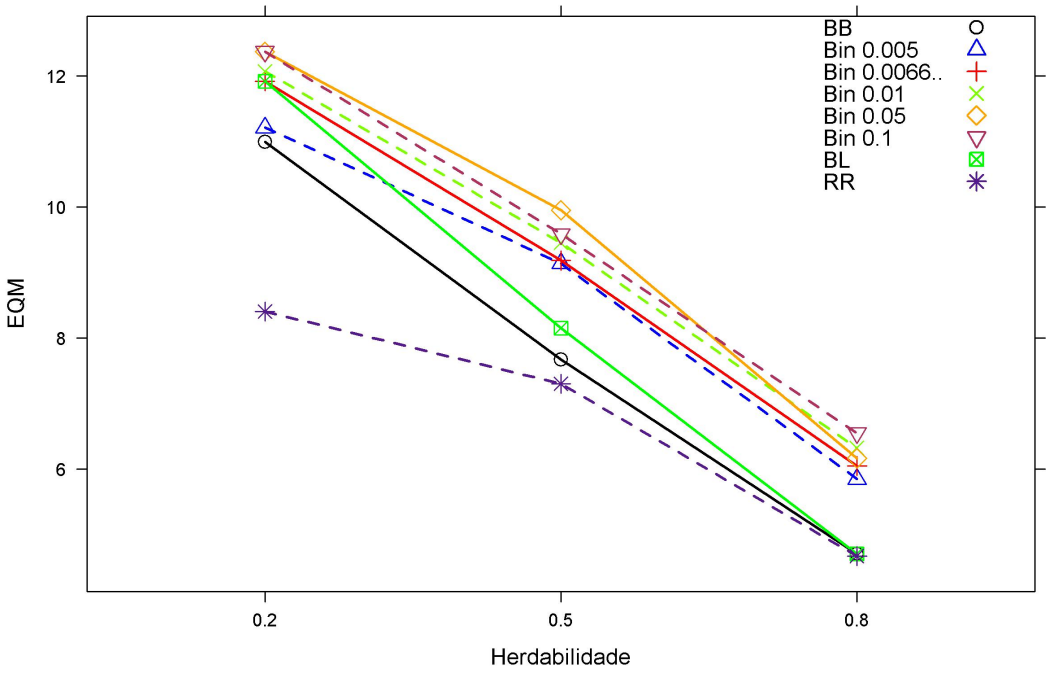


De acordo com a Figura 4.9, observou-se que grande parte dos QTLs de grande e médio efeito simulados não foram mapeados pelos métodos *bin* citados acima. Isto é, em comparação com os efeitos verdadeiros simulados, os efeitos preditos pelos modelos, na sua maioria, não mostraram padrão semelhante (ver apêndice). Os efeitos preditos do RR-BLUP (RR), Bayes B (BB) e Lasso Bayesiano (BL) (Figura 4.9) mostraram captar melhor os sinais de QTLs do que modelos *bins* (Figura 4.10). Verificou-se que alguns dos efeitos preditos pelos modelos *bins*,

para esse cenário, foram maiores do que os efeitos dos QTLs simulados, provavelmente, captou sinal de dois ou mais QTLs próximos e, com isso, gerou picos de maior efeito .

Na Figura 4.11 têm-se os valores de EQM dos modelos RR, BB e BL com as cinco configurações de janelas (Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005) para o cenário infinitesimal.

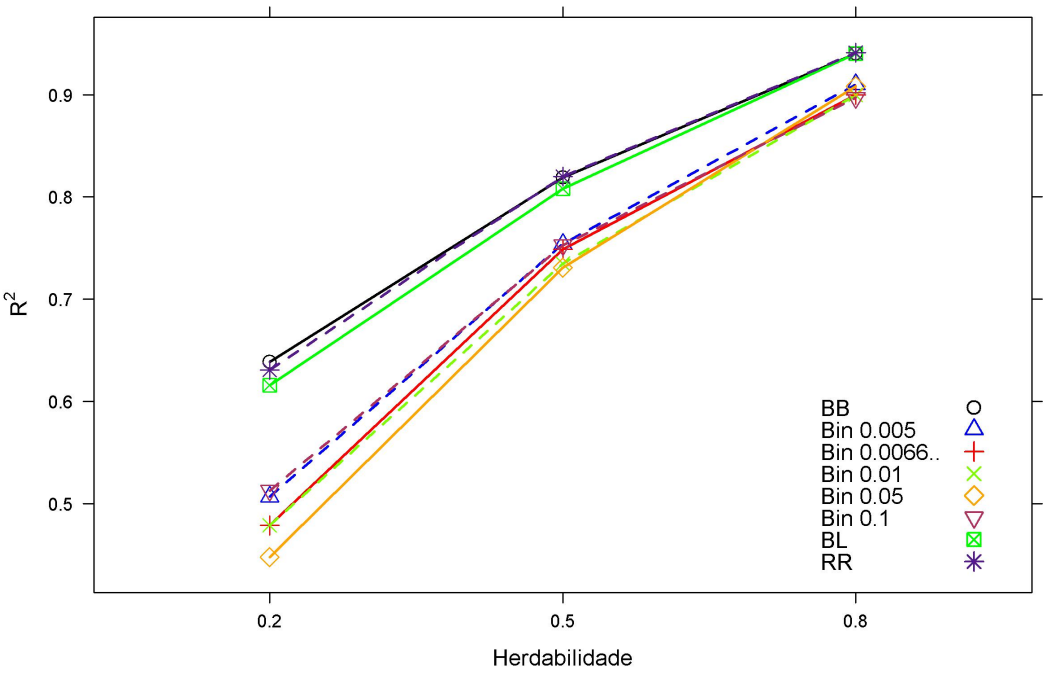
Figura 4.11 – Erro quadrático médio (EQM) no cenário infinitesimal com herdabilidades (0,2; 0,5 e 0,8) nos métodos RR BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.



No cenário infinitesimal (Figura 4.11), o modelo RR foi superior aos demais nas herdabilidades 0,2 e 0,5 e na herdabilidade 0,8 o mesmo foi equivalente ao BB e BL, sendo esses três com maior capacidade preditiva do que os modelos *bin*. Note que, para alta herdabilida, nas diversas configurações *bin* , o EQM não varia muito. Contudo, a maior configuração (Bin 0.005) obteve maior capacidade preditiva que as outras configurações e, esta foi superior a BL na herdabilidade 0,2.

Na Figura 4.12 têm-se os valores do coeficiente de determinação (R^2) dos diferentes modelos RR, BB e BL com as cinco configurações de janelas para o cenário infinitesimal.

Figura 4.12 – Coeficiente de determinação (R^2) no cenário infinitesimal com herdabilidades (0,2; 0,5 e 0,8) nos métodos RR-BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.



Os valores do coeficiente de determinação no cenário infinitesimal para os diferentes modelos (Figura 4.12) indicam que os modelos BB, RR e BL têm maiores valores de R^2 do que os modelos *bin*. Note que, todos os modelos parecem seguir um padrão, em que crescem quase que paralelos à medida que aumenta a herdabilidade. Observou-se que os modelos padrão convergem para o mesmo valor para h^2= 0,8; o mesmo ocorre com os modelos *bin*.

5 DISCUSSÃO

Hu, Wang e Xu (2012) desenvolveram um novo método capaz de lidar com grandes conjuntos de dados, com finalidade de predição do valor genético genômico, através da divisão de todo genoma em pequenos intervalos denominados *bin*. Os resultados obtidos por esses autores foram muito superiores aos métodos tradicionais do alfabeto bayesiano (eBayes, Bayes B1, Bayes B2 e Lasso) e de modelos mistos (G-Blup). Contudo, o modelo proposto no presente estudo obteve destaque nos cenários oligogênico e poligênico, sendo, portanto, mais efetivo se a proporção de QTLs for menor que 1%. Já no cenário infinitesimal, o uso do modelo *bin* pode apresentar perda de informação.

A hipótese inicial contida neste trabalho era que o número de *bin* pudesse ser tão grande quanto se queira (Figura 3.1), para que as estimativas dos valores de $\gamma(\lambda)$ fossem as mais precisas possíveis e, consequentemente, a integral em (3.2) também. No entanto, essa estrutura puramente matemática não é rigorosamente viável em um contexto genético, já que quanto mais *bins*, menos marcas por *bin* e, portanto, menor o desequilíbrio de ligação entre marcas e provável QTL (HU; WANG; XU, 2012). Devido ao baixo ou nenhum desequilíbrio de ligação, QTLs e marcadores têm baixa associação, sendo, assim, menos provável detectar QTLs. Com isso, a configuração máxima de *bin* analisada no presente estudo foi de 1215 *bin* (10 marcas por *bin*).

Contudo, os *bins* adotados no presente estudo foram fixados sem nenhum critério a *priori*, podendo os mesmos terem sido mal escolhidos no que diz respeito ao LD (desequilíbrio de ligação). Entretanto, ao analisar o apêndice B percebe-se que, a partir dessa análise preliminar, pode-se construir *bins* de tal forma que todos os marcadores contidos em um *bin* específico, estejam em LD. Em palavras, pode-se pensar que os picos de maiores frequências, sejam os pontos de interrupção (*Breakpoints*) utilizados por (XU, 2013).

Para avaliar a efetividade do modelo proposto, foram comparadas algumas configurações de *bin* com os métodos denominados padrão (RR, BB e BL) em três cenários, conforme três herdabilidades (0,2; 0,5 e 0,8). Dessa forma, constatou-se, para o cenário oligogênico (Figura 4.3), que a maioria das configurações de modelo *bin* obteve menor EQM que os modelos concorrentes. A configuração Bin 0.005 (maior *bin* do presente estudo) obteve melhor capacidade preditiva do que o BL na herdabilidade 0,2 em todos os cenários.

Entre os modelos concorrentes, o RR mostrou destacada capacidade preditiva em relação aos outros dois em todos os cenários, seguido pelo BB. Entretanto, ao contrário dos resultados

obtidos por Hu, Wang e Xu (2012), em que BL foi o melhor modelo, o presente estudo mostrou-se que foi o pior dentre os modelos concorrentes. Todavia, é importante ressaltar que os dados simulados em questão seguem uma distribuição normal multivariada com média zero e variância comum, que é parte das pressuposições do método RR, podendo o mesmo ter sido favorecido nesse ponto.

De acordo com Van Den Berg et al. (2015), para o cenário infinitesimal o modelo de regressão aleatória GBLUP (teoricamente equivalente a RR-BLUP) tem maior precisão do que os modelos bayesianos e esses, por sua vez, têm vantagens no cenário oligogênico. Isto pode ser justificado pelo fato de que o RR-BLUP pressupõe que há um número grande de genes que estão uniformemente distribuídos em todo genoma e que eles contribuem igualmente para a característica de interesse. Diante disso, o RR-BLUP é eficiente para o cenário infinitesimal.

Apesar do exposto acima, segundo os autores Daetwyler et al. (2010) e Hu, Wang e Xu (2012), quantificar a superioridade de um modelo em relação a outros quanto à capacidade preditiva, depende de atributos como a estrutura populacional (por exemplo, o tamanho n) e a arquitetura genética do caráter em estudo (por exemplo, número de QTLs e herdabilidade). Além disso, de acordo com Hu, Wang e Xu (2012), o tamanho do *bin* também influencia no valor do EQM.

A principal vantagem de modelos funcionais (metodologia *bin*) é que não são simplesmente preditivos, mas sim passíveis de interpretação genética . Além disso, de acordo com Hu, Wang e Xu (20120), quaisquer modelos de regressão penalizados podem ser adotados ao modelo *bin*. Deste modo, apesar de a capacidade preditiva do modelo funcional (*bin*) ser menor no cenário infinitesimal, pode ocorrer que, ao longo das gerações, modelos *bin* podem ser mais eficientes do que modelos meramente preditivos. Ademais, de acordo com Zhang et al. (2010 apud HAYES; GODDARD, 2001), estudos de mapeamento de QTLs têm mostrado que a maioria das características quantitativas é afetada de forma significativa por um número finito de genes e não são uniformemente distribuídas e nem contribuem igualmente para a característica de interesse.

6 CONCLUSÃO

Diante do exposto, conclui-se que a técnica de predição (Modelos funcionais: Metodologia *bin*), proposta neste estudo, é recomendada no cenário oligogênico, pois para esse cenário obteve-se destacada capacidade preditiva em relação aos modelos RR-BLUP, Bayes B e Bayes Lasso.

Como em muitos casos o objetivo é detectar genes que contribuem para o caráter quantitativo e identificar sua respectiva posição no cromossomo, modelos funcionais (metodologia *bin*) podem se tornar uma ferramenta para avaliar a arquitetura genética dos traços de interesse. Dessa forma, estudos futuros poderão ser realizados usando esta metodologia em GWAS (*Genome-Wide Association Studies*) para identificar locos associados com as características de interesse. Também se pode pensar em atribuir ao método *bin* "*prioris*" dos métodos alfabeto Bayesianos e verificar se há ganho em relação ao método original.

REFERÊNCIAS

[1] BERNARDO, R. **Breeding for quantitative traits in plants**. 2 ed. Woodbury: Stemma Press. 2010. 390 p.

[2] CARDOT, H.; FERRATY, F.; SARDA, P. Spline estimators for the functional linear model. **Statistica Sinica**, Elmsford, v. 13, p. 571-591, 2003.

[3] CARDOT, H.; SARDA, P. Estimation in generalized linear models for functional data via penalized likelihood. **Journal of Multivariate Analysis**, New York, v. 92, n. 1, p. 24-41, Jan. 2005.

[4] DAETWYLER, H. D. et al. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, Austin, v. 185, n. 3, p. 1021-31, July 2010.

[5] EDWARDS, M. D.; STUBER, C. W.; WEDEL, J. F. Molecular-marker-facilitated investigations of quantitative-trait locos in maize: I. numbers, **genomic distribution and types of gene action**. Austin, Oxford, v. 116, n. 1, p. 113-125, May 1987.

[6] ENDELMAN, J. B. Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome**, Oxford, v. 4, n. 3, p. 250-255, May 2011.

[7] HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. **Biometrika**, London, v. 57, n. 1, p. 97-109, Apr. 1970.

[8] HAYES, B; GODDARD, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. **Genetics, Selection, Evolution**, London, v. 33, n. 3, p. 209-229, May/June 2001.

[9] HU, Z.; WANG, Z.; XU, S. An infinitesimal model for quantitative trait genomic value prediction.**PLoS One**, San Francisco, v. 7, n. 7, p. 1-13, 2012.

[10] HU, Z.; WANG, Z.; XU, S. An infinitesimal model for quantitative trait genomic value prediction.**PLoS One**, San Francisco, v. 7, n. 7, p. 1-13, 2012.

[11] JAMES, G. M.; WANG, J.; ZHU, J. Functional linear regression that's interpretable.**The Annals of Statistics**, Amsterdam, v. 37, n. 5, p. 2083–2108, 2009.

[12] JOEHANES, R.; NELSON, J. C. QGene 4.0, an extensible Java QTL-analysis platform. **Bioinformatics**, Oxford, v. 24, n. 23, p. 2788-2789, Dec. 2008.

[13] KIM, J. S.; MAITY, A.; STAICU, A. M. Generalized functional concurrent model. **Biometrics**, Washington, v. 1, p. 1-25, 2015.

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

[15] METRÓPOLIS, N. et al. Equation of state calculations by fast computing machines. **The Journal of Chemical Physics**, New York, v. 21, n. 4, p. 1087-1092, 1953.

[16] MEUWISSEN, T. H. E.; HAVES, B. J.; GOODARD, M. E. Prediction of total genetic value using genome-wide dense marker. **Genetics**, Austin, v. 157, n. 4, p. 1819−1829, Apr. 2001.

[17] MULLER, H. G.; STADTMULLER, U. Generalized functional linear models. **The Annals of Statistics**, Amsterdam, v. 33, n. 2, p. 774−805, 2005.

[18] PÉREZ, P.; DE LOS CAMPOS, G. Genome wide regression & prediction with BGLR Statistical Package. **Genetics**, Austin, v. 198, n. 2, p. 483−495, Oct. 2014.

[19] R CORE TEAM. **R**: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2015. Disponível em: <https://www.R-project.org/>. Acesso em: 1 jun. 2016.

[20] SINGH, B. D.; SINGH, A. K. **Marker-assisted plant breeding**: principles and practices. New York: Springer, 2015. 187 p.

[21] TOLEDO, E. R. et al. Mapeamento de QTLs: uma abordagem bayesiana. **Revista Brasileira de Biometria**, São Paulo, v. 26, n. 2, p. 107−114, 2008.

[22] VAN DEN BERG, S. et al. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. **BMC Genetics**, Oxford, v. 16, n. 1, p. 1-12, Dec. 2015.

[23] WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetical Research**, Cambridge, v. 75, n. 2, p. 249−252, Apr. 2000.

[24] XU, S. Estimating polygenic effects using markers of the entire genome. **Genetics**, Austin, v. 163, n. 2, p. 789−801, Feb. 2003.

[25] XU, S. Genetic mapping and genomic selection using recombination breakpoint data.**Genetics**, Austin, v. 195, n. 3, p. 1103-1115, Nov. 2013.

[26] ZHANG, Z. et al. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix. **PLoS One**, San Francisco, v. 5, n. 9, p. 12648−12656, Sept. 2010.

APÊNDICE A –

A seguir são apresentados os gráficos que ilustram os efeitos absolutos verdadeiro dos QTLs e os efeitos absolutos preditos, com suas respectivas escalas, em todos os cenários, com herdabilidades (0,2; 0,5 e 0,8) aos métodos, respectivamente, RR-BLUP, Bayes B, Lasso Bayesiano, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005.

Figura 1 – Efeitos verdadeiros e estimados no cenário oligogênico aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano. Pontos coloridos representam os 12 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.

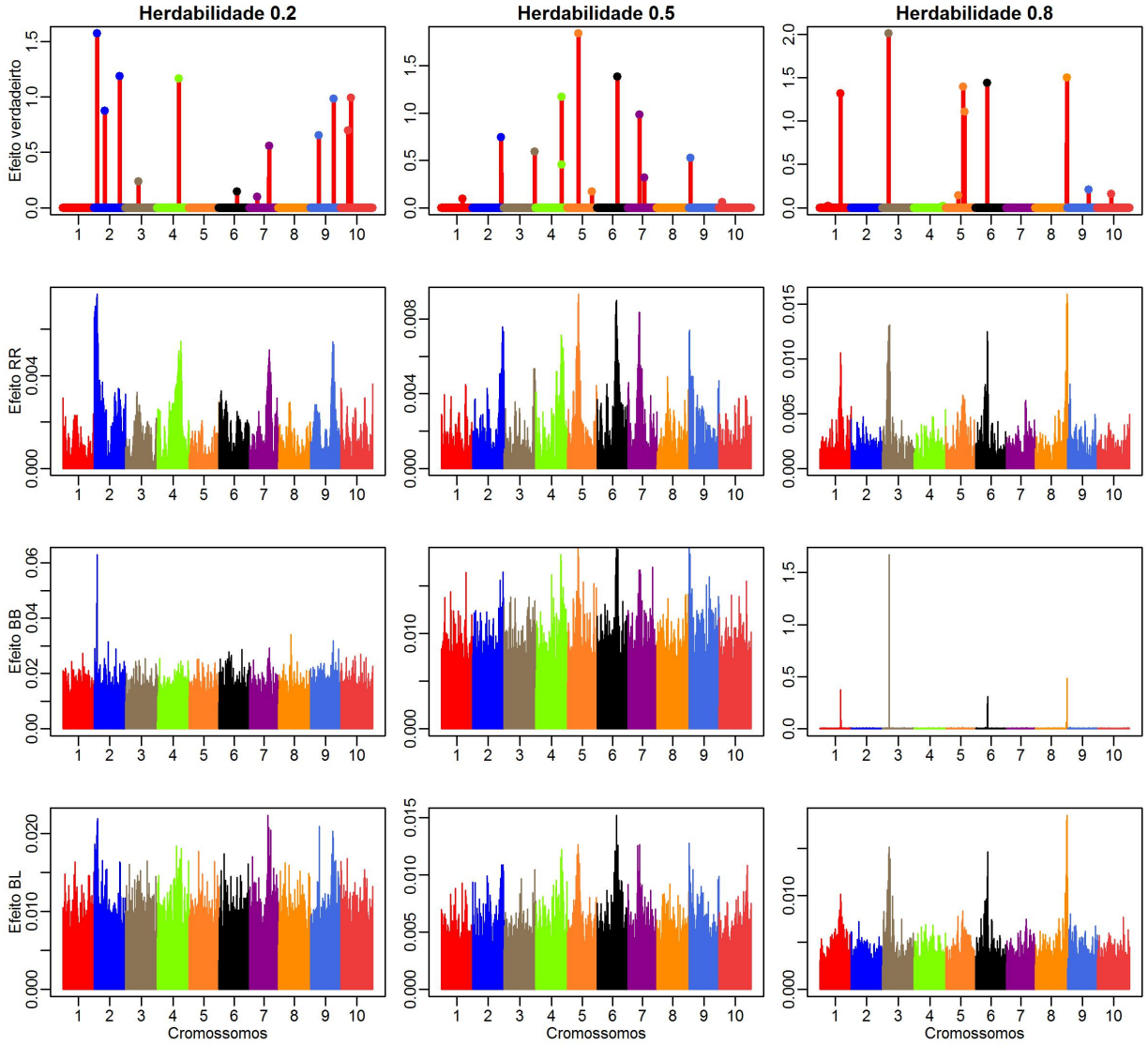


Figura 2 – Efeitos verdadeiros e estimados no cenário oligogênico, aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin 0.01, Bin 0.0066 e Bin 0.005. Pontos coloridos representam os 12 QTLs distribuídos em 12150 SNPs em 10 grupos de ligação.

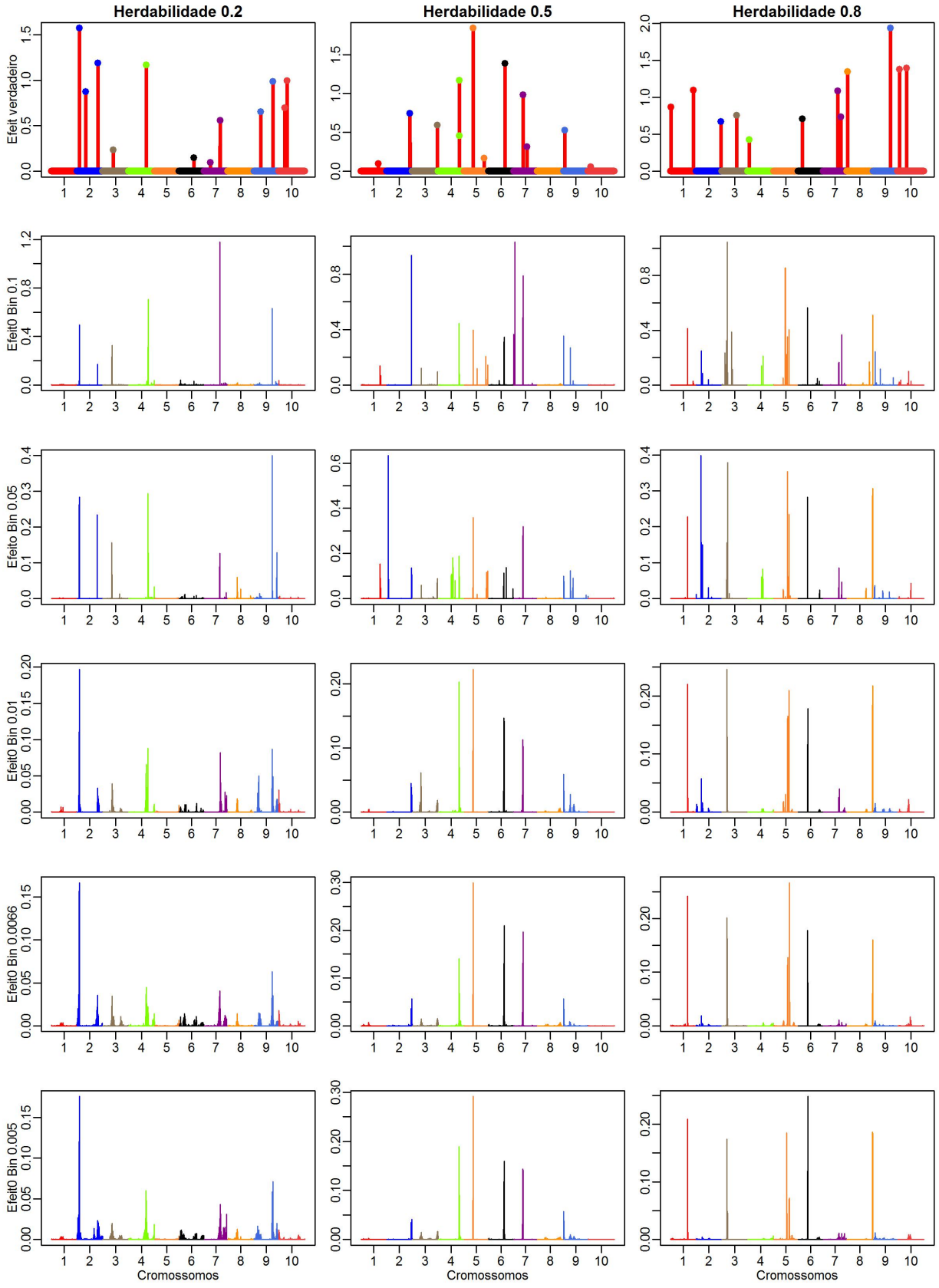


Figura 3 – Efeitos verdadeiros e estimados no cenário poligênico aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano. Pontos coloridos representam os 120 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.

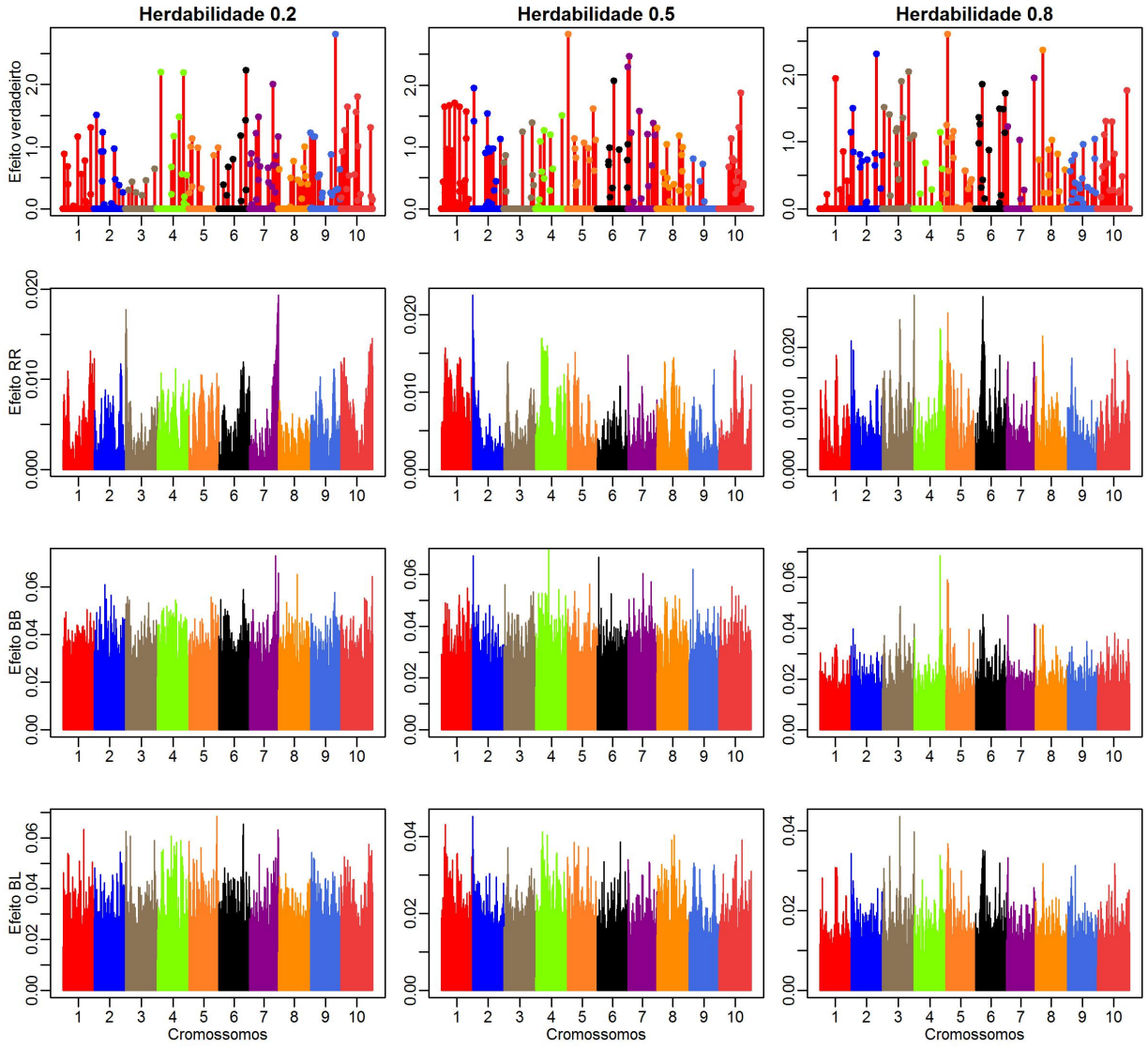


Figura 4 – Efeito no cenário poligênico com herdabilidades (0,2; 0,5 e 0,8) aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin .01, Bin 0.0066 e Bin 0.005. Pontos coloridos representam os 120 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.

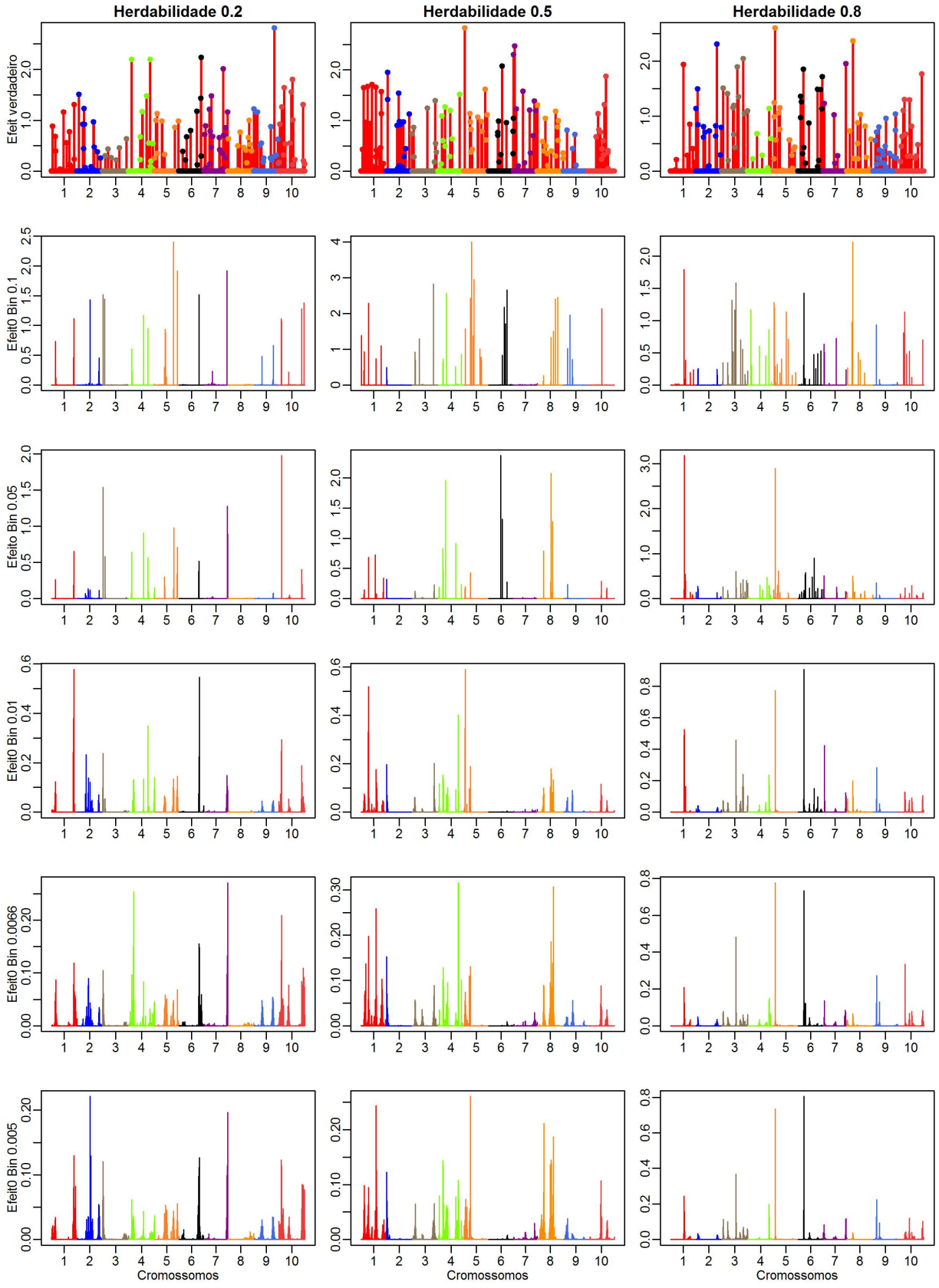


Figura 5 – Efeitos verdadeiros e estimados no cenário infinitesimal aos métodos, respectivamente, RR-BLUP, Bayes B e Lasso Bayesiano. Pontos coloridos representam os 600 QTLs distribuídos em 12150 SNPs distribuídos em 10 grupos de ligação.

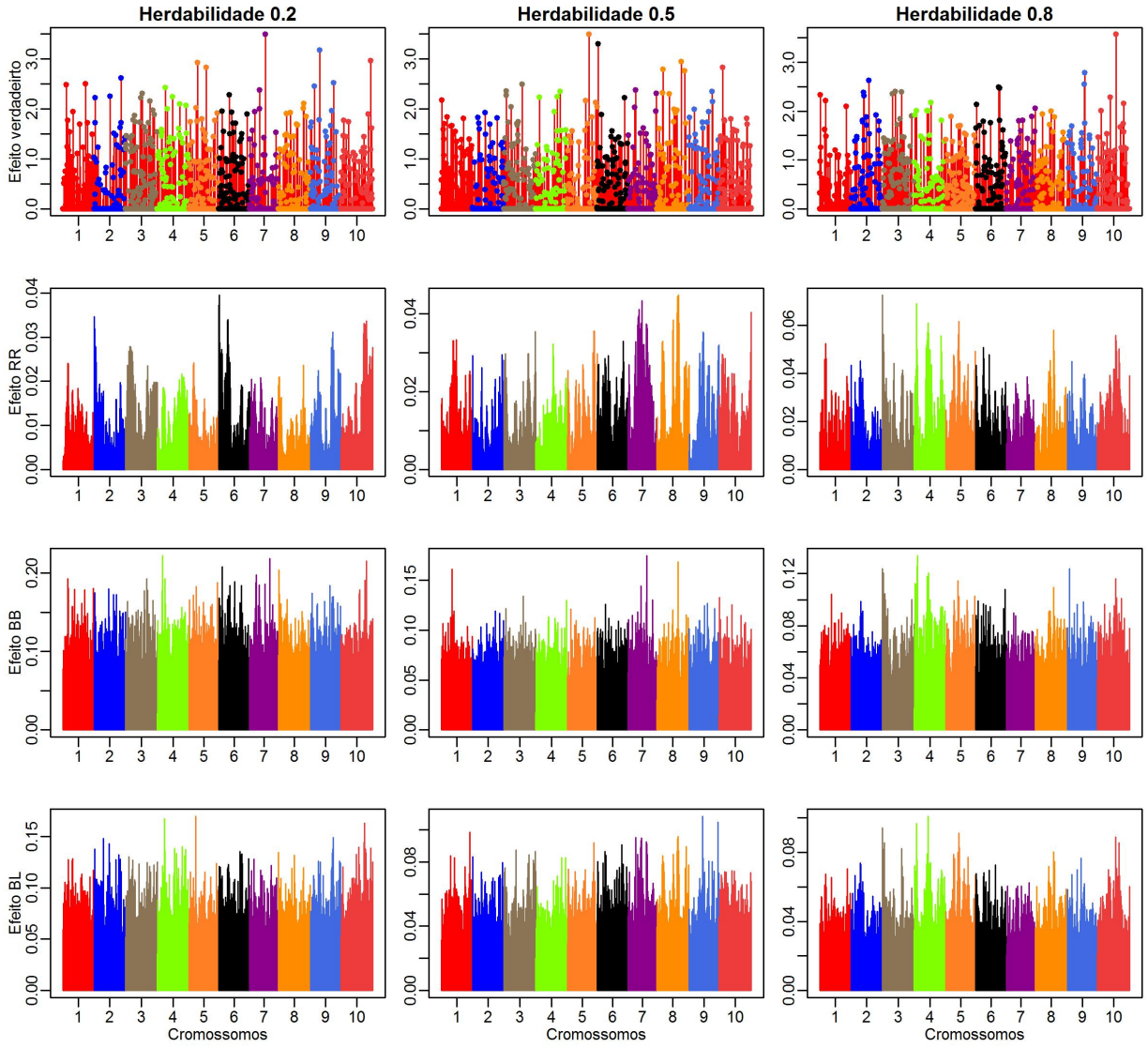
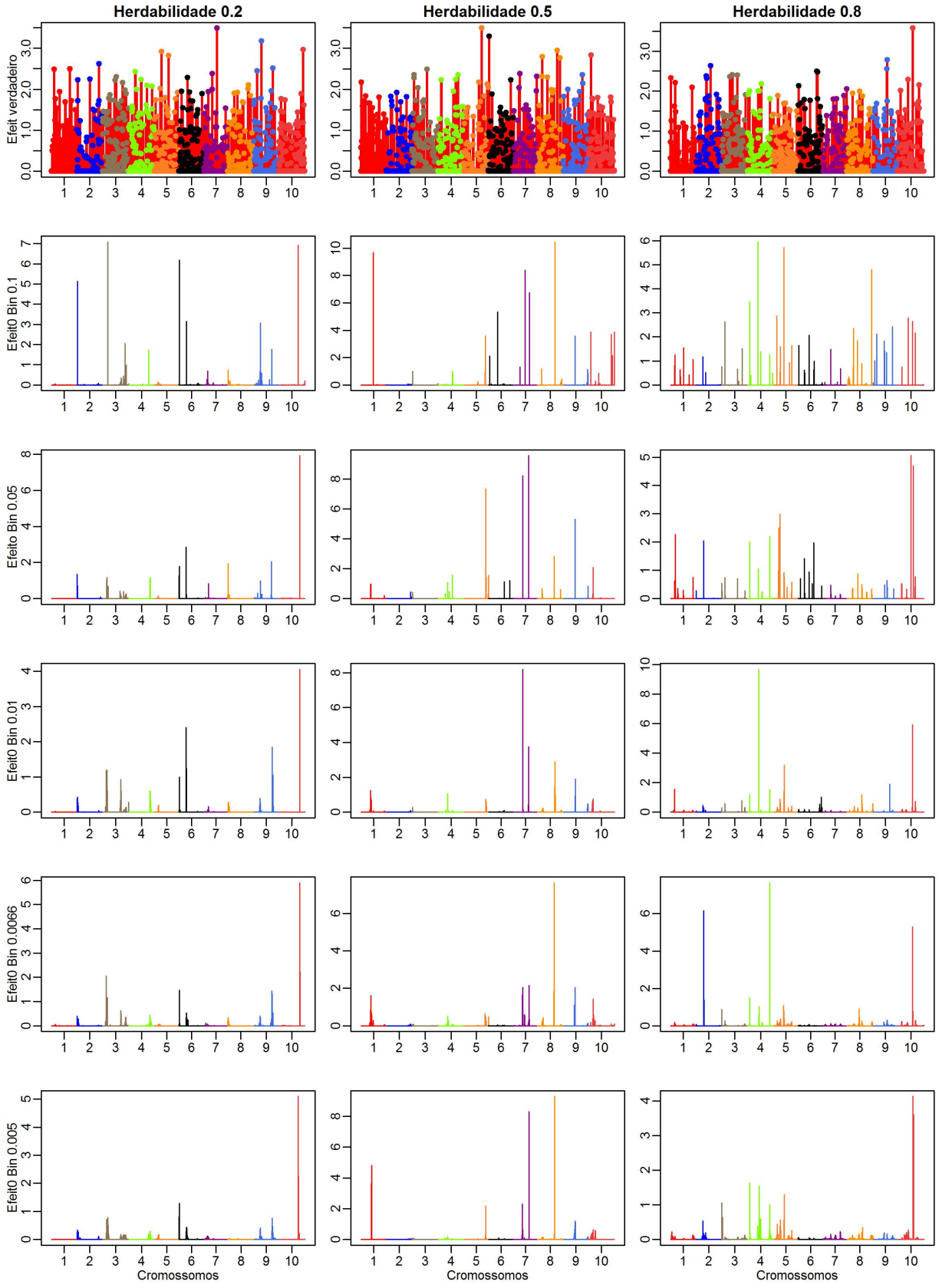


Figura 6 – Efeito no cenário infinitesimal com herdabilidades (0,2; 0,5 e 0,8) aos modelos, respectivamente, Bin 0.1, Bin 0.05, Bin0.01 Bin 0.0066 e Bin 0.005. Pontos coloridos representam os 600 QTLs distribuídos em 12150 SNPs em 10 grupos de ligação.



APÊNDICE B –

A seguir são apresentados os gráficos de frequências relativas dos marcadores que foram visitados pelo algoritmo, no cenário oligogênico, no modelo Bin 0.1 e Bin 0.005, que são, respectivamente, o menor e a maior configuração de modelos bins no presente estudo.

Figura 7 – Frequência relativa do modelo Bin 0.1, no cenário oligogênico.

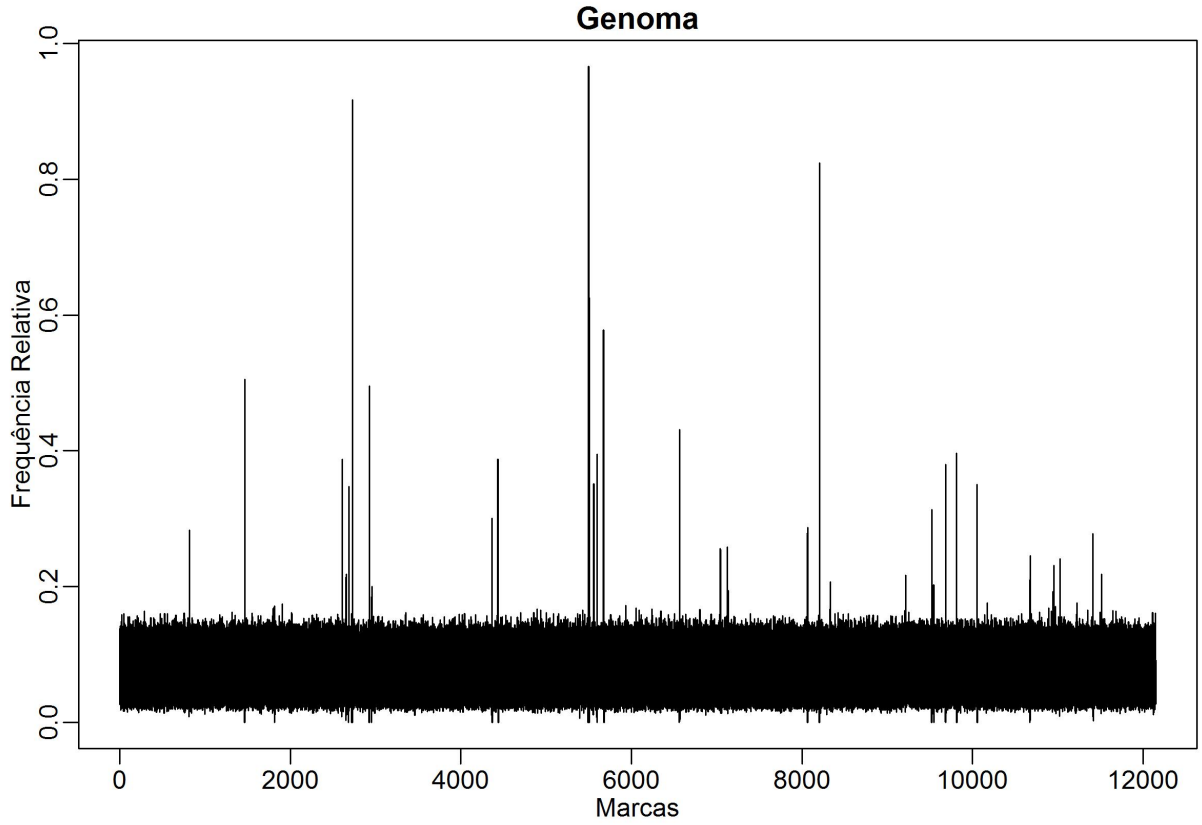


Figura 8 – Frequência relativa dos bins (1; 294; 550 e 1661) do modelo Bin 0.1, no cenário oligogênico.

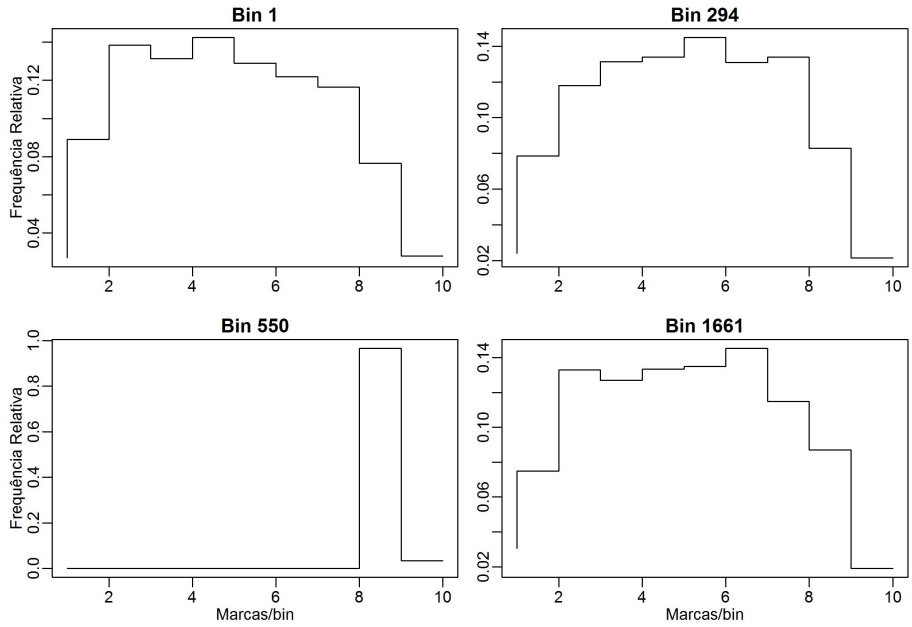


Figura 9 – Frequência relativa do modelo Bin 0.005, no cenário oligogênico.

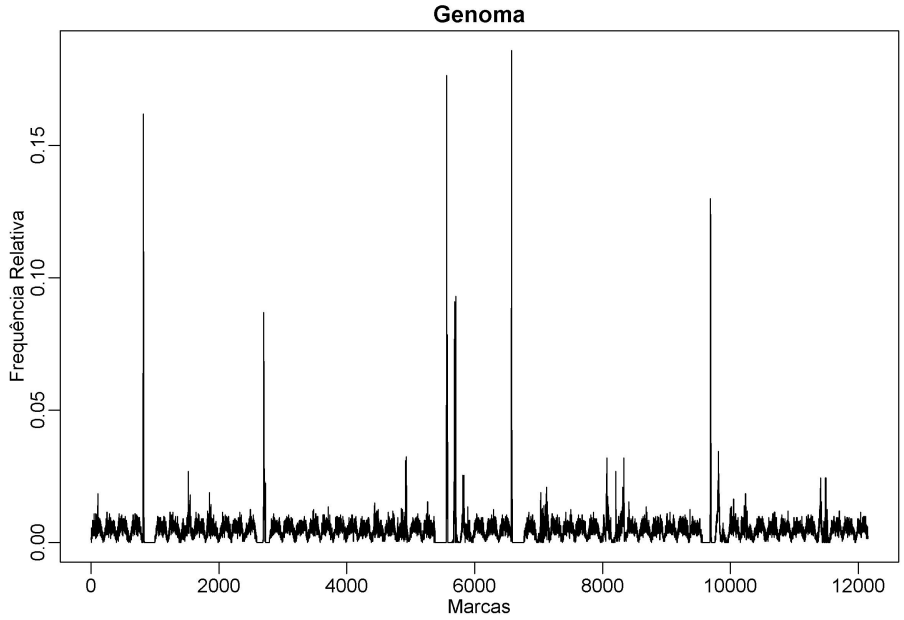


Figura 10 – Frequência relativa dos bins (1; 15; 30 e 59) do modelo Bin 0.1, no cenário oligogênico.