ANDREZZA KÉLLEN ALVES PAMPLONA

SPLINES E MODELO FUNCIONAL EM *BINS*:
ABORDAGENS INTEGRADAS À SELEÇÃO GENÔMICA

LAVRAS - MG

2018

ANDREZZA KÉLLEN ALVES PAMPLONA

***SPLINES* E MODELO FUNCIONAL EM *BINS*:** ABORDAGENS INTEGRADAS À SELEÇÃO GENÔMICA

Tese apresentada 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 Doutora.

Dr. Júlio Sílvio de Sousa Bueno Filho
Orientador

Dr. Marcio Balestre
Coorientador

LAVRAS - MG
2018

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

Pamplona, Andrezza Kéllen Alves.

Splines e modelo funcional em *bins* : abordagens integradas à seleção genômica / Andrezza Kéllen Alves Pamplona. - 2018.
143 p. : il.

Orientador(a): Júlio Sílvio de Sousa Bueno Filho.
Coorientador(a): Marcio Balestre.
Tese (doutorado) - Universidade Federal de Lavras, 2018.
Bibliografia.

1. Modelo funcional bayesiano. 2. Funções B-Spline. 3. Modelo genoma contínuo. I. Bueno Filho, Júlio Sílvio de Sousa. II. Balestre, Marcio. III. Título.

ANDREZZA KÉLLEN ALVES PAMPLONA

***SPLINES* E MODELO FUNCIONAL EM *BINS*:** ABORDAGENS INTEGRADAS À SELEÇÃO GENÔMICA

SPLINES AND FUNCTIONAL MODEL IN BINS: INTEGRATED APPROACHES TO GENOMIC SELECTION

Tese apresentada 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 Doutora.

APROVADA em 28 de junho de 2018.

Dr. Daniel Furtado Ferreira	UFLA-MG
Dr. José Airton Rodrigues Nunes	UFLA-MG
Dra. Maria Imaculada de Sousa Silva	UFU-MG
Dr. Antonio Augusto Franco Garcia	ESALQ-SP

Dr. Júlio Sílvio de Sousa Bueno Filho
Orientador

Dr. Marcio Balestre
Coorientador

LAVRAS - MG
2018

À minha família, pelo apoio e carinho em todas as etapas e por ser minha referência de vida.

Aos meus avós, Maria Emília e Waldiner (In Memorian), por sempre acreditarem em mim e serem um exemplo de amor.

DEDICO

AGRADECIMENTOS

A Deus, o maior mestre, que permitiu que tudo isso acontecesse, me dando força para superar as dificuldades e me enviando boas vibrações.

Aos meus pais, *Waldemar Pamplona da Silva* e *Rozâna Alves da Silva Pamplona*, e à minha irmã, G*reicy Kelly Alves Pamplona*, pelo amor, incentivo e apoio incondicionais, por estarem sempre a postos quando preciso e por todo o esforço que me permitiu estar aqui.

À minha avó-madrinha, *Maria Emília Ramos da Silva*, que, mesmo distante, me acompanhou nesta jornada com orações, me alimentando de certezas, força e paciência. Ao meu avô, *Waldiner Alves da Silva* (*In Memorian*), que, infelizmente, não presenciou a finalização desta etapa, mas sempre esteve ao meu lado transmitindo pensamentos positivos e transbordando seu amor de avô-padrinho.

Ao meu irmão, *Christian Darwin Alves Pamplona* (*In Memorian*), e ao meu avô, *José Pamplona da Silva* (*In Memorian*), que mesmo ausentes fisicamente, acredito estarem sempre comigo espiritualmente.

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

Aos meus amigos de longe, em especial *Weila Freitas* e *Rogério Reis dos Anjos*, pela preocupação demonstrada por meio de ligações, orações e emails.

Aos meus amigos e colegas acadêmicos, em especial *Carlos Pereira da Silva*, *Luciano Antonio de Oliveira*, *Fernando Ribeiro Cassiano*, *Ernandes Guedes Moura*, *Guilherme de Jong* e *Joel Jorge Nuvunga*, pelas amizades, apoios e ajudas em todos os momentos.

Ao meu orientador, *Júlio Sílvio de Sousa Bueno Filho*, por me orientar e proporcionar diversos conhecimentos.

Ao meu coorientador, *Marcio Balestre*, por se dispor a trabalhar comigo e pelo enorme auxílio neste trabalho no pouco tempo que lhe coube.

Aos membros da banca, *Daniel Ferreira Furtado*, *José Airton Rodrigues Nunes*, *Maria Imaculada de Sousa Santos* e *Antonio Augusto Franco Garcia*, pelas disponibilidades e contribuições oferecidas neste trabalho.

Aos diversos *professores* do Departamento de Estatística da UFLA, que fizeram parte da minha formação acadêmica.

Aos funcionários do Departamento de Estatística, em especial à secretária de Pós-Graduação *Nádia Ferreira*, pela amizade e ajuda neste processo.

À Universidade Federal de Lavras (UFLA) e ao Departamento de Estatística (DES), pela oportunidade de realização do doutorado.

À CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - pela concessão da bolsa de estudos.

Ao Instituto Federal do Triângulo Mineiro - *campus* Uberaba, pelo apoio e incentivo a fim de finalizar esta etapa da minha vida.

A todos que, diretamente ou indiretamente, me apoiaram e ajudaram nesta jornada.

MUITO OBRIGADA !

"O saber a gente aprende com os mestres e os livros. A sabedoria se aprende é com a vida e com os humildes." (Cora Coralina)

RESUMO

O modelo genoma contínuo para redução de dimensionalidade e multicolinearidade é baseado na divisão do genoma em janelas (*bins*) e assume-se um modelo funcional no qual a expressão gênica em um marcador é função desconhecida da posição no genoma (função sinal). O desafio principal é obter uma forma funcional relacionando fenótipos a genótipos de marcadores (vistos como milhares de covariáveis) e a valores genéticos. A abordagem mais simples é utilizar a média do estado genotípico do marcador dentro dos *bins* como medida de informação para prever valores genômicos de novos indivíduos. Duas alternativas foram propostas neste trabalho: primeira, incorporar a ideia de pesos aos efeitos dentro de *bins* usando a frequência relativa com que cada marcador é amostrado em uma cadeia de Markov, juntamente com o modelo funcional em *bins*, aos métodos de seleção genômica clássicos; segunda, obter uma expressão polinomial, por meio de técnicas de análise de dados funcionais, que represente a expressão gênica na seleção genômica. Ilustrou-se a adaptação dos métodos RR-BLUP, Bayes A e Bayes B sob a versão bayesiana do modelo funcional, sendo que se utilizaram suas formas originais como padrões para comparações. Além disso, foram utilizadas funções B-*Spline* para estimar a função sinal. Ambas alternativas apresentaram resultados satisfatórios, retornando análises em menor tempo computacional comparado aos originais. Modelos funcionais são muito atrativos e podem ser utilizados como princípios unificadores para seleção e localização de genes.

Palavras-chave: Inferência Bayesiana. Modelo funcional bayesiano. Funções B-Spline. Modelo genoma contínuo.

ABSTRACT

The continuous genome model for dimensionality and multicollinearity reduction is based on the dividing genome into windows (bins) and then it assumes a functional model in which the gene expression in a marker is an unknown function of the position in the genome (signal function). The main challenge is to obtain a functional form by relating phenotypes to marker genotypes (seen as thousands of covariates) and to genetic values. The simplest approach is to use the average genotype status of the marker within the bins as an information measure to predict genomic values of new individuals. This study proposed two alternatives: the first one was to incorporate the idea of weights into the effects within bins using the relative frequency with which each marker is sampled in a Markov chain, along with the functional model in bins, to classic genomic selection methods; the second one was to obtain a polynomial expression, by means of functional data analysis techniques, that represents the gene expression in the genomic selection. The adaptation of the RR-BLUP, Bayes A, and Bayes B methods was illustrated under the Bayesian version of the functional model, and their original forms were used as standards for comparisons. In addition, B-Spline functions were used to estimate the signal function. Both alternatives presented satisfactory results, returning analyzes in less computational time compared to the originals. Functional models are very attractive and can be used as unifying principles for selection and localization of genes.

Keywords: Bayesian Inference. Bayesian functional model. B-Spline Functions. Continuous genome model.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Verifica-se que, na seleção genômica, os modelos de regressão apresentam problemas com o uso de grandes painéis de marcadores, já que quanto maior o número de marcadores disponíveis, mais severos os problemas de multicolinearidade e dimensionalidade e maior a demanda computacional na estimação dos efeitos. Hu, Wang e Xu (2012) propuseram o modelo genoma contínuo com *bins*, cuja ideia é dividir o genoma em um número finito de intervalos e obter um genótipo médio representativo dentro de cada um destes intervalos. Isto levou à redução da dimensionalidade do modelo e dos efeitos da multicolinearidade.

A técnica de Hu, Wang e Xu (2012) e suas propriedades foram investigadas por Xu (2013a). Neste modelo, assume-se que a expressão gênica de um marcador é uma função (sinal) desconhecida da posição no genoma (quantidade contínua). O interesse, então, é estimar tal função e usá-la para predizer o valor genético genômico de novos indivíduos. No entanto, em vez de buscar a função, Hu, Wang e Xu (2012) utilizaram médias aritméticas dos *bins* como medida de informação. Visto que o efeito do k-ésimo *bin* é a soma dos efeitos de todos os marcadores dentro deste *bin*, duas hipóteses são necessárias para o modelo genoma contínuo ser trabalhado: alto desequilíbrio de ligação e efeitos homogêneos de marcadores dentro do *bin*. Se uma das condições não é satisfeita, o modelo apresentado é problemático e necessita-se de adaptá-lo, como apresentado em Hu, Wang e Xu (2012).

Uma alternativa foi proposta por Moura (2017), na qual se atribuem pesos para os efeitos dentro de um *bin* por meio da frequência relativa com que cada marcador é assumido possuir efeito dentro de um processo estocástico Monte Carlo via Cadeias de Markov (MCMC - *Monte Carlo Markov Chain*). Neste caso, o modelo genoma contínuo pode lidar, também, com populações em baixo desequi-

líbrio de ligação e efeitos heterogêneos dentro do *bin* (já que a ponderação pelos pesos evitaria o cancelamento de efeitos em direções opostas). Assume-se que os marcadores são variáveis aleatórias discretas e um marcador é selecionado dentro de cada *bin*, a cada passo do método estocástico. Isto permite obter um tipo de média ponderada, por meio dos *bins*, em vez de um efeito médio nos *bins*. Os resultados apresentados pelo autor quanto à capacidade preditiva desta técnica foram satisfatórios, o que levou à ideia de estendê-la aos métodos de seleção genômica.

Mais ainda, sabendo que a ideia de modelos funcionais em seleção genômica consiste no pressuposto de que a expressão gênica segue uma série espacial, técnicas de análise de dados funcionais podem ser utilizadas para estimar a função sinal. Quando a forma da função subjacente aos dados é complicada, é difícil aproximá-la por meio de um único polinômio. Nesse caso, *Spline* é uma das funções de aproximação mais apropriadas (BOOR, 1978; DIERCKX, 1993).

Assim, com base no exposto, o presente trabalho tem os seguintes objetivos:

a) Propor a adaptação dos métodos RR-BLUP, Bayes A e Bayes B sob o modelo funcional em *bins*.

b) Comparar os métodos adaptados com suas versões originais quanto à capacidade preditiva.

c) Avaliar os métodos adaptados quanto à capacidade em detectar QTL (*Quantitative Trait Loci*).

d) Utilizar curvas B-*Splines* para estimar a função sinal do modelo funcional.

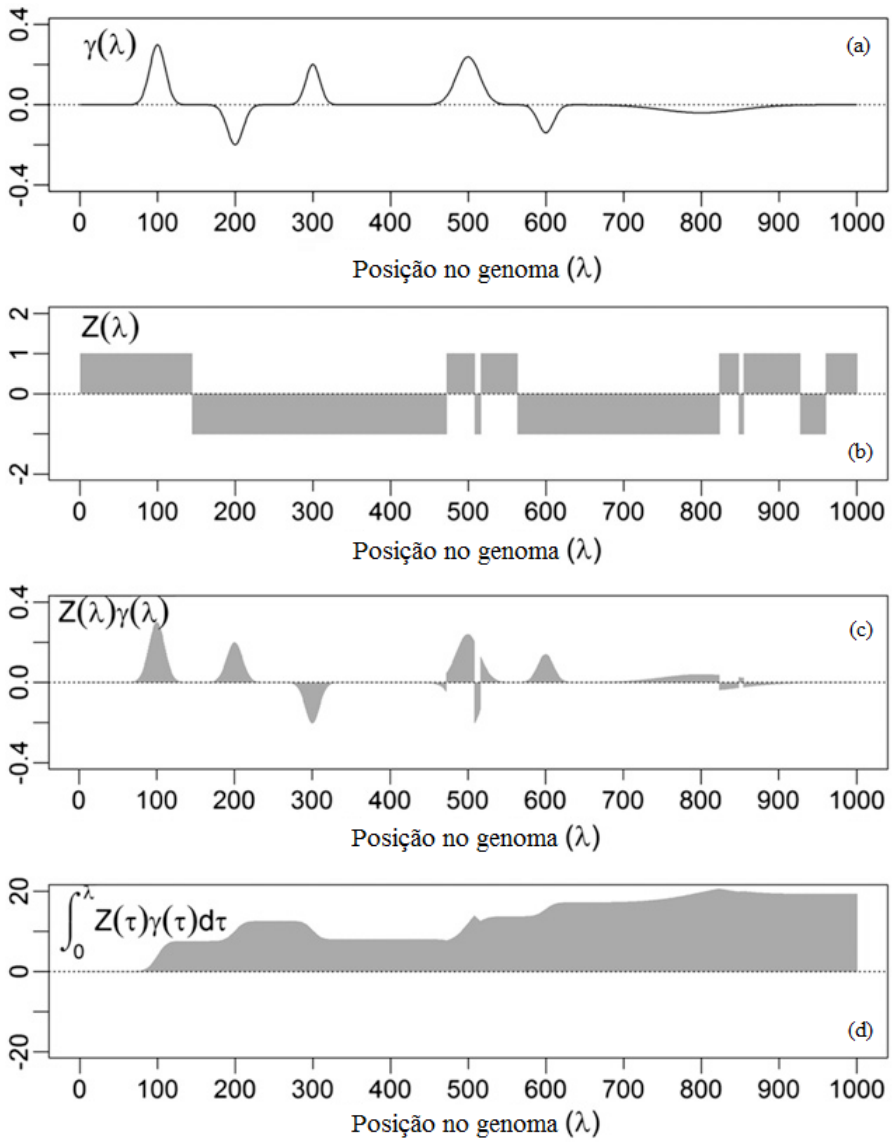
2 JUSTIFICATIVA

A aplicação de modelos funcionais à seleção genômica parte da ideia de que um cromossomo pode ser visto como uma sequência no espaço (genoma). Em virtude da grande quantidade de marcadores distribuídos no genoma, suas posições podem ser consideradas variáveis contínuas. Assim, as expressões gênicas dos marcadores podem ser tratadas como dados funcionais que variam de acordo com suas posições no genoma.

Hu, Wang e Xu (2012) e Xu (2013a) trataram os efeitos genéticos dos marcadores como medidas ruidosas de uma função subjacente (função sinal desconhecida) que varia ao longo do genoma. Em outras palavras, $f(\lambda) = \gamma$, sendo λ a posição no genoma e γ o sinal da expressão do gene em dada posição. O interesse destes autores foi estimar essa função sinal e, a partir dela, predizer o valor genético genômico de novos indivíduos. O modelo definido por estes autores é um tipo de modelo linear funcional em que a variável resposta é um escalar e a covariável é uma função. Na Figura 2.1 apresenta-se um exemplo das formas de algumas variáveis expressas como funções da posição no genoma (XU, 2013a).

A abordagem com modelo funcional fornece resultados que estão mais próximos da realidade biológica pela incorporação de princípios biológicos por trás das funções. Portanto, a possibilidade de tratar efeitos genéticos de marcadores como dados funcionais deve ser considerada em modelos de predição de valor genômico.

Figura 2.1 – Exemplo das formas de algumas variáveis expressas como funções da posição λ no genoma. (a) sinal do efeito genético do marcador; (b) genótipo do marcador, neste caso $\{-1, 1\}$; (c) multiplicação da função sinal com o genótipo do marcador; (d) valor genômico de um indivíduo.



Fonte: Xu (2013a).

3 REFERENCIAL TEÓRICO

Uma melhoria genética eficiente das populações de melhoramento equilibra ganho genético com seu tempo e custo. A seleção genômica oferece novos caminhos para aumentar a eficiência dos programas de melhoramento por apresentar a capacidade de selecionar indivíduos com os maiores valores genômicos sem a necessidade de coletar fenótipos pertencentes a esses indivíduos. Isto pode reduzir o ciclo de melhoramento e aumentar a resposta à seleção no tempo, já que a informação genômica pode ser utilizada para otimizar e, por conseguinte, reduzir o volume de dados fenotípicos que são necessários para tomar decisões precisas de seleção (HICKEY et al., 2014).

Entretanto, esse esquema de seleção genômica utiliza alta densidade de marcadores moleculares no genoma e poucos indivíduos da população em estudo, acarretando alguns problemas de modelagem, o que levou ao desenvolvimento de diversas metodologias estatísticas para lidar com eles.

3.1 Métodos de seleção genômica

A seleção genômica (GS - *Genomic Selection*) foi proposta por Meuwissen, Hayes e Goddard (2001) e representa uma forma de seleção assistida por marcadores na qual a informação de dados fenotípicos e genotípicos de indivíduos de uma população escolhida é utilizada para predizer valores genéticos genômicos de indivíduos que foram genotipados, mas não fenotipados.

A seleção genômica permite aos melhoristas selecionarem indivíduos, a partir do desempenho predito em vez de observá-lo, o que pode, por exemplo, reduzir potencialmente os custos de um programa de melhoramento animal em torno de 90%, segundo Schaeffer (2006). Além disso, de acordo com Zhou et

al. (2017), os benefícios dessas predições incluem melhorar a eficiência no melhoramento de plantas e animais para aumentar a sustentabilidade agrícola. Outra grande vantagem da utilização da GS, na concepção de melhoramento de plantas, é a aceleração da melhoria genética por unidade de tempo por meio da redução do tempo necessário para completar os ciclos de reprodução (HICKEY et al., 2014).

Para conduzir uma GS, primeiramente, indivíduos de uma "população de treinamento" são genotipados e fenotipados e os efeitos dos QTL associados com todos os marcadores do genoma são estimados com base nesses dados. Em seguida, estes efeitos estimados são utilizados para predizer os valores genéticos genômicos (VGG) de indivíduos de uma "população reprodutora" que foram genotipados, mas não fenotipados. Os indivíduos/linhagens superiores são selecionados da população reprodutora com base nos valores genéticos genômicos. O VGG de um indivíduo é a soma dos efeitos de todos os QTL associados aos marcadores presentes no genoma do indivíduo e incluídos no modelo GWS (*Genome Wide Selection*) aplicado à população sob seleção. A eficiência da seleção genômica depende da correlação entre o valor genotípico predito e o valor genotípico verdadeiro correspondente (COMBS; BERNARDO, 2013; GODDARD; HAYES, 2007; SINGH; SINGH, 2015).

Como o número de efeitos (preditores) é muito maior que o número de observações, os graus de liberdade disponíveis não são suficientes para ajustá-los, simultaneamente, utilizando regressão linear padrão (mínimos quadrados) e, ainda, pode haver um elevado grau de correlação ou multicolinearidade entre tais efeitos (JANNINK; LORENZ; IWATA, 2010; MEUWISSEN; HAYES; GODDARD, 2001; SINGH; SINGH, 2015). Para resolver estes problemas, métodos de predição que diferem com relação às hipóteses sobre os efeitos dos marcadores têm sido propostos, tais como os relacionados a seguir.

3.1.1 RR-BLUP

O método *ridge regression* (RR) foi proposto por Whittaker, Thompson e Denham (2000) para seleção assistida por marcadores em populações biparentais, na qual assume-se que os efeitos dos QTL associados aos marcadores têm variâncias iguais. Meuwissen, Hayes e Goddard (2001) propuseram modelar os efeitos como aleatórios e calcular seus BLUPs (*best linear unbiased predictors*), o que foi denominado de RR-BLUP.

Neste método, os efeitos são considerados como pertencentes a uma distribuição normal com média zero e variância comum, $N(0, \sigma_a^2)$, sendo que σ_a^2 é obtida dividindo a variância genética (conhecida) pelo número de efeitos. O modelo básico é:

$$\mathbf{y} = \mathbf{Z}\mathbf{a} + \mathbf{e}, \tag{3.1}$$

em que $\mathbf{y}$ é o vetor de observações, $\mathbf{a} \sim N(0, \mathbf{I}\sigma_a^2)$ é o vetor de efeitos dos marcadores, $\mathbf{Z}$ é a matriz de estado genotípico dos marcadores; $\mathbf{e} \sim N(0, \sigma_e^2)$ é o vetor de erros. A solução BLUP para os efeitos é escrita como $\hat{\mathbf{a}} = (\mathbf{Z}'\mathbf{Z} + \lambda\mathbf{I})^{-1}\,\boldsymbol{Z'y}$, com $\lambda = \frac{\sigma_e^2}{\sigma_a^2}$ (fator de encolhimento).

Note que assumir que todos os efeitos seguem a mesma distribuição não significa que eles são todos iguais, mas que eles são igualmente encolhidos para próximos de zero. O principal problema com BLUP é que marcadores com efeitos grandes são muito fortemente encolhidos para próximo de zero. Outro fato relevante é que a hipótese de variâncias iguais explicadas por todos os marcadores é biologicamente incorreta, mas torna a estatística robusta por limitar o número de parâmetros desconhecidos (JANNINK; LORENZ; IWATA, 2010; MEUWISSEN; HAYES; GODDARD, 2001; SINGH; SINGH, 2015; ZHANG; ZHANG et al., 2016).

Percebendo que existem muitos marcadores com efeitos pequenos e raros marcadores com efeitos grandes para certos caracteres, ou seja, que eles não contribuem igualmente para a variância genética (HAYES, GODDARD, 2001), seria adequado especificar distribuições prioritárias para os efeitos, de modo que os pequenos efeitos (ignoráveis) sejam levados a zero e somente aqueles com maiores efeitos sobre o fenótipo sejam ajustados no modelo.

3.1.2 Bayes A

A suposição de variâncias genéticas comuns não foi satisfatória e Meuwissen, Hayes e Goddard (2001) resolveram relaxar tal suposição usando análise bayesiana. Neste caso, assume-se que os marcadores podem contribuir diferencialmente para a variância genética e, assim, parece ser uma boa ideia estimar variâncias específicas para os efeitos em vez de adotar uma variância comum.

No método Bayes A, a partir do modelo (3.1), assume-se uma priori condicional Normal para cada efeito, todas com média zero e variância específica, $a_k|\sigma^2_{a_k} \sim N(0,\sigma^2_{a_k})$, com k = 1, ..., p marcadores. Já para as variâncias associadas a esses efeitos, assume-se distribuição qui-quadrado escalada invertida, $\sigma^2_{a_k}|\nu_{a_k},S^2_{a_k} \sim \chi^{-2}_{esc}\left(\nu_{a_k},S^2_{a_k}\right)$, sendo $S^2_{a_k}$ um parâmetro de escala e ν_{a_k} o número de graus de liberdade, conhecidos e arbitrários. A distribuição a priori para a variância residual é $\sigma^2_e \sim \chi^{-2}_{esc}\left(\nu_e = -2,S^2_e = 0\right)$, que resulta em uma priori uniforme e gera resultados similares à análise de máxima verossimilhança (MEUWISSEN; HAYES; GODDARD, 2001; XU; HU, 2010). Com isso, a distribuição a posteriori para $\sigma^2_{a_k}$ é uma qui-quadrado escalada invertida, $\sigma^2_{a_k}|a_k,\nu_a,S^2_a \sim \chi^{-2}_{esc}\left(\nu_a + 1,S^2_a + a^2_k\right)$, assim como a distribuição a posteriori para $\sigma^2_e \rightarrow \sigma^2_e|\mathbf{e} \sim \chi^{-2}_{esc}\left(\nu_e + n,\mathbf{e}'\mathbf{e}\right)$.

Entretanto, a priori adotada para σ^2_e não é correta por duas razões, de

acordo com Gianola (2013): primeira, uma distribuição qui-quadrado escalada invertida existe somente se ν_e e S_e^2 são maiores que zero; segunda, uma distribuição qui-quadrado escalada invertida com um parâmetro de escala nulo não é um modelo de probabilidade, uma vez que não transmite a incerteza, já que para qualquer valor da variância residual atribui-se uma densidade zero, anterior e posterior aos dados de observação.

3.1.3 Bayes B

O método Bayes B foi proposto por Meuwissen, Hayes e Goddard (2001), pois afirmaram que, na realidade, existem muitos marcadores sem variância genética (não segregantes) e poucos com variância genética. É um método de seleção de variáveis, já que excluem do modelo aqueles marcadores sem efeito, diminuindo o tempo computacional quando comparado ao do Bayes A (MIAR; PLASTOW; WANG, 2015; THAVAMANIKUMAR; DOLFERUS; THUMMA, 2015). Neste caso, assumem-se, por meio de um modelo de mistura, uma priori que tem alta densidade, π, em $\sigma^2_{a_k} = 0$ e uma distribuição qui-quadrado escalada invertida para $\sigma^2_{a_k} > 0$. Assim,

$$\begin{aligned} \sigma^2_{a_k} &= 0 \qquad\qquad\quad , \quad \text{com probabilidade } \pi; \\ \sigma^2_{a_k} &\sim \chi^{-2}_{esc}\left(\nu_a, S^2_a\right), \quad \text{com probabilidade } (1-\pi), \end{aligned} \tag{3.2}$$

em que $\nu_a = 4{,}234$ e $S^2_a = 0{,}0429$ (MEUWISSEN; HAYES; GODDARD, 2001), resultando na média e na variância de $\sigma^2_{a_k}$ dado que $\sigma^2_{a_k} > 0$. O valor de π é arbitrário e conhecido; se $\pi = 0$, este método retorna o Bayes A.

Meuwissen, Hayes e Goddard (2001) assumiram a priori que $\sigma^2_{a_k} = 0$ implica $a_k = 0$. Entretanto, afirmar que um parâmetro tem variância zero a priori não significa necessariamente que ele tem valor zero: ele poderia ter qualquer

valor, mas conhecido com certeza. Assim, Gianola et al. (2009) criticaram o Bayes B quanto à essa formulação e propuseram colocar uma mistura baseada nos efeitos e não em suas variâncias:

$$a_k|\sigma^2_{a_k} \sim \begin{cases} \text{massa de pontos em } c\,, & \text{se } \sigma^2_{a_k} = 0, \\ N\left(0,\sigma^2_{a_k}\right) \quad , & \text{se } \sigma^2_{a_k} > 0. \end{cases} \tag{3.3}$$

$$\sigma^2_{a_k}|\pi \sim \begin{cases} 0 \quad , & \text{com probabilidade } \pi; \\ \chi^{-2}_{esc}\left(\nu_a,S^2_a\right), & \text{com probabilidade } 1-\pi. \end{cases} \tag{3.4}$$

Marginalmente, depois de integrar em relação à $\sigma^2_{a_k}$, a priori tem a forma descrita em (3.5), sendo c uma constante. Gianola (2013) e Habier et al. (2011) consideraram $c = 0$.

$$p\left(a_k|\pi\right) = \begin{cases} a_k = c \quad , & \text{com probabilidade } \pi; \\ t\left(0,\nu_a,S^2_a\right), & \text{com probabilidade } 1-\pi. \end{cases} \tag{3.5}$$

Mais ainda, Gianola (2013), Gianola et al. (2009) e Habier et al. (2011) comentaram sobre os inconvenientes estatísticos de Bayes A e Bayes B que se centram em torno dos hiperparâmetros das prioris para os efeitos dos marcadores. Pelo fato da condicional a posteriori da variância específica de cada marcador ter somente um grau de liberdade a mais que sua priori em Bayes A (como descrita acima) e a priori em Bayes B ser mais precisa em virtude de a fração π reduzir a variância a priori comparada com a de Bayes A, nenhum dos dois métodos permite aprendizado bayesiano sobre essas variâncias. Como consequência, o encolhimento produzido nos efeitos dos marcadores dependerá fortemente desses hiperparâmetros (de acordo com suposições baseadas na teoria da genética quantitativa).

Para contornar tais inconvenientes, alguns autores sugeriram modificações na metodologia: primeira, assumir variância comum simples para os efeitos de todos os marcadores em vez de variância específica para cada um, o que proporcionaria menor influência dos hiperparâmetros; segundo, tratar a probabilidade π como desconhecida e estimada dos dados, já que depende da arquitetura genética do caráter em estudo (GIANOLA et al., 2009; HABIER et al., 2011; KIZILKAYA; FERNANDO; GARRICK, 2010).

3.2 Modelos funcionais

Dados funcionais vêm em muitas formas, mas a definição é que eles consistem de funções - muitas vezes, mas nem sempre - curvas suaves. O termo foi cunhado por Ramsay e Dalzell (1991) e, conceitualmente, são definidos continuamente. É claro que, na prática, muitas vezes, as funções são amostradas discretamente, mas isso não altera o modo de pensar subjacente (RAMSAY; SILVERMAN, 2002). Isso significa que existe um conjunto finito de pontos distintos com valores funcionais correspondentes amostrados que devem ser usados para estimar a função subjacente.

Existem três casos de modelos funcionais: a variável resposta, as covariáveis ou ambas são curvas (funções).

a) Modelos com resposta funcional e covariáveis escalares: $\mathbf{Z}$ é a matriz de números reais conhecidos e $\mathbf{a}$ é o parâmetro (função) a ser estimado. Sendo s a variável independente, o modelo (3.1) é reescrito por:

$$\mathbf{Y}(s) = \mathbf{Z}\mathbf{a}(s) + \boldsymbol{\varepsilon}(s), \tag{3.6}$$

Exemplos em que se aplica o modelo (3.6) são curvas de temperaturas mé-

dias ou níveis de precipitações medidas, em certas regiões, onde os dados são disponíveis continuamente no tempo, mas são coletados diária ou mensalmente (SILVEIRA NETO, 2012; ZOGLAT, 2008).

b) Modelos com resposta escalar e covariáveis funcionais: é o caso em que uma variável funcional é utilizada para explicar uma variável escalar; com isso, necessita-se acrescentar uma integral para resumir as informações de uma função em um único número (CARDOT; FERRATY; SARDA, 1999). Sendo s a variável independente, o modelo (3.1) é, então, reescrito da seguinte forma:

$$\mathbf{Y} = \int \mathbf{Z}(s)\mathbf{a} = (s)\ ds + \boldsymbol{\varepsilon}, \tag{3.7}$$

Como $\mathbf{Z}(s)$ é um objeto funcional, $\mathbf{a}(s)$ também será. O que $\mathbf{a}(s)$ faz é ponderar o quanto cada ponto de $\mathbf{Z}(s)$ contribuirá para a integral e, com isso, para a estimativa de $\mathbf{Y}$. Muitas vezes, $\mathbf{a}(s)$ traz uma importante informação sobre quais pontos, intervalos ou regiões influenciam na variável resposta. Um exemplo encontrado em Silveira Neto (2012) foi o uso do logaritmo da precipitação anual em 35 estações como variável resposta e a curva de temperatura média diária como covariável, ou seja, o objetivo foi verificar o quanto a temperatura, ao longo dos dias do ano, conseguia explicar a precipitação anual.

c) Modelos totalmente funcionais: este é o caso mais complexo e geral, sendo que uma variável funcional ($\mathbf{Z}$) é utilizada para explicar uma resposta funcional ($\mathbf{Y}$). Neste caso, $\mathbf{a}$ (função) é bivariada, ou seja, uma superfície nos domínios de $\mathbf{Z}$ e $\mathbf{Y}$, e o objetivo principal é estimá-la (CHIOU; MÜLLER; WANG, 2004). Sendo s e t as variáveis independentes, o modelo em (3.1) é

reescrito da seguinte forma:

$$\mathbf{Y}(t) = \int \mathbf{Z}(s)\,\mathbf{a}(s,t)\;\, ds + \boldsymbol{\varepsilon}(t)\,, \tag{3.8}$$

Um exemplo de aplicação é utilizar as curvas de temperatura média diária, ao longo do ano, como variável explicativa para as curvas de precipitação, ao longo do ano, ou seja, tanto a variável resposta quanto a variável independente são funcionais (SILVEIRA NETO, 2012).

3.2.1 Modelo genoma contínuo

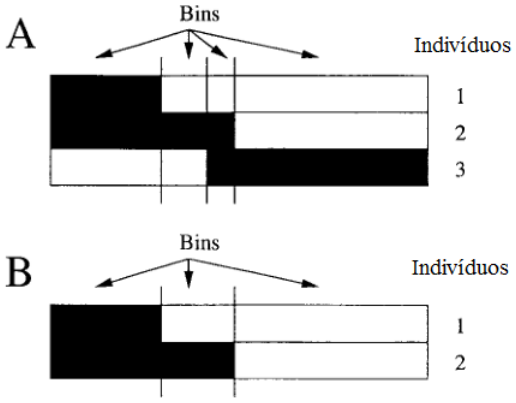
Em razão do desenvolvimento de tecnologias de sequenciamento, pode-se obter uma quantidade muito grande de marcadores. Se cada marcador é tratado como um QTL, o modelo é virtualmente um modelo infinitesimal, não sendo possível estimar todos os efeitos conjuntamente num único modelo. Estratégias para redução da dimensionalidade do modelo são necessárias para facilitar as análises, tanto do ponto de vista estatístico quanto biológico.

Uma abordagem mais recente de redução de dimensão de modelo foi proposta por Hu, Wang e Xu (2012). Os autores desenvolveram um modelo infinitesimal baseado em marcadores e utilizam a informação sobre a segregação de cada marcador no genoma: particiona-se o genoma em um número finito de *bins*, tornando o problema de análise de marcadores em análise *bin*, tal que a dimensão do modelo diminui de um número virtual infinito para um número finito de *bins*. Com isso, pode-se manusear eficientemente um número ilimitado de marcadores sem fazer seleção. Segundo os autores, essa abordagem nunca foi proposta antes para a predição genômica.

A primeira definição de *bin* foi dada por Vision et al. (2000), como sendo um intervalo ao longo de um grupo de ligação limitado por *breakpoints* (pontos

de interrupção) de recombinação em pelo menos um indivíduo (Figura 3.1), sendo que todos os marcadores dentro do mesmo *bin* têm exatamente os mesmos genótipos (XU, 2013b).

Figura 3.1 – Representação esquemática do conceito de *bin*. Limites entre áreas sombreadas e não sombreadas representam pontos de interrupção. **(A)** A inclusão dos três indivíduos quebra o intervalo em quatro *bins* de comprimentos desiguais. **(B)** Remover o terceiro indivíduo provoca a perda do *bin* menor, mas o comprimento máximo do *bin* permanece inalterado.



Fonte: Vision et al. (2000).

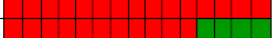
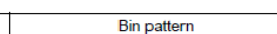
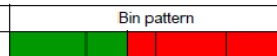
Como não há *breakpoints* permitidos dentro de um *bin*, os tamanhos de *bins* naturais variam aleatoriamente de muito pequeno a muito grande, dependendo do tamanho da amostra. Os *bins* naturais também são específicos da amostra. Introduzir (ou excluir) um indivíduo na amostra atual pode produzir novos *breakpoints* e, com isso, novos *bins*. Na Figura 3.1 mostra-se que enquanto a exclusão de um ou mais indivíduos pode resultar em menos *bins*, isso não precisa resultar em um aumento no tamanho máximo do *bin*. Assume-se que os indivíduos são haploides (VISION et al., 2000).

Xu (2013a) estendeu a definição de *bin* para permitir que *breakpoints*

aconteçam dentro de *bins*, os chamados *bins* artificiais (Figura 3.2-b). Os tamanhos dos *bins* artificiais podem ser arbitrariamente ajustados de acordo com a preferência do investigador. Além disso, a adição de novos indivíduos não mudará os *bins* previamente definidos. Portanto, de acordo com o autor, a análise de *bins* artificiais pode facilitar a seleção genômica.

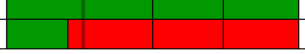
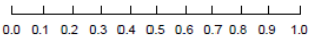
Figura 3.2 – O padrão de *breakpoints* de recombinação de um genoma hipotético de 1,0 M de comprimento em uma população consistindo de 15 linhas. **(a)** Quinze *bins* naturais e os respectivos genótipos. **(b)** Quatro *bins* artificiais igualmente espaçados e os respectivos genótipos.

(a)

Line	Bin pattern
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	

Line	Bin data														
1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1
3	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	1	1	1	1	1	1	0	1	
9	1	0	0	0	0	0	0	0	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	1	1	0	0	0	0	0	1	1	1	1	1	1	1	1
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0

(b)

Line	Bin pattern	Bin 1	Bin 2	Bin 3	Bin 4
1		1	0.54	0	0
2		0	0	0	0.82
3		1	1	0.36	0
4		0	0	0	0
5		0	0	0.08	1
6		1	1	1	1
7		1	0.04	0	0
8		0	0.72	1	0.72
9		0.16	0	0.9	1
10		1	1	0.66	0
11		0	0	0	0
12		0.52	0.16	1	1
13		0	0	0	0
14		1	1	1	1
15		0.76	0	0	0

Fonte: Xu (2013a).

Hu, Wang e Xu (2012) desenvolveram o modelo genoma contínuo (modelo infinitesimal), propondo que o efeito genético de um marcador é uma função da posição no genoma (quantidade contínua). O valor genético total de um indivíduo é a integral ponderada da função do efeito genético ao longo do genoma e, então, divide-se o genoma em *bins* para a abordagem da integração numérica para calcular a integral. Os efeitos de *bin* são estimados em vez dos efeitos dos marcadores. Cada *bin* pode conter muitos marcadores e o efeito *bin* representa os efeitos totais de todos os marcadores dentro dele.

O modelo genoma contínuo como especificado é um tipo de modelo linear funcional em que a variável resposta é um escalar e a covariável é uma função. Em vez de estimar os efeitos em todas as posições possíveis, o modelo genoma contínuo dividido em *bins* só precisa estimar alguns parâmetros que determinam a forma da função de um processo biológico particular. Esta abordagem não só facilita o processo de estimação de parâmetros genéticos, mas também fornece resultados que estão mais próximos da realidade biológica pela incorporação de princípios biológicos por trás das funções.

O modelo genoma contínuo de Hu, Wang e Xu (2012) surgiu do modelo infinitesimal clássico para populações com alto desequilíbrio de ligação e foi estendido (adaptado) para populações com baixo ou nenhum desequilíbrio de ligação. Será apresentado aqui apenas o primeiro. Seja y_j o valor fenotípico observado para o indivíduo *j* em uma população de tamanho *n*. O modelo linear usual é:

$$y_j = \beta + \sum_{k=1}^{p} Z_{jk}\gamma_k + \varepsilon_j, \qquad \forall j = 1, \ldots, n, \tag{3.9}$$

em que β é o intercepto, γ_k é o efeito do marcador k, Z_{jk} é o genótipo conhecido do marcador k do indivíduo j (definido em (3.10)) e ε_j é o erro normalmente distribuído com média zero e variância σ^2 desconhecida. Note que p é o número

de marcadores incluídos no modelo.

$$Z_{jk} = \begin{cases} 1, & \text{para } A_1A_1; \\ 0, & \text{para } A_1A_2; \\ -1, & \text{para } A_2A_2. \end{cases} \tag{3.10}$$

Quando $p \to \infty$, o modelo (3.9) torna-se o chamado modelo infinitesimal. O coeficiente γ_k não pode ser estimado porque o modelo tem tamanho infinito e apresenta alta multicolinearidade. Agora, substituindo k pela posição λ correspondente do marcador no genoma, que é contínua e varia de 0 a L, com L sendo o tamanho do genoma, o modelo infinitesimal pode ser substituído por:

$$y_j = \beta + \int_0^L Z_j(\lambda)\,\gamma(\lambda)\,d\lambda + \varepsilon_j, \tag{3.11}$$

com $Z_j(\lambda)$ conhecido para genoma saturado com marcadores e $\gamma(\lambda)$ sendo o efeito genético expresso como uma função desconhecida da posição no genoma. A proposta foi estimar $\gamma(\lambda)$ usando os dados, sendo que os parâmetros são β, σ^2 e $\gamma(\lambda)$.

A integral em (3.11) não tem expressão explícita e, então, uma integração numérica é necessária para aproximar a integral. Neste caso, dividiu-se o genoma todo em m *bins* (intervalos), com Δ_k sendo o tamanho do k-ésimo *bin* (podendo ser o mesmo tamanho para todos os *bins* ou variar para diferentes *bins*). A aproximação numérica do modelo (3.11) é:

$$y_j = \beta + \sum_{k=1}^{m} \bar{Z}_j(\lambda_k)\,\bar{\gamma}(\lambda_k)\,\Delta_k + \varepsilon_j, \tag{3.12}$$

sendo λ_k a posição do ponto médio do k-ésimo *bin* no genoma, $\bar{Z}_j(\lambda_k)$ o valor

médio de Z_j para todos os marcadores cobertos pelo k-ésimo *bin*, $\bar{\gamma}(\lambda_k)$ o efeito médio de todos os QTL naquele *bin* e Δ_k o tamanho deste *bin*.

O número de *bins* (m) depende do tamanho amostral (n) e do nível de desequilíbrio de ligação (LD). Um tamanho amostral grande permite uma quantidade de *bins* grande. Pelo fato de $\bar{Z}_j(\lambda_k)$ ser o valor médio de todos os marcadores dentro do *bin*, marcadores do genoma todo são utilizados.

Sejam $Z_j(h)$ e $\gamma(h)$ o genótipo e o efeito, respectivamente, do marcador h no *bin* k, para $h = 1, \ldots, p_k$, em que p_k é o número de marcadores no *bin* k. Na análise de dados reais, o tamanho do *bin* Δ_k é substituído pelo número de marcadores no *bin* k e o tamanho do genoma é substituído pelo número total de marcadores p. O modelo a ser trabalhado para m *bins* torna-se:

$$y_j = \beta + \sum_{k=1}^{m} Z_{jk}\gamma_k + \varepsilon_j, \tag{3.13}$$

sendo $Z_{jk} = \bar{Z}_j(\lambda_k) = \frac{1}{p_k}\sum_{h=1}^{p_k} Z_j(h)$ e $\gamma_k = \bar{\gamma}(\lambda_k)\, p_k = \sum_{h=1}^{p_k} \gamma(h)$.

Pelo modelo (3.13), visto que o efeito do k-ésimo *bin* é a soma dos efeitos de todos os marcadores dentro deste *bin*, duas hipóteses são necessárias para este modelo ser trabalhado: alto desequilíbrio de ligação e efeitos homogêneos de marcadores dentro do *bin*. De acordo com os autores, se pelo menos uma dessas condições não é satisfeita, o modelo é problemático e necessita-se adaptá-lo.

O modelo adaptativo, denominado modelo infinitesimal adaptativo, pode homogeneizar os efeitos dos marcadores dentro de um *bin* para que ele possa lidar com populações em baixo ou nenhum desequilíbrio de ligação. Define-se, então, a média ponderada de Z_j para todos os marcadores no *bin* k por:

$$Z^*_{jk} = \frac{1}{p_k}\sum_{h=1}^{p_k} w_h Z_j(h) = \frac{1}{p_k}\sum_{h=1}^{p_k} Z^*_j(h), \tag{3.14}$$

em que p_k é o número de marcadores dentro do *bin* k e w_h é um peso assumido para o marcador h dentro do *bin* k.

O efeito médio ponderado para este *bin* é definido como:

$$\gamma_k^* = \sum_{h=1}^{p_k} w_h^{-1}\gamma(h) = \sum_{h=1}^{p_k} \gamma^*(h). \tag{3.15}$$

A ideia dos autores foi escolher um peso tal que os efeitos dentro do *bin* fossem homogeneizados. Existem muitas maneiras de escolher o peso, mas eles propuseram utilizar o seguinte:

$$w_h = \frac{p_k \hat{b_h}}{\sum_{h=1}^{p_k} \left|\hat{b_h}\right|}\ , \tag{3.16}$$

em que $\hat{b_h}$ é a estimativa de mínimos quadrados do h-ésimo marcador em uma análise de marcas simples.

3.2.2 Modelo funcional bayesiano

Uma alternativa ao método de Hu, Wang e Xu (2012) foi apresentada por Moura (2017), em que são introduzidos pesos para cada efeito dentro de um *bin* por meio da frequência relativa com que cada marcador é assumido possuir efeito dentro de um processo MCMC. Neste caso, o modelo genoma contínuo pode lidar, também, com populações em baixo desequilíbrio de ligação, sem precisar usar métodos adaptativos.

De acordo com os resultados apresentados pelo autor, o modelo funcional bayesiano forneceu capacidade preditiva maior ou similar, quando comparado com os métodos clássicos de regressão em cenários reais e simulados, fornecendo informações adicionais sobre os *breakpoints* de desequilíbrio de ligação. Em geral, o modelo funcional bayesiano também obteve maior eficiência computacional, uma

vez que o número de marcadores de referência que estavam relacionados a *bins* foi menor que o tamanho do vetor de informações fenotípicas.

O modelo em *bins* de Moura (2017) superou os modelos de comparação RR-BLUP, Bayes B e Lasso Bayesiano na predição de valores faltantes em cenários de desbalanceamento simulado. Em quatro cenários de herdabilidade, o modelo em *bins* supera todos os modelos de comparação, o que sugere sua aplicabilidade em GS para predizer genótipos não testados.

A principal vantagem desta técnica é que representa uma abordagem mais genética de métodos de predição genômica dado que se buscam regiões causais capazes de predizer o valor genético genômico de forma acurada. Este método mostrou-se viável para predição de valores genéticos genômicos em dados de SNP (*Single Nucleotide Polymorphism*) via genotipagem por sequenciamento, unindo novamente os modelos preditivos e os modelos de buscas de regiões causais e descrição da arquitetura genética em razão da característica funcional da análise (MOURA, 2017).

3.3 *Spline* e funções base B-*Spline*

Como, na maioria das vezes, os dados funcionais são coletados discretamente, faz-se necessária a passagem dos dados da forma discreta para a forma de funções, para se ter valores das unidades amostrais para quaisquer valores num intervalo definido. Assim, devido à presença de ruído (erro) na maioria dos conjuntos de dados, a representação funcional geralmente envolve suavização e, portanto, algumas técnicas de suavização são necessárias. Técnicas de ajuste de curvas polinomiais são comumente usadas neste caso, já que tais funções são avaliadas, diferenciadas e integradas facilmente e em finitos passos (ULBRICHT, 2004). Porém, as funções polinomiais também apresentam algumas limitações desagradá-

veis: com um aumento crescente no grau do polinômio surgem fortes oscilações; são inflexíveis para capturar características locais; são muito sensíveis à escolha dos pontos de parada para o ajuste; dependem globalmente das propriedades locais; e surgem problemas de sobreajuste (SAHA et al., 2013; ULBRICHT, 2004).

Apesar disso, os polinômios com baixa ordem parecem funcionar bem em intervalos suficientemente pequenos. Isso sugere que, para alcançar uma classe de funções aproximadas com maior flexibilidade, o intervalo a ser analisado deve ser subdividido em partes menores e a função de interesse pode ser aproximada por polinômios definidos localmente de grau relativamente baixo (chamados funções base). O uso de funções base geralmente resulta em curvas suaves e bom ajuste aos dados, reduzindo a complexidade dos modelos ajustados (MICHNA et al., 2016). Esta é a motivação por trás de polinômios por partes, *Splines*, B-*Splines*, etc.

Uma curva *Spline* é uma sequência de segmentos de curva que estão conectados para formar uma única curva. Essas curvas são construídas dividindo o intervalo de observações em subintervalos com limites em pontos chamados nós (*knots*). Sobre qualquer intervalo, esta função é um polinômio de grau fixo, mas sua natureza muda quando se passa para o próximo subintervalo (RAMSAY; HOOKER; GRAVES, 2009).

De acordo com Wegman e Wright (1983), existem três principais abordagens para o ajuste de curvas utilizando *Splines* e esses métodos podem ser diferenciados pelo modo como os resíduos são tratados. O primeiro método utiliza mínimos quadrados penalizados, sendo que o problema consiste em minimizar as distâncias dos erros para ajustar o modelo mais aderente aos dados. A segunda abordagem consiste em utilizar o método que estima intervalos com 100% de confiança para cada ponto dos dados. O terceiro método, abordado neste trabalho, é a regressão por *Spline*, que também faz uso da estimação pelo método de mínimos

quadrados, porém para cada subintervalo (nós) da função será ajustado um polinômio com parâmetros diferentes, forçando o encontro das funções subjacentes nos nós.

Qualquer método de regressão tem dois objetivos principais: fornecer ao pesquisador informações a respeito da relação entre as variáveis em estudo e fornecer predições para valores que ainda serão observados. Para o primeiro propósito, um método de estimação não paramétrico é o ideal, pois permite que o modelo seja versátil e tenha bom ajuste aos dados. Assim, o uso de funções *Splines* como suavizadoras da curva de regressão é interessante, pois é uma técnica bastante flexível e de simples implementação computacional (SILVERMAN, 1985; UTPOTT, 2015).

Geralmente, uma *Spline* de grau M associada com K nós é uma função contínua, com $(M-1)$ derivadas contínuas e a M-ésima derivada é constante entre os nós, cuja equação é dada por (BONNER, 1992):

$$f(x) = \beta_0 + \sum_{m=1}^{M} \beta_m x^m + \sum_{k=1}^{K} b_k (x - \xi_k)_+^M \ , \tag{3.17}$$

em que

$$(x - \xi_k)_+^M = \begin{cases} 0 & , \quad x < \xi_k; \\ (x - \xi_k)^M & , \quad x \geq \xi_k. \end{cases}$$

é chamado de polinômio truncado de grau M e ξ_k é o k-ésimo nó, com $\xi_k \leq \xi_{k+1}$.

Aumentar o grau da *Spline* faz com que a função ajustada pareça mais suave, mas também aumenta a complexidade dos segmentos polinomiais em cada intervalo. Na prática, a escolha mais comum é a *Spline* cúbica $(M = 3)$. Alternativamente, pode-se escrever uma *Spline* em termos de um conjunto de funções de base normalizadas. Uma base muito popular é a B-*Spline* (BOOR, 1978), pois

suas funções base são estritamente locais e a diferença é que nos nós a função é suavizada, ou seja, não ocorre uma mudança abrupta da função entre um nó e outro, como acontece na *Spline*.

B-*Splines* são definidas pela ordem n e pelos números de k nós dentro do intervalo especificado, chamados nós internos. Os dois pontos extremos, início e fim do intervalo, também são considerados nós, então o número total de nós é $k + 2$. O grau do polinômio B-*Spline* é $m = n - 1$. Seja uma sequência não decrescente de nós (números reais) tal que $a = \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{K+1} = b$, em um intervalo $[a, b]$. Definindo um conjunto de nós aumentados $\xi_{-m} = \cdots = \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{k+1} = \cdots = \xi_{k+m+1}$, em que os limites *a* e *b* são repetidos *m* vezes, as funções base B-*Spline* $B_{i,j}$ são representadas recursivamente por:

$$\mathrm{B}_{i,j+1}(x) = \alpha_{i,j+1}(x)\,\mathrm{B}_{i,j}(x) + [1 - \alpha_{i+1,j+1}(x)]\,\mathrm{B}_{i+1,j}(x), \tag{3.18}$$

com

$$\mathrm{B}_{i,0}(x) = \begin{cases} 1, & \xi_i \leq x < \xi_{i+1}; \\ 0, & \text{caso contrário.} \end{cases} \tag{3.19}$$

e

$$\alpha_{i,j} = \begin{cases} \dfrac{x - \xi_i}{\xi_{i+j} - \xi_i}, & \text{se } \xi_{i+j} - \xi_i \neq 0; \\ 0 \qquad\quad , & \text{caso contrário.} \end{cases} \tag{3.20}$$

Com isso, uma *Spline* de grau m pode ser representada por uma combinação linear de bases B-*Spline* $B_{i,m}$ dada por:

$$f(x) = \sum_{i=0}^{k+m} \beta_i\,\mathrm{B}_{i,m}(x) \qquad x \in [a,b] \tag{3.21}$$

sendo que os β_i são chamados de pontos de controle. Para uma B-*Spline* de grau m contendo k nós interiores, existem $k + m + 1$ (um, se o intercepto é incluído) pontos de controle.

O grau m e o número de nós k geralmente são definidos pelo pesquisador. Já a posição dos nós ξ_i pode ser fixa ou livre. Geralmente, a escolha dos nós (posição) e o grau do polinômio são as duas maiores dificuldades reportadas por autores na aplicação dessa técnica (UTPOTT, 2015; WEGMAN; WRIGTH, 1983).

4 CONCLUSÃO GERAL

Neste trabalho, os métodos RR-BLUP Bins, Bayes A Bins e Bayes B Bins apresentaram resultados mais satisfatórios que suas respectivas versões originais para o cenário com poucos QTL, pois obtiveram melhores acurácias preditivas. Em cenários com muitos QTL, apenas o método RR-BLUP Bin apresentou melhores resultados de acurácia. Mais ainda, os métodos adaptados mostraram ser mais apropriados para capturar o sinal de um QTL. Assim, a aproximação da integral no modelo genoma contínuo por meio da divisão em *bins* surge como um método promissor, apresentando ótimos resultados em menor tempo computacional e merece mais estudos.

O uso de curvas B-*Spline* apresentou resultados similares ao RR-BLUP em relação à acurácia preditiva. Mais ainda, este método obteve o menor custo computacional tanto para os dados simulados quanto para os dados reais. A forma funcional das curvas polinomiais por partes (*Splines*) pode lidar com um número ilimitado de marcadores e permitir novos tipos de análises.

REFERÊNCIAS

BONNER, S. **An introduction to splines.** [S.l.: s.n.], 1992. 54 p. Disponível em: <http://people.stat.sfu.ca/∼cschwarz/Consulting/Trinity/Phase2/TrinityWorkshop/Workshop-material-Simon/Intro_to_splines/intro_to_splines_notes.pdf>. Acesso em: 19 de Julho de 2017.

BOOR, C. de. **A practical guideto splines.** New York: Springer, 1978. 392 p.

CARDOT, H.; FERRATY, F.; SARDA, P. Functional linear model. **Statistics and Probability Letters**, New York, v. 45, n. 1, p. 11-22, Oct. 1999.

CHIOU, J. M.; MÜLLER, H. G.; WANG, J. L. Functional response models. **Statistica Sinica**, Taipei, v. 14, p. 675-693, 2004.

COMBS, E.; BERNARDO, R. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. **The Plant Genome**, New York, v. 6, n. 1, p. 1-7, Mar. 2013.

DIERCKX, P. **Curve and surface fitting with splines.** Clarendon: University Press, 1993. 304 p.

GIANOLA, D. et al. Additive genetic variability and the Bayesian alphabet. **Genetics**, Austin, v. 183, p. 347-363, Sept. 2009.

GIANOLA, D. Priors in whole-genome regression: the Bayesian alphabet returns. **Genetics**, Austin, v. 194, n. 3, p. 573-596, July 2013.

GODDARD, M. E.; HAYES, B. J. Genomic selection. **Journal of Animal Breeding and Genetics**, Hamburg, v. 124, n. 6, p. 323-330, Dec. 2007.

HABIER, D. et al. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, London, v. 12, p. 186, May 2011.

HAYES, B. E. N.; 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, May/June 2001.

HICKEY, J. M. et al. Evaluation of genomic selection training population designs and genotyping strategies in plant breeding programs using simulation. **Crop Science**, Amsterdam, v. 54, n. 4, p. 1476-1488, July/Aug. 2014.

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.

JANNINK, J. L.; LORENZ, A. J.; IWATA, H. Genomic selection in plant breeding: from theory to practice. **Briefings in Functional Genomics**, Oxford, v. 9, n. 2, p. 166-177, Mar. 2010.

KIZILKAYA, K.; FERNANDO, R. L.; GARRICK, D. J. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. **Journal of Animal Science**, Champaign, v. 88, n. 2, p. 544-551, Feb. 2010.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, v. 157, n. 4, p. 1819-1829, Apr. 2001.

MIAR, Y.; PLASTOW, G.; WANG, Z. Genomic selection, a new era for pork quality improvement. **Springer Science Reviews**, Basel, v. 3, n. 1, p. 27-37, June 2015.

MICHNA, A. et al. Natural cubic spline regression modeling followed by dynamic network reconstruction for the identi?cation of radiation-sensitivity gene association networks from time-course transcriptome data. **PLoS One**, San Francisco, v. 11, n. 8, p. e0160791, 2016.

MOURA, E. G. **Aplicação de modelos funcionais na seleção genômica ampla.** 2017. 54 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2017.

RAMSAY, J. O.; DALZELL, C. Some tools for functional data analysis (with discussion). **Journal of the Royal Statistical Society**. **Series B**, London, v. 53, n. 3, p. 539-572, 1991.

RAMSAY, J. O.; HOOKER, G.; GRAVES, S. **Functional data analysis with R and MATLAB.** New York: Springer Science & Business Media, 2009. 202 p.

RAMSAY, J. O.; SILVERMAN, B. W. **Applied functional data analysis:** methods and case studies. New York: Springer, 2002. 191 p.

SAHA, S. et al. Missing value estimation in DNA microarrays using B-Splines. **Journal of Medical and Bioengineering**, New York, v. 2, n. 2, p. 88-92, June 2013.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, Berlin, v. 123, n. 4, p. 218-223, Aug. 2006.

SILVEIRA NETO, P. C. **Suavização não-paramétrica e análise de variância funcional.** 2012. 46 p. Monografia (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2012.

SILVERMAN, B. W. Some aspects of the spline smoothing approach to non-parametric regression curve fitting. **Journal of the Royal Statistical Society**. **Series B**, London, v. 47, n. 1, p. 1-52, 1985.

SINGH, B. D.; SINGH, A. K. **Marker-assisted plant breeding:** principles and practices. New Delhi: Springer, 2015. 514 p.

THAVAMANIKUMAR, S.; DOLFERUS, R.; THUMMA, B. R. Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. **G3:** genes, genomes, genetics, Bethesda, v. 5, n. 10, p. 1991-1998, July 2015.

ULBRICHT, J. **Representing functional data as smooth functions.** 2004. 72 p. Thesis (Master of Science) - Institute of Statistics and Econometrics, Humboldt University, Berlin, 2004.

UTPOTT, N. M. **Regressão logística utilizando b-splines:** uma maneira de lidar com relações não lineares. 2015. 75 p. Monografia - (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015.

VISION, T. J. et al. Selective mapping: a strategy for optimizing the construction of high-density linkage maps. **Genetics**, Austin, v. 155, n. 1, p. 407-420, May 2000.

WEGMAN, E. J.; WRIGHT, I. W. Splines in statistics. **Journal of the American Statistical Association**, Washington, v. 78, n. 382, p. 351-365, 1983.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetics Research**, London, v. 75, n. 2, p. 249-252, Apr. 2000.

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

XU, S.; HU, Z. Methods of plant breeding in the genome era. **Genetics Research**, London, v. 92, n. 5/6, p. 423-441, Dec. 2010.

XU, S. Mapping quantitative trait loci by controlling polygenic background effects. **Genetics**, Austin, v. 195, n. 4, p. 1209-1222, 2013b.

ZHANG, X. et al. Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. Frontiers in Genetics, Lausanne, v. 7, p. 151, Aug. 2016.

ZHOU, Y. et al. Systematic bias of correlation coefficient may explain negative accuracy of genomic prediction. **Briefings in Bioinformatics**, London, v. 18, n. 5, p. 744-753, Sept. 2017.

ZOGLAT, A. Functional analysis of variance. **Applied Mathematical Sciences**, Boca Raton, v. 2, n. 23, p. 1115-1129, Jan. 2008.

SEGUNDA PARTE

Nesta parte do trabalho, são apresentados dois artigos que abordam a ideia do modelo genoma contínuo, sendo o primeiro baseado nas adaptações dos métodos de seleção genômica e o segundo baseado na curva B-*Spline*.

ARTIGO 1 Métodos de estimação baseados em *bins* para seleção genômica

RESUMO

Verifica-se que, em um painel com grande quantidade de marcadores, dois problemas clássicos dificultam a seleção genômica: dimensionalidade e multicolinearidade. Assim sendo, métodos de redução dimensional têm se tornado bastante atrativos. O modelo genoma contínuo baseia-se na divisão do genoma em blocos de alto desequilíbrio de ligação (denominados *bins*) e utiliza a média dos marcadores dentro desses *bins* como medida de informação para predizer o valor genético genômico de novos indivíduos. Este trabalho tem por objetivos: desenvolver uma abordagem baseada em *bins* (janelas) e introduzi-la em métodos de seleção genômica, juntamente com o modelo funcional bayesiano; examinar os benefícios potenciais do uso dos métodos adaptados com relação às suas respectivas formas originais, sob alguns cenários de arquiteturas simuladas; avaliar se a escolha de diferentes tamanhos de *bins* fixos pode impactar o desempenho dos métodos adaptados; ajudar a melhorar a eficiência computacional dos métodos de seleção genômica. Populações F_2 de 300 indivíduos genotipados com 12150 marcadores e F_{10} de 320 indivíduos genotipados com 10010 marcadores foram simuladas para três cenários (oligogênico, poligênico I e poligênico II) com herdabilidades 0,2; 0,5 e 0,8, sendo adotadas as configurações de 10, 30, 90 e 150 *bins*. Para o cenário oligogênico, os métodos adaptados em *bins* foram mais acurados que os respectivos métodos originais. Para o cenário poligênico I, apenas Bayes B original foi mais acurado que os métodos *bins*; adaptações *bins* de RR-BLUP e Bayes A foram mais acurados em quase todas as configurações. Já para o cenário poligênico II, apenas os RR-BLUP Bins foram equivalentes ao original; os métodos Bayes A e Bayes B originais foram mais acurados que suas respectivas adaptações. O modelo funcional bayesiano mostrou-se atrativo na identificação de regiões causais e alta flexibilidade de análise. Esta abordagem unifica os modelos preditivos e de busca de QTL para representar a arquitetura genética com base na relação funcional entre o efeito de um marcador e sua posição.

Palavras-chave: Redução de dimensionalidade. Modelo genoma contínuo. Valor genômico. Genoma em bins.

1 INTRODUÇÃO

A introdução da seleção genômica (GS - *Genomic Selection*) (MEUWISSEN; HAYES; GODDARD, 2001) proporcionou acesso a dados com alta densidade de marcadores. Entretanto, esses conjuntos de dados também apresentam grandes desafios. Primeiro, o tamanho das amostras, geralmente, é limitado a poucas centenas de indivíduos e o número de marcadores pode chegar a centenas de milhares ou até vários milhões (p muito maior que n, ou seja, o número de marcadores é muito maior do que o número de indivíduos). Segundo, os marcadores que são fisicamente próximos uns dos outros estão, frequentemente, em desequilíbrio de ligação, isto é, correlacionados.

Para lidar com estes problemas, diversos métodos estatísticos têm sido desenvolvidos. Estimadores penalizados tais como RR-BLUP (*Ridge Regression Best Linear Unbiased Prediction*), Bayes A e Bayes B são muito utilizados em GS e estudos de associação genômica ampla (GWAS - *Genome-wide Association Studies*) em melhoramento de plantas e animais. O principal objetivo de GWAS é identificar os *loci* de caracteres quantitativos (QTL - *Quantitative Trait Loci*) com maiores efeitos, enquanto GS se concentra em usar informação genômica para melhor avaliar o potencial genético de indivíduos.

Os métodos consideram diferentes hipóteses sobre a distribuição dos efeitos de QTL. O método RR-BLUP assume variâncias iguais para todos os marcadores. Esta hipótese é biologicamente incorreta, mas faz com que a estatística seja robusta, limitando o número de parâmetros desconhecidos (MEUWISSEN; HAYES; GODDARD, 2001). Os métodos bayesianos Bayes A e Bayes B assumem heterogeneidade de variância dos efeitos de marcadores, com ênfase naqueles com efeitos principais. A consequência de utilizar estimadores penalizados (*shrinkage*) é a capacidade de lidar com modelos de alta dimensão. Entretanto, a implementa-

ção de modelos bayesianos hierárquicos normalmente requer demandas intensivas de computação, já que a inferência a posteriori é tipicamente baseada em técnicas de Monte Carlo via Cadeia de Markov (MCMC - *Monte Carlo Markov Chain*).

Como um genoma com grande quantidade de marcadores em alto desequilíbrio de ligação (LD - *linkage disequilibrium*) contém vários grupos de marcadores que explicam quantidades similares de variação genética em um determinado caráter, é razoável aplicar um procedimento de redução de dimensão para excluir as informações redundantes dos dados e, também, reduzir o custo computacional (LI et al., 2018). De acordo com Zhang et al. (2016), reduzir as dimensões dos dados de marcadores com base nas informações biológicas é crucial e, como tal, deve ser o primeiro passo.

Uma abordagem baseada em intervalos (janelas) agrega informações de vários marcadores correlacionados e usa algumas estatísticas resumidas para substituir os dados originais. Hu, Wang e Xu (2012) foram pioneiros em uma metodologia inovadora baseada em janelas (*bins*) para mapeamento de QTL. Primeiro, o cromossomo era dividido em muitos *bins* naturais (selecionados com base em pontos de interrupção - *breakpoints* - por meio do desequilíbrio de ligação). Segundo, uma abordagem de integração numérica era usada para agregar os dados dos marcadores em todos os *bins*, por meio do cálculo do valor médio dos genótipos de vários marcadores dentro dos *bins*. De acordo com Li et al. (2018), esta abordagem está relacionada ao "teste de sobrecarga" (*burden test*), inicialmente proposto em genética humana (MORGENTHALER, THILLY, 2007), para testar um grupo de marcadores como uma unidade biológica significativa, como um gene ou uma via bioquímica.

Entretanto, visto que o efeito do k-ésimo *bin* é a soma dos efeitos de todos os marcadores dentro deste *bin*, duas hipóteses são necessárias para este modelo

ser trabalhado: alto desequilíbrio de ligação e efeitos homogêneos de marcadores dentro do *bin*. De acordo com Hu, Wang e Xu (2012), se pelo menos uma dessas condições não é satisfeita, o modelo é problemático e necessita-se adaptá-lo. No modelo adaptativo, denominado modelo infinitesimal adaptativo, define-se a média ponderada de Z_j para todos os marcadores no *bin* k tal que o peso é obtido dos coeficientes de regressão calculados por meio da análise de marcadores simples. Esta técnica pode homogeneizar os efeitos dos marcadores dentro de um *bin* para que o modelo possa lidar com populações em baixo ou nenhum desequilíbrio de ligação.

Xu (2013) investigou as propriedades da metodologia de Hu et al. (2012) e afirmou que utilizar os denominados *bins* artificiais (em que os pesquisadores fixam os tamanhos dos *bins*) é mais conveniente para a seleção genômica, mas não para detecção de QTL. Uma alternativa à abordagem de integração numérica de Hu, Wang e Xu (2012) foi apresentada por Moura (2017), que distribui pesos para efeitos dentro de *bins* usando a frequência relativa com que cada marcador é amostrado em uma cadeia de Markov. O autor propôs uma abordagem mais direta de modelos funcionais e integração numérica da função genômica e relaxou a suposição de blocos com alto LD nos *bins*. Isto permitiu obter um tipo de média ponderada nos *bins* em vez de um efeito médio.

A principal vantagem de aplicar o método *bin* associado aos modelos funcionais na GS é que ele coloca os métodos GS de volta ao cenário genético, uma vez que as posições dos marcadores e as informações de LD são novamente levadas em conta na busca por regiões causais, cuja identificação pode contribuir para a predição mais precisa dos valores genômicos. Além disso, qualquer modelo de regressão penalizada ou estrutura do alfabeto bayesiano pode ser adaptado ao modelo *bin*, permitindo uma análise rápida.

Com essa ideia, existem quatro objetivos gerais neste trabalho. O primeiro objetivo é desenvolver uma abordagem baseada em *bins* (janelas) e introduzi-la em métodos de seleção genômica, juntamente com o modelo funcional bayesiano. O segundo objetivo é examinar os benefícios potenciais do uso dos métodos adaptados com relação às suas respectivas formas originais, sob alguns cenários de arquiteturas simuladas. O terceiro objetivo é avaliar se a escolha de diferentes tamanhos de *bins* fixos pode impactar o desempenho dos métodos adaptados. O quarto objetivo é ajudar a melhorar a eficiência computacional dos métodos de seleção genômica.

2 MATERIAL E MÉTODOS

2.1 Dados simulados

Para representar uma população com alto desequilíbrio de ligação, simulou-se um conjunto de dados composto de 300 indivíduos pertencentes a uma população F_2, genotipados com 12150 marcadores SNPs (*Single Nucleotide Polymorphism*) com distância média de 0,001 cM no genoma, distribuídos em dez cromossomos com tamanho de 120 cM cada. Para uma população com baixo desequilíbrio de ligação, simulou-se um conjunto de dados composto de 320 indivíduos pertencentes a uma população F_{10}, genotipados com 10020 marcadores SNPs com distância média de 0,001 cM no genoma, distribuídos em dez grupos de ligação com tamanho de 120 cM cada (software QGenes - JOEHANES; NELSON, 2008).

Nove cenários foram considerados envolvendo diferentes números de QTL (6, 15 e 60) e diferentes níveis de herdabilidade (0,2; 0,5 e 0,8). Todos os QTL foram selecionados dentre os marcadores e seus efeitos foram amostrados de uma distribuição normal com média zero e variância um. Para todos os cenários, foram adotadas quatro configurações *bins*:

a) Bin01: 10 *bins* com 1215 marcadores por *bin*;

b) Bin02: 30 *bins* com 405 marcadores por *bin*;

c) Bin03: 90 *bins* com 135 marcadores por *bin*;

d) Bin04: 150 *bins* com 81 marcadores por *bin*.

2.2 Modelo funcional bayesiano

Neste trabalho, foi utilizado o modelo funcional bayesiano proposto por Moura (2017), que é uma abordagem diferente daquela apresentada por Hu, Wang e Xu (2012). É importante ressaltar que o modelo funcional bayesiano, diferentemente do modelo proposto por Hu, Wang e Xu (2012), não é um modelo de análise *bin;* os *bins* são utilizados somente como estratégia de integração com o propósito de aumentar a chance de encontrar máximos locais e globais.

O modelo funcional será incorporado aos métodos RR-BLUP, Bayes A e Bayes B, a fim de adaptá-los ao conceito de genoma contínuo. O método Bayes B foi utilizado, nesta seção, para demonstrar a técnica de *bins*, sendo que os outros métodos estão apresentados no Apêndice B.

O modelo funcional genômico fundamenta-se na ideia de que o sinal de expressão de um gene pode ser descrito por uma função espacial do genoma. Por não ser possível verificar todas as posições no genoma, pois se trata de uma variável contínua, podemos utilizar as posições dos marcadores como pseudoposições (marcador M_m tem a posição λ_m). Em outras palavras, $f(\lambda) = \gamma$, sendo λ a posição no genoma e γ o sinal de expressão do gene em dada a posição. O interesse é estimar esta função sinal e, a partir dela, predizer o valor genético genômico de novos indivíduos.

Tem-se que $f(\lambda)$ e γ são desconhecidas e só podem ser estimadas pela informação referente ao fenótipo de um indivíduo (y) e as posições λ dos marcadores do genoma. Dado que os domínios das funções $\gamma(\lambda)$ e y são diferentes ($f(\lambda) := \{\lambda_j | \lambda_j \in \Omega \equiv [0, L], \forall j\}$, em que L é o comprimento do cromossomo em pares de base - pb), deve-se atribuir uma função de ligação do domínio de $\gamma(\lambda)$ para o domínio de y. Isto pode ser realizado pela matriz de estado genotípico $\mathbf{Z}_i(\lambda)$, que é condicional a λ. Ou seja, assumindo $f(\lambda) = \gamma$, tem-se a seguinte

igualdade: $\mathbf{Z}(\lambda)f(\lambda) = \mathbf{Z}(\lambda)\gamma$. Tomando $\mathbf{Z}(\lambda)\gamma$ como a predição do valor genômico $\hat{g}$, ou seja, a combinação linear do estado genotípico do marcador na posição λ com o efeito deste marcador nesta posição, tem-se que $\mathbf{Z}(\lambda)f(\lambda) = \hat{g} = y + \varepsilon$. Nesse caso, vale a equivalência $\mathbf{Z}(\lambda)f(\lambda) \equiv \int_0^L \mathbf{Z}_i(\lambda)\gamma(\lambda)d\lambda = y + \varepsilon$.

Com isso, seja y_i o valor fenotípico do indivíduo *i*, para $i = 1, \ldots, n$. O modelo genoma contínuo, considerando um cromossomo, é dado por:

$$y_i = \mu + \int_0^L Z_i(\lambda)\,\gamma(\lambda)\,d\lambda + \varepsilon_i, \tag{2.1}$$

em que

μ é a média geral;

λ é a posição do marcador no cromossomo;

L é o tamanho do cromossomo;

$\gamma(\lambda)$ é o efeito aditivo do marcador na posição λ (expresso como uma função desconhecida);

ε_i é o erro para o indivíduo i, sendo $\varepsilon_i \sim N\left(0,\sigma_e^2\right)$, com σ_e^2 desconhecida;

$Z_i(\lambda)$ é o genótipo do marcador na posição λ para o indivíduo i definido como

$$Z_i(\lambda) = \begin{cases} 2, & \text{para homozigoto dominante;} \\ 1, & \text{para heterozigoto;} \\ 0, & \text{para homozigoto recessivo.} \end{cases} \tag{2.2}$$

Assim, para C cromossomos no genoma, o modelo (2.1) torna-se:

$$y_i = \mu + \sum_{t=1}^{C} \int_0^L Z_{it}(\lambda)\,\gamma(\lambda)\,d\lambda + \varepsilon_i, \qquad \forall i = 1, \ldots, n \tag{2.3}$$

em que o somatório descreve a descontinuidade da função ao longo dos cromossomos.

Dado um marcador candidato amostrado dentro do k-ésimo *bin*, o modelo funcional parcial a ser considerado, durante o Monte Carlo via Cadeia de Markov (MCMC - *Monte Carlo Markov Chain*) com Metropolis-Hasting, é descrito praticamente como um modelo linear dado por:

$$y_i = \mu + \sum_{j=1}^{k} Z_{ij}\gamma_j + \varepsilon_i, \quad \forall i = 1,...,n, \tag{2.4}$$

em que k é o número de *bins* e j refere-se ao j-ésimo marcador candidato amostrado estocasticamente dentro do intervalo $\left[\lambda_{j(\min)}, \lambda_{j(\max)}\right]$.

2.2.1 Modelo hierárquico e distribuições a priori

As variáveis observáveis são os valores fenotípicos ($\mathbf{y}$), os genótipos dos marcadores ($\mathbf{Z}$) e as posições λ (que são as posições dos marcadores). As variáveis não observáveis são os coeficientes de regressão ($\boldsymbol{\mu}$ e $\boldsymbol{\gamma}$) e as variâncias residual e dos efeitos aditivos $\left(\sigma_e^2 \text{ e } \sigma_\gamma^2\right)$. Neste trabalho, utilizou-se uma formulação do Bayes B apresentada em Gianola (2013) e Gianola et al. (2009). Assim, o modelo hierárquico pode ser descrito por:

$$y_i|\mu,\sigma^2 \sim N\left(\mu + \sum_{t=1}^{C}\int_0^L Z_{it}\left(\lambda\right)\gamma\left(\lambda\right)d\lambda, I\sigma_e^2\right); \tag{2.5}$$

$$\gamma_j|\sigma^2_{\gamma_j},\pi \sim IID \begin{cases} N\left(0,\omega\right) & , \quad \text{com probabilidade } \pi; \\ N\left(0,\sigma^2_{\gamma_j}\right), & \quad \text{com probabilidade } 1-\pi; \end{cases} \tag{2.6}$$

$$\sigma^2_{\gamma_j}|\pi \sim \begin{cases} 0 & , \quad \text{com probabilidade } \pi; \\ \chi^{-2}_{esc}\left(\nu,S^2\right), & \quad \text{com probabilidade } 1-\pi; \end{cases} \tag{2.7}$$

com $j = 1, \ldots, M$ marcadores, $t = 1, \ldots, C$ cromossomos, $\nu = 1$ e $S^2 = \frac{\sigma_y^2 \cdot 0{,}005}{M}$ sendo, respectivamente, o grau de liberdade e o parâmetro de escala para a variância do marcador.

A priori é uma mistura de uma massa de pontos em 0 (zero) com uma distribuição Normal em que $\omega = 10^{-8}$, sendo π e $(1 - \pi)$ as probabilidades de mistura, respectivamente, com π assumido ser conhecido e arbitrariamente especificado (neste trabalho, assumiu-se $\pi = 0{,}95$). As prioris para μ e σ_e^2 são $p(\mu) \propto 1$ e $p\left(\sigma_e^2\right) \propto \frac{1}{\sigma_e^2}$, respectivamente.

A posição λ varia dentro do k-ésimo *bin* e assume-se que ela coincide com a posição de um marcador M dentro deste *bin*. Assim, considerando Δ_k como o tamanho do *bin* k, utilizamos a priori de que a posição é uniformemente distribuída neste *bin*, ou seja, $p(\lambda_k) = \frac{1}{\Delta_k}$.

Para simplificar a notação, tomando $\boldsymbol{\gamma} = \{\gamma_j\}$, $\boldsymbol{\sigma^2_\gamma} = \left\{\sigma^2_{\gamma_j}\right\}$ e $\boldsymbol{\lambda} = \{\lambda_j\}$, para $j = 1, \ldots, M$ marcadores, a priori conjunta das variáveis não observáveis é dada por:

$$p\left(\boldsymbol{\mu}, \sigma_e^2, \boldsymbol{\gamma}, \boldsymbol{\sigma^2_\gamma}\right) \propto p(\boldsymbol{\mu})\, p\left(\sigma_e^2\right) \prod_{j=1}^{m} p\left(\boldsymbol{\gamma}_j | \lambda\right) p\left(\boldsymbol{\sigma^2_\gamma}\right) \tag{2.8}$$

2.2.2 Verossimilhança conjunta

A verossimilhança, para os dados fenotípicos, considerando o modelo completo, pode ser descrita por:

$$\begin{aligned} p\left(\mathbf{y} | \boldsymbol{\mu}, \sigma_e^2, \boldsymbol{\gamma}, \boldsymbol{\sigma^2_\gamma}\right) &= \prod_{i=1}^{n} p\left(y_i | \mu, \sigma_e^2, \boldsymbol{\gamma}, \boldsymbol{\sigma^2_\gamma}\right) \propto \\ &\propto \left(\sigma_e^2\right)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{t=1}^{C} \int_0^L Z_{it}(\lambda)\, \gamma(\lambda)\, d\lambda \right)^2 \right\} \\ &\propto \left(\sigma_e^2\right)^{-n/2} \exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{i=1}^{n} \left(y_i - \mu - \mathbf{Z}_i \mathbf{P}_\lambda \boldsymbol{\gamma} \right)^2 \right\} \end{aligned} \tag{2.9}$$

em que $\mathbf{P}_\lambda$ é a função peso dada pela frequência com que λ foi visitada, ao longo de uma cadeia de tamanho N, obtida pelo processo MCMC.

2.2.3 Distribuição a posteriori conjunta

A distribuição a posteriori conjunta pode ser descrita por:

$$p\left(\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}}|\mathbf{y},\lambda\right) \propto p\left(\mathbf{y}|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}}\right) p\left(\lambda|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}},\mathbf{y}\right) p\left(\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}}\right) \tag{2.10}$$

em que $p\left(\lambda|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}},\mathbf{y}\right)$ é uma função desconhecida, porém condicional a todos os parâmetros do modelo.

2.2.4 MCMC para modelos funcionais genômicos

Como a função $p\left(\lambda|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^2_{\boldsymbol{\gamma}},\mathbf{y}\right)$ é desconhecida, a integral em (2.3) não é explícita e, então, uma forma de integração é necessária. Existem vários algoritmos para realizar a integração numérica. Neste estudo, adotou-se a estratégia de divisão cromossômica em *bins*. Dentro de cada *bin*, a função contínua desconhecida $f(\lambda) = \gamma$ pode ser estimada empiricamente usando o par $(\lambda_j;\gamma_j)$, quando a posição λ_j é conhecida para um marcador específico. Neste caso, $\gamma(\lambda) \cong (\lambda_j;\gamma_j)$, ou para cada posição discretizada λ_j há uma resposta funcional γ_j.

Dado que o efeito do par $(\lambda_j;\gamma_j)$ é desconhecido, ele só pode ser obtido usando a posição observada λ_j e o resultado escalar y. Então, $(\lambda_j;\gamma_j)$ pode ser considerado uma variável aleatória cuja distribuição a posteriori pode ser obtida por $p(\gamma_j|\lambda_j,y) \propto p(\gamma_j)p(y|\gamma_j)p(\lambda_j|\gamma_j,y) = p(\gamma_j|y)p(\lambda_j|\gamma_j,y)$. O par $(\lambda_j;\gamma_j)$ pode ser estimado como $\hat{\gamma}_j|\lambda_j \overset{\wedge}{=} \underset{\gamma_j\in\Re}{\arg.\max}\left[p(\gamma_j|y)\right]p(\lambda_j|\gamma_j,y)$. No entanto,

$p(\lambda_j|\gamma_j,y) := \left\{\lambda_j|\lambda_j \in S \equiv \left[\lambda_{j(\min)},\lambda_{j(\max)}\right],\forall j\right\}$ é desconhecido no domínio do j-ésimo *bin* $\left[\lambda_{j(\min)},\lambda_{j(\max)}\right]$.

Sob alto desequilíbrio de ligação entre marcadores, em uma janela genômica, pode-se considerar $p(\lambda|\gamma,y)$ constante, uma vez que $\mathbf{Z}(\lambda_j)$ é constante, pelo qual se pode justificar os modelos que foram propostos por Hu et al. (2012) e Xu (2013), em que cada *bin* artificial corresponde a um *bin* natural. Além disso, se y não for informativo sobre γ, a distribuição a posteriori $p(\lambda|y,\gamma)$ converge para uma distribuição uniforme discreta, isto é, este *bin* artificial não é causal. Assim, a probabilidade condicional $p(\lambda_j|\gamma_j,y)$ pode ser integrada numericamente por meio do algoritmo MCMC usando o método Metropolis-Hastings. Assim, dentro de cada *bin*, uma posição λ_j foi sorteada, de modo que a dimensão do modelo foi restrita ao número de k *bins* (a depender das configurações descritas) , durante o processo MCMC, cujos passos são descritos a seguir.

1 Inicialização: Os parâmetros μ e σ_e^2 foram inicializados com a média e a variância dos dados fenotípicos, respectivamente; o vetor $\boldsymbol{\gamma}$ será inicializado com o valor zero e dimensão k, em que k é o número de *bins* do modelo. A matriz do estado genotípico $\boldsymbol{Z}_{\lambda_j}$ de dimensão (n x k) será amostrada da matriz completa de marcadores $\mathbf{Z}$ de dimensão (n x m), em que o índice λ_j correspondente à posição inicial do marcador amostrado no k-ésimo *bin*. Assim, $\boldsymbol{Z}_{\lambda_j}$ foi inicialmente amostrado com base na posição mediana λ_j do k-ésimo *bin*. As variâncias dos efeitos dos marcadores $\boldsymbol{\sigma^2_\gamma}$ serão assumidas inicialmente como não nulas (0,5).

$$I^{(0)} = \left[\mu^{(0)},{\gamma_1}^{(0)},\ldots,{\gamma_j}^{(0)},{\sigma_e^2}^{(0)},{\sigma^2_{\gamma_1}}^{(0)},\ldots,{\sigma^2_{\gamma_j}}^{(0)},\mathbf{Z}_{\lambda_j}\right] \tag{2.11}$$

2 Amostrar μ utilizando a seguinte condicional:

$$\mu|\ldots \sim N\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} Z_{\lambda_{j(i)}}\gamma_{\lambda_j}\right)/n \ \ , \ \frac{\sigma_e^2}{n}\right] \tag{2.12}$$

Note que o índice sobre o somatório é dado por k e não m. Isso é justificado pelo processo Bernoulli, observado em buscas estocásticas, durante o MCMC, dado que, dentro de um determinado *bin*, quando λ_j é amostrada na t-ésima iteração, $p_{\lambda_j^t} = 1$ e $p_{\lambda_{-j}^t} = 0$; assim, para a *t*-ésima iteração, $\mathbf{Z}_{\lambda^t}\mathbf{P}_{\lambda^t}\hat{\gamma}_{\lambda^t} = \sum_{j=1}^{k} Z_{\lambda_j}\gamma_{\lambda_j}$, em que $\mathbf{P}_{\lambda^t}$ é uma matriz diagonal (m x m) indicadora de qual marcador dentro do *bin* está sendo avaliado na respectiva iteração. Esse processo torna o modelo em *bins* mais rápido.

3 A distribuição condicional completa a posteriori referente aos efeitos dos marcadores $\left(\gamma_{\lambda_j^t}\right)$, dada a posição λ_j amostrada da *t*-ésima iteração, é uma normal com média e variância especificadas abaixo:

$$\overline{\gamma}_{\lambda_j^t} = \left(\sum_{i=1}^{n} Z_{\lambda_{j(i)}^t}^2 + \frac{\sigma_e^2}{\eta}\right)^{-1} \sum_{i=1}^{n} Z_{\lambda_{j(i)}^t}\left(y_i - \mu - \sum_{m\neq j}^{k} Z_{\lambda_{m(i)}^t}\gamma_{\lambda_m^t}\right) \tag{2.13}$$

$$s^2_{\gamma_{\lambda_j^t}} = \left(\sum_{i=1}^{n} Z_{\lambda_{j(i)}^t}^2 + \frac{\sigma_e^2}{\eta}\right)^{-1}\sigma_e^2 \tag{2.14}$$

em que $\eta = \delta\sigma^2_{\gamma_{\lambda_j^t}} + (1-\delta)\,\omega$, com $\delta : \{1{,}0\}$.

4 Como mencionado anteriormente, $p\left(\lambda|\boldsymbol{\mu}{,}\sigma_e^2{,}\boldsymbol{\gamma}{,}\boldsymbol{\sigma^2_\gamma}{,}\mathbf{y}\right)$ não é conhecida e o algoritmo Metropolis-Hastings (HASTINGS, 1970; METROPOLIS et al.,

1953) pode ser utilizado dado que não exige que o parâmetro tenha uma função de probabilidade fechada. Para isso, faz-se uso de uma função auxiliar possível de ser amostrada, retirando valores candidatos que podem ser aceitos com α de probabilidade. Neste caso, utilizou-se uma distribuição uniforme como gerdadora de candidatos para λ_j, que foi amostrada em cada *bin* em todos os cromossomos, sob um intervalo delimitado por $max(LI_j$, $\lambda_j - c)$ e $min\,(LS_j,\lambda_j + c)$, sendo c uma constante discreta que define o caminhamento (salto) dentro do j-ésimo *bin*, normalmente fixado um valor de 10 ou 20% do número de posições alocadas dentro *bin*. Esta função é denotada por $u\left(\lambda_j^*,\lambda_j^{(t)}\right)$, para o j-ésimo *bin* e a nova posição λ_j^* é aceita na t-ésima iteração com $min\,(1,\alpha_j)$ de probabilidade. Assim, se α_j for aceito, uma nova posição é estabelecida λ_j^* e o estado do marcador $Z_{\lambda_j^*}$ é amostrado do painel completo. A regra de decisão para mudança da posição do marcador dentro do *bin* é dada por:

$$\alpha_j = \frac{p\left(\lambda_j^*|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma^2_\gamma},\mathbf{y}\right) u\left(\lambda_j^*,\lambda_j^{(t)}\right)}{p\left(\lambda_j^{(t)}|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma^2_\gamma},\mathbf{y}\right) u\left(\lambda_j^{(t)},\lambda_j^*\right)} \tag{2.15}$$

em que

$$p\left(\lambda^*|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma^2_\gamma},\mathbf{y}\right) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{k-1} Z_{\lambda_{j(i)}}\gamma_{\lambda_j} - Z_{\lambda^*}\gamma_{\lambda^*}\right)^2\right\} \tag{2.16}$$

e

$$p\left(\lambda_j^{(t)}|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma^2_\gamma},\mathbf{y}\right) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{k} Z_{\lambda_{j(i)}}\gamma_{\lambda_j}\right)^2\right\} \tag{2.17}$$

A correção de Hastings é dada por:

$$u\left(\lambda_j^{(t)},\lambda^*\right) = \begin{cases} \dfrac{1}{2c} & , \quad \text{se } \lambda_j^{(t)} + c \leq LS_j \text{ e } \lambda_j^{(t)} - c \geq LI_j \\ \dfrac{1}{c + \lambda_j^{(t)} - LI_j}, & \text{se } \lambda_j^{(t)} + c < LS_j \text{ e } \lambda_j^{(t)} - c < LI_j \\ \dfrac{1}{c + LS_j - \lambda_j^{(t)}}, & \text{se } \lambda_j^{(t)} + c > LS_j \text{ e } \lambda_j^{(t)} - c > LI_j \end{cases} \tag{2.18}$$

$$u\left(\lambda^*,\lambda_j^{(t)}\right) = \begin{cases} \dfrac{1}{2c} & , \quad \text{se } \lambda^* + c \leq LS_j \text{ e } \lambda^* - c \geq LI_j \\ \dfrac{1}{c + \lambda^* - LI_j}, & \text{se } \lambda^* + c < LS_j \text{ e } \lambda^* - c < LI_j \\ \dfrac{1}{c + LS_j - \lambda^*}, & \text{se } \lambda^* + c > LS_j \text{ e } \lambda^* - c > LI_j \end{cases} \tag{2.19}$$

5 A condicional para a variância residual, após aceitar a j-ésima posição no genoma, é dada por:

$$\sigma_e^2 | \ldots \sim \chi_{esc}^{-2}\left(n + \nu, FQ\right) \tag{2.20}$$

em que $FQ = \sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{k} Z_{\lambda_{j(i)}}\gamma_{\lambda_j}\right)^2$.

6 Finalmente, amostrar a variância da *j*-ésima marca $\sigma^2_{\gamma_{\lambda_j}}$ utilizando:

$$\sigma^2_{\gamma_{\lambda_j}} | \ldots \sim \chi_{esc}^{-2}\left(\upsilon + 1, \gamma^2_{\lambda_j} + S^2\right) \tag{2.21}$$

Repete-se a sequência de 1 a 6 até a convergência da cadeia para uma distribuição estacionária. Na cadeia final, adotou-se $f(\lambda|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma^2_\gamma},\mathbf{y}) = P_\lambda$ como a frequência de visitas realizadas na posição λ_j dentro de um *bin* específico.

Contudo, a integral $\int_0^L Z_i(\lambda)\gamma(\lambda)d\lambda$ para recompor o valor genético genômico não é conhecida e $\gamma(\lambda)$ também não. Hu et al. (2012) utilizaram o efeito médio dos *bin* como a esperança de $p\left(\lambda|\boldsymbol{\mu},\sigma_e^2,\boldsymbol{\gamma},\boldsymbol{\sigma}^{\mathbf{2}}_{\boldsymbol{\gamma}},\mathbf{y}\right)$. Neste estudo, relaxa-se a suposição do efeito médio e, como para cada λ_j pode-se atribuir a frequência do número de vezes que o modelo visitou o marcador correspondente, é possível substituir $\int_0^L Z_i(\lambda)\gamma(\lambda)d\lambda$ por $\mathbf{Z}_i P_\lambda \hat{\boldsymbol{\gamma}}$. Assim, para um intervalo de *bin* $\left[\lambda_{j(\min)},\lambda_{j(\max)}\right]$, cada λ pode ser considerado um elemento discreto e $\lim_{N\to\infty}\left[p\left(\lambda|\mu,\sigma_e^2,\gamma,\sigma^{\mathbf{2}}_{\gamma},y\right) \triangleq f(\lambda|\mu,\sigma_e^2,\gamma,\sigma^{\mathbf{2}}_{\gamma},y)\right]$, isto é, a função de probabilidade contínua pode ser aproximada pela distribuição de frequências dada a taxa de visita para λ ao longo das N execuções do algoritmo MCMC. Assim, $f(\lambda|\mu,\sigma_e^2,\gamma,\sigma^{\mathbf{2}}_{\gamma},y) = P_\lambda$ e a relação descrita, anteriormente, é aproximada por $\hat{\gamma}_j|\lambda_j \triangleq \underset{\gamma_j\in\Re}{\arg\,.\max}\left[p(\gamma_j|y)\right]p(\lambda_j|\gamma_j,y) \doteq \underset{\gamma_j\in\Re}{\arg\,.\max}\left[p(\gamma_j|y)\right]P_\lambda$.

Uma forma mais rápida seria utilizar a média a posteriori das cadeias dos efeitos dos marcadores sem criar a função peso. Isso é facilmente obtido se na *t*-ésima iteração, caso a marca não for visitada, atribuir seu efeito como nulo $(\hat{\gamma}=0)$. Tomando ψ como a média da Cadeia de Markov após a convergência, temos $\psi = \frac{\sum_{l=1}^{N}\hat{\gamma}_l}{N} = \frac{n\sum_{l=1}^{n}\hat{\gamma}_l + n(N-n)\times 0}{nN} = \frac{n\bar{\gamma}}{N} = \hat{P}_\lambda\bar{\gamma}$, em que N é o tamanho total da Cadeia de Markov, n o número de vezes em que a marca foi selecionada, durante o processo MCMC, e $\hat{P}_\lambda = \frac{n}{N}$. Sob uma distribuição a posteriori gaussiana, tem-se que $\psi = \hat{P}_\lambda\hat{\gamma} = P_\lambda \underset{\gamma_j\in\Re}{\arg\,.\max}\left[p(\gamma_j|y)\right]$. Assim, a predição do valor genético genômico final foi dada por $Z\psi = \hat{g}$.

2.3 Capacidade preditiva

Os métodos adaptados foram implementados no software R (R CORE TEAM, 2018) e comparados com seus respectivos originais. Os métodos originais utilizados foram: RR-BLUP, por meio da função **mixed.solve** (ENDELMAN, 2011) do pacote *rr-BLUP*, Bayes A e Bayes B, por meio da função **BGLR** contida no pacote *BGLR* (PÉREZ; DE LOS CAMPOS, 2014), sendo todos bibliotecas do software R.

Três critérios foram utilizados para comparar os métodos: (1) Erro Quadrático Médio - EQM; (2) coeficiente de determinação (R^2) entre valor genético predito $\hat{g}$ e verdadeiro (g), por meio de regressão linear; (3) capacidade de detectar QTL simulado, feito pela inspeção visual de gráficos denominados Manhattan *plots* (não foi utilizado nenhum teste para isto porque o objetivo foi identificar a capacidade dos novos métodos no rastreamento dos QTL simulados, desconsiderando a significância dos seus efeitos sobre o caráter).

3 RESULTADOS

Pelo fato dos resultados das populações F_2 e F_{10} terem sido similares, foram mostrados apenas os relacionados à F_2 nesta seção. Em Apêndice A, estão fornecidos os valores de EQM e R^2 para a população F_{10}, para todos os cenários avaliados (Tabelas 5.1-5.3) e os Manhattan *plots* dos efeitos simulados e preditos pelos métodos (Figuras 5.1-5.9).

3.1 Cenário Oligogênico

Na Tabela 3.1 apresentam-se os erros quadráticos médios (EQM) e os coeficientes de determinação (R^2) entre os valores genéticos genômicos (VGG) simulados e preditos dos métodos avaliados, para as três herdabilidades, no cenário oligogênico (seis QTL).

Os métodos RR-BLUP-Bins foram iguais ou mais acurados que o RR-BLUP tradicional nas herdabilidades 0,2 e 0,8. Na herdabilidade 0,5, os métodos RR-BLUP Bin03 e RR-BLUP Bin04 foram menos acurados em relação ao RR-BLUP original, mesmo que muito próximos. Note que RR-BLUP Bin01 e RR-BLUP Bin02 apresentaram valores de R^2 bem maiores que o correspondente método tradicional, em todas as herdabilidades. Isso mostra que, num cenário oligogênico, utilizar o método RR-BLUP com poucos marcadores por iteração no processo MCMC é mais eficiente que com todo o painel de marcadores.

Em relação ao método Bayes A tradicional, todas as configurações *bins* concorrentes foram mais acuradas, apresentando valores de R^2 muito maiores. Com isso, sugere-se o uso do método Bayes A incorporado ao modelo funcional *bin* em vez do método tradicional, considerando um cenário oligogênico. Com o método Bayes B também se obteve alta capacidade preditiva das adaptações bins.

Apenas o Bayes B Bin04 foi considerado menos acurado que o Bayes B original, na herdabilidade 0,2, apesar de serem resultados bem próximos. Observe, ainda que, utilizando modelo funcional *bin* com apenas 10 ou 30 marcadores no processo MCMC (Bin01 e Bin02, respectivamente), gerou aumento significativo na acurácia preditiva de todos os métodos.

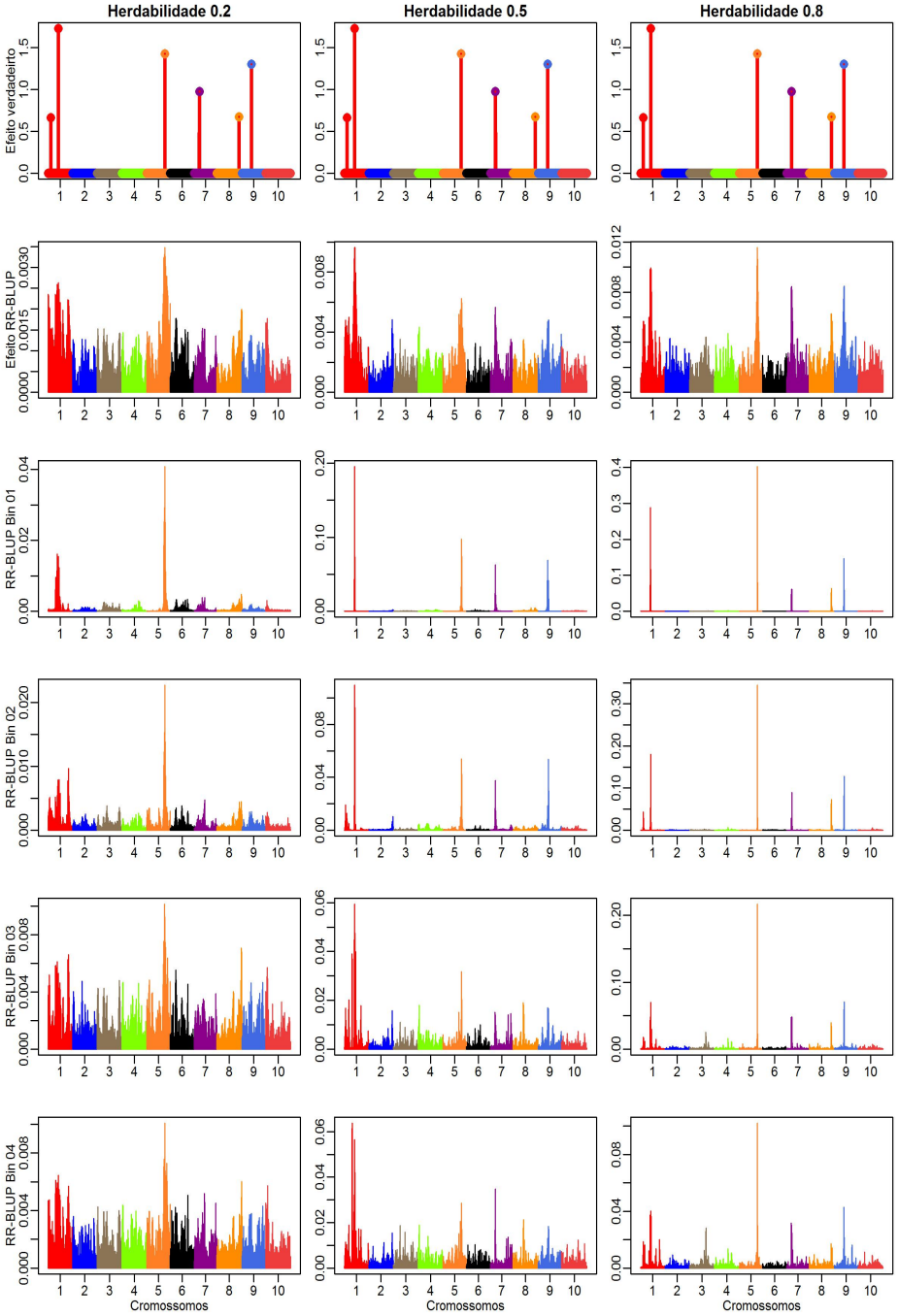
Tabela 3.1 – Erro Quadrático Médio (EQM) e Coeficiente de determinação (R^2 - em porcentagem) dos métodos RR-BLUP, Bayes A, Bayes B e suas respectivas adaptações em *bins*, para as três herdabilidades no cenário oligogênico (6 QTL).

Modelos	Herdabilidades					
	0,2		0,5		0,8	
	EQM	R^2 (%)	EQM	R^2 (%)	EQM	R^2 (%)
RR-BLUP	**668,3**[a]	**58,8**	**221,8**	**85,5**	**92,3**	**94,2**
RR-BLUP Bin01	535,8	**67,1**[b]	117,0	**92,4**	68,4	**95,7**
RR-BLUP Bin02	591,9	**62,0**	137,0	**91,0**	30,9	**98,1**
RR-BLUP Bin03	636,7	58,3	250,7	84,3	70,7	**95,4**
RR-BLUP Bin04	641,0	58,0	297,1	82,1	86,8	**94,3**
Bayes A	**650,9**	**57,5**	**229,5**	**85,1**	**83,3**	**94,8**
Bayes A Bin01	582,0	**63,3**	103,4	**93,3**	69,1	**95,6**
Bayes A Bin02	568,0	**63,9**	54,2	**96,5**	19,7	**98,8**
Bayes A Bin03	589,5	**61,8**	123,5	**92,0**	21,6	**98,7**
Bayes A Bin04	583,4	**62,8**	122,6	**92,1**	19,4	**98,8**
Bayes B	**658,0**	**56,9**	**204,9**	**86,6**	**79,5**	**95,1**
Bayes B Bin01	659,2	**58,8**	130,0	**91,6**	71,7	**95,4**
Bayes B Bin02	639,9	**58,9**	129,5	**91,6**	19,9	**98,8**
Bayes B Bin03	630,5	**59,2**	133,2	**91,3**	27,9	**98,3**
Bayes B Bin04	696,7	54,4	141,8	**90,8**	19,3	**98,8**

[a] Em negrito são os valores de EQM e R^2 para os modelos tradicionais avaliados nas três herdabilidades. Deve-se compará-los com suas respectivas adaptações *bins*. [b] Em vermelho são os valores de R^2 para os *bins* considerados melhores que ou iguais aos respectivos métodos tradicionais. Fonte: Do autor (2018).

Para obter uma ideia geral do genoma e da distribuição espacial dos sinais, na Figura 3.1 mostram-se Manhattan *plots* dos efeitos absolutos simulados e estimados pelos métodos RR-BLUP e RR-BLUP-Bins, para as três herdabilidades.

Figura 3.1 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos RR-BLUP, RR-BLUP Bin01 (10 *bins*), RR-BLUP Bin02 (30 *bins*), RR-BLUP Bin03 (90 *bins*) e RR-BLUP Bin04 (150 *bins*), no cenário oligogênico. Os pontos coloridos representam os seis QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



Fonte: Do autor (2018).

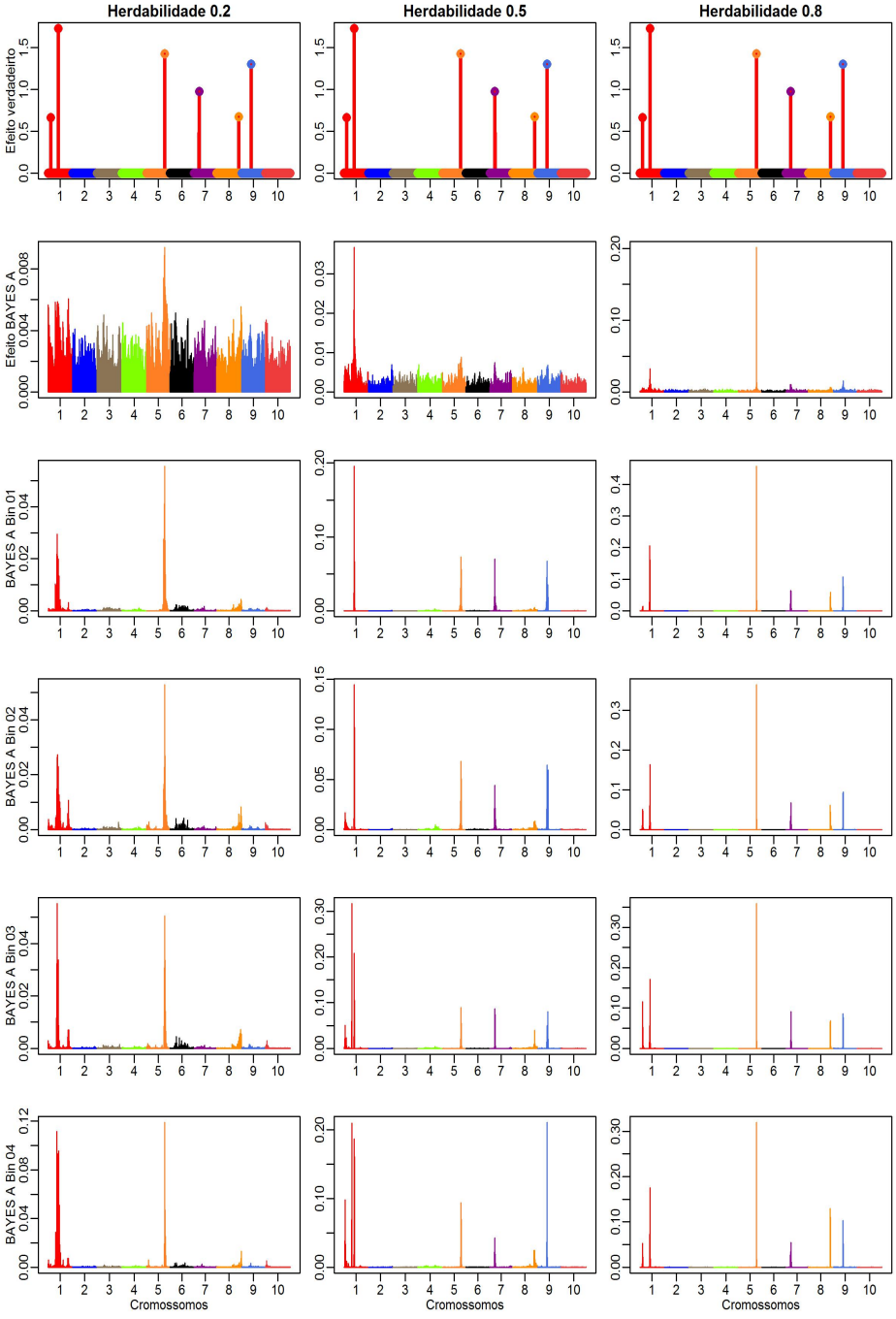
Observa-se que a maioria dos QTL foi mapeada pelos cinco métodos, descrevendo o perfil genômico, principalmente nos métodos *bins*. Mais ainda, analisando o gráfico referente ao RR-BLUP Bin01 juntamente com sua acurácia preditiva (Tabela 3.1), observa-se que utilizar 10 marcadores por iteração no processo MCMC é suficiente para descrever corretamente o cenário oligogênico.

Os métodos RR-BLUP-Bins apresentaram melhor resolução que o método tradicional. O método RR-BLUP reduziu a diferença entre efeitos de SNPs grandes e pequenos, o que levou à pior resolução do perfil genômico, com efeitos bem menores em comparação com o simulado. Como esperado, com o aumento da herdabilidade, a resolução dos métodos em definir regiões causais também aumenta.

Na Figura 3.2 mostram-se os efeitos absolutos dos QTL simulados e os efeitos absolutos que foram estimados pelos métodos Bayes A e suas configurações *bins*, para as três herdabilidades.

Observa-se que a maioria dos QTL foi mapeada pelos métodos *bins*, descrevendo claramente o padrão dos efeitos. Já o método Bayes A tradicional identifica apenas uma ou duas regiões causais. O método Bayes A também apresenta forte encolhimento dos efeitos, porém os métodos Bayes A Bins ainda obtiveram magnitudes dos efeitos maiores que as do Bayes A tradicional. Observe que, novamente, com apenas o uso do Bayes A Bin01 já se consegue boa resolução do perfil genômico.

Figura 3.2 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, no cenário oligogênico. Os pontos coloridos representam os seis QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.

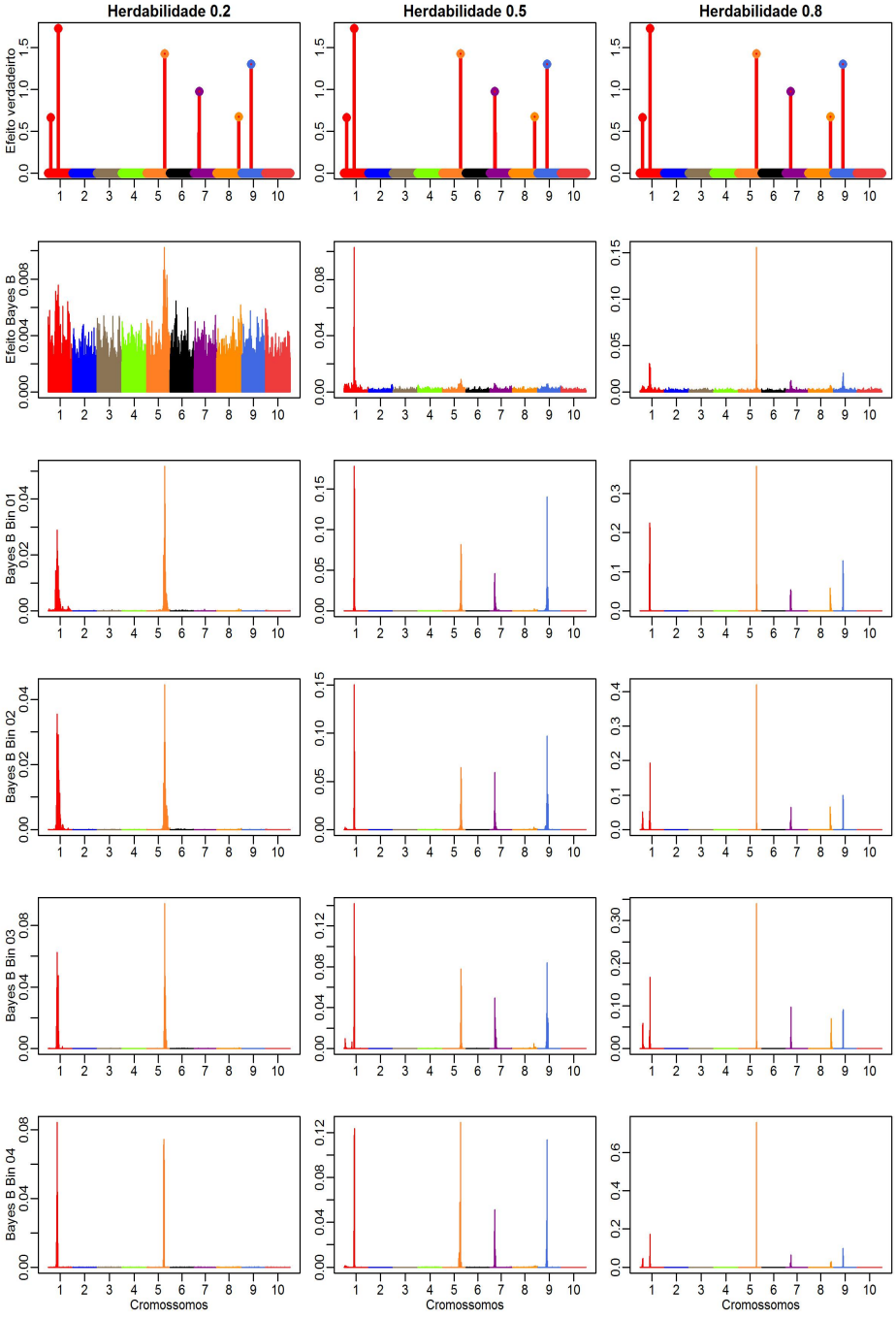


Fonte: Do autor (2018).

Na Figura 3.3 apresentam-se os efeitos absolutos dos QTL simulados e os efeitos absolutos estimados pelos métodos Bayes B tradicional, Bayes B Bin01, Bayes B Bin02, Bayes B Bin 03 e Bayes B Bin04, para as três herdabilidades.

Observa-se que a maioria dos QTL foi mapeada pelos métodos *bins*, descrevendo, novamente, o padrão dos efeitos à medida que aumenta a herdabilidade. Os métodos Bayes B Bins apresentaram melhor resolução que o respectivo método tradicional. Assim como os métodos acima, Bayes B também apresenta forte encolhimento dos efeitos, o que já era esperado pela característica *shrinkage* destes métodos. Entretanto, os métodos Bayes B Bins ainda obtiveram magnitudes dos efeitos maiores que as do método tradicional.

Figura 3.3 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes B, Bayes B Bin01, Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04, no cenário oligogênico. Os pontos coloridos representam os seis QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



Fonte: Do autor (2018).

3.2 Cenário Poligênico I

Na Tabela 3.2 apresentam-se os erros quadráticos médios (EQM) e os coeficientes de determinação (R^2) entre os valores genéticos genômicos (VGG) simulados e preditos dos métodos avaliados, para as três herdabilidades, no cenário poligênico I (15 QTL). Já se percebe bastante diferença comparando-a com a Tabela 3.1.

Tabela 3.2 – Erro Quadrático Médio (EQM) e Coeficiente de determinação (R^2 - em porcentagem) dos métodos RR-BLUP, Bayes A, Bayes B e suas respectivas adaptações em *bins*, para as três herdabilidades no cenário poligênico I (15 QTL).

Modelos	Herdabilidades					
	0,2		0,5		0,8	
	EQM	R^2 (%)	EQM	R^2 (%)	EQM	R^2 (%)
RR-BLUP	**407,1**[a]	**69,9**	**229,7**	**82,8**	**103,4**	**92,3**
RR-BLUP Bin01	361,5	**72,7**[b]	250,3	81,3	152,5	88,8
RR-BLUP Bin02	402,0	**71,4**	212,4	**84,1**	75,4	**94,4**
RR-BLUP Bin03	564,6	66,4	243,0	81,7	89,0	**93,3**
RR-BLUP Bin04	622,8	64,9	285,8	79,1	101,4	**92,3**
Bayes A	**477,1**	**68,4**	**250,7**	**81,1**	**85,5**	**93,7**
Bayes A Bin01	323,7	**76,9**	263,1	80,3	159,0	88,3
Bayes A Bin02	286,1	**78,4**	242,1	**81,8**	75,2	**94,4**
Bayes A Bin03	308,3	**76,6**	268,4	79,8	59,5	**95,6**
Bayes A Bin04	314,1	**76,2**	248,3	**81,3**	61,1	**95,5**
Bayes B	**463,6**	**69,3**	**227,8**	**82,8**	**81,4**	**94,0**
Bayes B Bin01	459,2	67,8	279,1	79,1	161,8	88,1
Bayes B Bin02	448,6	68,6	291,2	78,1	79,0	**94,1**
Bayes B Bin03	520,4	61,5	314,9	76,2	68,6	**94,9**
Bayes B Bin04	478,3	63,8	299,5	77,3	69,2	**94,8**

[a] Em negrito são os valores de EQM e R^2 para os modelos tradicionais avaliados nas três herdabilidades. Deve-se compará-los com suas respectivas adaptações *bins*. [b] Em vermelho são os valores de R^2 para os *bins* considerados melhores que ou iguais aos respectivos métodos tradicionais. Fonte: Do autor (2018).

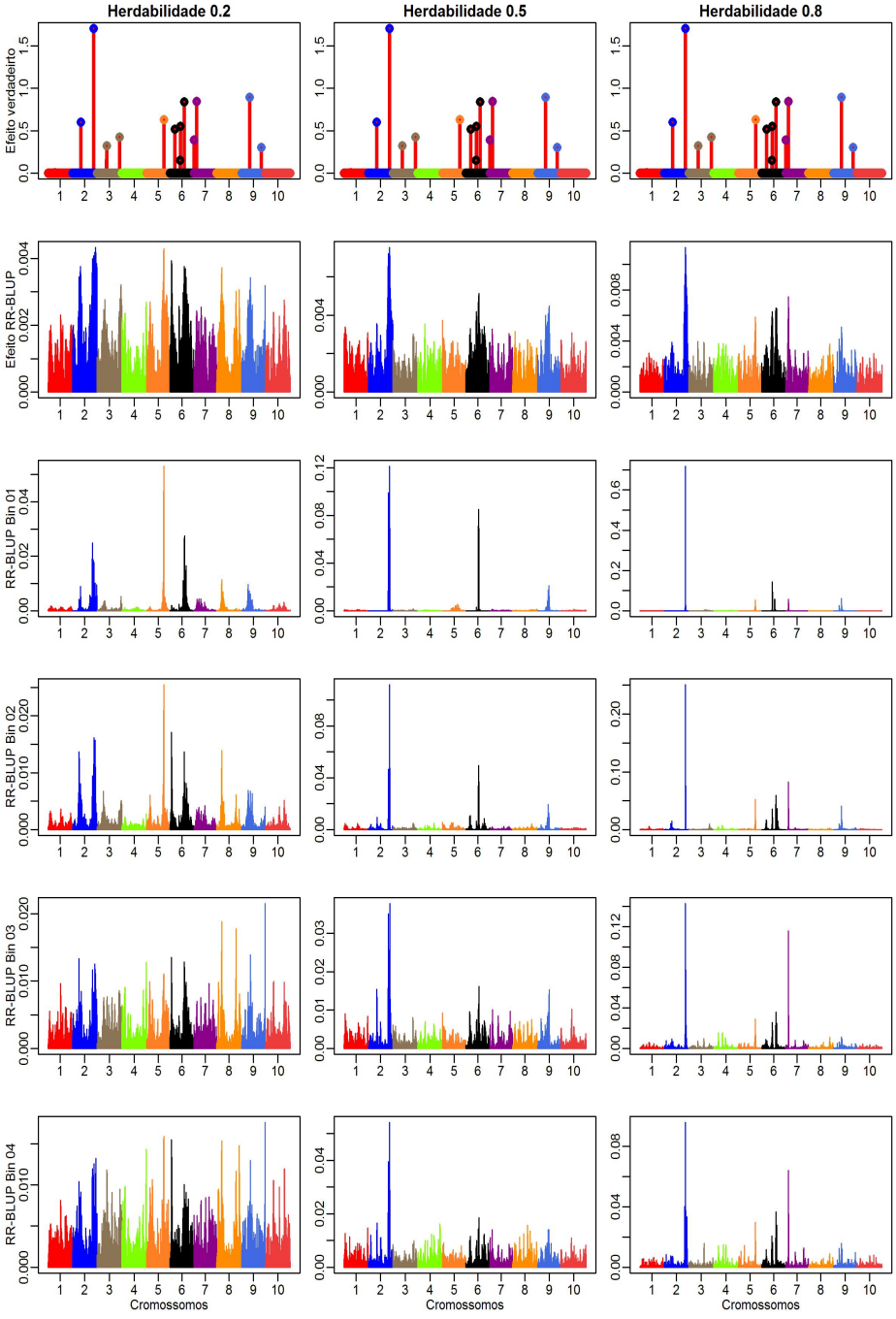
Neste cenário, os métodos *bins* começam a ser menos acurados, princi-

palmente para os métodos Bayes B Bins. Entretanto, verifica-se pouca diferença nas acurácias obtidas pelos métodos originais e seus respectivos métodos *bins*. Observa-se que os métodos RR-BLUP-Bins avaliados tiveram desempenhos muito diferentes, de acordo com as herdabilidades, quando comparados ao RR-BLUP tradicional: RR-BLUP Bin01 e RR-BLUP Bin02 foram mais acurados na herdabilidade 0,2; apenas o RR-BLUP Bin02 obteve melhor acurácia preditiva na herdabilidade 0,5; e somente o RR-BLUP Bin 01 foi o menos acurado na herdabilidade 0,8. Com isso, o RR-BLUP Bin02 foi o único que apresentou melhor desempenho preditivo nas três herdabilidades, ou seja, considerando 30 *bins* no modelo dentro de cada passo do MCMC.

De acordo com o apresentado na Tabela 3.2, em relação ao método Bayes A tradicional, a maioria das configurações *bins* concorrentes foi mais acurada, apresentando valores de R^2 muito maiores na herdabilidade 0,2. Observe que os valores de R^2, para os dois *bins* menos acurados na herdabilidade 0,5, estão muito próximos ao do método RR-BLUP. Já os métodos Bayes B Bins não obtiveram bons desempenhos no cenário poligênico, pois foram considerados todos menos acurados que o Bayes B original nas herdabilidades 0,2 e 0,5. Na herdabilidade 0,8, os métodos Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04 tiveram acurácia equivalentes ao Bayes B original.

Na Figura 3.4 mostram-se os Manhattan *plots* com os efeitos absolutos dos QTL simulados e os efeitos absolutos que foram estimados pelos métodos RR-BLUP, RR-BLUP Bin01 (10 *bins*), RR-BLUP Bin02 (30 *bins*), RR-BLUP Bin03 (90 *bins*) e RR-BLUP Bin04 (150 *bins*), para as herdabilidades 0,2; 0,5 e 0,8. Observa-se que a maioria dos QTL foi mapeada pelos cinco métodos, descrevendo o padrão dos efeitos principalmente nos métodos *bins*.

Figura 3.4 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos RR-BLUP, RR-BLUP Bin01, RR-BLUP Bin02, RR-BLUP Bin03 e RR-BLUP Bin04, no cenário poligênico I. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



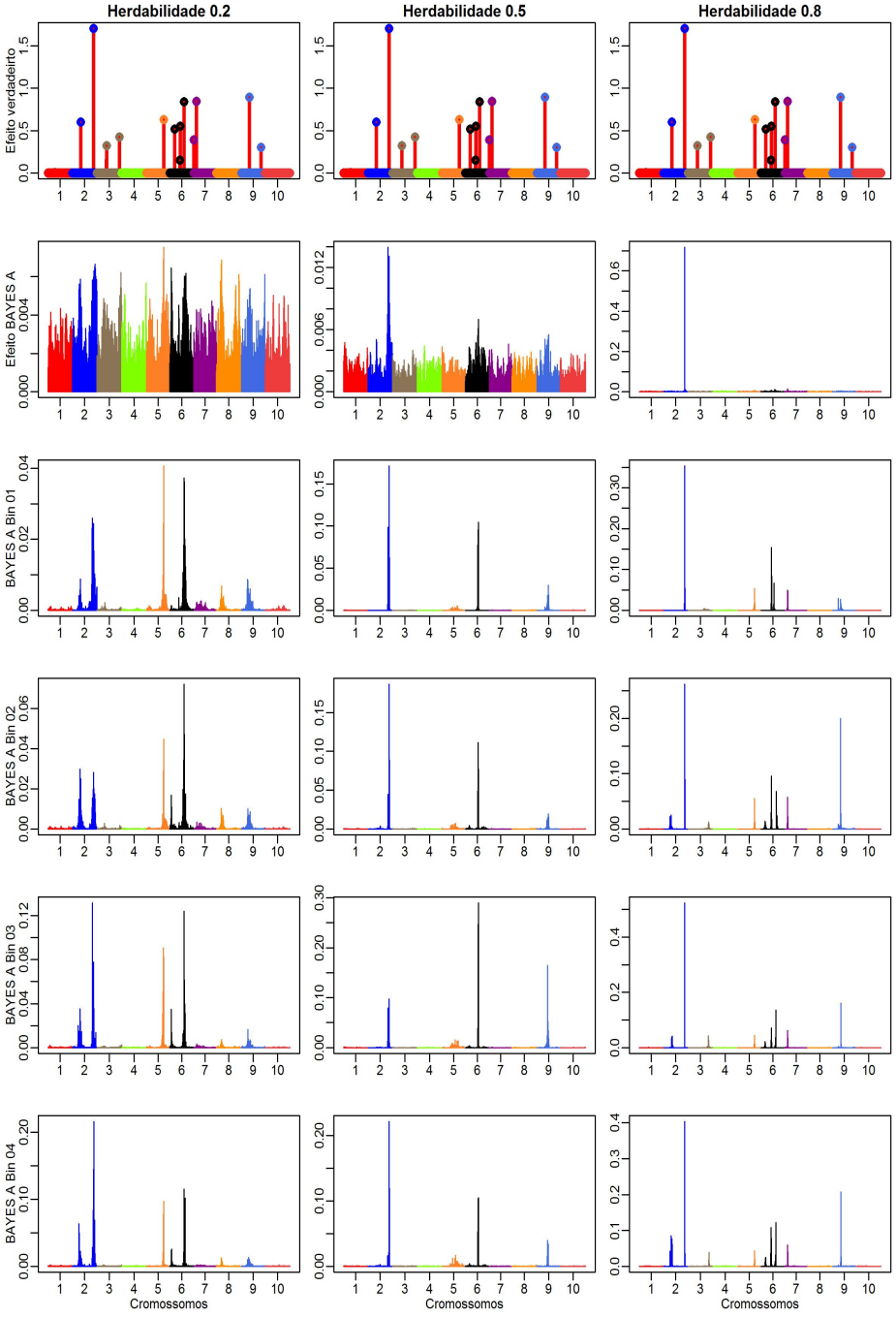
Fonte: Do autor (2018).

Os métodos RR-BLUP-Bins apresentaram melhor resolução que o método tradicional; entretanto, em geral, houve forte efeito de encolhimento, subestimando os efeitos simulados. Como esperado, com o aumento da herdabilidade, a resolução dos métodos em definir regiões causais também aumenta. Note que utilizar o método RR-BLUP Bin02 apresentou melhor acurácia preditiva e melhor resolução do perfil genômico. Isto implica que utilizar 30 marcadores nas iterações do processo MCMC gera melhores resultados que utilizar todos os 12150 marcadores. Ou seja, para o cenário poligênico simulado com 15 QTL, foram necessários apenas 30 marcadores para conseguir reproduzi-lo melhor.

Na Figura 3.5 mostram-se os efeitos absolutos dos QTL simulados e os efeitos absolutos que foram estimados pelos métodos Bayes A e suas configurações *bins*, para as três herdabilidades. Observa-se que a maioria das regiões dos QTL foi mapeada pelos métodos *bins*, descrevendo o perfil genômico. Já o método Bayes A tradicional identifica apenas uma ou duas regiões causais.

Assim como no RR-BLUP, o método Bayes A apresenta forte encolhimento dos efeitos, porém os métodos Bayes A Bins ainda obtiveram magnitudes dos efeitos maiores. Observe que, novamente, com apenas o uso do Bayes A Bin01 já se consegue melhor resolução do perfil genômico que o Bayes A. Entrentao, os métodos *bins* agruparam alguns QTL e os identificaram como apenas um.

Figura 3.5 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, no cenário poligênico I. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.

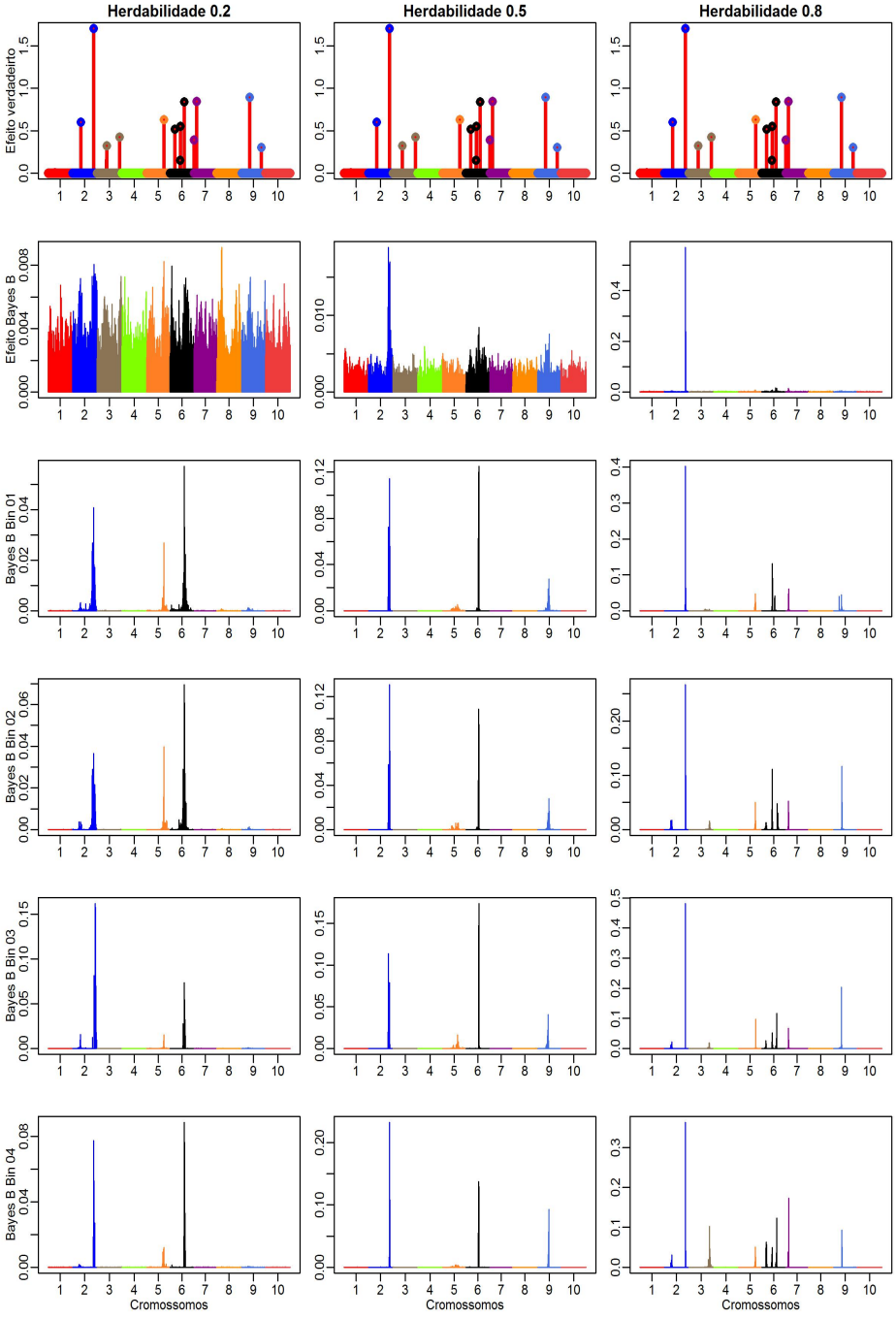


Fonte: Do autor (2018).

Na Figura 3.6 apresentam-se os efeitos absolutos dos QTL simulados e os efeitos absolutos estimados pelos métodos Bayes B e suas configurações *bins*, para as três herdabilidades. Neste cenário, apenas o QTL de maior efeito absoluto foi identificado por todos os métodos.

Assim como os métodos acima, Bayes B também apresenta forte encolhimento dos efeitos, o que já era esperado pela característica *shrinkage* destes métodos. Observa-se que a maioria das regiões causais não foi mapeada pelos métodos *bins*. Nas herdabilidades 0,2 e 0,5, os QTL de efeitos menores foram identificados como um único QTL de efeito maior. Apenas na herdabilidade 0,8, os métodos *bins* apresentaram melhor resolução que o respectivo método tradicional.

Figura 3.6 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes B, Bayes B Bin01, Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04, no cenário poligênico I. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



Fonte: Do autor (2018).

3.3 Cenário Poligênico II

Na Tabela 3.3 apresentam-se os erros quadráticos médios (EQM) e os coeficientes de determinação (R^2) entre os valores genéticos genômicos (VGG) simulados e preditos dos métodos avaliados, para as três herdabilidades, no cenário poligênico II (60 QTL).

Tabela 3.3 – Erro Quadrático Médio (EQM) e Coeficiente de determinação (R^2 - em porcentagem) dos métodos RR-BLUP, Bayes A, Bayes B e suas respectivas adaptações em *bins*, para as três herdabilidades no cenário poligênico II (60 QTL).

Modelos	Herdabilidades					
	0,2		0,5		0,8	
	EQM	R^2 (%)	EQM	R^2 (%)	EQM	R^2 (%)
RR-BLUP	**5600,7**[a]	**54,1**	**2248,1**	**81,6**	**964,9**	**92,1**
RR-BLUP Bin01	6341,8	47,9	3623,4	70,2	3017,1	75,6
RR-BLUP Bin02	5610,7	**54,9**[b]	2127,6	**82,5**	819,7	**93,3**
RR-BLUP Bin03	5908,6	**54,0**	2485,2	80,3	866,3	**92,9**
RR-BLUP Bin04	5910,8	53,6	3097,2	77,4	963,7	**92,2**
Bayes A	**5972,8**	**53,5**	**2328,7**	**81,1**	**955,5**	**92,2**
Bayes A Bin01	7607,4	37,6	3803,4	68,8	3141,2	74,8
Bayes A Bin02	7467,9	39,0	2721,9	77,7	1072,7	91,3
Bayes A Bin03	7788,7	37,0	2851,4	76,6	1057,6	91,3
Bayes A Bin04	7662,4	37,2	2918,7	76,0	1061,9	91,3
Bayes B	**5800,3**	**53,6**	**2230,7**	**81,7**	**960,3**	**92,2**
Bayes B Bin01	11680,1	20,1	4273,4	65,1	3678,6	70,5
Bayes B Bin02	11786,8	20,1	3933,3	67,8	2140,3	82,6
Bayes B Bin03	12012,4	20,4	4597,9	62,3	2307,5	81,1
Bayes B Bin04	12086,1	29,8	4144,7	65,9	2239,6	81,7

[a] Em negrito são os valores de EQM e R^2 para os modelos tradicionais avaliados nas três herdabilidades. Deve-se compará-los com suas respectivas adaptações *bins*. [b] Em vermelho são os valores de R^2 para os *bins* considerados melhores que ou iguais aos respectivos métodos tradicionais. Fonte: Do autor (2018).

Observa-se que os métodos RR-BLUP-Bins avaliados tiveram desempenhos muito diferentes, de acordo com as herdabilidades, quando comparados ao

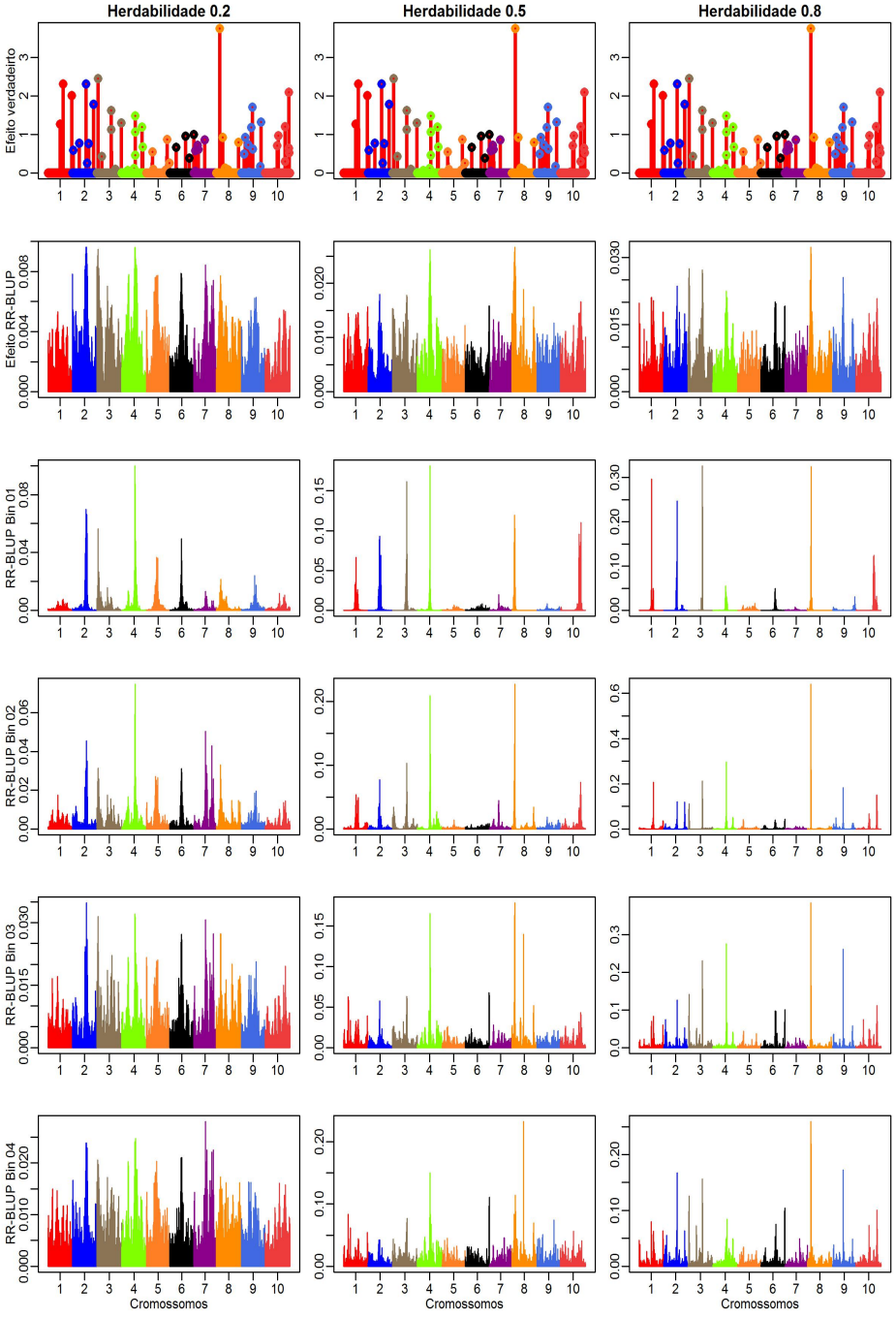
RR-BLUP tradicional: RR-BLUP Bin02 e RR-BLUP Bin03 tiveram o mesmo desempenho que o RR-BLUP original na herdabilidade 0,2; apenas o RR-BLUP Bin02 obteve melhor acurácia preditiva na herdabilidade 0,5; e somente o RR-BLUP Bin 01 foi o menos acurado na herdabilidade 0,8. Com isso, o RR-BLUP Bin02 foi o único que apresentou melhor desempenho preditivo nas três herdabilidades, ou seja, considerando 30 *bins* no modelo dentro de cada passo do MCMC.

Para os métodos Bayes A e Bayes B, tanto Bayes A Bins quanto Bayes B Bins não foram considerados melhores. Isso mostra que os métodos RR-BLUP Bins são favorecidos pelas suposições acerca de uma arquitetura infinitesimal.

Na Figura 3.7 mostram-se os efeitos absolutos dos QTL simulados e os efeitos absolutos que foram estimados pelos métodos RR-BLUP, RR-BLUP Bin01 (10 *bins*), RR-BLUP Bin02 (30 *bins*), RR-BLUP Bin03 (90 *bins*) e RR-BLUP Bin04 (150 *bins*), para as herdabilidades 0,2; 0,5 e 0,8.

Os métodos RR-BLUP Bin03 e RR-BLUP Bin04 apresentaram melhores resoluções que o método tradicional. Note que, na herdabilidade 0,8, ao utilizar os métodos RR-BLUP Bin03 e RR-BLUP Bin04, boa parte dos QTL foi identificada. Entretanto, em geral, houve forte efeito de encolhimento, subestimando os efeitos simulados. Como esperado, com o aumento da herdabilidade, a resolução dos métodos em definir regiões causais também aumenta. O QTL de maior efeito foi identificado em todos os métodos *bins* nas herdabilidades 0,5 e 0,8.

Figura 3.7 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos RR-BLUP, RR-BLUP Bin01, RR-BLUP Bin02, RR-BLUP Bin03 e RR-BLUP Bin04, no cenário poligênico II. Os pontos coloridos representam os 60 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.

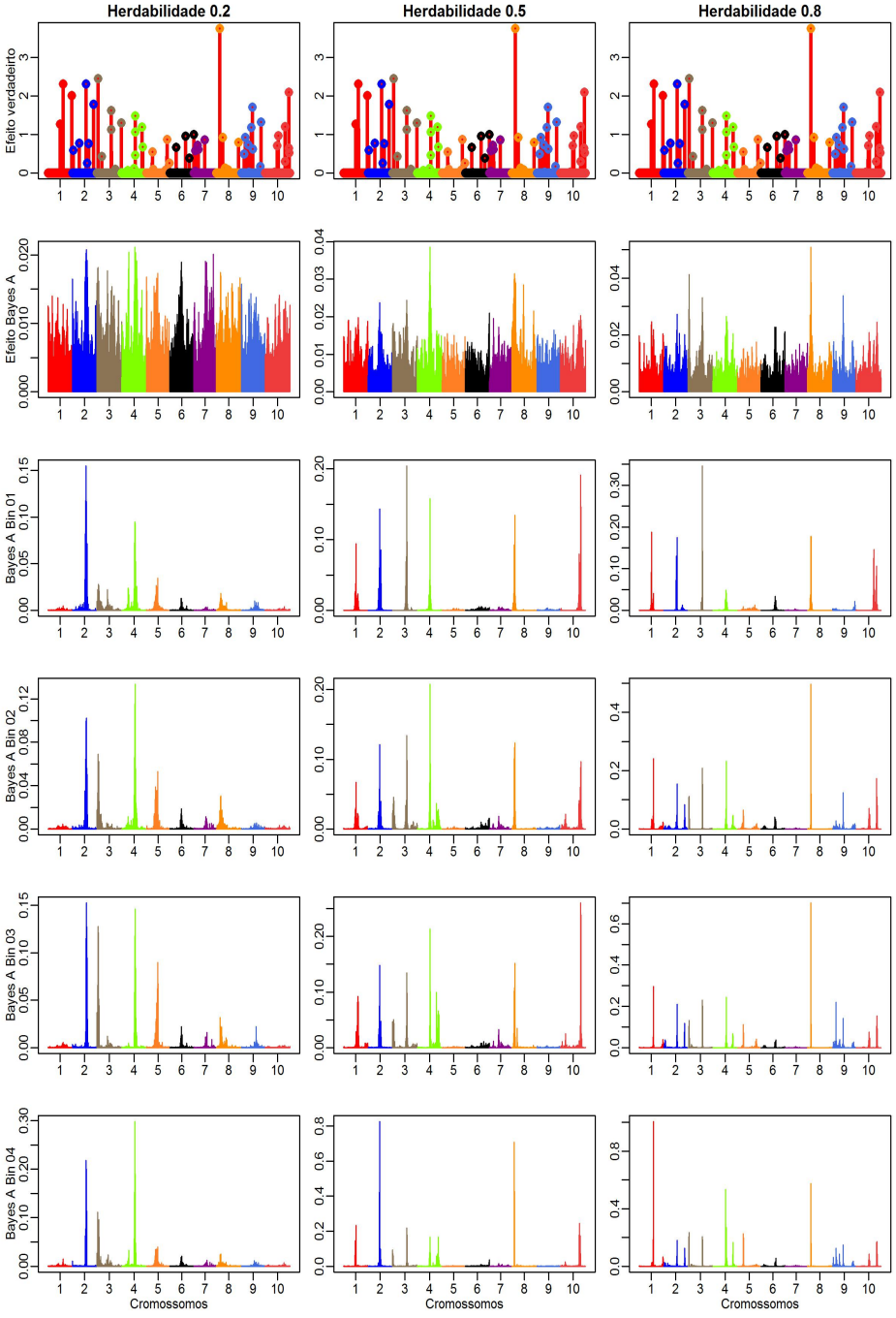


Fonte: Do autor (2018).

Na Figura 3.8 mostram-se os efeitos absolutos dos QTL simulados e os efeitos absolutos que foram estimados pelos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, para as três herdabilidades.

Observa-se que grande parte dos QTL não foi mapeada pelos métodos *bins*, sendo os QTL encolhidos próximos ao valor zero. Já o método Bayes A tradicional identifica apenas uma ou duas regiões causais. Assim como no RR-BLUP, o método Bayes A apresenta forte encolhimento dos efeitos, porém os métodos Bayes A Bins ainda obtiveram magnitudes dos efeitos maiores que as do Bayes A tradicional. Observe que, novamente, com apenas o uso do Bayes A Bin01 já se consegue boa resolução do perfil genômico.

Figura 3.8 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, no cenário poligênico II. Os pontos coloridos representam os 60 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.

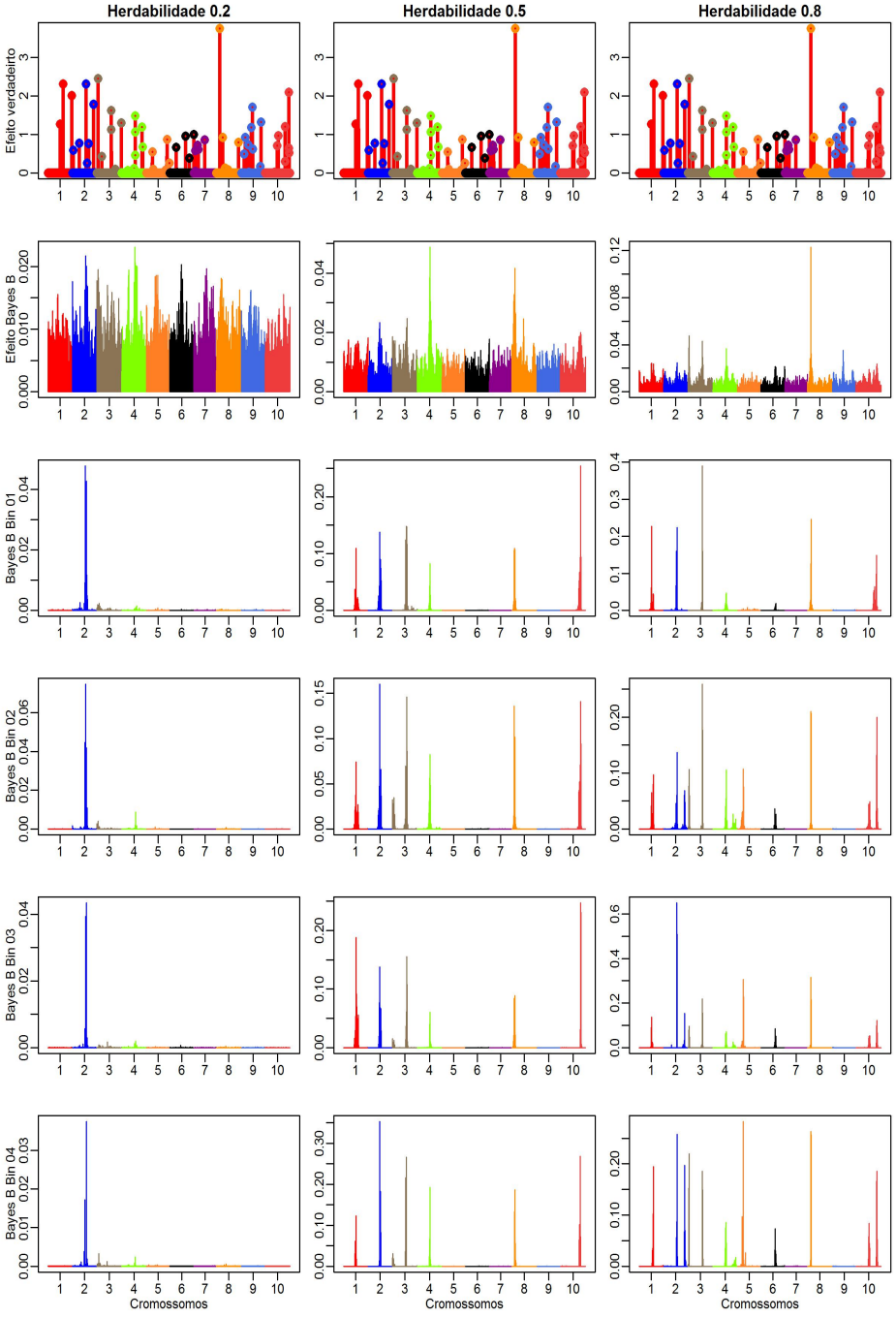


Fonte: Do autor (2018).

Na Figura 3.9 apresentam-se os efeitos absolutos dos QTL simulados e os efeitos absolutos estimados pelos métodos Bayes B e suas configurações *bins*, para as três herdabilidades.

Observa-se que a maioria das regiões causais não foi mapeada pelos métodos *bins*. O QTL de maior efeito foi identificado pelos métodos *bins* nas herdabilidades 0,5 e 0,8. Entretanto, na herdabilidade 0,2, poucos QTL foram identificados, sendo localizados apenas no cromossomo 2. Este fato corrobora com a baixa acurácia apresentada na Tabela 3.3.

Figura 3.9 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes B, Bayes B Bin01, Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04, no cenário poligênico II. Os pontos coloridos representam os 60 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



Fonte: Do autor (2018).

4 DISCUSSÃO

Hu, Wang e Xu (2012) e Xu (2013) trabalharam a ideia de que os efeitos genéticos dos marcadores têm uma estrutura funcional intrínseca e que marcadores em LD podem ser agrupados em *bins*. Intuitivamente, essa abordagem faz sentido na análise genética. Por exemplo, assuma que o caráter é afetado por um QTL causal. Então, os marcadores em torno deste QTL podem ter impacto no caráter, em função do desequilíbrio de ligação. Assim, tratá-los como um único pseudomarcador que representa o bloco genômico faz sentido. Fan et al. (2013) consideraram a ideia de utilizar modelo funcional para testar associações entre QTL e variantes genéticas em genoma humano. Seus resultados mostraram que modelos lineares funcionais de efeito fixo e seus testes são robustos e podem ser úteis em GWAS, enquanto os modelos funcionais de efeitos mistos podem ser úteis em análises de genes candidatos. Mais ainda, o desempenho superior dos modelos funcionais é mais provável em virtude do uso tanto da ligação genética quanto da informação LD de múltiplas variantes genéticas em uma região do genoma humano.

No presente estudo, introduziu-se a adaptação dos métodos RR-BLUP, Bayes A e Bayes B com o uso da abordagem de modelos funcionais e integração numérica com *bins*, proposta por Moura (2017), a fim de identificar se é possível aumentar a capacidade preditiva destes métodos por meio de uma pré-seleção de marcadores representativos (ou seja, *bins*). A resposta é: sim, é possível. Por exemplo, no cenário oligogênico, os métodos adaptados em *bins*, quase em sua totalidade, foram mais acurados que suas respectivas formas originais, independente do tipo de população analisada (Tabelas 3.1 e Tabela 5.1). Com isso, os milhares de marcadores das duas populações investigadas neste estudo são totalmente bem representados por 10, 30, 90 ou 150 *bins*. Isto corrobora os resultados encontrados por Hu, Wang e Xu (2012), Moura (2017) e Xu (2013) de que a capacidade pre-

ditiva do modelo funcional *bin* foi maior que aqueles dos modelos tradicionais, na maioria dos casos. Ou seja, a incorporação da técnica *bin* aos métodos de seleção é uma ótima abordagem de redução de dados sem perda de informação (LI et al., 2018; XU, 2013).

Em um cenário poligênico/infinitesimal, os métodos RR-BLUP-Bins foram mais acurados que as versões *bins* dos métodos bayesianos investigados. Por outro lado, os métodos Bayes A Bins e Bayes B Bins apresentaram vantagens no cenário oligogênico. Isto é consistente com estudos anteriores que indicam que métodos baseados em GBLUP (equivalente ao RR-BLUP) são melhores, quando o número de QTL é maior (pois assume um modelo infinitesimal), enquanto métodos bayesianos têm a vantagem de seleção quando caracteres são influenciados por poucos QTL principais (DAETWYLER et al., 2010; VAN DEN BERG; CALUS; WIENTJES, 2015; ZHANG et al., 2016).

Hu, Wang e Xu (2012) e Xu (2013) utilizaram a média dos efeitos dos marcadores dentro de um *bin* para representar seu efeito. O problema que surge é se os marcadores dentro do *bin* estão em baixo LD, efeitos contrários são cancelados um com o outro, levando a um efeito zero. No contexto do nosso trabalho, os pesos utilizados para os efeitos dos marcadores foram baseados na frequência relativa com que a posição do marcador foi selecionada dentro do processo MCMC. Esta abordagem não exige alto LD na população, como no modelo de Hu, Wang e Xu (2012), já que não ocorre cancelamento dos efeitos em razão dos pesos diferentes. Pelo fato do método ser estocástico, a estrutura LD não influenciou a análise, pois os resultados de capacidade preditiva dos métodos *bins* para as duas populações avaliadas F_2 (alto LD) e F_{10} (baixo LD) foram similares (Tabelas 3.1-3.3, 5.1-5.3). Diferente do caso de Hu, Wang e Xu (2012), no presente trabalho, os *bins* não são assumidos em função do LD.

Zhang et al. (2016) também utilizaram a técnica de divisão em *bins* (o que eles chamam de janelas genômicas) para atribuir pesos aos SNP, na matriz de parentesco genômico, e avaliaram predições genômicas em animais. Os autores apresentaram dois procedimentos para calcular pesos de SNP individualmente e três procedimentos para pesos de SNP nas janelas genômicas. De acordo com os resultados destes autores, utilizar janelas para ponderar SNP resultou em melhor desempenho, quando comparado com SNP ponderados individualmente, porque a incerteza foi menor (SU et al., 2014), para cenários com poucos QTL e, também, evitavam valores extremamente pequenos para pesos de SNP. Estes resultados são parecidos com os encontrados no presente trabalho, no contexto avaliado, já que no cenário oligogênico os métodos adaptados com *bins* foram iguais ou mais acurados que os respectivos originais.

Alguns autores (COMBS; BERNARDO, 2013; DAETWYLER et al., 2010; DESTA, ORTIZ, 2014; SU et al., 2017; YU et al., 2011) afirmam que a eficiência de cada modelo na seleção genômica depende de vários atributos interconectados, como a arquitetura genética (por exemplo, número de QTL, decaimento do desequilíbrio de ligação, herdabilidade, etc.) ou estrutura populacional (por exemplo, tamanho da amostra, densidade de marcadores e número de indivíduos fenotipados). Sob estas circunstâncias, as hipóteses do modelo são um importante fator. Outro problema é que a arquitetura genética é desconhecida e uma escolha incorreta do número de *bins* pode diminuir o desempenho dos métodos *bin* em relação aos respectivos originais.

Em geral, em nossos resultados, houve forte efeito de encolhimento para todos os métodos, subestimando os efeitos simulados; porém, o encolhimento foi maior para o RR-BLUP. Levando isso em consideração, o modelo funcional reduziu o efeito de encolhimento. Sob o ponto de vista da identificação de QTL, ana-

lisando todos os cenários, usar os métodos RR-BLUP-Bin, Bayes A Bin e Bayes B Bin produziu Manhattan *plots* mais claros (maior resolução, menos ruído) que os métodos originais, para o cenário oligogênico, mesmo que com escala diferente dos efeitos de marcadores. Entretanto, os resultados mostraram que o uso de *bins*, quando o caráter é influenciado por vários QTL (cenários poligênicos I e II), não fornece bom desempenho na resolução do Manhattan *plot*.

Do ponto de vista computacional, a abordagem baseada em *bins* tem uma vantagem distinta sobre os métodos com SNPs individuais. Por exemplo, no caso dos dados simulados no cenário oligogênico, para a população F_2, os 12150 marcadores originais podem ser resumidos com apenas 10 ou 30 *bins* em um processo MCMC (Tabela 3.1). Isso leva a uma redução substancial na complexidade computacional. Usando nossa própria rotina R, realizar uma análise clássica RR-BLUP leva cerca de 100 dias usando 10000 iterações em um computador desktop Windows 7 de 64 bits com uma CPU Intel (i5) de 1,6 GHz e 6 GB de memória instalada (as estimativas do tempo computacional para todos os métodos foram implementadas em um único núcleo). Usando a mesma configuração, a análise RR-BLUP Bins levou apenas 35 minutos para 10 *bins*, 38 minutos para 30 *bins*, 1,27 horas para 90 *bins* e 2,40 horas para 150 *bins*. Métodos Bayesianos para predição genômica geralmente dependem da amostragem MCMC para inferência estatística. O tempo computacional para esses métodos aumenta linearmente com o número de marcadores e o número de observações (ZENG et al., 2018). A abordagem Bayes A Bins consumiu o mínimo de 1,0 hora para o menor número de *bins* e o máximo de 4,0 horas para a maior quantidade de *bins*, enquanto o método original pode levar vários meses para os dados SNPs completos. Com o método Bayes B original, o tempo de análise pode durar mais de um ano, enquanto Bayes B Bins, apresentado neste trabalho, variou de 3,0 a 6,0 horas da menor quantidade de *bins*

à maior.

A questão da demanda computacional deve ser ressaltada, pois está inversamente relacionada ao número de bins, ou seja, quanto menor o número de bins, mais rápido o processo MCMC e menor a demanda computacional (MOURA, 2017). Uma estratégia para reduzir o tempo computacional é usar computação paralela. Cheng et al. (2014) paralelizaram o MCMC usando amostragens independentes do Metropolis-Hastings. Fernando, Dekkers e Garrick (2014) distribuíram o cálculo para a amostragem de Gibbs em núcleos e nodes. No nosso trabalho, cada *bin* pode ser processado independentemente, paralelizando em muitos clusters do computador.

Na GWAS, abordagens baseadas em *bins* com tamanhos iguais têm sido criticadas, porque eles podem dividir, acidentalmente, uma região significativa em *bins* adjacentes separados, resultando potencialmente na perda de poder de detecção de QTLs (BEISSINGER et al., 2015; LI et al., 2018). Determinar o tamanho adequado do *bin* é tipicamente subjetivo, sendo apenas uma escolha do pesquisador. De acordo com Beissinger et al. (2015), uma definição subjetiva do tamanho do *bin* normalmente leva ao uso de um tamanho uniforme ao longo do genoma, o que não é apropriado, já que vários parâmetros genéticos, incluindo taxa de recombinação e LD, variam ao longo de cada cromossomo.

Alguns estudos têm sido realizados para determinar tamanhos diferentes de *bins*. Beissinger et al. (2015) introduziram um método para definir *bins* com base em pontos de interrupção guiados estatisticamente nos dados. Esse método envolve primeiro ajustar uma *Spline* de suavização cúbica aos dados e, em seguida, identificar os pontos de inflexão da *Spline* ajustada, que servem como limites dos *bins* adjacentes, permitindo assim a não uniformidade dos tamanhos dos *bins*. Outra maneira de definir os *bins* é utilizando o desequilíbrio de ligação entre os mar-

cadores, de tal maneira que marcadores consecutivos com LD maior que um limiar podem ser combinados juntos dentro de um único *bin* (LI et al., 2018; XU, 2013; ZHANG et al., 2016). Entretanto, Moura (2017) mostrou que o modelo funcional bayesiano com *bins* artificiais automaticamente define os tamanhos de *bins* naturais (blocos LD) por meio da frequência com que cada marcador é amostrado no processo MCMC. Ou seja, os tamanhos dos *bins*, na abordagem deste trabalho, são irrelevantes, já que eles são usados somente para otimizar o processo estocástico (MOURA, 2017).

Uma direção para pesquisas futuras é permitir que os dados ajudem a determinar as melhores opções, para o número e a posição dos *bins*, por meio do uso de saltos reversíveis (*Reversible Jump*) no processo MCMC, por exemplo. Uma vantagem da abordagem bayesiana é que essas variáveis podem ser tratadas como parâmetros adicionais a serem estimados. Outro ponto é estender a abordagem atual para análises de efeitos epistáticos, já que exigem altas demandas computacionais. Como a abordagem apresentada neste trabalho é capaz de reduzir a dimensionalidade dos dados, reduziria o tempo computacional destes tipos de análises.

Outra direção de pesquisa é utilizar a técnica para outros tipos de dados de marcadores, por exemplo, baseados em Sequências Simples Repetidas (SSR - *Simple Sequence Repeats*), os chamados microssatélites, já que a abordagem apresentada neste trabalho não exige pessoal especializado e equipamento sofisticado para sequenciamento automático. É necessário apenas o mapa genético ou informações físicas dos marcadores para sua obtenção, o que pode ser feito por meio de métodos de *cluster* e frequência de recombinação entre marcadores dois a dois, ou anelamento simulado, por exemplo.

Por tudo isso, o uso de informações genômicas em modelos de predição

é uma possível solução para um melhoramento genético mais efetivo. Procedimentos que consideram pesos específicos, para os marcadores em um *bin* (janela), podem ser uma escolha melhor, comparada aos marcadores únicos. Ainda, considerar um grupo de marcadores em uma mesma janela genômica pode ser mais apropriado para capturar o sinal de um QTL desconhecido. O modelo funcional bayesiano em *bins*, para métodos de seleção genômica, fornece a máxima preditividade de valor genético quando um caráter é influenciado por poucos QTL. A abordagem baseada em *bins*, também, pode ser usada, para outros modelos GS, computacionalmente intensivos, para melhorar sua eficiência computacional.

5 CONCLUSÃO

A abordagem baseada em *bins* mostrou ser atrativa em métodos de seleção genômica, pois os métodos adaptados obtiveram melhores acurácias em comparação às suas versões originais. Mais ainda, obtiveram melhor desempenho na identificação de regiões causais, diminuindo os ruídos da resolução dos Manhattan *plots*. Observou-se que os diferentes tamanhos de *bins* influenciaram no desempenho dos métodos adaptados, mostrando que a partir de certa quantidade de *bins* o desempenho diminui, igualando-se aos originais em alguns cenários. Do ponto de vista computacional, a abordagem baseada em *bins* teve vantagem distinta sobre os métodos originais, reduzindo bastante o tempo de realização da análise.

Por tudo isso, quanto ao desempenho geral, a recomendação é o uso do método RR-BLUP Bin02 (a versão RR-BLUP funcional bayesiana com 30 *bins*), pois foi o que apresentou desempenho satisfatório em todos os cenários simulados.

REFERÊNCIAS

BEISSINGER, T. M. et al. Defining window-boundaries for genomic analyses using smoothing spline techniques. **Genetics Selection Evolution**, New York, v. 47, n. 1, p. 30, Apr. 2015.

CHENG, H.; GARRICK, D. J.; FERNANDO, R. L. Parallel computing to speed up whole-genome analyses using independent Metropolis-Hastings sampling. In: WORLD CONGRESS ON GENETICS APPLIED TO LIVESTOCK PRODUCTION, 10., 2014, Vancouver. **Proceedings...** Vancouver: American Society of Animal Science, 2014. p. 1-3.

COMBS, E.; BERNARDO, R. Accuracy of genomewide selection for different traits with constant population size, heritability, and number of markers. **The Plant Genome**, New York, v. 6, n. 1, p. 1-7, Mar. 2013.

DAETWYLER, H. D. et al. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, Austin, v. 185, n. 3, p. 1021-1031, July 2010.

DESTA, Z. A.; ORTIZ, R. Genomic selection: genome-wide prediction in plant improvement. **Trends in Plant Science**, Kidlington, v. 19, n. 9, p. 592-601, Sept. 2014.

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

FAN, R. et al. Functional linear models for association analysis of quantitative traits. **Genetic Epidemiology**, New York, v. 37, n. 7, p. 726-742, Nov. 2013.

FERNANDO, R. L.; DEKKERS, J. C.; GARRICK, D. J. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. **Genetics Selection Evolution**, London, v. 46, p. 50, Sept. 2014.

GIANOLA, D. et al. Additive genetic variability and the Bayesian alphabet. **Genetics**, Austin, v. 183, n. 1, p. 347-363, Sept. 2009.

GIANOLA, D. Priors in whole-genome regression: the Bayesian alphabet

returns. **Genetics**, Austin, v. 194, n. 3, p. 573-596, July 2013.

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

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

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.

LI, Z. et al. Linkage disequilibrium clustering?based approach for association mapping with tightly linked genomewide data. **Molecular Ecology Resources**, Oxford, v. 18, n. 4, p. 809-824, July 2018.

METROPOLIS, N. et al. Equation of state calculations by fast computing machines. **The Journal of Chemical Physics**, London, v. 21, p. 1087-1092, 1953.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, v. 157, n. 4, p. 1819-1829, Apr. 2001.

MORGENTHALER, S.; THILLY, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis**, Amsterdã, v. 615, n. 1, p. 28-56, Feb. 2007.

MOURA, E. G. **Aplicação de modelos funcionais na seleção genômica ampla.** 2017. 54 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2017.

PÉREZ, P.; DE LOS CAMPOS, G. Genome wide regression and prediction with BGLR Statistical Package. **Genetics**, Austin, v. 144, n. 2, p. 164-442, Oct. 2014.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 21 jan. 2018.

SU, C. et al. High density linkage map construction and mapping of yield trait QTLs in maize (Zea mays) using the genotyping-by-sequencing (GBS) technology. **Frontiers in Plant Science**, Lausanne, v. 8, p. 706, May 2017.

_____. Comparison of genomic predictions using genomic relationship matrices built with different weighting factors to account for locus-specific variances. **Journal of Dairy Science**, Lancaster, v. 97, n. 10, p. 6547-6559, Oct. 2014.

VAN DEN BERG, S.; CALUS, M. P.; WIENTJES, Y. C. J. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. **BMC Genetics**, London, v. 16, n. 1, p. 146-157, 2015.

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

YU, H. et al. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. **PloS One**, San Francisco, v. 6, n. 3, p. e17595, Mar. 2011.

ZENG, J. et al. A nested mixture model for genomic prediction using whole-genome SNP genotypes. **PloS One**, San Francisco, v. 13, n. 3, p. e0194683, Mar. 2018.

ZHANG, X. et al. Weighting strategies for single-step genomic BLUP: an iterative approach for accurate calculation of GEBV and GWAS. **Frontiers in Genetics**, Lausanne, v. 7, p. 151, Aug. 2016.

APÊNDICE A - Resultados para População F_{10}

Nesta seção estão apresentados os resultados referentes à população F_{10} simulada.

Tabela 5.1 – Erro Quadrático Médio (EQM) e Coeficiente de determinação (R^2 - em porcentagem) dos métodos RR-BLUP, Bayes A, Bayes B e suas respectivas adaptações em *bins*, para as três herdabilidades no cenário oligogênico (6 QTL).

Modelos	Herdabilidades					
	0,2		0,5		0,8	
	EQM	R^2 (%)	EQM	R^2 (%)	EQM	R^2 (%)
RR-BLUP	**390,5**[a]	**55,3**	**149,3**	**79,4**	**72,6**	**90,1**
RR-BLUP Bin01	203,6	**74,1**[b]	46,1	**93,7**	36,4	**95,1**
RR-BLUP Bin02	324,8	**62,9**	61,8	**91,5**	31,7	**95,7**
RR-BLUP Bin03	548,1	52,3	127,0	**83,2**	44,7	**93,9**
RR-BLUP Bin04	662,3	49,2	161,2	**79,7**	54,5	**92,5**
Bayes A	**408,7**	**55,0**	**82,0**	**89,2**	**34,2**	**95,3**
Bayes A Bin01	146,4	**80,5**	39,7	**94,5**	36,2	**95,1**
Bayes A Bin02	171,6	**77,9**	38,4	**94,7**	24,8	**96,6**
Bayes A Bin03	200,5	**75,3**	37,2	**94,9**	23,0	**96,9**
Bayes A Bin04	253,0	**69,8**	37,6	**94,8**	21,6	**97,1**
Bayes B	**416,7**	**55,2**	**58,9**	**91,9**	**39,7**	**94,6**
Bayes B Bin01	119,7	**83,6**	43,0	**94,1**	51,5	93,0
Bayes B Bin02	117,9	**83,9**	46,6	**93,6**	25,7	**96,5**
Bayes B Bin03	202,0	**73,5**	43,8	**94,0**	25,6	**96,5**
Bayes B Bin04	177,8	**76,7**	44,6	**93,9**	32,0	**95,6**

[a] Em negrito são os valores de EQM e R^2 para os modelos tradicionais avaliados nas três herdabilidades. Deve-se compará-los com suas respectivas adaptações *bins*. [b] Em vermelho são os valores de R^2 para os *bins* considerados melhores que ou iguais aos respectivos métodos tradicionais. Fonte: Do autor (2018).

Tabela 5.2 – Erro Quadrático Médio (EQM) e Coeficiente de determinação (R^2 - em porcentagem) dos métodos RR-BLUP, Bayes A, Bayes B e suas respectivas adaptações em *bins*, para as três herdabilidades no cenário poligênico I (15 QTL).

Modelos	Herdabilidades					
	0,2		0,5		0,8	
	EQM	R^2 (%)	EQM	R^2 (%)	EQM	R^2 (%)
RR-BLUP	**3146,0**[a]	**58,5**	**1561,9**	**77,9**	**588,2**	**91,6**
RR-BLUP Bin01	2603,4	**66,5**[b]	1536,5	**78,3**	892,4	87,3
RR-BLUP Bin02	2781,4	**62,4**	1162,0	**83,5**	367,8	**94,8**
RR-BLUP Bin03	3151,7	55,1	1488,7	**78,7**	463,5	**93,5**
RR-BLUP Bin04	3198,2	54,5	1750,6	**75,5**	547,5	**92,4**
Bayes A	**3137,5**	**55,1**	**1563,9**	**77,6**	**503,5**	**92,8**
Bayes A Bin01	3113,6	**57,8**	1626,4	77,1	896,5	87,2
Bayes A Bin02	2985,3	**58,9**	1176,4	**83,4**	401,1	**94,3**
Bayes A Bin03	3130,9	**56,7**	1238,1	**82,4**	294,6	**95,8**
Bayes A Bin04	3064,6	**57,8**	1253,0	**82,1**	242,4	**96,6**
Bayes B	**3064,8**	**56,3**	**1375,9**	**80,5**	**412,2**	**94,1**
Bayes B Bin01	4107,1	41,8	1907,9	73,4	938,5	86,6
Bayes B Bin02	4098,0	41,7	1493,6	79,2	464,6	93,4
Bayes B Bin03	3956,8	43,6	1452,3	79,3	396,3	**94,4**
Bayes B Bin04	4044,4	42,3	1569,2	77,6	452,3	93,6

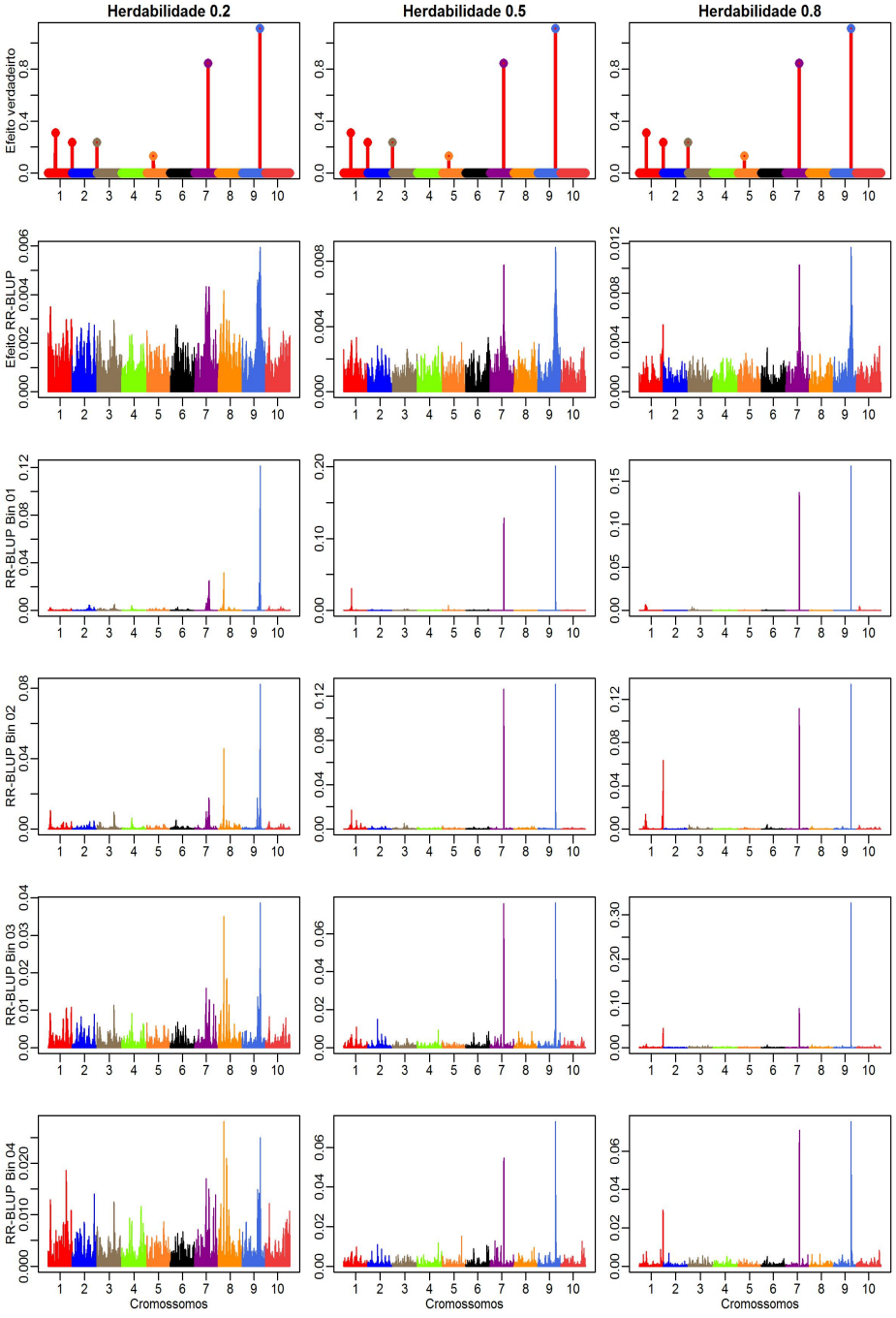
[a] Em negrito são os valores de EQM e R^2 para os modelos tradicionais avaliados nas três herdabilidades. Deve-se compará-los com suas respectivas adaptações *bins*. [b] Em vermelho são os valores de R^2 para os *bins* considerados melhores que ou iguais aos respectivos métodos tradicionais. Fonte: Do autor (2018).

Tabela 5.3 – Erro Quadrático Médio (EQM) e Coeficiente de determinação (R^2 - em porcentagem) dos métodos RR-BLUP, Bayes A, Bayes B e suas respectivas adaptações em *bins*, para as três herdabilidades no cenário poligênico II (60 QTL).

Modelos	Herdabilidades					
	0,2		0,5		0,8	
	EQM	R^2 (%)	EQM	R^2 (%)	EQM	R^2 (%)
RR-BLUP	**8160,2**[a]	**58,3**	**3303,8**	**79,6**	**1614,4**	**90,0**
RR-BLUP Bin01	9265,1	48,2	5428,1	66,5	4493,8	74,1
RR-BLUP Bin02	6608,7	**60,7**[b]	3347,6	**79,6**	1614,6	**90,1**
RR-BLUP Bin03	6683,0	**59,5**	3457,6	**79,9**	1441,4	**91,1**
RR-BLUP Bin04	6585,8	**59,6**	3702,4	**79,0**	1536,2	**90,6**
Bayes A	**6555,2**	**60,1**	**3310,6**	**79,7**	**1604,0**	**90,1**
Bayes A Bin01	11262,3	31,6	6016,3	63,3	5177,5	69,0
Bayes A Bin02	10991,7	32,0	4013,2	75,2	1971,3	88,1
Bayes A Bin03	11019,0	31,8	3914,7	75,8	1603,8	**90,3**
Bayes A Bin04	10668,4	34,0	4096,7	74,9	1632,4	**90,0**
Bayes B	**6629,1**	**60,5**	**3290,7**	**79,7**	**1646,2**	**89,8**
Bayes B Bin01	15267,6	17,2	8365,1	48,4	5418,3	68,4
Bayes B Bin02	15443,5	16,3	8235,1	49,1	3049,9	81,5
Bayes B Bin03	15839,1	15,4	8034,7	50,3	3125,8	80,9
Bayes B Bin04	15993,2	25,0	8667,1	46,4	2958,1	81,9

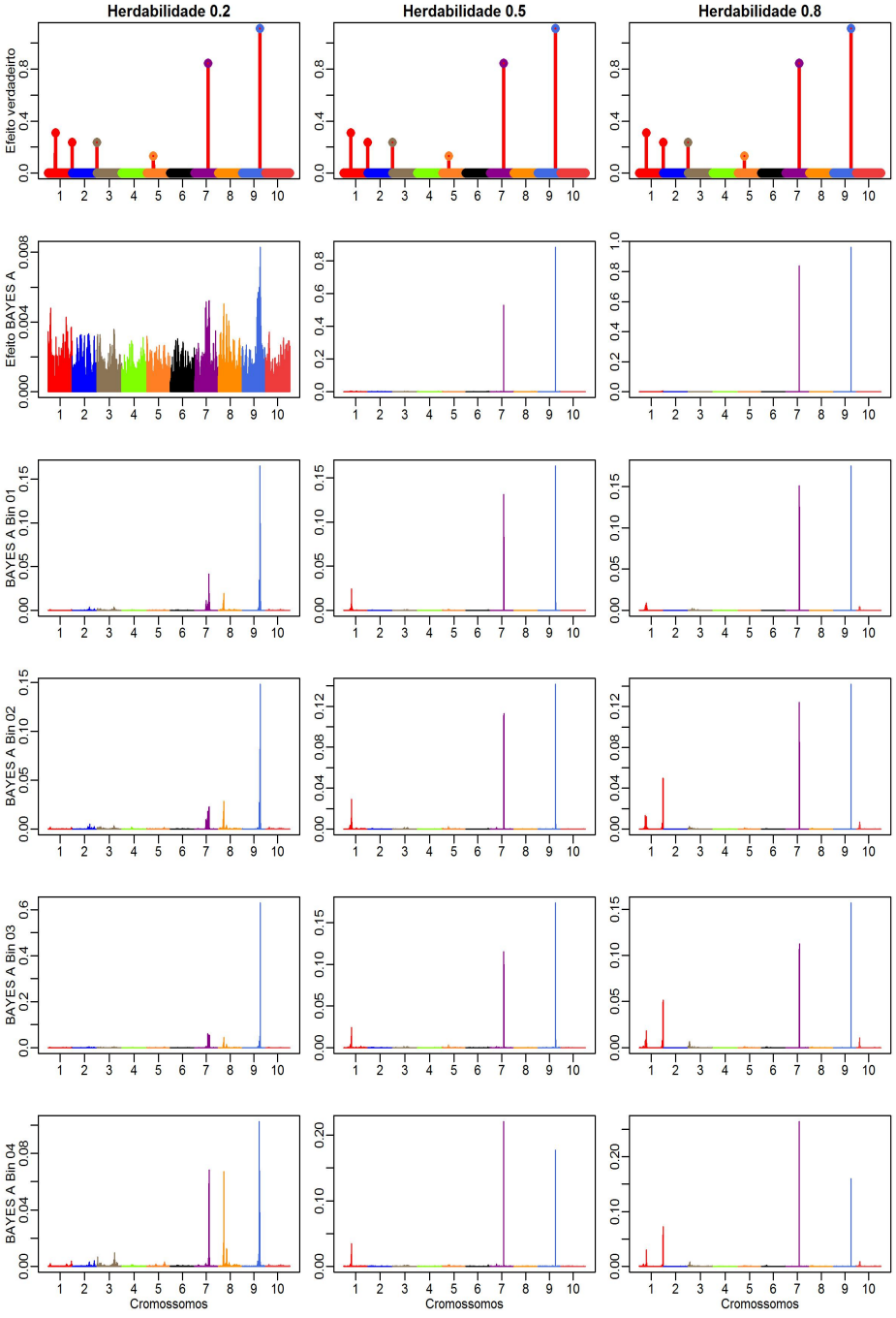
[a] Em negrito são os valores de EQM e R^2 para os modelos tradicionais avaliados nas três herdabilidades. Deve-se compará-los com suas respectivas adaptações *bins*. [b] Em vermelho são os valores de R^2 para os *bins* considerados melhores que ou iguais aos respectivos métodos tradicionais. Fonte: Do autor (2018).

Figura 5.1 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos RR-BLUP, RR-BLUP Bin01, RR-BLUP Bin02, RR-BLUP Bin03 e RR-BLUP Bin04, no cenário oligogênico. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



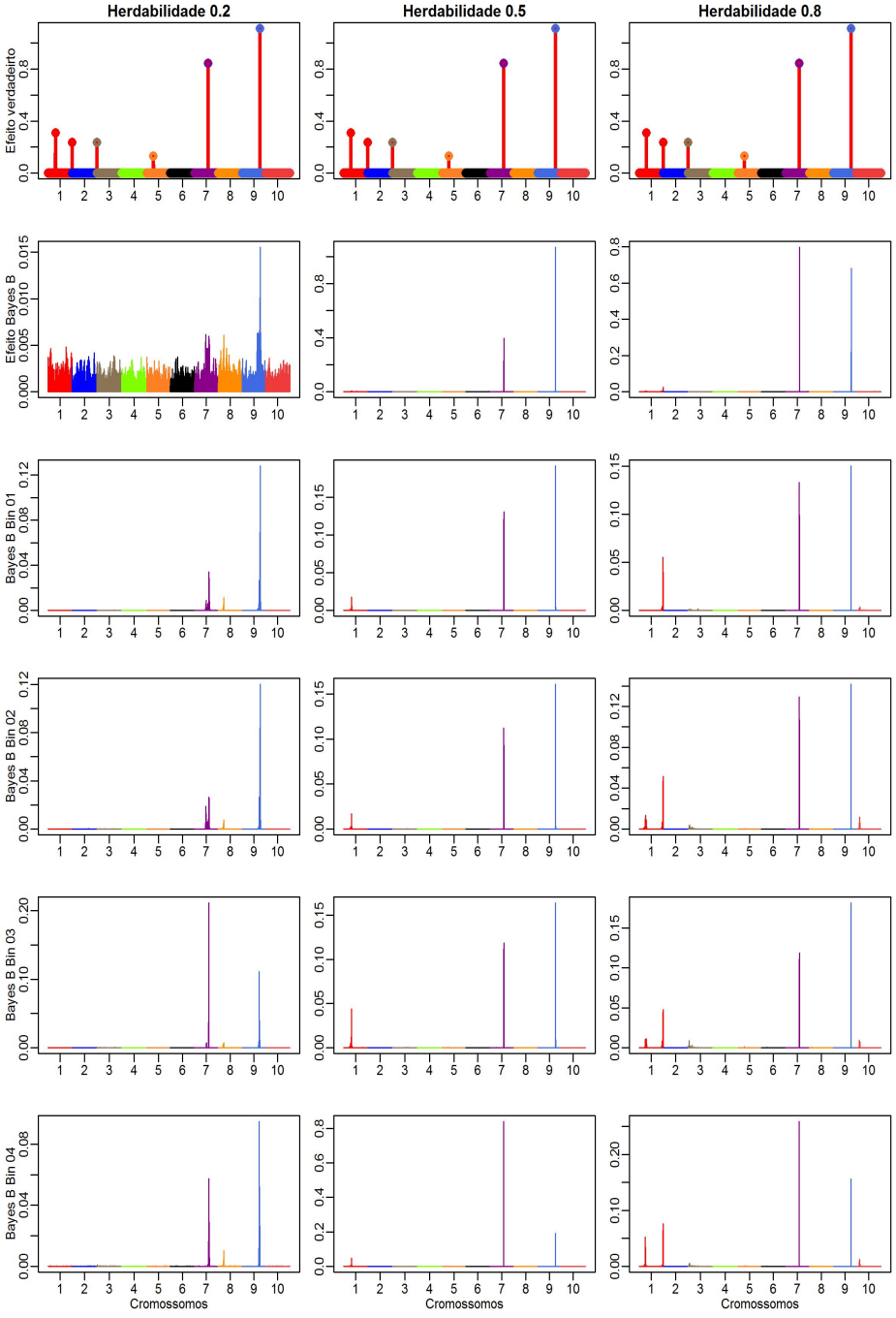
Fonte: Do autor (2018).

Figura 5.2 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, no cenário oligogênico. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



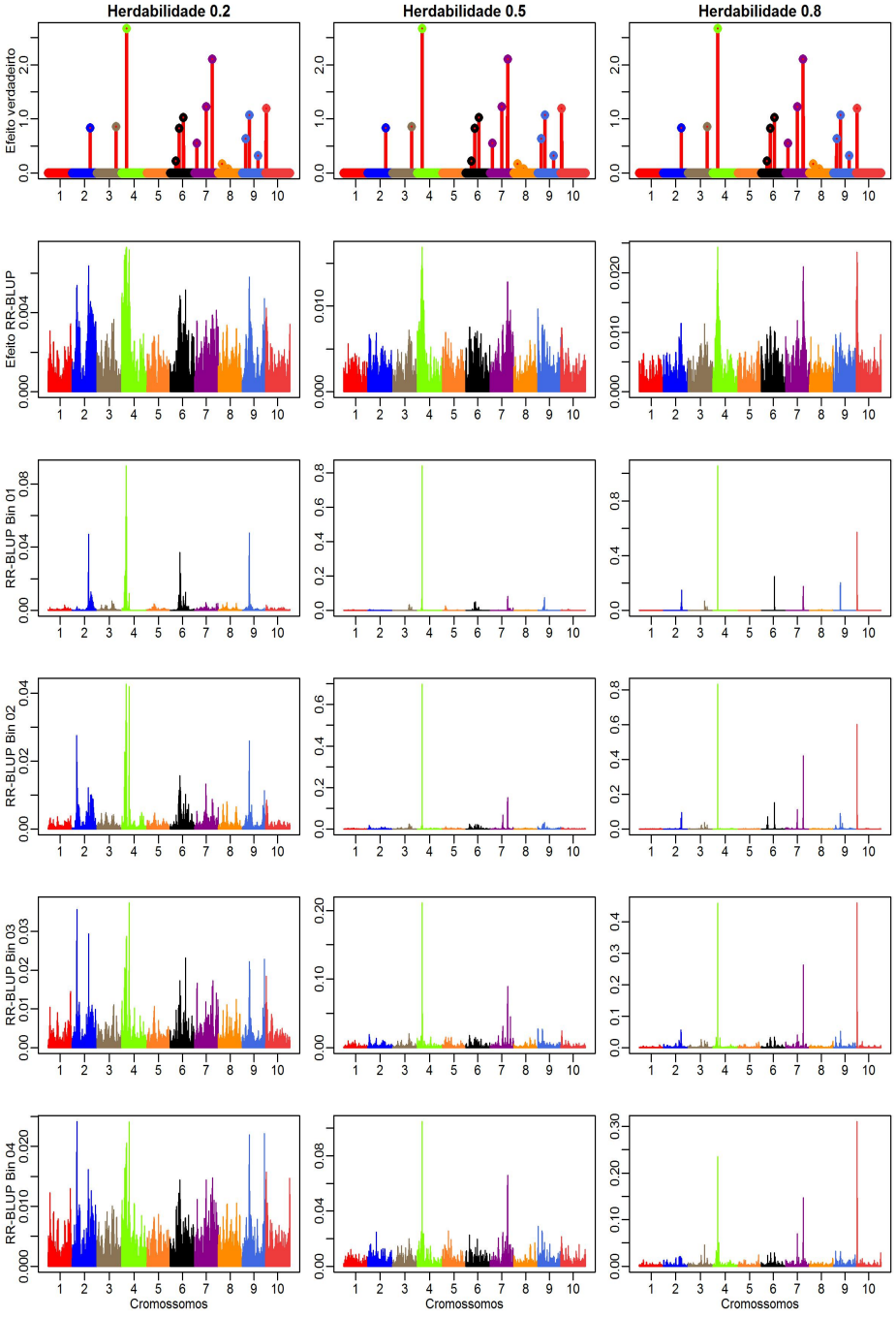
Fonte: Do autor (2018).

Figura 5.3 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes B, Bayes B Bin01, Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04, no cenário oligogênico. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



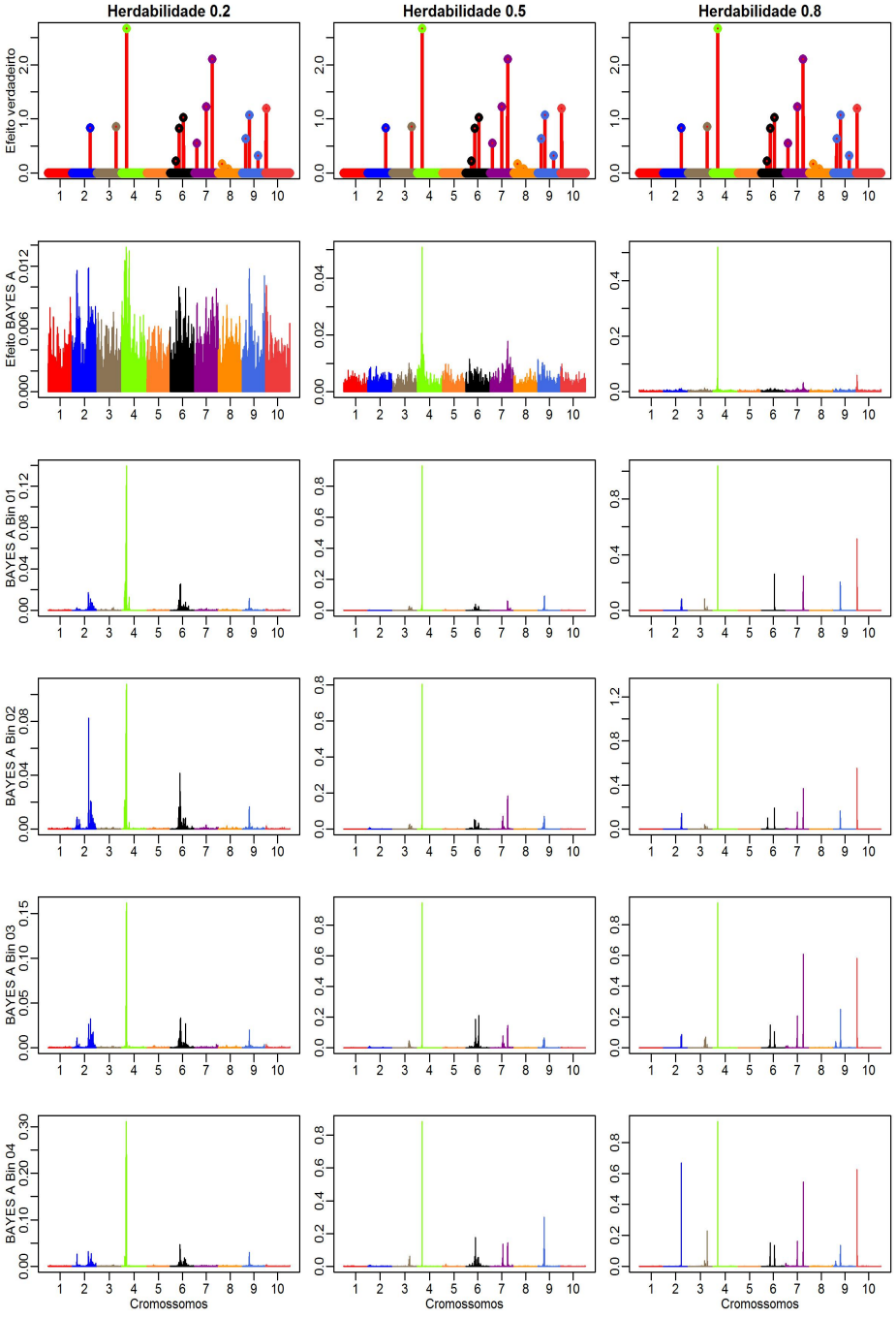
Fonte: Do autor (2018).

Figura 5.4 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos RR-BLUP, RR-BLUP Bin01, RR-BLUP Bin02, RR-BLUP Bin03 e RR-BLUP Bin04, no cenário poligênico I. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



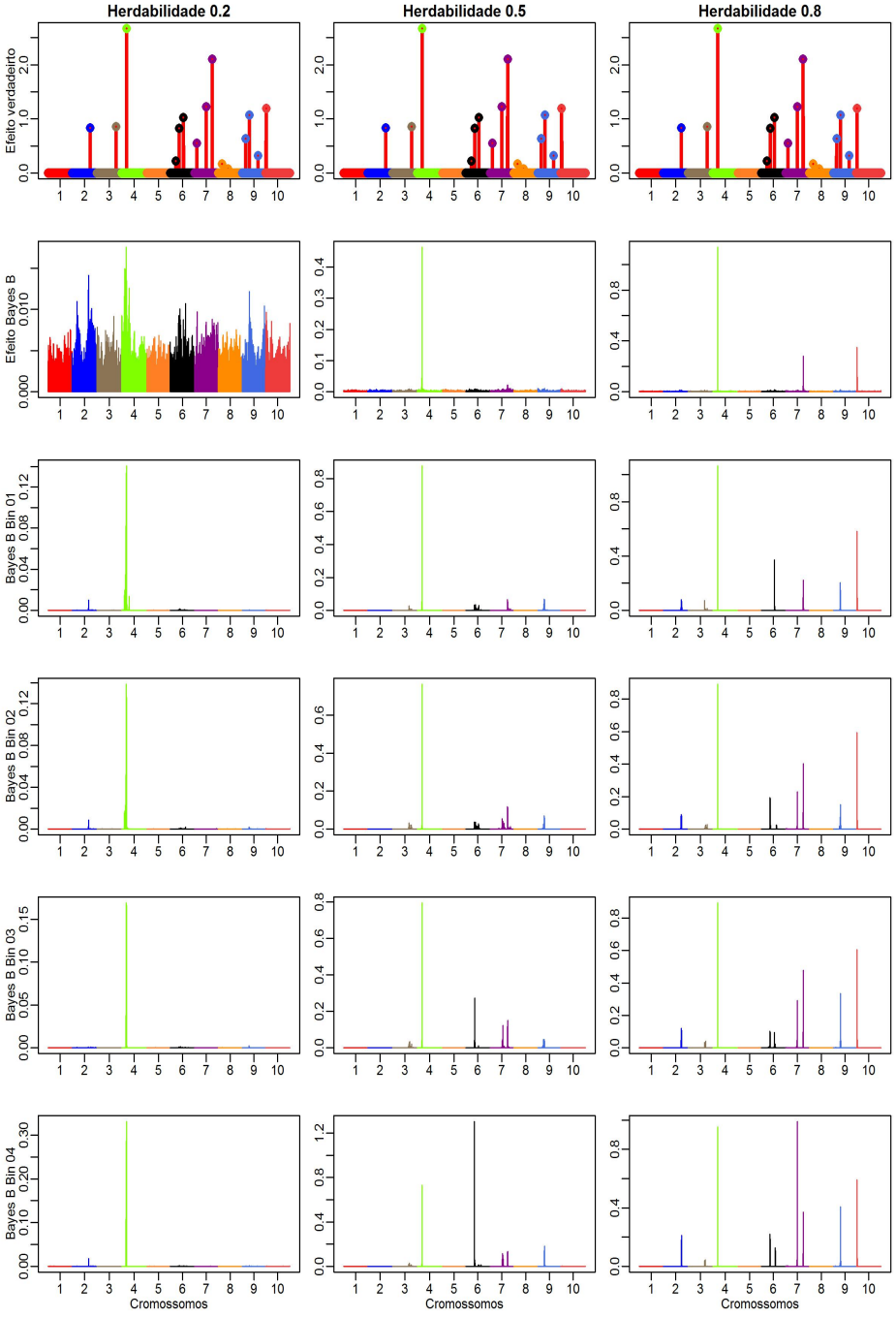
Fonte: Do autor (2018).

Figura 5.5 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, no cenário poligênico I. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



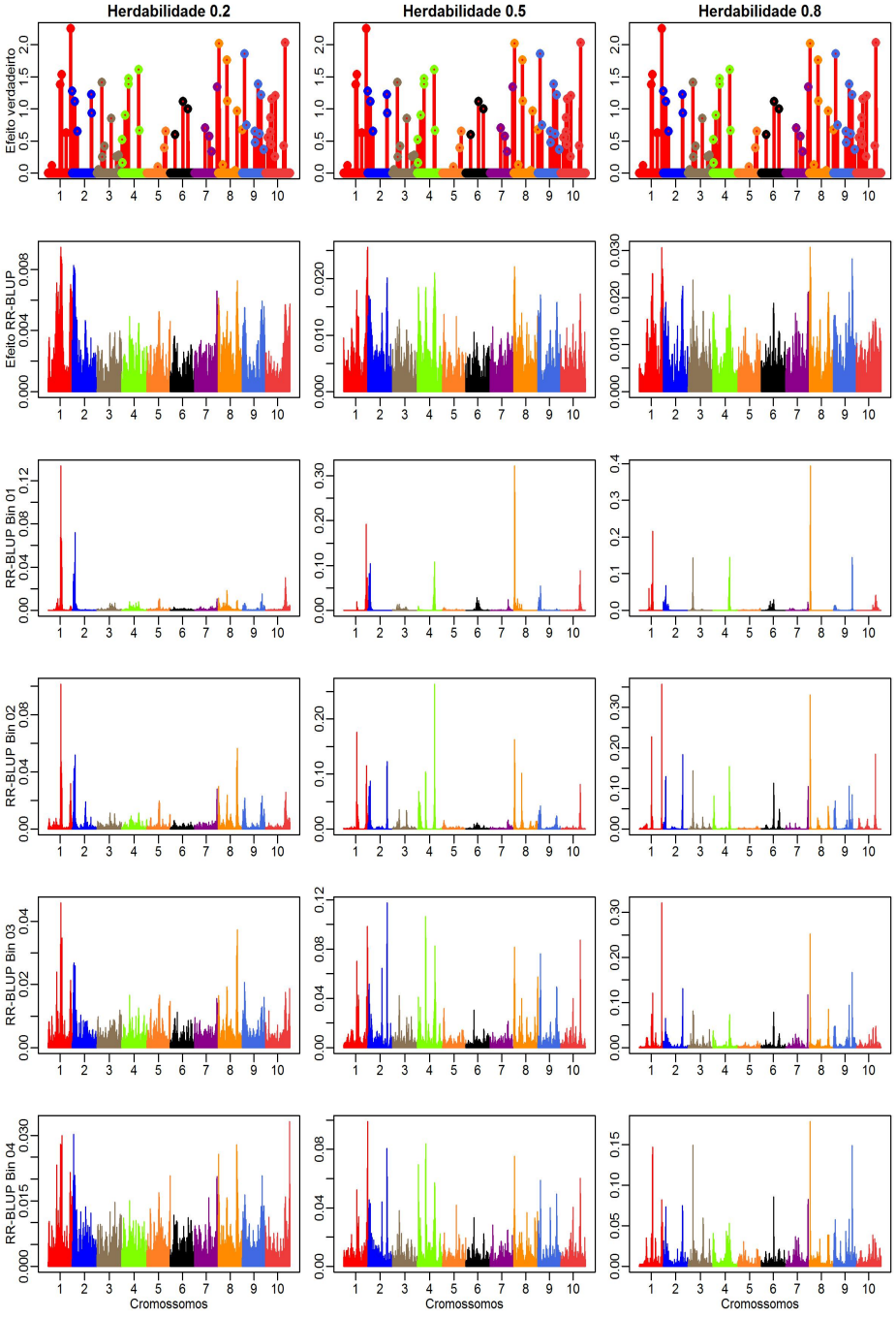
Fonte: Do autor (2018).

Figura 5.6 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes B, Bayes B Bin01, Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04, no cenário poligênico I. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



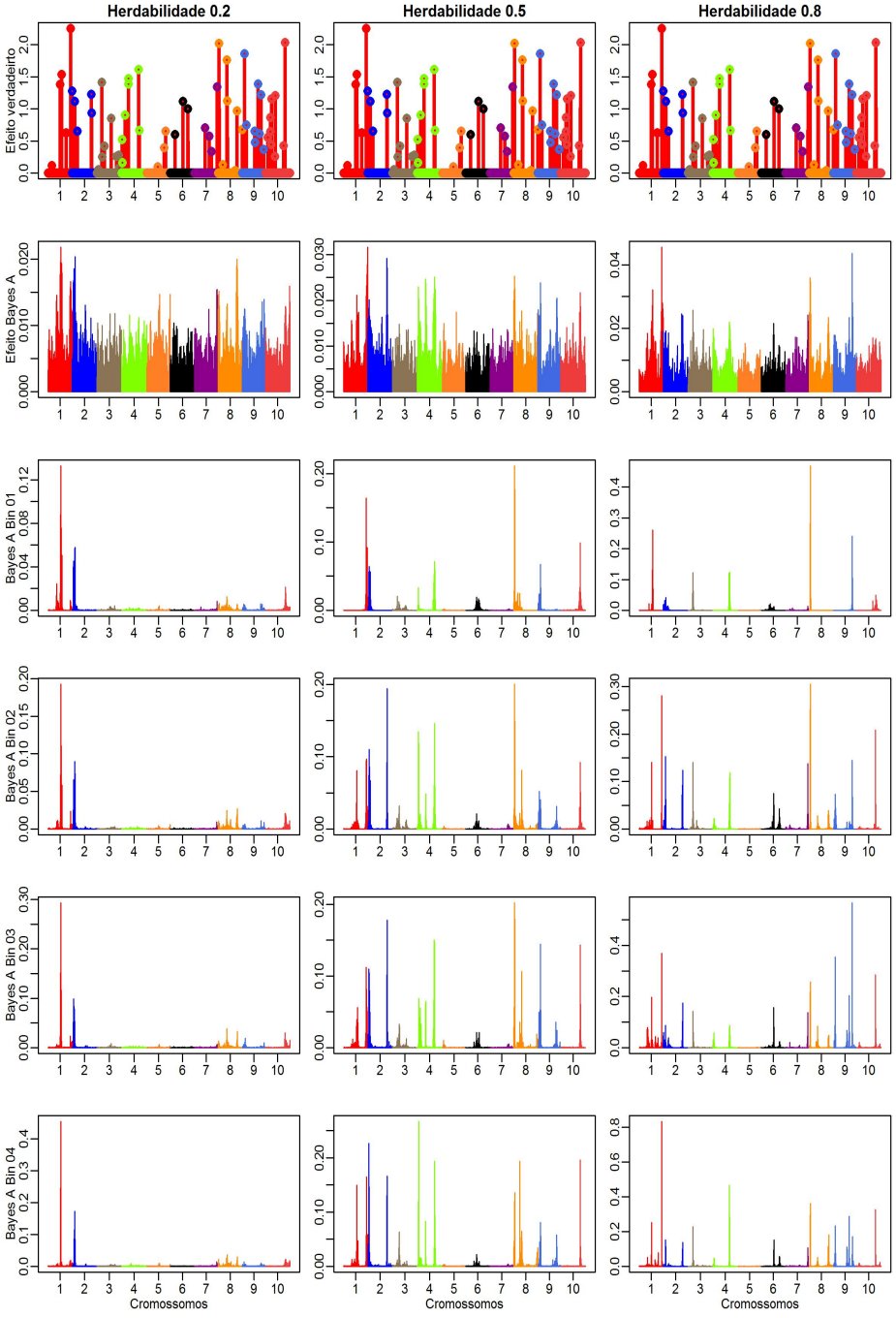
Fonte: Do autor (2018).

Figura 5.7 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos RR-BLUP, RR-BLUP Bin01, RR-BLUP Bin02, RR-BLUP Bin03 e RR-BLUP Bin04, no cenário poligênico II. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



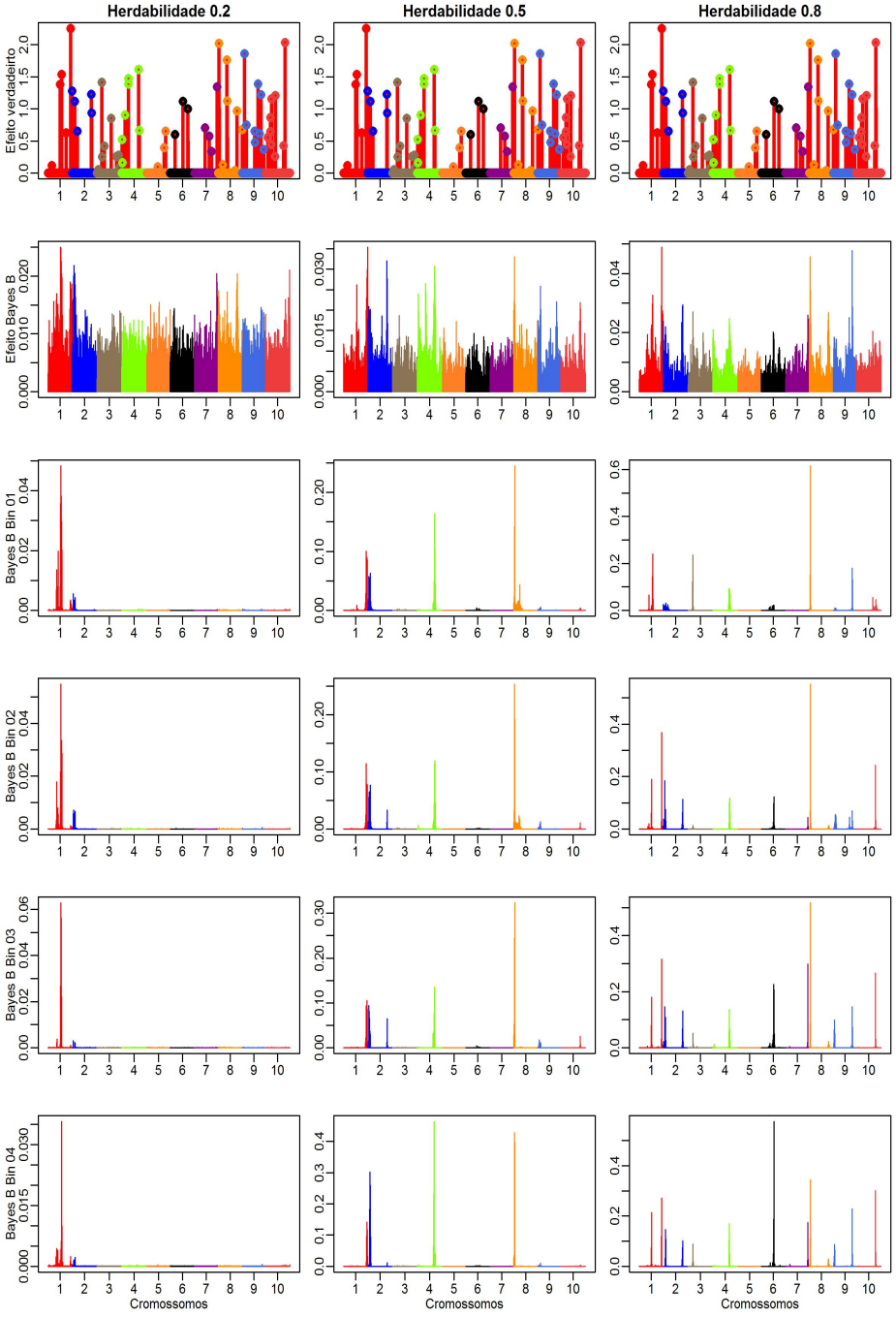
Fonte: Do autor (2018).

Figura 5.8 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes A, Bayes A Bin01, Bayes A Bin02, Bayes A Bin03 e Bayes A Bin04, no cenário poligênico II. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



Fonte: Do autor (2018).

Figura 5.9 – Efeitos simulados absolutos de QTL e estimativas absolutas dos métodos Bayes B, Bayes B Bin01, Bayes B Bin02, Bayes B Bin03 e Bayes B Bin04, no cenário poligênico II. Os pontos coloridos representam os 15 QTL simulados que estão distribuídos ao longo dos 12150 SNPs em 10 cromossomos.



Fonte: Do autor (2018).

APÊNDICE B - Distribuições para RR-BLUP e Bayes A

Nesta seção estão apresentadas as distribuições a priori e condicionais completas a posteriori para os métodos Bayes A e RR-BLUP adaptados ao modelo genoma contínuo em *bins*. Com isso, basta substituir as respectivas distribuições dadas na seção Material e Métodos por essas.

O modelo adotado é:

$$y_i|\mu,\sigma^2 \sim N\left(\mu + \sum_{t=1}^{C}\int_0^L Z_{it}\left(\lambda\right)\gamma\left(\lambda\right)d\lambda, I\sigma_e^2\right) \tag{5.1}$$

RR-BLUP

- ***Distribuições a priori:***

As prioris para μ e σ_e^2 são $p\left(\mu\right) \propto 1$ e $p\left(\sigma_e^2\right) \propto \frac{1}{\sigma_e^2}$, respectivamente. As distribuições a priori para o efeito de marcador γ e sua variância σ_γ são:

$$\begin{aligned} p\left(\gamma_j\right) &\propto N\left(0, \sigma_{\gamma_j}^2\right), \\ p\left(\sigma_\gamma^2\right) &\propto \frac{1}{\sigma_\gamma^2} \end{aligned} \tag{5.2}$$

com $\nu = 1$ e $S^2 = \frac{\sigma_y^2 \cdot 0{,}005}{M}$ sendo, respectivamente, o grau de liberdade e o parâmetro de escala para a variância do marcador.

Considerando Δ_k como o tamanho do *bin* k, utilizamos a priori de que a posição é uniformemente distribuída neste *bin*, ou seja, $p\left(\lambda_k\right) = \frac{1}{\Delta_k}$.

- Distribuições condicionais completas a posteriori:

Distribuição condicional completa para μ:

$$\mu|\ldots \sim N\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} Z_{\lambda_{j(i)}}\gamma_{\lambda_j}\right)/n \ \ , \frac{\sigma_e^2}{n}\right] \tag{5.3}$$

Distribuição condicional completa para os efeitos dos marcadores γ_{λ_j}:

$$\overline{\gamma}_{\lambda_j} = \left(\sum_{i=1}^{n} Z_{\lambda_{j(i)}}^2 + \frac{\sigma_e^2}{\sigma_\gamma^2}\right)^{-1} \sum_{i=1}^{n} Z_{\lambda_{j(i)}}\left(y_i - \mu - \sum_{m\neq j}^{k} Z_{\lambda_{m(i)}}\gamma_{\lambda_m}\right) \tag{5.4}$$

$$s_{\gamma_{\lambda_j}}^2 = \left(\sum_{i=1}^{n} Z_{\lambda_{j(i)}}^2 + \frac{\sigma_e^2}{\sigma_\gamma^2}\right)^{-1} \sigma_e^2 \tag{5.5}$$

Distribuição condicional para a variância comum do marcador σ_γ^2:

$$\sigma_\gamma^2|\ldots \sim \chi_{esc}^{-2}\left(\nu + n_1, \gamma_{\lambda_j}^2 + S^2\right) \tag{5.6}$$

sendo n_1 o número de marcadores dentro do *bin*.

Distribuição condicional para a variância residual σ_e^2:

$$\sigma_e^2|\ldots \sim \chi_{esc}^{-2}\left(n + \nu, FQ\right) \tag{5.7}$$

em que $FQ = \sum_{i=1}^{n}\left(y_i - \mu - \sum_{j=1}^{k} Z_{\lambda_{j(i)}}\gamma_{\lambda_j}\right)$.

BAYES A

- ***Distribuições a priori:***

As prioris para μ e σ_e^2 são $p(\mu) \propto 1$ e $p\left(\sigma_e^2\right) \propto \frac{1}{\sigma_e^2}$, respectivamente. As distribuições a priori para o efeito de marcador γ e sua variância σ_γ são:

$$\begin{aligned} p(\gamma_j) &\propto N\left(0,\sigma_{\gamma_j}^2\right), \\ p\left(\sigma_\gamma^2\right) &\propto \chi_{esc}^{-2}\left(\nu, S^2\right) \end{aligned} \tag{5.8}$$

com $\nu = 1$ e $S^2 = \frac{\sigma_y^2 \cdot 0{,}005}{M}$ sendo, respectivamente, o grau de liberdade e o parâmetro de escala para a variância do marcador.

Considerando Δ_k como o tamanho do *bin* k, utilizamos a priori de que a posição é uniformemente distribuída neste *bin*, ou seja, $p(\lambda_k) = \frac{1}{\Delta_k}$.

- ***Distribuições condicionais completas a posteriori:***

Distribuição condicional completa para μ:

$$\mu|\ldots \sim N\left[\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{k} Z_{\lambda_{j(i)}}\gamma_{\lambda_j}\right)/n \ , \ \frac{\sigma_e^2}{n}\right] \tag{5.9}$$

Distribuição condicional completa para os efeitos dos marcadores γ_{λ_j}:

$$\overline{\gamma}_{\lambda_j} = \left(\sum_{i=1}^{n} Z_{\lambda_{j(i)}}^2 + \frac{\sigma_e^2}{\sigma_{\gamma_{\lambda_{j(i)}}}^2}\right)^{-1} \sum_{i=1}^{n} Z_{\lambda_{j(i)}}\left(y_i - \mu - \sum_{m\neq j}^{k} Z_{\lambda_{m(i)}}\gamma_{\lambda_m}\right) \tag{5.10}$$

$$s^2_{\gamma_{\lambda_j}} = \left(\sum_{i=1}^{n} Z^2_{\lambda_{j(i)}} + \frac{\sigma^2_e}{\sigma^2_{\gamma_{\lambda_{j(i)}}}} \right)^{-1} \sigma^2_e \tag{5.11}$$

Distribuição condicional para a variância específica do marcador $\sigma^2_{\gamma_{\lambda_j}}$*:*

$$\sigma^2_{\gamma_{\lambda_j}} | \ldots \sim \chi^{-2}_{esc} \left(\nu + 1, \gamma^2_{\lambda_j} + S^2 \right) \tag{5.12}$$

Distribuição condicional para a variância residual σ^2_e*:*

$$\sigma^2_e | \ldots \sim \chi^{-2}_{esc} \left(n + \nu, FQ \right) \tag{5.13}$$

em que $FQ = \sum_{i=1}^{n} \left(y_i - \mu - \sum_{j=1}^{k} Z_{\lambda_{j(i)}} \gamma_{\lambda_j} \right)$.

ARTIGO 2 Uso de B-*Spline* para estimar a função sinal no modelo genoma contínuo

RESUMO

Com o desenvolvimento de técnicas estatísticas e computacionais de métodos de seleção genômica, tornou-se possível lidar com a alta dimensão de marcadores disponíveis para diversos organismos. Embora úteis, estes métodos negligenciam as características genéticas de formação do caráter. Por exemplo, a expressão gênica de um QTL pode ser representada como uma função de sua posição no genoma. Uma abordagem aproximada para a estimação de efeitos genéticos dependentes da posição é apresentada com o modelo genoma contínuo. Acreditando-se que os efeitos genéticos dos marcadores têm uma estrutura funcional intrínseca e adotando técnicas de análise de dados funcionais, a disponibilidade de alta densidade de marcadores pode ser usada para estimar a função sinal por meio de funções *Spline*. Com isso, o objetivo deste trabalho é utilizar o modelo genoma contínuo e buscar uma expressão polinomial via sistema de bases *Spline* (B-*Spline*) que represente a expressão gênica. Além disso, busca-se verificar se tal abordagem apresenta vantagens preditivas com relação a alguns métodos clássicos de seleção genômica. Uma população F_2 de 300 indivíduos genotipadados com 10010 marcadores foi simulada para dois cenários (oligogênico e poligênico) e herdabilidades 0,2; 0,5 e 0,8. Já os dados reais consistem de 610 híbridos de eucalipto genotipados com 15104 marcadores. Para o cenário oligogênico, o método B-*Spline* foi considerado o menos acurado nas herdabilidades 0,2 e 0,8 e foi equivalente aos métodos Bayes A e Bayes B na herdabilidade 0,5. Para o cenário poligênico, a capacidade preditiva do B-*Spline* foi análoga à do RR-BLUP. Para os dados reais, os métodos bayesianos foram mais acurados que RR-BLUP e B-*Spline*. Entretanto, o custo computacional do método B-spline foi bem menor que de todos os outros. A forma funcional das curvas polinomiais por partes (*Splines*) pode permitir novos tipos de análise. Este estudo mostra que a abordagem com B-*Spline* pode lidar com um número ilimitado de marcadores e com baixo custo computacional.

Palavras-chave: B-*Spline*. Modelo genoma contínuo. Modelo funcional. Análise de dados funcionais.

1 INTRODUÇÃO

A maior motivação para pesquisas em genética molecular é a perspectiva de que a informação em nível de DNA (*Deoxyribonucleic Acid*) conduza ao ganho genético mais rápido e eficaz do que o alcançado com base somente em dados fenotípicos (MEUWISSEN; HAYES; GODDARD, 2001). A seleção genômica (GS - *Genomic Selection*) fundamenta-se na alta densidade de marcadores genéticos moleculares que, com os avanços na biotecnologia molecular, estão se tornando disponíveis para muitas espécies de animais e plantas.

Modelos paramétricos e métodos estatísticos de GS modelam as associações entre os marcadores e um fenótipo com foco no cálculo de um valor genético para um indivíduo. Assim sendo, há explicação de grande parte da variação genética do caráter quantitativo (MEUWISSEN; HAYES; GODDARD, 2001). Para isso, a condição essencial é que haja desequilíbrio de ligação entre alelos dos marcadores e QTL (*Quantitative Trait Loci*) (CALUS et al., 2008). Em tese, basicamente, a GS utiliza regressões lineares dos marcadores sobre os valores fenotípicos usando o artifício de *shrinkage* para contornar problemas de superparametrização (MOURA, 2017).

Para os autores De Los Campos et al. (2009), Hu, Wang e Xu (2012), Tempelman (2015) e Thavamanikumar, Dolferus e Thumma (2015), o aspecto central na GS é, justamente, como lidar com a alta dimensionalidade (grande número de marcadores e pequeno número de observações) e a multicolinearidade. Devido a essa particularidade, métodos de regressão fixa simples que utilizam mínimos quadrados na seleção genômica são proibitivos, a menos que se faça uma redução na dimensão do modelo. Várias metodologias para contornar tais situações vêm sendo propostas para realizar a seleção genômica, por exemplo, métodos lineares por meio de modelos mistos GBLUP (*Genomic Best Linear Unbiased Prediction*)

e RR-BLUP (*Ridge Regression Best Linear Unbiased Prediction*), métodos não lineares de seleção de variáveis (métodos bayesianos), dentre outros.

Hu, Wang e Xu (2012) desenvolveram um modelo infinitesimal que denominaram de modelo genoma contínuo, em que se utiliza aproximação de modelos funcionais. Nesse estudo, os autores dividiram o genoma em intervalos de alto desequilíbrio de ligação (LD - *linkage disequilibrium*) que denominaram de *bins* e utilizaram a média dos *bins* como medida de informação. Embora *bins* naturais tenham sido utilizados com bastante êxodo, essa técnica pode não ser diretamente aplicada em algumas situações específicas, devido ao número de marcadores e ao tamanho amostral (XU, 2013).

Com o propósito de contornar esse problema, Xu (2013) estabelece o conceito de *bins* artificiais, em que podem ocorrer pontos de interrupção (*breakpoints*) ao longo dos *bins* naturais. Neste caso, como os *bins* artificiais podem ser arbitrariamente definidos de acordo com a preferência do investigador, não dependendo do número de marcadores nem do tamanho amostral. Contudo, os dados dos *breakpoints* devem ser convertidos em dados *bin* antes da análise, o que pode não ser uma boa opção, uma vez que os dados originais não são usados diretamente para análise, mas sim a média dos *bins*.

Com essa preocupação, Moura (2017) propôs uma abordagem mais direta de modelos funcionais e integração numérica da função genômica e relaxou a suposição de blocos com alto LD nos *bins*. O autor foi um pouco além e assumiu os marcadores como variável aleatória dentro dos *bins*, cuja função sinal de expressão gênica é desconhecida. Nessa metodologia, ao invés da média do *bin* como informação, o autor utilizou métodos bayesianos Monte Carlo Cadeias de Markov via algoritmo Metropolis-Hasting para realização do processo de amostragem por importância e integração numérica. A análise *bin* proposta mostrou melhora

significativa na capacidade preditiva em relação aos modelos tradicionais de GS.

A ideia de modelos funcionais em GS consiste no pressuposto de que a expressão gênica segue uma série espacial em que os picos da expressão gênica são funções da posição no genoma. Quando a forma da função subjacente aos dados é complicada, é difícil aproximá-la através de um único polinômio. Nesse caso, *Spline* é uma das funções de aproximação mais apropriadas (BOOR, 1978; DIERCKX, 1993).

Uma curva *Spline* é uma sequência de curvas polinomiais por partes que estão conectadas para formar uma única curva. A ideia é dividir o intervalo de interesse em K regiões distintas. Dentro de cada região, uma função polinomial é ajustada aos dados. No entanto, esses polinômios são restritos para que eles se juntem suavemente nos limites da região, ou nós (*knots*). Desde que o intervalo seja dividido em *knots* suficientes, isso pode produzir um ajuste extremamente flexível. Sobre qualquer intervalo, esta função é um polinômio de grau fixo, mas sua natureza muda quando se passa para o próximo subintervalo (RAMSAY; HOOKER; GRAVES, 2009). O uso de polinômios de baixo grau resulta em curvas suaves, evitando os problemas de sobreajuste causados por um polinômio de alto grau (MICHNA et al., 2016). Na literatura sobre modelagem de séries temporais de expressão gênica, por exemplo, a capacidade de descrever dados em termos de funções matemáticas é um ingrediente-chave em algoritmos que inferem mecanismos reguladores entre genes (BAR-JOSEPH et al., 2003; KONOPKA, 2011).

B-*Spline* é uma combinação linear de um conjunto de funções de base que são determinadas pelo número e pela localização de *knots*, bem como o grau do polinômio. Foi proposto inicialmente por Boor (1978) e vem sendo usado nas mais variadas áreas da ciência, produzindo excelentes resultados em termo de suavização e, além disso, utiliza um número menor de funções base (BOOR, 2001).

Com isso, menos complexa será a estimação e menos recurso computacional será exigido. De acordo com Zhang et al. (2003), baseia-se no mesmo princípio que a regressão polinomial: escolhe-se uma base e, em seguida, ajusta-se a curva usando a regressão de mínimos quadrados.

Baseado no exposto acima, o presente estudo tem como objetivo utilizar modelos funcionais e buscar uma expressão polinomial via sistema de bases *Spline* (B-*Spline*) que represente a expressão gênica na seleção genômica. Além disso, busca-se verificar se tal abordagem apresenta vantagens preditivas com relação a alguns métodos usuais de seleção genômica.

2 MATERIAL E MÉTODOS

2.1 Dados simulados

Utilizando o software QGenes (JOEHANES; NELSON, 2008), foram simulados cinco cromossomos com tamanho de 120 cM cada e distância média de 0,001 cM no genoma, em 300 indivíduos pertencentes a uma população F_2, totalizando 10010 marcadores SNP (*Single Nucleotide Polymorphism*). Dez marcadores foram assumidos como QTL para representar o cenário oligogênico e seus efeitos amostrados de uma distribuição normal com média igual a zero e desvio padrão igual a um. Para o cenário poligênico, 100 marcadores foram assumidos como QTL e seus efeitos, também, foram amostrados de uma distribuição $N(0,1)$. Os valores genotípicos dos indivíduos foram construídos pela combinação linear dos efeitos dos QTL com seus genótipos, submetidos a três níveis de herdabilidades: 0,2; 0,5 e 0,8.

2.2 Dados reais

Os dados consistem numa população de híbridos da empresa Fibria S.A., derivados do cruzamento entre plantas pré-selecionadas de E. *grandis*, E. *urophylla*, E. *globulus* e E. *camaldulensis*. Foram avaliados 12 indivíduos por combinação de híbridos e, devido às perdas, foram obtidos, ao final, 610 indivíduos. A fenotipagem das plantas ocorreu em três épocas distintas: aos 24 meses de idade (Época 1), aos 36 meses de idade (Época 2) e aos 63 meses de idade (Época 3). As características mensuradas, em cada idade da planta, utilizadas neste trabalho foram: cap1 - circunferência à altura do peito (em centímetros) na Época 1; lig1 - teor de lignina (em porcentagem) na Época 1; cap2 - circunferência à altura do peito (em centímetros) na Época 2; alt2 - altura de planta (em metros) na Época 2; epc2 -

espessura da casca (em milímetros) na Época 2; cap3 - circunferência à altura do peito (em centímetros) na Época 3; alt3 - altura de planta (em metros) na Época 3; epc3 - espessura da casca (em milímetros) na Época 3.

Concomitantemente à tomada dos dados fenotípicos, aos dois anos de idade, foi extraído o DNA de todas as 610 plantas dessa população para genotipagem via GBS, método proposto por Elshire et al. (2011). A genotipagem foi feita com 15104 marcadores DArT-seq.

2.3 Estimativa de efeito de marcadores

Este método envolve duas etapas. Primeiro, cria-se uma matriz **B** de bases *Spline* (B-*Spline*) utilizando a posição de cada marcador. Segundo, obtém-se uma matriz **W** como combinação linear da matriz **B** com a matriz **Z** de estado genotípico. O modelo estatístico para estimar efeitos de p marcadores é definido por:

$$y_i = \beta + \sum_{j=1}^{p} Z_{ij}\gamma_j + \varepsilon_i \tag{2.1}$$

em que β é o intercepto, γ_j é o efeito do marcador j, Z_{ij} é o estado genotípico do marcador j para o indivíduo i e $\varepsilon_i \sim N(0,\sigma^2)$ é o erro com média zero e variância desconhecida σ^2. O estado do genótipo do j-ésimo marcador para o indivíduo i é definido como:

$$Z_{ij} = \begin{cases} 2, & \text{para homozigoto dominante;} \\ 1\,, & \text{para heterozigoto;} \\ 0\,, & \text{para homozigoto recessivo;} \end{cases} \tag{2.2}$$

para SNP ou,

$$Z_{ij} = \begin{cases} 1\,, & \text{para homozigoto dominante/heterozigoto;} \\ 0\,, & \text{para homozigoto recessivo;} \end{cases} \tag{2.3}$$

para DArT.

Quando $p \to \infty$, o modelo em (2.1) é denominado de modelo infinitesimal. Nesse modelo, há alguns problemas de estimação, devido à dimensionalidade e à multicolinearidade.

2.4 Modelo funcional

A ideia de modelo funcional baseia-se no genoma como uma série espacial em que o sinal relacionado à expressão gênica $\gamma(\lambda)$ é uma função desconhecida e pode ser aproximada por $f(\lambda)$ via escaneamento genômico ao longo da posição λ, dado que se conhece $\mathbf{y}$, λ e $Z(\lambda)$, em que $Z(\lambda)$ é o estado genotípico na posição λ. A ligação de $\gamma(\lambda)$ para $\mathbf{y}$ é realizada através da matriz $Z_i(\lambda)$ que descreve o estado genotípico do indivíduo i no domínio de $\gamma(\lambda)$. Assim, tomando $f(\lambda) \approx \gamma$, tem-se a equação $Z(\lambda) f(\lambda) \approx Z(\lambda)\gamma$. Dessa forma, assumindo $Z(\lambda)\gamma$ como a predição do valor genético genômico $(\hat{g})$, por transitividade matemática, temos $Z(\lambda) f(\lambda) = \hat{g} = y + \varepsilon$. Logo, tem-se a equivalência $Z(\lambda) f(\lambda) \equiv \int_0^L Z_i(\lambda)\gamma(\lambda)\, d\lambda = y + \varepsilon$.

Considere n indivíduos que estão sequenciados em uma região genômica que tem m marcadores. Nós assumimos que os p marcadores estão localizados em uma região com locais físicos ordenados $0 \leq \lambda_1 < \cdots < \lambda_p = L$, em que L é o tamanho do cromossomo, em centiMorgan (cM). Logo, sendo y_i o valor fenotípico para o indivíduo i, o modelo linear funcional é:

$$y_i = \mu + \int_0^L Z_i(\lambda)\,\gamma(\lambda)\,d\lambda + \varepsilon_i, \tag{2.4}$$

em que μ é a média geral, λ é a posição no cromossomo expressa como uma quantidade contínua, L é o tamanho do cromossomo, $Z_i(\lambda)$ é o estado genotípico do marcador na posição λ para o indivíduo i, $\gamma(\lambda)$ é o efeito genético do marcador em função da posição λ, $\varepsilon_i \sim N(0,\sigma^2)$ é o erro para o indivíduo i.

Se existem C cromossomos no genoma, o modelo (2.4) pode ser reescrito como:

$$y_i = \mu + \sum_{c=1}^{C} \int_0^L Z_{ic}(\lambda)\,\gamma(\lambda)\,d\lambda + \varepsilon_i, \tag{2.5}$$

em que o somatório descreve a descontinuidade da função ao longo dos cromossomos.

A integral em (2.5) resulta no valor genético genômico para o indivíduo i. Este é o modelo genoma contínuo proposto por Hu, Wang e Xu (2012). Dado que $\gamma(\lambda)$ é desconhecida, não existe uma expressão explícita para resolver a integral. Hu, Wang e Xu (2012) dividiram o genoma em blocos de alto LD que denominaram de *bins* e tomaram o efeito médio dos *bins* como informação. No presente estudo, será utilizado um sistema de bases *Spline* que, combinado com a matriz de estado genotípica $\mathbf{Z}$, pode estimar $\hat{\gamma}(\lambda) = f(\lambda)$ e, consequentemente, resolver a integral para recuperar o valor genético genômico.

Por não ser possível verificar todas as posições no genoma, pois se trata de uma variável contínua, podemos utilizar as posições dos marcadores como pseudoposições (por exemplo, marcador M_1 tem a posição λ_1, marcador M_2 tem a posição λ_2, etc.).

2.4.1 B-*Spline*

A ideia de regressão via *Spline*, também denominada de regressão local, consiste basicamente em um problema de interpolação. Dessa forma, uma função *Spline* pode ser analisada como um conjunto de funções polinomiais, conectadas em determinados *knots*, utilizados para ajustar uma curva aos dados. A quantidade de *knots*, suas posições e o grau do polinômio a ser ajustado são determinados, a priori, com critérios definidos pelo pesquisador (os *knots* devem estar contidos no domínio a ser analisado). A diferença entre os limites de dois *knots* é denominada de janela ou *bin*. Embora o *Spline* linear permita ajustar várias relações funcionais, não apresenta suavização nos *knots* e, portanto, não reflete associações curvilíneas, muitas vezes encontradas em situações práticas. Por isso, é mais comum ajustar B-*Splines*, que possuem vantagens de suavização nos *knots*.

No contexto de seleção genômica é necessário assumir o genoma como uma série espacial em que o sinal da expressão gênica $\gamma(\lambda)$ é uma função do escaneamento genômico ao longo da posição λ. Pode-se ajustar $\gamma(\lambda)$ assumindo qualquer base funcional, *Spline*, B-*Spline*, Série de Fourier, *Wavelet*, dentre outras. Neste trabalho, funções B-*Spline* serão utilizadas, pois são melhores quando empregadas para suavizar curvas em que não há periodicidade específica, como é o caso da função sinal da expressão gênica.

B-*Splines* são definidas pela ordem q e pelos números de k *knots* dentro do intervalo especificado, chamados *knots* internos. Os dois pontos extremos, início e fim do intervalo, também são considerados *knots*, então o número total de *knots* é $k+2$. O grau do polinômio B-*Spline* é $m = q-1$. Seja uma sequência não decrescente de *knots* (números reais) tal que $a = \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{k+1} = b$, em um intervalo $[a, b]$. Define-se um conjunto de *knots* aumentados $\xi_{-m} = \cdots = \xi_0 \leq \xi_1 \leq \cdots \leq \xi_{k+1} = \cdots = \xi_{k+m+1}$, em que os limites a e b são repetidos

m vezes. De acordo com Boor (1978), o t-ésimo B-*Spline* de ordem q pode ser calculado por:

$$\mathrm{B}_{t,q}(x) = \alpha_{t,q}(x)\,\mathrm{B}_{t,q-1}(x) + [1 - \alpha_{t+1,q}(x)]\,\mathrm{B}_{t+1,q-1}(x), \qquad (2.6)$$

com

$$\mathrm{B}_{t,1}(x) = \begin{cases} 1, & \xi_t \leq x < \xi_{t+1}; \\ 0, & \text{caso contrário;} \end{cases} \qquad (2.7)$$

sendo

$$\alpha_{t,q}(x) = \begin{cases} \dfrac{x - \xi_t}{\xi_{t+q-1} - \xi_t}, & \text{se } \xi_{t+q-1} - \xi_t \neq 0; \\ 0, & \text{caso contrário.} \end{cases} \qquad (2.8)$$

Assim, considerando B-*Splines* de ordem q com k *knots* interiores, é possível escrever a função γ como:

$$\gamma(\lambda) = \sum_{t=1}^{k+q} B_{t,q}(\lambda)\,\phi_t(\lambda) = \mathbf{B\Phi}, \qquad \text{com } \lambda \in [0,L]. \qquad (2.9)$$

Por sua vez, a estimativa suave de γ é dada por:

$$\hat{\gamma}(\lambda) = \sum_{t=1}^{k+q} B_{t,q}(\lambda)\,\hat{\phi}_t(\lambda). \qquad (2.10)$$

Substituindo (2.8) em (2.5) e considerando C cromossomos, o modelo pode ser reescrito como:

$$y_i = \mu + \sum_{c=1}^{C} \sum_{t=1}^{k+q} Z_{ic}(\lambda) B_{t,q}(\lambda) \phi_t(\lambda), \tag{2.11}$$

com $\lambda \in [0,L]$.

Um polinômio de grau $m = 2$ (ou seja, ordem $q = 3$) foi escolhido para ser ajustado em cada intervalo de *knots*, pois se supõe que entre dois *knots* adjacentes não existe mais que um QTL com efeito expressivo. Logo, $k + q = k + 3$ parâmetros a serem estimados em cada cromossomo. Com C cromossomos, obtêm-se $b = C \cdot (k + q) = C(k + 3)$ parâmetros no genoma todo. Em forma matricial, temos:

$$\begin{aligned} \mathbf{y}_{n\text{x}1} &= \boldsymbol{\mu}_{n\text{x}1} + \mathbf{Z}_{n\text{x}p}\boldsymbol{\gamma}_{p\text{x}1} + \boldsymbol{\varepsilon}_{n\text{x}1} = \\ &= \boldsymbol{\mu}_{n\text{x}1} + \mathbf{Z}_{n\text{x}p}\mathbf{B}_{p\text{x}b}\boldsymbol{\Phi}_{b\text{x}1} + \boldsymbol{\varepsilon}_{n\text{x}1}, \end{aligned} \tag{2.12}$$

em que $\boldsymbol{\mu}$ é o vetor da média geral, $\mathbf{Z}$ é a matriz de estado genotípico dos marcadores, $\mathbf{B}$ é a matriz de bases B-*Spline*, Φ é o vetor de coeficientes a serem estimados e ε é o vetor dos erros.

Tomando $\mathbf{ZB} = \mathbf{W}$, o modelo a ser trabalhado é:

$$\mathbf{y}_{n\text{x}1} = \boldsymbol{\mu}_{n\text{x}1} + \mathbf{W}_{n\text{x}b}\boldsymbol{\Phi}_{b\text{x}1} + \boldsymbol{\varepsilon}_{n\text{x}1}. \tag{2.13}$$

Dado que sempre se pode obter $b < n$, pode-se estimar os parâmetros pelo método de mínimos quadrados ordinários. Logo,

$$\hat{\boldsymbol{\Phi}} = \left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\left(\mathbf{y} - \mu\right) \tag{2.14}$$

2.5 Implementação da análise

Para o modelo proposto, a matriz **B** de bases B-*Spline* foi obtida utilizando a função **bSpline** disponível pela biblioteca *splines2* (WANG; YAN, 2017) do software R (R CORE TEAM, 2018). Para os modelos concorrentes, os dados foram analisados utilizando os modelos Bayes A, Bayes B e Lasso Bayesiano, por meio da função **BGLR** contida no pacote *BGLR* (PEREZ; DE LOS CAMPOS, 2014) e o modelo RR-BLUP com a função **mixed.solve** do pacote *RR-BLUP* (ENDELMAN, 2011).

2.6 Acurácia preditiva

Para avaliar a acurácia preditiva dos modelos, obteve-se o coeficiente de correlação de Pearson (r) entre o valor genético genômico predito pelos métodos e o observado. Quanto maior o valor de r, melhor é a predição do modelo.

Na análise de dados reais, o tamanho dos *knots* foi determinado usando a validação cruzada 10-*fold*. Com esse método, uma parte dos dados (10%) é removida e uma B-*Spline* é ajustada com certo número de *knots* aos dados restantes. Em seguida, utiliza-se B-*Spline* para fazer predições para a parte retida. Realiza-se esta sequência várias vezes até que cada observação tenha sido omitida uma vez e, em seguida, calcula-se a correlação de Pearson (r) global. Este procedimento foi repetido para um *grid* de *knots*, a fim de determinar o *knot* que fosse considerado ótimo para cada fenótipo, sendo aquele que forneceu a maior correlação, ou seja, a maior acurácia. Uma vez determinados tais *knots*, realizaram-se as análises finais com eles para, então, fazer a comparação dos métodos.

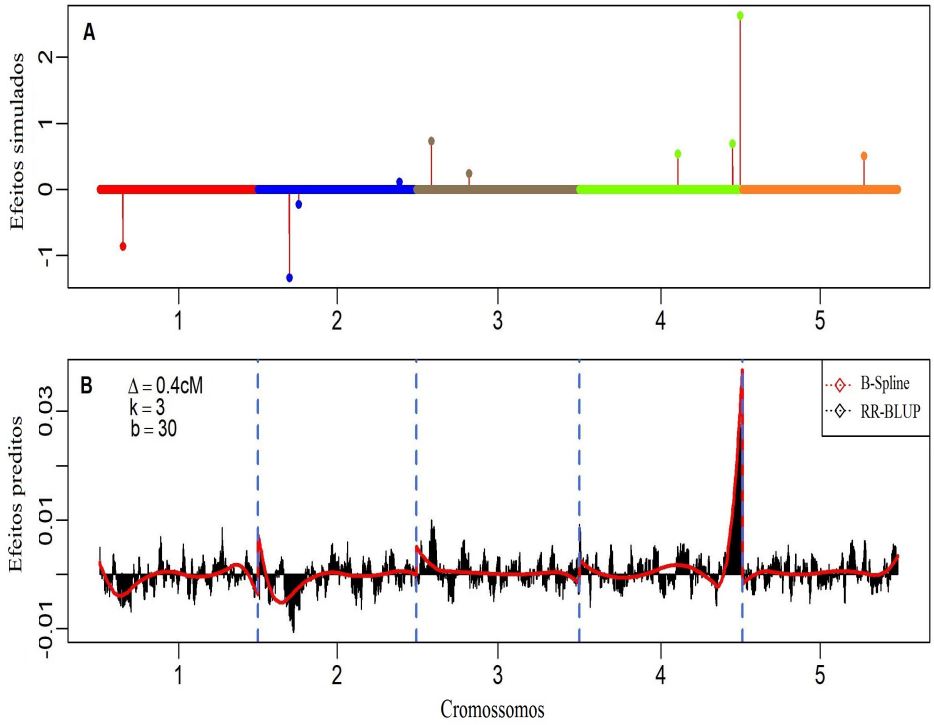
3 RESULTADOS

3.1 Análise de dados simulados

Cenário I: Modelo oligogênico. Na Figura 3.1 estão representados os efeitos dos QTL simulados (painel A) e os efeitos preditos pelos métodos RR-BLUP (em preto) e B-*Spline* (em vermelho) no painel B, considerando a herdabilidade 0,8. Para suavizar a curva B-*Spline*, foram adotados $k = 3$ *knots* por cromossomos. Logo, o número de bases funcionais por cromossomo é $k + m + 1 = 3 + 2 + 1 = 6$, perfazendo, portanto, um total de $b = 30$ bases funcionais (número total de parâmetros a serem estimados) no genoma, haja vista que foram simulados cinco cromossomos. O número de *knots* foi determinado mediante uma busca em *grid* pela maximização na acurácia de predição, o que não é nenhum desafio computacional.

Observa-se que enquanto os efeitos QTL simulados variaram de -1,33 a 2,63 (Figura 3.1-A), os efeitos de marcadores estimados variaram de -0,01 a 0,03 para RR-BLUP e -0,005 a 0,04 para B-*Spline* (Figura 3.1-B). A maioria dos segmentos contendo grandes QTL foi mapeada por ambos os métodos, isto é, os efeitos preditos mostraram padrões semelhantes, porém com efeitos viesados para baixo. A resolução do B-*Spline* foi melhor que a do RR-BLUP e apresentou um pico expressivo no cromossomo 4, região em que foi simulado QTL de maior efeito.

Figura 3.1 – **(A)** Efeitos verdadeiros dos 10 QTL simulados ao longo do genoma representado; **(B)** Efeitos estimados pelos métodos RR-BLUP (em preto) e B-*Spline* (em vermelho), para a herdabilidade 0,8.



O símbolo Δ é o tamanho dos *knots*, k é a quantidade de *knots* por cromossomos e b é a quantidade total de bases funcionais (parâmetros). Linha tracejada em azul separa os cromossomos. Fonte: Do autor (2018).

Na Tabela 3.1 estão apresentadas as correlações de Pearson (r) entre valores genômicos simulados e preditos pelos diferentes métodos ao variar as herdabilidades, a fim de avaliar suas capacidades preditivas. O método Bayes B foi o mais acurado nas três herdabilidades e obteve a mesma acurácia que o Bayes A na herdabilidade 0,8. O modelo B-*Spline* foi menos acurado que RR-BLUP apenas na herdabilidade 0,2 e mais acurado que o Lasso Bayesiano em todas as herdabilidades.

Tabela 3.1 – Correlação de Pearson (r) dos métodos RR-BLUP, Bayes A, Bayes B, Lasso Bayesiano e B-*Spline* para as herdabilidades 0,2; 0,5 e 0,8, no cenário oligogênico.

Métodos	Herdabilidades		
	0,2	0,5	0,8
RR-BLUP	0,841	0,933	0,979
Bayes A	0,813	0,937	0,986
Bayes B	0,855	0,947	0,986
Lasso Bayesiano	0,818	0,935	0,979
B-*Spline*	0,821	0,945	0,980

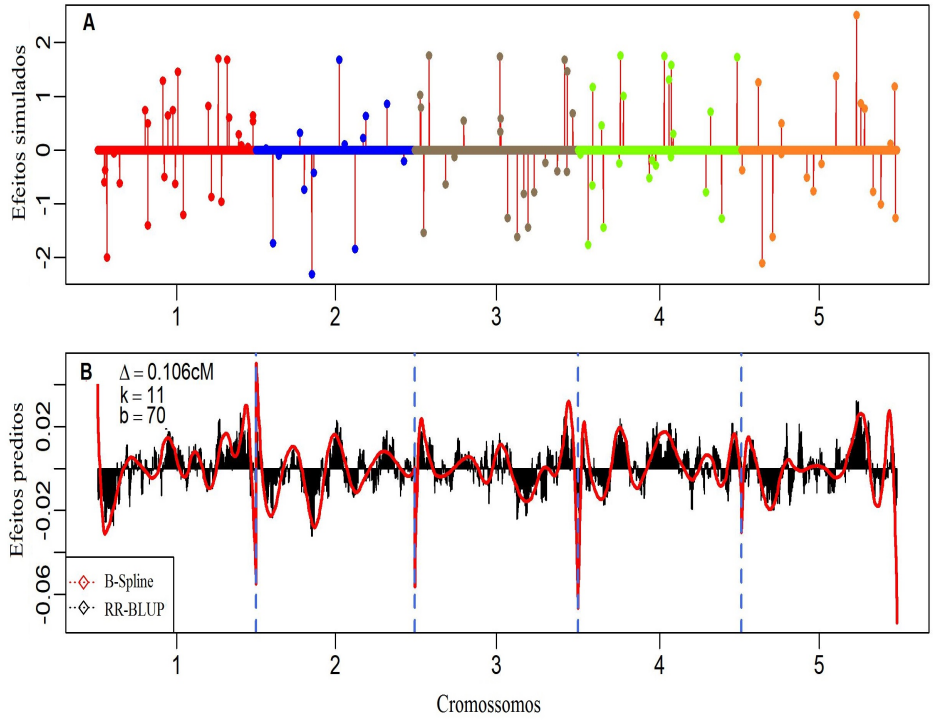
Fonte: Do autor (2017).

Pelos resultados apresentados na Figura 3.1 e na Tabela 3.1, observa-se que a capacidade preditiva do método B-*Spline* é melhor que dos métodos RR-BLUP, Bayes A e Lasso Bayesiano em quase todos os cenários avaliados.

Cenário II: Modelo poligênico. Na Figura 3.2 estão apresentados os efeitos dos QTL simulados (Painel A) e os efeitos preditos pelos métodos RR-BLUP (em preto) e pelo B-*Spline* (em vermelho) no Painel B, considerando a herdabilidade 0,8. Para suavizar a curva B-*Spline*, foram adotados $k = 11$ *knots* por cromossomo. Logo, o número de bases funcionais por cromossomo é $k+m+1 = 11+2+1 = 14$, em que k é o número de *knots* e m é o grau do polinômio suavizado, perfazendo, portanto, um total de $b = 70$ bases funcionais (número total de parâmetros a serem estimados) no genoma, haja vista que foram simulados cinco cromossomos. O número de *knots* também foi escolhido mediante uma busca em *grid* pela maximização na na acurácia preditiva.

Uma observação interessante é que, neste cenário, em que há dez vezes mais QTL simulados do que no cenário anterior, o número de bases funcionais foi pouco mais que o dobro da quantidade utilizada no cenário anterior.

Figura 3.2 – **(A)** Efeitos verdadeiros dos 100 QTL simulados ao longo do genoma representados; **(B)** efeitos estimados pelos métodos RR-BLUP (em preto) e B-*Spline* (em vermelho), para a herdabilidade 0,8.



O símbolo Δ é o tamanho dos *knots*, k é a quantidade de *knots* por cromossomo e b é a quantidade total de bases funcionais (parâmetros). Linha tracejada em azul separa os cromossomos. Fonte: Do autor (2018).

Observa-se que enquanto os efeitos QTL simulados variaram de -2,31 a 2,51 (Figura 3.2-A), os efeitos de marcador estimados variaram de -0,03 a 0,03 para RR-BLUP e de -0,07 a 0,05 para B-*Spline* (Figura 3.2-B). A maioria dos segmentos contendo grande QTL foi mapeada por ambos os métodos, isto é, os efeitos preditos por ambos os métodos mostraram padrão semelhante, porém, em geral, apresentaram viés para baixo devido ao efeito de *shirinkage*. No presente cenário, a resolução do RR-BLUP foi melhor que a do B-*Spline*.

Na Tabela 3.2 estão representadas as correlações de Pearson (r) entre valores genômicos simulados e preditos para os diferentes métodos ao variar a herdabilidade. O método RR-BLUP foi o mais acurado nas três herdabilidades e obteve a mesma acurácia que o método Bayes B na herdabilidade 0,5 e que o Bayes A na herdabilidade 0,8. O método B-*Spline* foi considerado o menos acurado nas três herdabilidades, seguido pelo Lasso Bayesiano.

Tabela 3.2 – Correlação de Pearson (r) dos métodos RR-BLUP, Bayes A, Bayes B, Lasso Bayesiano e B-*Spline* para as herdabilidades 0,2; 0,5 e 0,8, no cenário poligênico.

Métodos	Herdabilidades		
	0,2	0,5	0,8
RR-BLUP	0,822	0,939	0,971
Bayes A	0,805	0,938	0,971
Bayes B	0,812	0,939	0,970
Lasso Bayesiano	0,799	0,935	0,970
B-*Spline*	0,784	0,925	0,961

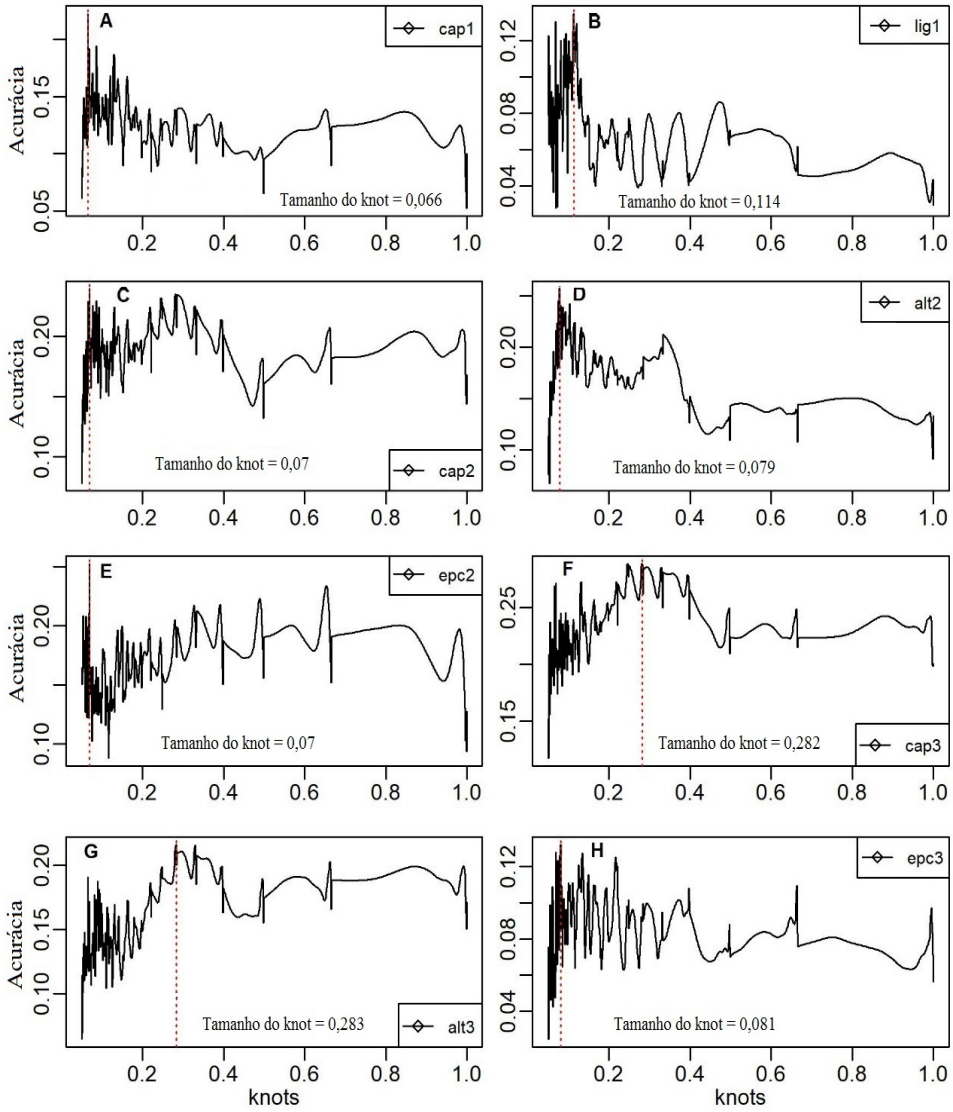
Fonte: Do autor (2018).

Pelos resultados apresentados na Figura 3.2 e na Tabela 3.2, observa-se que a capacidade preditiva do método B-*Spline* é próxima, porém inferior aos outros métodos analisados. Note que o foco está na inferência do valor genético, em vez da detecção de QTL.

3.2 Análise de dados reais

Na Figura 3.3, do painel A-H estão apresentados gráficos com o valor da acurácia em função de um *grid* de candidatos (*knots*) que "varre" todo o espaço das posições.

Figura 3.3 – Acurácias obtidas a partir de um *grid* de *knots* do modelo Spline proposto para oito fenótipos de eucalipto: **(A)** cap1 - circunferência à altura do peito na Época 1; **(B)** lig1 - lignina na Época 1; **(C)** cap2 - circunferência à altura do peito na Época 2; **(D)** alt2 - altura de planta na Época 2; **(E)** epc2 - espessura da casca na Época 2; **(F)** cap3 - circunferência à altura do peito na Época 3; **(G)** alt3 - altura de planta na Época 3; **(H)** epc3 - espessura da casca na Época 3.



A linha tracejada vermelha representa o *knot* que maximiza a acurácia. Fonte: Do autor (2018).

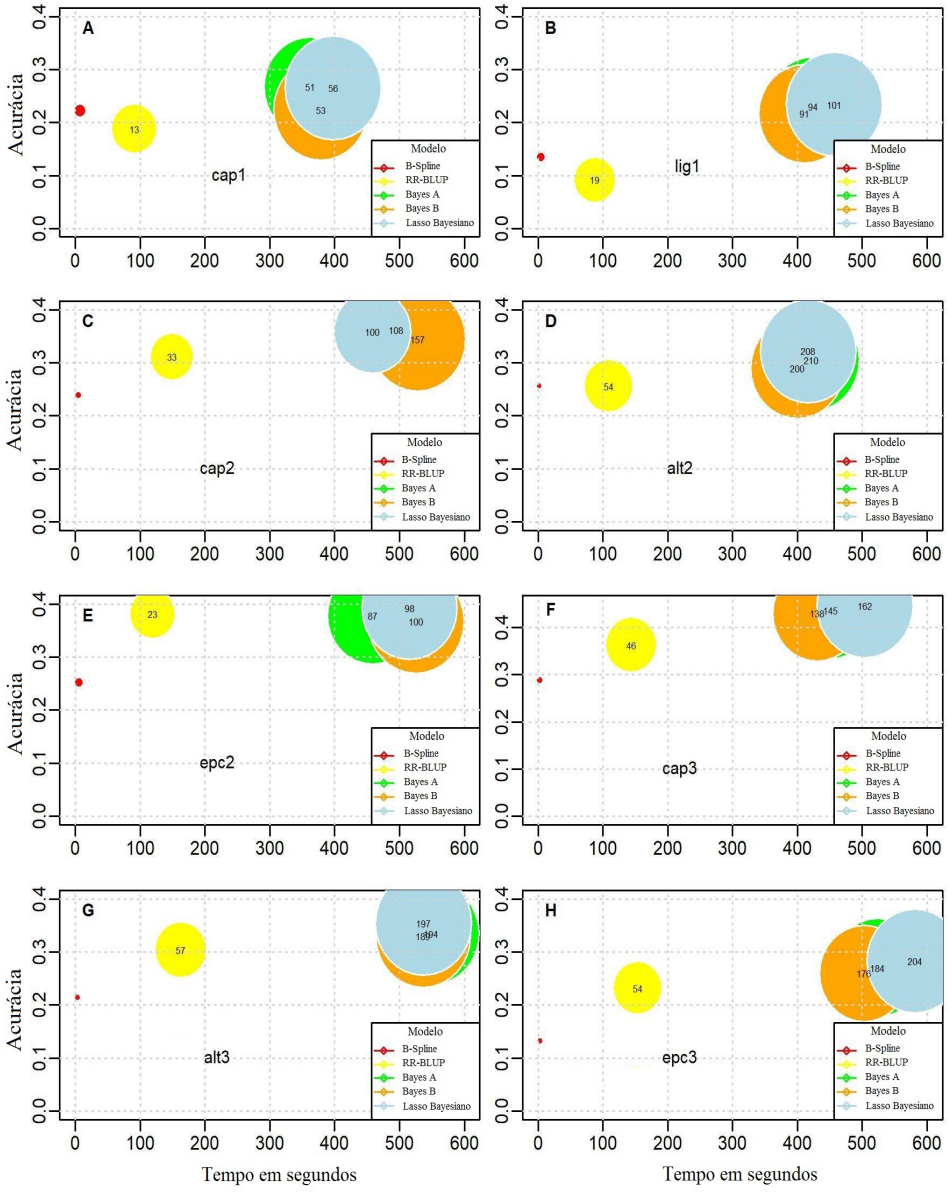
Os valores encontrados de tamanhos de *knots* ótimo para cada fenótipo foram: 0,066 para circunferência à altura do peito na Época 1 (cap1); 0,114 para lignina na Época 1 (lig1); 0,07 para circunferência à altura do peito na Época 2 (cap2); 0,079 para altura de planta na Época 2 (alt2); 0,07 para espessura da casca na Época 2 (epc2); 0,282 para circunferência à altura do peito na Época 3 (cap3); 0,283 para altura de planta na Época 3 (alt3); 0,081 para espessura da casca na Época 3 (epc3).

3.2.1 Acurácia *versus* tempo de análise

O tamanho da amostra foi $n = 610$ e o número de marcadores DArT foi $p = 15104$. Na Figura 3.4 estão apresentadas as acurácias para os oito fenótipos (do painel de A-H) em relação ao tempo de análise referentes aos quatros modelos concorrentes (RR-BLUP, Bayes A, Bayes B e Lasso Bayesiano) e o modelo proposto. Nesta figura, os diâmetros dos círculos são proporcionais ao tempo de análise do modelo B-*Spline*. Com isso, podemos perceber, por exemplo, que para cada segundo gasto pela análise via B-*Spline*, o RR-BLUP gastou 13 segundos (valor expresso no centro do círculo), Bayes A, 51 segundos, Bayes B, 53 segundos, e Lasso Bayesiano, 56 segundos. Além disso, o tempo de análise dos três modelos bayesianos é muito próximo, por isso, às vezes há sobreposição de círculos.

Observa-se que para todos os fenótipos, métodos bayesianos foram considerados mais acurados que RR-BLUP e B-*spline*. Já os métodos RR-BLUP e B-*Spline* apresentaram resultados alternados, sendo que B-*Spline* foi mais acurado para os fenótipos cap1 e lig1, os dois métodos obtiveram mesma acurácia (0,256) para o fenótipo alt2 e, B-*Spline* foi menos acurado para os outros fenótipos. Entretanto, em todos os cenários, B-*Spline* obteve o menor tempo computacional.

Figura 3.4 – Acurácia em relação ao tempo de análise obtida a partir da validação cruzada 10-*fold*, usando os cinco modelos avaliados para os fenótipos: **(A)** cap1 - circunferência à altura do peito na Época 1; **(B)** lig1 - lignina na Época 1; **(C)** cap2 - circunferência à altura do peito na Época 2; **(D)** alt2 - altura de planta na Época 2; **(E)** epc2 - espessura da casca na Época 2; **(F)** cap3 - circunferência à altura do na Época 3; **(G)** alt3 - altura de planta na Época 3; **(H)** epc3 - espessura da casca na Época 3.



Os números no interior de cada circunferência representam o tempo em segundos (aproximado para inteiro) para cada segundo que levaria o B-*spline* (em vermelho).
Fonte: Do autor (2018).

4 DISCUSSÃO

Neste trabalho, acredita-se que os efeitos genéticos dos marcadores têm uma estrutura funcional intrínseca. Embora os dados genéticos sejam sempre discretos, os vemos como realizações de uma função sinal γ contínua nas posições λ. Aqui, λ é simplesmente a localização do marcador no genoma, que é uma variável contínua, e não discreta. O desafio principal é o de postular uma forma funcional relacionando fenótipos a genótipos de marcadores (vistos como milhares de covariáveis altamente colineares) e a valores genéticos. Utilizando técnicas de análise de dados funcionais, a disponibilidade de alta densidade de marcadores pode ser usada para estimar a função sinal por meio de funções *Spline*, B-*Spline* ou Fourier (BOOR, 2001; FAN et al., 2013; HORVÁTH; KOKOSZKA, 2012; RAMSAY; HOOKER; GRAVES, 2009; RAMSAY; SILVERMAN, 1996). Então, a função sinal estimada é usada no modelo de regressão linear funcional para se conectar ao ajuste do fenótipo.

Na análise dos dados simulados no presente estudo, todos os métodos avaliados tiveram correlações superiores a 90%, sendo que B-*Spline* obteve desempenho equivalente (ora mais acurado, ora menos acurado) aos métodos concorrentes avaliados (Tabelas 3.1 e 3.2). Fan et al. (2013) também desenvolveram um estudo em que consideraram o genoma humano como uma função estocástica e utilizaram modelos lineares funcionais de efeitos fixos e mistos para testar associação entre QTL e variantes genéticas. Uma das abordagens foi estimar a função sinal por meio de funções base Fourier e B-*Spline* e comparar com os testes de associação baseado em *kernel* de Lee et al. (2012) e Wu et al. (2011). Os autores mostraram que os resultados dos testes de modelos funcionais de efeitos fixos utilizando B-*Spline* ou Fourier foram muito melhores que os testes concorrentes avaliados. Já os resultados dos testes de modelos funcionais de efeitos mistos foram um pouco

misturados, ou seja, às vezes obteve desempenho pior, às vezes desempenho melhor que os testes concorrentes.

Analisando os resultados referentes aos dados reais, os três métodos bayesianos foram mais acurados que os demais. Estes resultados corroboram com os apresentados por Gianola, Fernando e Stella (2006) e Howard, Carriquiry e Beavis (2014) que avaliaram métodos paramétricos clássicos e métodos não paramétricos e semiparamétricos. De acordo com os autores, se o objetivo for o melhoramento genético e a arquitetura genética subjacente for conhecida como aditiva, os métodos GS paramétricos fornecerão melhores predições para a seleção (HOWARD; CARRIQUIRY; BEAVIS, 2014). Mais ainda, os resultados apresentados por Howard, Carriquiry e Beavis (2014) sugerem que, se o objetivo da pesquisa é predizer com precisão o valor genotípico de um indivíduo, particularmente para fins de seleção, e se a arquitetura genética subjacente dos caracteres não for conhecida, então é melhor usar o método não paramétrico.

Os desempenhos de diversos métodos de seleção genômica foram amplamente comparados quanto à capacidade preditiva em diversos estudos, por exemplo, Calus (2010), Daetwyler et al. (2010), Howard, Carriquiry e Beavis (2014), Meuwissen, Hayes e Goddard (2001), Moser et al. (2009), Thavamanikumar, Dolferus e Thumma (2015), Van den Berg, Calus e Wientjes (2015) e Zhang et al. (2010). A conclusão é que parece existir uma incerteza na determinação de qual método é o mais apropriado, haja vista que, segundo esses autores, a eficiência de cada método depende, entre outros fatores, da arquitetura genética e da densidade do marcador. Além disso, como destacam Gianola (2013) e Gianola et al. (2009), os métodos Bayesianos (apelidados por Gianola et al. (2009) de "alfabeto bayesiano") são influenciados pela suposição acerca dos hiperparâmetros das prioris, de modo que as prioris podem exercer mais influência na estimativa das componentes

de variância do que os próprios dados.

Gianola, Fernando e Stella (2006) afirmaram que as abordagens paramétricas para GS têm vários inconvenientes. Os pressupostos do modelo paramétrico nem sempre são válidos (por exemplo, normalidade, variáveis explicativas independentes, linearidade). Dessa forma, as críticas a essa abordagem foram seguidas por uma defesa a abordagens alternativas baseadas em técnicas de aprendizado de máquina (*machine learning*) ou até mesmo abordagens não paramétricas (TEMPELMAN, 2015). Uma observação importante é que $\hat{\mathbf{\Phi}}$ depende da matriz transformada $\mathbf{W}$, mas não sofre influência diretamente de quão grande é p. O que se quer mostrar com isso é que mesmo com $p \to \infty$, não é necessário estimar uma infinidade de parâmetros, ou seja, encontra-se um modelo que converte problemas genômicos de alta dimensão em um modelo de dimensão finita b representado na equação (2.11).

Analisando o tempo computacional, os métodos bayesianos exigiram mais demandas; já o método B-*Spline* apresentou o menor custo computacional (Figura 3.4). Note na Figura 3.4-D que apesar dos métodos bayesianos terem sido mais acurados que RR-BLUP e B-*Spline*, o tempo computacional gasto foi de 200 a 210 vezes maior que o tempo do B-*Spline* (ou seja, a cada 1 segundo gasto pelo B-*Spline*, os métodos bayesianos gastaram de 200 a 210 segundos). Este resultado notável faz com que o uso de *Splines* seja relevante em análises de seleção genômica.

É bem conhecido que o ponto chave no uso de uma *Spline* é a determinação das posições dos *knots*, tendo que ser colocados o mais precisamente possível (DIERCKX, 1993; UTPOTT, 2015; WEGMAN; WRIGTH, 1983). No presente estudo, o tamanho ideal dos *knots*, para a análise dos dados reais, foi escolhido por meio de um escaneamento em *grid*. Este procedimento exigiu trabalho computaci-

onal intensivo, pois para cada tamanho de *knot* testado realizou-se uma análise de acurácia preditiva; ao final de todo o escaneamento, obteve-se o tamanho correspondente à maior acurácia. Outra maneira seria permitir que os dados ajudassem a determinar as "melhores" opções para o número e a posição dos *knots*.

Biller (2000) introduziu uma abordagem totalmente bayesiana de regressão *Splines* com seleção automática de *knots* em modelos lineares generalizados usando B-*Spline*. O número e a localização dos *knots* são escolhidos por métodos MCMC de salto reversível, juntamente com os coeficientes desconhecidos que determinam a forma da *Spline*. Estes procedimentos permitem a amostragem a partir da posteriori do conjunto de modelos, consistindo em diferentes números e posicionamentos de *knots*, e correspondente dimensionalidade diferente dos parâmetros de *Spline*. Wallstrom, Liebner e Kass (2008) também utilizou o método MCMC de salto reversível para executar a regressão não paramétrica generalizada baseada em splines. Para o modelo genoma contínuo apresentado neste trabalho, pesquisas futuras poderão abordar o uso de técnicas de *knots* livres para testar a eficiência do modelo em seleção genômica.

5 CONCLUSÃO

A forma funcional das curvas polinomiais por partes (*Splines*) pode permitir novos tipos de análise. Este estudo mostra que a abordagem com B-*Spline* pode lidar com um número ilimitado de marcadores e com custo computacional muito menor do que modelos tradicionais de seleção genômica. Assim, na genômica, a forma funcional pode descrever uma região em oposição a marcadores individuais e é, portanto, um método promissor e merece mais exploração.

REFERÊNCIAS

BAR-JOSEPH, Z. et al. Continuous representations of time-series gene expression data. **Journal of Computational Biology**, New York, v. 10, n. 3/4, p. 341-356, 2003.

BILLER, C. Adaptive bayesian regression splines in semiparametric generalized linear models. Journal of Computational and Graphical Statistics, Alexandria, v. 9, n. 1, p. 122-140, 2000.

BOOR, C. de. **A practical guide to splines.** New York: Springer, 1978. 392 p.

_____. **A practical guide to splines.** New York: Springer, 2001. 348 p.

CALUS, M. P. L. Genomic breeding value prediction: methods and procedures. **Animal**, Cambridge, v. 4, n. 2, p. 157-164, Feb. 2010.

CALUS, M. P. L. et al. Accuracy of genomic selection using different methods to de?ne haplotypes. **Genetics**, Austin, v. 178, n. 1, p. 553-561, Jan. 2008.

DAETWYLER, H. D. et al. The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, Austin, v. 185, n. 3, p. 1021-1031, July 2010.

DE LOS CAMPOS, G. et al. Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, Austin, v. 182, n. 1, p. 375-385, May 2009.

DIERCKX, P. **Curve and surfasse ?tting with splines.** Clarendon: University Press, 1993. 304 p.

ELSHIRE, R. J. et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS One**, San Francisco, v. 6, n. 5, p. 1-10, May 2011.

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

FAN, R. et al. Functional linear models for association analysis of quantitative

traits. **Genetic Epidemiology**, New York, v. 37, n. 7, p. 726-742, Nov. 2013.

GIANOLA, D. Priors in whole-genome regression: the Bayesian alphabet returns. **Genetics**, Austin, v. 194, n. 3, p. 573-596, July 2013.

GIANOLA, D. et al. Additive genetic variability and the Bayesian alphabet. **Genetics**, Austin, v. 183, n. 1, p. 347-363, Sept. 2009.

GIANOLA, D.; FERNANDO, R. L.; STELLA, A. Genomic assisted prediction of genetic value with semi-parametric procedures. **Genetics**, Austin, v. 173, n. 3, p. 1761-1776, July 2006.

HORVÁTH, L.; KOKOSZKA, P. **Inference for functional data with applications.** New York: Springer, 2012. 422 p.

HOWARD, R.; CARRIQUIRY, A. L.; BEAVIS, W. D. Parametric and nonparametric statistical methods for genomics selection of traits with additive and epistatic genetic architectures. **G3:** genes, genomes, genetics, Bethesda, v. 4, n. 6, p. 1027-1046, Apr. 2014.

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

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.

KONOPKA, T. Automated analysis of biological oscillator models using mode decomposition. **Bioinformatics**, Oxford, v. 27, n. 7, p. 961-967, Apr. 2011.

LEE, S. et al. Optimal uni?ed approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. **The American Journal of Human Genetics**, Baltimore, v. 91, n. 2, p. 224-237, Aug. 2012.

MEUWISSEN, T. H. E.; HAYES, B. J.; GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, Austin, v. 157, n. 4, p. 1819-1829, Apr. 2001.

MICHNA, A. et al. Natural cubic spline regression modeling followed by

dynamic network reconstruction for the identi?cation of radiation-sensitivity gene association networks from time-course transcriptome data. **PLoS One**, San Francisco, v. 11, n. 8, p. e0160791, 2016.

MOSER, G. et al. A comparison of ?ve methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics, Selection, Evolution**, London, v. 41, p. 56, Dec. 2009.

MOURA, E. G. **Aplicação de modelos funcionais na seleção genômica ampla.** 2017. 54 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2017.

PÉREZ, P.; DE LOS CAMPOS, G. Genome wide regression and prediction with BGLR Statistical Package. **Genetics**, Austin, v. 144, n. 2, p. 164-442, Oct. 2014.

R CORE TEAM. **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Disponível em: <https://www.R-project.org/>. Acesso em: 21 jan. 2018.

RAMSAY, J. O.; HOOKER, G.; GRAVES, S. **Functional data analysis with R and MATLAB.** New York: Springer Science & Business Media, 2009. 202 p.

RAMSAY, J. O.; SILVERMAN, B. W. **Functional data analysis.** New York: Springer, 1996. 426 p.

TEMPELMAN, R. J. Statistical and computational challenges in whole genome prediction and genome-wide association analyses for plant and animal breeding. **Journal of Agricultural, Biological and Environmental Statistics**, Alexandria, v. 20, n. 4, p. 442-466, Dec. 2015.

THAVAMANIKUMAR, S.; DOLFERUS, R.; THUMMA, B. R. Comparison of genomic selection models to predict flowering time and spike grain number in two hexaploid wheat doubled haploid populations. **G3:** genes, genomes, genetics, Bethesda, v. 5, n. 10, p. 1991-1998, July 2015.

UTPOTT, N. M. **Regressão logística utilizando b-splines:** uma maneira de lidar com relações não lineares. 2015. 75 p. Monografia - (Bacharelado em Estatística) - Universidade Federal do Rio Grande do Sul, Porto Alegre, 2015.

VAN DEN BERG, S.; CALUS, M. P.; WIENTJES, Y. C. J. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. **BMC Genetics**, London, v. 16, n. 1, p. 146-157, 2015.

WALLSTROM, G.; LIEBNER, J.; KASS, R. E. An implementation of Bayesian Adaptive Regression Splines (BARS) in C with S and R Wrappers. **Journal of Statistical Software**, Los Angeles, v. 26, n. 1, p. 1-21, June 2008.

WANG, W.; YAN, J. **Splines2:** regression spline functions and classes. R package, version 0.2.7. [S.l.: s.n.], 2017. Disponível em: <https://CRAN.R-project.org/ package=splines2>. Acesso em: 15 mar. 2018.

WEGMAN, E. J.; WRIGHT, I. W. Splines in statistics. **Journal of the American Statistical Association**, Washington, v. 78, n. 382, p. 351-365, 1983.

WU, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. **The American Journal of Human Genetics**, Baltimore, v. 89, n. 1, p. 82-93, July 2011.

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

ZHANG, X. et al. Integration of association statistics over genomic regions using Bayesian adaptive regression splines. **Human Genomics**, London, v. 1, n. 1, p. 20-29, Nov. 2003.

ZHANG, Z. et al. Best linear unbiased prediction of genomic breeding values using a trait-speci?c marker-derived relationship matrix (T Mailund, Ed.). **PLoS One**, San Francisco, v. 5, n. 9, p. e12648, Sept. 2010.