ERNANDES GUEDES MOURA

ANÁLISE GENÔMICA POR JANELAS CROMOSSÔMICAS: REGRESSÃO FUNCIONAL E SALTOS REVERSÍVEIS

LAVRAS – MG
2022

ERNANDES GUEDES MOURA

ANÁLISE GENÔMICA POR JANELAS CROMOSSÔMICAS: REGRESSÃO FUNCIONAL E SALTOS REVERSÍVEIS

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 Doutor.

Orientador

Dr. Márcio Balestre

(*in memoriam*)

Coorientador

Júlio Sílvio de Sousa Bueno Filho

Coorientador

Carlos Pereira da Silva

LAVRAS - MG
2022

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

Moura, Ernandes Guedes.
Análise genômica por janelas cromossômicas : Regressão funcional e saltos reversíveis / Ernandes Guedes Moura. - 2021.
93 p. : il.

Orientador(a): Márcio Balestre.
Coorientador(a): Júlio Sílvio de Sousa Bueno Filho, Carlos Pereira da Silva.
Tese (doutorado) - Universidade Federal de Lavras, 2021.
Bibliografia.

1. B-Splines. 2. MCMC. 3. Modelos funcionais. I. Balestre, Márcio. II. Bueno Filho, Júlio Sílvio de Sousa. III. da Silva, Carlos Pereira. IV. Título.

ERNANDES GUEDES MOURA

ANÁLISE GENÔMICA POR JANELAS CROMOSSÔMICAS: REGRESSÃO FUNCIONAL E SALTOS REVERSÍVEIS

GENOMIC ANALYSIS THROUGH CHROMOSOMAL WINDOWS: FUNCTIONAL REGRESSION AND REVERSIBLE JUMPS

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 Doutor.

Aprovado em 26 de fevereiro de 2021.

Dr. Carlos Pereira da Silva UFLA

Dr. Antônio Augusto Franco Garcia USP

Dra. Andrezza Kellen Alves Pamplona IFTM

Dr. José Lindenberg Rocha Sarmento UFPI

Júlio Sílvio de Sousa Bueno Filho

Coorientador

Carlos Pereira da Silva

Coorientador

LAVRAS - MG
2022

AGRADECIMENTOS

Agradeço primeiramente a Deus, pela oportunidade de fazer o Doutorado na área de Estatística e Experimentação Agropecuária.
À Universidade Federal de Lavras, especialmente ao Departamento de Estatística, pela oportunidade.
Ao CNPq, pela concessão da bolsa de doutorado por um período.
Ao Instituto Federal do Maranhão – IFMA pela Liberação para Capacitação.
Ao Instituto Federal do Piauí – IFPI.
Aos professores Márcio Balestre (*In memoriam*) e Júlio Sílvio de Sousa Bueno Filho, pela orientação, paciência e disposição para contribuir e ajudar da melhor forma possível. A todos os professores do departamento de Estatística - DES pelos valiosos ensinamentos ao longo do doutorado, em especial aos professores Renato Ribeiro de Lima e Joel Augusto Muniz por serem tão solícitos sempre que precisei.
Aos professores José Lindenberg Rocha Sarmento, Antônio Augusto Franco Garcia e Andrezza Kellen Alves Pamplona por participarem da defesa da tese .
Ao Carlos Pereira da Silva ("programador chefe") pela grande contribuição nesta tese.
A Nádia Ferreira Secretária do Departamento de Estatística –DES.
A todos os colegas do setor do Departamento de Estatística, pois muitos se tornaram amigos!
Aos amigos que integraram o Grupo de Pesquisa e Orientação do Professor Márcio Balestre, os quais mais tive contato diretamente. (Cristian Tiago Mendes, Andrezza Kellen Alves Pamplona, Taís Alvarenga, Luciano Oliveira, Carlos Pereira da Silva, Michele Barbosa, Indalécio Cunha, Joel Jorge Nuvunga e Wilson Sanches Mateus).
Quero agradecer aos amigos e companheiros que também fizeram parte dessa caminhada: Jodean Alves da Silva, Rodrigo Fonseca da Silva, José Ferreira Lustosa Filho, Fabrício Andrade, Kelly Pereira de Lima, Henrique José de Paula Alves, Sérgio Domingos Simão, Elliezer De Almeida Melo, Ismael Simão.
Aos meus pais, Francisco de Assis Guedes Lima e Francisca Maria de Moura pelo amor e apoio incondicional, em todas as minhas decisões nas diferentes etapas da minha vida e aos meus irmãos Francis James Guedes Lima e Juliana Guedes Moura e toda minha família que me apoiaram de alguma forma nessa caminhada, em especial o meu tio Pe. Cícero de Moura Sobrinho por ser sempre presente.

RESUMO

Há muitos métodos para a seleção genômica que tratam dos problemas da multicolinearidade e da alta dimensionalidade, dentre os quais destacam-se na literatura o rr-BLUP e o Bayes B. Métodos de genoma contínuo e regressão funcional em janelas cromossômicas (*bins*) foram recentemente propostos para melhor utilizar o desequilíbrio de ligação entre SNP (*Single Nucleotide Polymorphism*) e potenciais QTLs (*Quantitative Trait Loci*). Uma das estratégias propostas é utilizar funções polinomiais ou trigonométricas em versões ajustadas para janelas cromossômicas (*bins*). Nesse caso, um fator complicador é a potencial má especificação do número e tamanho dos *bins*, com potencial ampliação do erro de predição. Neste trabalho nós investigamos as vantagens de fazer a inferência conjunta *a posteriori* do número, do tamanho e dos efeitos das marcas nos *bins* em um processo de amostragem por saltos reversíveis. Este tipo de técnica era de difícil implementação para modelos anteriores que levavam em conta a distância entre marcas e QTLs, mas pode ser muito simplificado em modelos de regressão simples típicos da genômica moderna (em que cada SNP segrega potencialmente como um QTL). Estudamos as duas estratégias e suas consequências imediatas para a seleção genômica. Uma revisão de literatura inicial sobre os métodos foi feita como subsídio para dois artigos. No primeiro, avaliamos a implementação de modelos funcionais de *bins* com o uso de séries de Fourier e de B-Splines. No segundo, introduzimos o RJ-MCMC (*Reverse Jump Markov Chain Monte Carlo*) para modelos funcionais em que cada *bin* é representado na amostragem por uma só de suas marcas. Os modelos considerados foram comparáveis aos métodos mais utilizados quanto à predição (Bayes B, rr-BLUP) e são adequados para a seleção genômica. Como potencial subproduto do da tese, os resultados para estudos de associação são também interessantes, embora não tenha sido nosso objetivo principal avaliá-los.

Palavras-chave: *B-Splines*, MCMC, Modelos funcionais, Séries de *Fourier*.

ABSTRACT

There are many methods for genomic selection that address the problems of multicollinearity and high dimensionality, among which the rr-BLUP and Bayes B stand out in the literature. Methods of continuous genome and functional regression in chromosomal windows (bins) were recently proposed to better utilize the linkage disequilibrium between SNP (Single Nucleotide Polymorphism) and potential QTLs (Quantitative Trait Loci). One of the proposed strategies is to use polynomial or trigonometric functions in bins-fitted versions. In this case, a complicating factor is the potential misspecification of the number and the sizes of the bins, with a potential increase in the prediction error. In this thesys we investigate the advantages of making inference in the joint posterior distribution for the number, the size and the effects of marks in bins in a reversible jump sampling process. This type of technique was difficult to implement for previous models that took into account the distance between marks and QTLs, but it can be greatly simplified in simple regression models typical of modern genomics (where each SNP potentially segregates as a QTL). We study the two strategies and their immediate consequences for genomic selection. A basic review of the literature methods was used to subsidize two original papers. In the first one, we evaluated the implementation of functional models of bins using Fourier series and B-Splines. In the second, we introduce the RJ-MCMC (Reverse Jump Markov Chain Monte Carlo) for functional models in which each bin is represented in the sampling by only one of its marks. The models considered were comparable to the most used for prediction (Bayes B, rr-BLUP) and are suitable for genomic selection. As a potential by-product of the thesys, the results for association studies are also interesting, despite not being our main goal to evaluate them.

Keywords: B-Splines, MCMC, Functional models, Fourier series.

SUMÁRIO

PRIMEIRA PARTE

1 INTRODUÇÃO

A disponibilidade de marcadores moleculares densos tornou possível o uso da seleção genômica ampla (GWS) para o melhoramento de animais e plantas (MEUWISSEN; HAYES; GOODARD, 2001). O objetivo da seleção genômica ampla é predizer os efeitos de marcadores densos que estão distribuídos em todo o genoma e, em seguida, utilizar esses efeitos de preditos para obter o valor genético genômico estimados para indivíduos genotipados (ZHANG et al., 2011; WANG et al., 2012). Para isso, de acordo com Calus et al. (2008), uma condição essencial é que haja desequilíbrio de ligação (LD) entre alelos dos marcadores e locus reguladores de caracteres quantitativos (ou QTL, do Inglês *Quantitative Trait Loci*).

Em geral, os modelos de seleção genômica (GS, do Inglês *Genomic Selection*) enfrentam os problemas de dimensionalidade e multicolinearidade. Os primeiros decorrem de que o número de preditores (que descrevem estados dos marcadores) ser muito maior que o número de observações fenotípicas. Os últimos aparecem devido ao desequilíbrio de ligação entre os marcadores. Nesses casos as estimativas de mínimos quadrados ordinários (OLS – do Inglês *Ordinary Least Squares*), são inviáveis (DESTA; ORTIZ, 2014; PÉREZ; DE LOS CAMPO, 2014; BALESTRE; DE SOUZA, 2016). Para superar esses problemas pode-se utilizar métodos estatísticos (regressões) que facilitem a seleção de covariáveis ou empreguem fatores de encolhimento. Os métodos BLUP (melhor predição linear não viesada) e métodos bayesianos são mais usados para estimar o mérito genético. Mais recentemente apareceram artigos com abordagens alternativas baseadas em técnicas de aprendizado de máquina (*machine learning*) ou até mesmo abordagens não paramétricas (TEMPELMAN, 2015).

Hu, Wang e Xu (2012) propuseram um novo método de GS de redução dimensional. Esse método, foi denominado de genoma contínuo, em que os autores assumiram que o efeito de expressão gênica de um locus é função de sua posição no genoma (uma quantidade contínua). A estratégia foi de dividir o genoma em subintervalos (janelas ou como os autores denominaram, *bins*) de alto LD. Desta forma seria possível modelar os efeitos médios dos *bins* ao invés dos marcadores individuais. Embora o método desses autores tenha apresentado melhoria expressiva em termos de acurácia preditiva em relação aos métodos tradicionais de GS, esse método pode não ser diretamente aplicado em algumas situações particulares, devido a dificuldades de especificação de janelas naturais de segregação conforme o tamanho amostral, o número de

marcadores, e a estrutura populacional (XU, 2013). O autor desenvolveu o conceito de "*bins* artificiais" em que se permite pontos de quebra (*breakpoints*) dentro do que inicialmente seriam janelas naturais. Desta forma, o número de *bins* resultante independe do tamanho amostral e do número de SNPs (XU, 2013). Com o uso de janelas de recombinação é possível estimar uma função que sinaliza a expressão gênica desconhecida em função da posição no genoma (e da segregação na janela). Tal função permite predizer o valor genético genômico de indivíduos que ainda não foram fenotipados, embora já tenham sido genotipados. Por simplicidade, os autores utilizaram a média da segregação genotípica nos marcadores da janela como medida de informação. Pode-se usar e já foram testadas com sucesso funções polinomiais ou senoidais, por exemplo. Os principais pressupostos para o sucesso do método utilizando médias diretas são o alto desequilíbrio de ligação e a presença de efeitos homogêneos de marcadores dentro de cada *bin*. Na violação destas pressuposições é preciso desenvolver modelos adaptativos com pesos heterogêneos ou funções mais complexas para os *bins*.

Uma abordagem alternativa foi proposta Moura, Pamplona e Balestre (2019), em que os autores assumiram os marcadores como variáveis aleatórias dentro de cada *bin* ao longo do genoma, em busca da função sinal desconhecida. Assim, buscou-se descobrir os pesos relativos das funções associadas (segregação de marcadores aleatoriamente escolhidos no *bin*) para descrever a resposta do gene baseado na posição. A inferência *a posteriori* foi implementada em um processo estocástico Monte Carlo via Cadeias de Markov (MCMC – do Inglês *Markov Chain Monte Carlo*). Os resultados obtidos por essa proposta foram satisfatórios em termos de acurácia preditiva. Constatou-se também que esse método é eficiente mesmo para populações com baixo desequilíbrio de ligação.

Pamplona (2018) realizou adaptações dos métodos rr-BLUP, Bayes A e Bayes B. Seu objetivo era identificar se é possível aumentar a capacidade preditiva desses métodos por meio de uma pré-seleção de marcadores representativos em janelas de segregação (*bins*). A autora concluiu que, na maioria das situações, os métodos adaptativos em *bins* foram mais acurados na predição de valores fenotípicos "futuros" do que suas respectivas formas originais, tanto em populações com baixo quanto alto desequilíbrio de ligação.

A presente tese está dividida nas seguintes seções: no capítulo II faremos uma revisão geral da literatura sobre modelos funcionais em genômica. No capítulo III apresentaremos um artigo sobre o emprego das duas funções base mais empregadas (transformadas de *Fourier* e *B*-

Splines). No capítulo IV apresentaremos em outro artigo um modelo de inferência bayesiana sobre a seleção automática de *bins* e compararemos sua eficiência à dos métodos tradicionais de seleção genômica. O capítulo V apresenta conclusões finais do estudo que gerou os dois trabalhos.

2 REFERENCIAL TEÓRICO

Um dos objetivos finais da pesquisa genômica é predizer fenótipos completos a partir de múltiplos genes (WANG et al., 2018). Essa linha de pesquisa é denominada seleção genômica. Com a disponibilidade de marcadores de alta densidade em todo o genoma, a seleção genômica tornou-se um método promissor para estimar o mérito genético e de importância econômica para espécies de animais e plantas (ZHANG et al., 2010).

A seleção genômica utiliza valores de melhoramento molecular derivados de marcadores densos em todo o genoma para a seleção de indivíduos jovens. A finalidade é usar marcadores genômicos para estimar os efeitos de todos os locus e, assim, calcular o valor genético genômico estimado (GEBV), de modo a obter uma seleção mais abrangente e confiável (WANG et al., 2018). Dessa forma, a seleção genômica tem o potencial para reduzir os custos de reprodução, eliminando indivíduos com menos potencial numa fase precoce (TEMPELMAN, 2015).

2.1 Análise de dados funcionais (FDA)

A análise de dados funcionais (FDA) é um campo de estudo da estatística que lida com a análise de dados cujas observação são funções (curvas) (MORRIS, 2015). A maioria dos trabalhos em análise de dados funcional é baseada em uma variante do Modelo Linear Funcional (FLM), introduzida primeiramente por (RAMSAY; DALZELL, 1991). Esse tipo análise utiliza combinações lineares de funções bases como principal método para representar funções. Assim, o termo funcional em referência aos dados observados refere-se à estrutura intrínseca dos dados e não à sua forma explícita. Na prática, os dados funcionais geralmente são observados e registrados discretamente como pares (t, x) (MONTESINOS-LÓPEZ et al., 2018). Todavia, assume-se que existe uma função suave *f* que deu origem aos dados observados. Existem uma gama de diferentes sistemas de funções de base, tais como funções de base polinomial, funções de base gaussiana, funções de base *Wavelet*. Duas das funções base mais populares são abordadas nesse estudo, *Fourier* e *B-Spline*.

2.1.1 *Fourier*

Uma série de *Fourier* é uma expansão de uma função periódica em termos de uma soma infinita de combinações de senos e cossenos da seguinte forma:

$$x(t) = c_0 + c_1 sen(wt) + c_2 \cos(2wt) + c_3 sen(2wt) + c_4 \cos(2wt) + \cdots \quad (1)$$

No contexto aqui, utiliza-se essa série truncada em algum valor m, de modo a não superestimar o modelo. Então, a equação (1) pode ser reescrita da seguinte maneira:

$$x(\mathrm{t}) = c_0 + \sum_{j=1}^{m} \left[c_{2j-1} sen(jwt) + c_{2j} \cos(jwt) \right] \quad (2)$$

em que $b = 1 + 2m$ é o número total de bases *Fourier*. A constante ω está relacionada ao período T pela relação $\omega = 2\pi / T$, ou seja, as primeiras parcelas de seno e cosseno oscilarão uma vez durante o domínio de $x(\mathrm{t})$, as parcelas associadas a c_3 e c_4 oscilarão duas vezes durante o domínio, e assim por diante e, T pode ser definido como próprio período (amplitude) do espaço da posição ou tempo (RAMSAY; HOOKER; GRAVES, 2009) e, além disso, note que para garantir a ortogonalidade, o número total de bases b é sempre ímpar (intercepto e sucessivos pares seno e cosseno). Note que, por causa de como definimos ω, cada função base se repete após T unidades de tempo decorridas. Por isso, é frequentemente o uso de análise de *Fourier* em dados com um certo grau de periodicidade.

De acordo Ramsay, Hooker e Graves (2009), apenas duas informações são necessárias para definir um sistema de base de *Fourier*:

1. O número de funções básicas K e
2. O período T.

2.1.2 B-Spline

Antes de definir *B-Spline*, é importante conhecermos uma *Spline*. Uma *Spline* nada mais é que um polinômio por partes, com limites em pontos chamados pontos de interrupção ou nós. De acordo com Ramsay, Hooker e Graves (2009) a função *Spline* em qualquer subintervalo é um polinômio de grau ou ordem fixa, mas a natureza do polinômio muda quando se passa para o próximo subintervalo. O termo grau se refere a maior potência no polinômio. Por exemplo,

uma parábola é definida por um polinômio de grau dois, já que sua maior potência é dois, mas é de ordem três, porque também tem um termo constante. São mais flexíveis que as séries de *Fourier* e é caracterizada por número de nós (pontos em que os segmentos se conectam), a ordem e o grau do polinômio. A equação para uma *Spline* de grau p com k nós é:

$$x(t) = c_0 + c_1 t + . \quad . \quad . + c_p t^p + \sum_{k=1}^{K} c_{pk} (t - \xi_k)^p_+ \tag{3}$$

sujeito a seguinte restrição:

$$(t - \xi_k)^p_+ = \begin{cases} (t - \xi_k)^p \; se \; t \geq \xi_k \\ 0 \qquad\quad se \; t < \xi_k \end{cases} \tag{4}$$

em que $\{\xi_1, ..., \xi_k\}$ é um conjunto de nós fixos. Com esta configuração, a estimativa de $x(t)$ pode ser alcançado através da estimativa dos coeficientes $\mathbf{c} = (c_0, ..., c_p, c_{p1}, ..., c_{pk})'$ da seguinte maneira. Dado $\mathbf{y} = (y_1, ..., y_n)'$ e a matriz de base para uma *Spline* de grau p com k nós é:

$$\mathbf{X} = \begin{bmatrix} 1 & t_1 \ldots t_1^p & (t_1 - \xi_1)^P_+ & \ldots & (t_1 - \xi_k)^p_+ \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & t_n \ldots t_n^p & (t_n - \xi_1)^P_+ & \ldots & (t_n - \xi_k)^p_+ \end{bmatrix}.$$

Se um número de bases for menor que o número de observações ($b < n$), pode-se obter as soluções das bases *Splines* pelo estimador clássico de mínimos quadrados ordinários.

$$\hat{\mathbf{c}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} \tag{5}$$

Todavia, caso o número de bases seja maior que número de observações, é necessário um método de penalização. Alguns trabalhos (CURRIE; DURBAN, 2002; HALL; OPSOMER, 2005) têm sido proposto para determinar o parâmetro de suavização α de *P-Spline*. Então, para um determinado parâmetro de suavização α, uma estimativa de mínimos

quadrados penalizados $\hat{\mathbf{c}}_\alpha$ pode ser obtido como o minimizador do seguinte critério de mínimos quadrados penalizados:

$$\|\mathbf{y}-\mathbf{Xc}\|+\alpha^{2p}\mathbf{c'Dc} \tag{6}$$

Pode-se mostrar que $\hat{\mathbf{c}}_\alpha$ admite a seguinte solução:

$$\hat{\mathbf{c}}_\alpha=(\mathbf{X'X}+\alpha^{2p}\mathbf{D})^{-1}\mathbf{X'y} \tag{7}$$

em que $\mathbf{D}=diag\left\{0_{(p+1)\text{x}(\text{p}+1)},1_{bxb}\right\}$ e os valores ajustados correspondentes $\hat{x}_\alpha=(\hat{x}_\alpha(t_1),...,\hat{x}_\alpha(t_n))'$ são dados por $\hat{\mathbf{x}}_\alpha=\mathbf{X}\hat{c}_\alpha$. Na prática, a qualidade de $\hat{x}_\alpha$ depende das escolhas de α e da quantidade e localização dos nós $\{\xi_1,...,\xi_k\}$. Por fim, o número b de funções bases em um sistema de base *Spline* é determinado pela seguinte relação:

$$Número\ de\ funções\ bases\ =\ ordem\ +\ número\ de\ nós\ internos$$

Em que nós internos são o conjunto de nós excluindo os dois nós extremos (primeiro e último nó). Como antes mencionado, existem uma família de sistemas de bases diferentes para a construção de funções *Splines*. Todavia, abordaremos a mais popular, ou seja, o sistema de base *B-Spline*. Assim como uma *Spline*, uma *B-Spline* é uma combinação linear de um conjunto de funções de base determinadas pelo número e a localização de nós, bem como pelo grau da curva.

Considere uma sequência não decrescente de nós, tal que $\min(x)=\xi_0\le\xi_1...\le\xi_k\le\xi_{k+1}=\max(x)$ de modo que haja k nós $\{\xi_1,...,\xi_k\}$. A determinação de uma *B-Spline* é baseada numa relação de recorrência proposta inicialmente por De Boor (1978) em que, o *j*-ésimo *B-Spline* de ordem 1 (constante por parte) é:

$$B_{j,1}(x)=\begin{cases}1, \xi_j\le x<\xi_{j+1}\\ 0,\ caso\,contrário\end{cases} \tag{8}$$

As *B-Splines* de ordem podem ser construídas pela seguinte relação de recorrência:

$$B_{j,m}(x) = \theta_{j,m} B_{j,(m-1)}(x) + [1 - \theta_{(j+1),m}(\mathrm{x})]\mathrm{B}_{(j+1),(m-1)}$$

em que

$$\theta_{j,m}(x) = \begin{cases} \dfrac{x - \xi_j}{\xi_{j+m-1} - \xi_j}, \; se \;\; \xi_{j+m-1} - \xi_j \neq 0 \\ 0, \; caso \; contrário \end{cases} \tag{9}$$

2.2 Modelo genoma contínuo

Apesar dos modelos funcionais como *Fourier* e *B-Spline* serem eficazes e constituírem apenas uma simplificação de polinômios ortogonais, Hu, Wang e Xu (2012) e Xu (2013) desenvolveram uma técnica alternativa de redução dimensional (o método genoma contínuo) em que o genoma é dividido em blocos em alto LD. Para isso, um modelo funcional foi assumido $y_i = \boldsymbol{\beta} + \int_0^L \mathbf{Z}_i(\lambda)\boldsymbol{\gamma}(\lambda)d\lambda + \varepsilon_i$ onde λ é a posição do SNP dentro do genoma (no intervalo de 0 a L), $\boldsymbol{\beta}$ o intercepto, $\boldsymbol{\gamma}(\lambda)$ é o efeito genético expresso como uma função desconhecida em função de λ, $Z_i(\lambda)$ é a variável indicadora de genótipo para o indivíduo i na posição λ e $\varepsilon_i \sim N(0, \sigma^2)$ é o erro *i* com uma variância desconhecida σ^2. Note que a função do efeito genético $\boldsymbol{\gamma}(\lambda)$ é desconhecida, e estimar essa função é o objetivo final da análise.

Dado que a função $\gamma(\lambda)$ é desconhecida, a integral para recuperar o valor genético genômico não é explicita, assim sendo, os autores mencionados acima, usaram a integração numérica para resolver esse problema. Em particular, eles buscaram descrever a região em desequilíbrio por meio da média do genótipo do SNP como medida de informação. Assim, dividiu-se o genoma em *m* intervalos (*bins*), sendo Δ_k o comprimento do *k*-ésimo *bin* (não necessariamente de tamanho uniformes). A aproximação numérica desse modelo é:

$$y_i \approx \beta + \sum_{k=1}^{m} \bar{Z}_i(\lambda_k)\bar{\gamma}(\lambda_k)\Delta_k + \varepsilon_i \qquad (10)$$

em que λ_k a posição do ponto médio do *k*-ésimo *bin* no genoma, $\bar{Z}_i(\lambda_k)$ o valor médio de Z_i para todos os marcadores cobertos pelo *k*-ésimo *bin*, $\bar{\gamma}(\lambda_k)$ o efeito médio de todos os QTL no referido *bin* e Δ_k é o comprimento deste *bin*.

É importante ressaltar que 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* maior. No entanto, Xu (2013) estendeu a abordagem de *bin* natural e abordou o conceito de *bins* artificiais. Assim sendo, o tamanho e a localização desses *bins* podem ser definidos a priori pelo pesquisador.

2.3 Modelo funcional Bayesiano

Uma proposta alternativa ao método genoma contínuo de (HU; WANG; XU, 2012; XU, 2013) foi abordada por (MOURA; PAMPLONA; BALESTRE, 2019). A ideia dos autores foi que, ao invés de tomar a média dos *bins* como informação, assumiu-se os marcadores como variáveis aleatórias dentro dos *bins* distribuídos uniformemente no genoma. Os resultados apresentados pelos autores, demonstraram que o modelo funcional bayesiano foi competitivo quando comparado com os métodos clássicos de regressão em cenários reais e simulados e, quando comparado ao método genoma contínuo (HU; WANG; XU, 2012). Em geral, o modelo funcional bayesiano também obteve maior eficiência computacional em relação aos modelos concorrentes.

Os resultados apresentados por Moura, Pamplona e Balestre (2019) indicam que o número ou tamanho dos *bins* pode influenciar nos resultados das análises, embora o perfil genômico entre os métodos não seja tão divergente. Uma maneira de determinar o número de *bins* seria utilizar o desequilíbrio de ligação (LD) como ponto de quebra e formar os chamados "*bins* naturais" (YU et al., 2011; XU, 2013) onde marcas contíguas que apresentam alto LD

formam um intervalo e possuem informações semelhantes. Dessa forma, a grande crítica dessa metodologia, foi que, na prática, não se sabe o tamanho ideal das janelas. Assim sendo, uma outra maneira de tratar esse método, seria assumir o número de *bins* também como variáveis aleatórias. É evidente que essa incerteza em relação ao número de *bins*, resulta em complicações na obtenção da amostra da distribuição conjunta *a posteriori*, dado que a dimensão do espaço do modelo pode variar (dimensionalidade do vetor de parâmetros não é fixa). No entanto, pode-se utilizar o método MCMC com Saltos Reversíveis (MCMC-SR), proposto por (GREEN, 1995).

3 CONCLUSÃO GERAL

De nossa revisão de literatura encontramos técnicas de representação do genoma em funções contínuas que se mostraram competitivas em relação aos modelos usuais de análise estatística para a GWS, tanto em termos de acurácia da predição quanto em termos de custo computacional. No que se segue, apresentamos dois trabalhos que tratam de alternativas de análise em GWS. No primeiro estuda-se duas alternativas de análise funcional (séries de *Fourier* e *B-Splines*). No segundo se estuda uma implementação de amostragem por saltos reversíveis (*Reversible Jump*) para a estimação do modelo genômico final. No final, é apresentado um apêndice com resultados utilizando uma outra estrutura de dados (uma população F_{10}), cujo objetivo foi verificar o comportamento do método proposto em cenários com baixo desequilíbrio de ligação (LD). A amostragem, portanto, é feita dividindo-se os cromossomos em *bins* de forma análoga aos modelos funcionais.

REFERÊNCIAS BIBLIOGRÁFICAS

BALESTRE, M.; SOUZA JÚNIOR, C. L. de. Bayesian reversible-jump for epistasis analysis in genomic studies. **BMC Genomics**, [S.l.], v. 17, n. 1012, 2016.

DE BOOR, C. **A pratical guide to spline.** Springer-Verlag, New York. 1978. 392p.

CALUS, M. P. L.; MEUWISSEN, T. H. E.; DE ROOS, A. P. W.; VEERKAMP, R. F. Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v.178, p.553–561, 2008.

CURRIE, I. D.; DURBAN, M. Flexible smoothing with P -splines : a unified approach. **Statistical Modelling**, v.2, n.4, p.333-349, 2002.

DESTA, Z. A.; ORTIZ, R. Genomic selection: Genome-wide prediction in plant improvement. **Trends Plant Sci.**, v.19, p.592–601, 2014.

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

HALL, P.; OPSOMER, J. D. Theory for penelized spline regression. **Biometrika**,v. 92, p. 105–118, 2005.

HU, Z.; WANG, Z.; XU, S. An infinitesimal model for quantitative trait genomic value prediction. **PLoS One**, v.7: p.1–14, 2012.

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

MONTESINOS-LOPEZ, A.; MONTESINOS-LOPEZ, O. A.; DE LOS CAMPOS, G.; CROSSA, J.; BURGUEÑO, J.; LUNA-VAZQUEZ, F. J. Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture. **Plant Methods**, v.14, n.46, p.1-17, 2018.

MORRIS, J.S. Functional regression. **Annu. Rev. Stat. Appl.** v.2, p.321–359, 2015.

MOURA, E. G.; PAMPLONA, A. K. A.; BALESTRE, M. Functional models in genome-wide selection. **Plos One**, v. 14, n. 10, p.1:27, 2019.

PAMPLONA, A. K. A. Iplines e modelo funcional em bins: abordagens integradas à seleção genômica. 2018. 142 p. Tese (Doutorado em Estatística e Experimentação Agropecuária)– Universidade Federal de Lavras, Lavras, 2018.

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

RAMSAY, J. O.; DALZELL, C. J. Some Tools for Functional Data Analysis. **J. R. Stat. Soc.** v.53, p.539–572, 1991.

RAMSAY, J.; HOOKER, G.; GRAVES, S. **Functional data analysis with R and MATLAB.** Springer Science & Business Media, New York, 2009. p. 1–19.

TEMPELMAN, R. J. Statistical and Computational Challenges in Whole Genome Prediction and Genome-Wide Association Analyses for Plant and Animal Breeding. **J. Agric. Biol. Environ. Stat.**, v.20, p.442–466, 2015.

WANG, C.L.; MA, P.P.; ZHANG, Z.; DING, X.D.; LIU, J.F.; FU, W.X.; WENG, Z.Q.; ZHANG, Q. Comparison of five methods for genomic breeding value estimation for the common dataset of the 15th QTL-MAS Workshop. **BMC Proc.**, v.6, n.2, 2012. doi: 10.1186/1753-6561-6-S2-S13.

WANG, J.; ZHOU, Z.; ZHANG, Z.; LI, H.; LIU, D.; ZHANG, Q.; BRADBURY, P.J.; BUCKLER, E.S.; ZHANG, Z. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. **Heredity**, v.121, p.648–662, 2018. https://doi.org/10.1038/s41437-018-0075-0

WANGA, X.; XU, Y.; HU, Z.; XU, C. Genomic selection methods for crop improvement: Current status and prospects. **Crop Journa**l, v.6, p.330–340, 2018.

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

ZHANG, Z.; LIU, J.; DING, X.; BIJMA, P.; DE KONING D. J. Best linear unbiased prediction of genomic breeding values using a trait-specific marker-derived relationship matrix (T Mailund, Ed.). **PLoS One**, v. 5, n. 9, p.12648−12656, 2010.

ZHANG, Z.; DING, X.; LIU, J.; DE KONING, D.J.; ZHANG, Q. Genomic selection for QTL-MAS data using a trait-specific relationship matrix. **BMC Proceedings**. v.5, n.3, 2011.

SEGUNDA PARTE

Nesta parte do trabalho, são apresentados dois artigos que abordam a ideia de modelo genoma contínuo, sendo o primeiro baseado em regressão funcional e o segundo trata-se de seleção automática de *bins*.

ARTIGO 1 Regressão funcional como análise estatística alternativa para seleção genômica

RESUMO

Com a disponibilidade de marcadores de alta densidade, a seleção genômica tornou-se um método promissor. A estimativa do valor genético genômico (*Genomic Breeding Value* – GBV) é o passo chave. Assim, dado que os polimorfismos do DNA são a fonte de variação do mérito genético, marcadores SNPs (*Single Nucleotide Polymorphism*) em desequilíbrio de ligação com locus de características quantitativas (*Quantitative Trait Loci* – QTL) são utilizados para predizer o GBV. Entre muitas abordagens, o método rr-BLUP (*Ridge Regression*) e os modelos Bayesianos, que tratam os efeitos de marcadores SNPs (*Single Nucleotide Polymorphism*) como aleatórios, são os mais comumente usados para estimar os valores de reprodução genômica. Contudo, esses métodos consideram uma distribuição discreta dos efeitos de marcadores ao longo do genoma, o que reflete em modelos superparametrizados. A distribuição dos efeitos pode ser modelada por uma função contínua por partes (havendo quebra entre os cromossomos), sendo esse processo denominado de genoma contínuo. Com isso, objetivou-se utilizar o modelo genoma contínuo e buscar uma curva, via sistemas de bases (*Fourier* e *B-Spline*), que represente a distribuição dos efeitos dos SNPs ao longo do genoma. Esse processo requer muito menos parâmetros do que modelos tradicionais de seleção genômica. Além disso, buscou-se verificar se tal abordagem apresenta vantagens preditivas com relação ao rr-BLUP e ao Bayes B. Para isso, utilizaram-se duas populações com diferentes estruturas de desequilíbrio de ligação: uma população F_2 composta de 300 indivíduos genotipados com 12150 marcadores, avaliada sob três cenários (oligogênico, poligênico I e poligênico II) e herdabilidades 0,2; 0,5 e 0,8; a segunda obtida por dados reais referente ao cruzamento de diferentes espécies de eucalipto, em que 610 indivíduos foram genotipados com 15104 marcadores Dart-Seq. Para os dados simulados, no cenário oligogênico, o Bayes B foi o mais acurado nas herdabilidades 0,5 e 0,8; todavia, foi o menos acurado na herdabilidade 0,2. Os métodos propostos com *Fourier* e *B-Spline* foram mais acurados que o rr-BLUP nas herdabilidades 0,2 e 0,5 e foram equivalentes ao rr-BLUP na herdabilidade 0,8. Para o cenário poligênico I, os métodos propostos foram equivalentes ao rr-BLUP e superaram o Bayes B nas herdabilidades 0,2 e 0,5; entretanto, Bayes B foi o mais acurado na herdabilidade 0,8. Para o cenário poligênico II, os métodos *P-Bspline* e *P-Fourier* foram mais acurados que rr-BLUP e Bayes B para herdabilidade 0,2 e igualmente acurado nas

herdabilidades 0,5 e 0,8. A seleção genômica usando modelos funcionais (*P-Fourier* e *P-Bspline*) foi comparada com rr-BLUP e Bayes B em uma gama de cenários, produzindo resultados similares ao do rr-BLUP e indicando que esses modelos podem lidar com um número ilimitado de marcadores em menor tempo computacional.

Palavras-chave: Modelos Funcionais, Modelo genoma contínuo.

1 INTRODUÇÃO

Os primeiros estudos sobre *Genome Wide Selection* (GWS) ou, simplesmente, *Genome Selection* (GS), foram realizados por Meuwissen, Hayes e Goodard (2001). Essa metodologia vem revolucionando a seleção de animais e plantas, culminando, portanto, em uma nova direção para seleção de caráter complexo de interesses econômicos. Essa abordagem utiliza alta densidade de marcadores SNP (*Single Nucleotide Polymorphism*) que cobrem todo o genoma a fim de maximizar as chances de que pelo menos um marcador esteja em desequilíbrio de ligação (LD) com um locus de características quantitativas (*Quantitative Trait Loci* – QTL) (GODDARD; HAYES, 2007; THAVAMANIKUMAR; DOLFERU; THUMMA, 2015). Dessa forma, utilizam-se regressões lineares sobre os valores fenotípicos por meio de artifícios de encolhimento (*shrinkage*) para predizer os valores genéticos genômicos (*Genomic Breeding Value* – GBV) de indivíduos previamente genotipados, incluindo aqueles que ainda não foram fenotipados.

Uma característica intrínseca que torna essas análises complexas é que, de modo geral, o número de marcadores é muito maior do que o número de indivíduos genotipados e fenotipados além de serem altamente correlacionados (GIANOLA et al., 2009; PÉREZ; DE LOS CAMPOS, 2014; MOURA; PAMPLONA; BALESTRE, 2019). Devido essa particularidade, métodos de regressão fixa simples que utilizam mínimos quadrados na seleção genômica são proibitivos. No entanto, métodos estatísticos que tratam os efeitos de marcadores como aleatórios, tais como regressão de cumeeira, a melhor predição linear não-viesada (rr-BLUP) e metodologias baseadas em inferência bayesiana, são usados na predição. Essa gama de métodos Bayesianos são adaptações do método proposto inicialmente por Meuwissen, Hayes e Goddard (2001) (diferem apenas por uma *priori*) e, por isso, foram apelidados por Gianola et al. (2009) de “alfabeto bayesiano”. Eles diferem principalmente nas suposições dos efeitos de

marcadores que contribuem para a variância total (THAVAMANIKUMAR; DOLFERU; THUMMA, 2015).

Em oposição a esses métodos tradicionais de seleção genômica, nos últimos tempos, têm-se proposto modelos de redução dimensional, como, por exemplo, o modelo genoma contínuo de Hu, Wang e Xu (2012) e Xu (2013), em que os autores dividiram o genoma em janelas (bloco em alto LD denominadas de *bins* naturais) no primeiro caso, e em janelas deslizantes (denominado de *bins* artificiais) no segundo caso. Nesse cenário, os autores tomaram o efeito médio das janelas como medida de informação para estimar a função sinal de efeito genético. Motivados por essa nova abordagem, Moura, Pamplona e Balestre (2019) propuseram uma metodologia alternativa em que se assumiram os marcadores como variável aleatória dentro dos *bins*, cuja função de sinal de expressão gênica é desconhecida. Com isso, ao invés de tomar a média do *bin* como informação, os autores utilizaram métodos bayesianos Monte Carlo Cadeia de Markov via algoritmo Metropolis-Hasting para realização do processo de amostragem por importância e integração numérica. Abordagem análoga foi utilizada no mapeamento de QTL em que a série espacial genômica era delimitada pelas marcas flanqueadoras em posições específicas no genoma (BALESTRE et al., 2012).

A ideia de divisão do genoma em *bins* vem sendo utilizada com êxito no mapeamento de QTL (HUANG et al., 2009; YU et al., 2011; XU, 2013; CHEN et al., 2014; SU et al., 2017). Em comparação com marcadores moleculares convencionais, como marcadores SNP únicos, os marcadores *bins* são um conjunto mais informativo e parcimonioso para uma determinada população (CHEN et al., 2014; SU et al., 2017). Os intervalos no genoma definidos pelo alto desequilíbrio de ligação foram denominados por Xu (2013) de *bins* naturais. Uma abordagem do tipo *Spline* foi adotada por Beissinger et al. (2015) para identificação de regiões de limiar considerando a endogamia do SNP.

A justificativa de se utilizar modelos funcionais em GWS é baseada no fato de que a expressão gênica pode ser pensada como uma série espacial em que os picos da expressão gênica representam uma função da posição no genoma. Dado que o objetivo é estimar a função sinal do efeito genético do modelo genoma contínuo, técnicas de ajuste de curvas polinomiais com baixa ordem podem ajustar bem em intervalos suficientemente pequenos, e até mesmo uma série de *Fourier* assumindo os intervalos como período de senoides e cossenoides.

Uma curva *Spline* é uma sequência de segmentos de curva que estão conectados juntos para formar uma única curva. Essas curvas são construídas dividindo o intervalo de observações (neste caso, o genoma) em subintervalos com limites em pontos chamados *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). A escolha do número ideal e as posições dos *knots* é uma tarefa complexa. *Knots* equidistantes podem ser usados, mas uma quantidade pequena de *knots* permite apenas um controle limitado sobre a suavidade e sobre o ajuste (EILERS; MARX, 1996). Assim sendo, realizou-se uma busca em grade para determinar o número ideal de *knots*, combinada com uma penalidade *Ridge Regression* (regressão de cumeeira), que também foi determinada via grade de busca mediante uma validação cruzada.

Diante do exposto, o presente trabalho tem por objetivo propor dois métodos alternativos de redução dimensional e averiguar a capacidade preditiva deles em relação a dois métodos tradicionais de seleção genômica.

2 MATERIAL E MÉTODOS

2.1 Dados simulados

Utilizando o software QGenes (JOEHANES; NELSON, 2008), foram simulados dez grupos de ligação de tamanho de 120 cM cada e distância média de 0,001 cM no genoma, em 300 indivíduos pertencentes a uma população F_2, totalizando 12150 marcadores SNP. Dentre os SNPs simulados, cinco 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, N(0;1). Para o cenário poligênico I, 20 marcadores foram assumidos como QTL e, finalmente, para o cenário poligênico II**,** 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 QTLs com seus l.

2.2 Dados reais

Os dados consistem numa população de genótipos de eucalipto da empresa Fibria S.A., derivados dos cruzamentos entre plantas de E. *grandis*, E. *urophylla,* E. *globulus* e E. *camaldulensis*. Foram avaliados 12 indivíduos por combinação de híbridos e, devido às perdas, foram obtidos no final 610 indivíduos.

As características mensuradas em três épocas diferentes foram: cap1- circunferência à altura do peito (em centímetros) na Época 1; lig1 - teor de lignina (em porcentagem) na Época 1; alt2 - altura de planta (em metros) na Época 2; cap3 - circunferência à altura do peito (em centímetros) 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 Dart-GBS, método proposto por Elshire et al. (2011). A genotipagem foi feita com 15104 marcadores Dart-Seq.

2.3 Estimativas de efeito de marcadores

Os dois métodos propostos alternativos envolvem duas etapas. Inicialmente, cria-se uma matriz **B** de bases utilizando a informação da posição de cada marcador como variável dependente. Há uma variabilidade de métodos de ajuste de curva que podem ser utilizados para gerar a matriz **B**. Aqui, utilizaram-se *B-Spline* e série de *Fourier*. Posteriormente, obtém-se uma matriz **W** como combinação linear da matriz **B** de bases com a matriz **Z** de estado genotípico, a qual será utilizada na análise.

O modelo estatístico tradicional para estimação de efeito de marcador pode ser escrito como:

$$y_i = \beta + \sum_{k=1}^{p} Z_{ik}\gamma_k + \varepsilon_i \qquad (1)$$

em que p são os marcadores, β é o intercepto, γ_k é o efeito do loco k, Z_{ik} é o estado genotípico para o indivíduo i no loco k e $\varepsilon_i \sim N(0,\sigma^2)$ é o erro com média zero e variância desconhecida σ^2. O estado do genótipo do marcador para o indivíduo i é definido como:

$$Z_{ik} = \begin{cases} 2, & \text{para homozigoto dominante} \\ 1, & \text{para heterozigoto} \\ 0, & \text{para homozigoto recessivo} \end{cases} \quad (2)$$

para SNP ou,

$$Z_{ik} = \begin{cases} 1, & \text{para homozigoto dominante/heterozigoto} \\ 0, & \text{para homozigoto recessivo} \end{cases} \quad (3)$$

para Dart-Seq.

Perceba que *p* é o número de efeitos a ser estimado no modelo e quando $p \to \infty$, o modelo em (1) é praticamente o modelo infinitesimal.

2.4 Modelo funcional

A ideia do modelo funcional genômico fundamenta-se na premissa que o sinal de expressão de um gene pode ser descrito por uma função espacial do genoma. Em outras palavras, $f(\lambda) = \gamma$ onde λ é a posição em pares de base no genoma e γ o sinal de expressão do gene dada a posição λ. Contudo $f(\lambda)$ e γ são desconhecidos e só podem ser estimados pela informação referente ao fenótipo de um indivíduo (*y*) e a posição dos marcadores do genoma λ. Mas, os domínios das funções $\gamma(\lambda)$ e y são diferentes ($f(\lambda) := \{\lambda_j \mid \lambda_j \in \Omega \equiv [1, L], \forall j\}$ 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 descreve o estado genotípico, que é condicional a λ. Ou seja, assumindo que $f(\lambda) = \gamma$ 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 $\hat{g}$, temos que $\mathbf{Z}(\lambda)f(\lambda) = \hat{g} = y + \varepsilon$. Nesse caso, têm-se a equivalência $\mathbf{Z}(\lambda)f(\lambda) \equiv \int_0^L \mathbf{Z}_i(\lambda)\gamma(\lambda)d\lambda = y + \varepsilon$.

Considere *n* indivíduos que estão genotipados em uma região genômica que tem *m* marcadores. Dentre esses *m* marcadores, assume-se que existem *p* marcadores localizados dentro de uma região com locais físicos ordenados $0 \le \lambda_1 < \cdots < \lambda_p = L$, 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:

$$y_i = \mu + \int_0^L Z_i(\lambda)\gamma(\lambda)d\lambda + \varepsilon_i \qquad (4)$$

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 convencional dado por $y_i = \beta + \sum_{j=1}^{p} Z_{ij}\gamma_j + \varepsilon_i$ pode ser descrito na sua forma funcional por intermédio de (4) como:

$$y_i = \mu + \sum_{c=1}^{C}\int_0^L Z_{ic}(\lambda)\gamma(\lambda)d\lambda + \varepsilon_i \qquad (5)$$

em que o somatório descreve a descontinuidade da função ao longo dos cromossomos. A integral em (4) resulta no valor genético genômico para o indivíduo *i*. Este é o modelo infinitesimal (modelo genoma contínuo) proposto por Hu, Wang e Xu (2012). Dado que $\gamma(\lambda)$ é desconhecida, não existe uma expressão explícita para resolvê-la. Hu, Wang e Xu (2012)

dividiram o genoma em blocos de alto LD que denominaram de *bins* e tomaram seus efeitos médios como informação. No presente estudo, será utilizado um sistema de bases (*B-Spline* e *Fourier*) que, combinado com a matriz de estado genotípico ***Z***, pode estimar $\hat{\gamma}(\lambda) = f(\lambda)$ e, consequentemente, resolver a integral em (04) para recuperar o valor genético genômico.

Para isto, $\gamma(\lambda)$ pode ser aproximado pela função série de Fourier $\hat{\gamma}(\lambda) \approx \sum_{j=0}^{b-1} \hat{\phi}_j \mathrm{F}_j(\lambda)$ ou por *Spline* $\hat{\gamma}(\lambda) = \sum_{t=1}^{k+q} B_{t,q}(\lambda)\hat{\phi}_t(\lambda)$, em que $\mathrm{F}_j(\lambda)$ são as bases Fourier e $\hat{\phi}_j$ é a solução da *j*-ésima base Fourier, $B_{t,q}(\lambda)$ são as bases *Splines* de *m*-ésimo grau e $\hat{\phi}_t(\lambda)$ a solução polinomial no *k*-ésimo *knot*. Os detalhes da construção dos modelos *Fourier* e *Spline* são dados a seguir.

2.4.1 Séries de *Fourier*

O sistema de bases de *Fourier* é mais utilizado em dados com periodicidade. No entanto, segundo Ramsay e Silverman (2005), são muito úteis por sua agilidade computacional e fácil adaptação a quaisquer espécies de dados. Tais sistemas decompõem as funções bases em uma combinação linear de senoide e cossenoide. A série é determinada pelo número de funções bases *K* e o período *T*. Suponha *n* valores de medições observados, $\{\lambda_i, y_i\}_{i=1}^{n}$, satisfazendo o modelo $y_i = x(\lambda_i) + \varepsilon_i$, em que $x(\lambda_i)$ é uma função de regressão desconhecida e os ε_i são erros independentes com variância constante σ^2. Dessa forma, $x(\lambda_i)$ pode ser modelada por um sistema de bases *Fourier*, truncado ao espaço de λ, da seguinte forma:

$x(\lambda) = \phi_0 + \phi_1 sen(w\lambda) + \phi_2 \cos(w\lambda) + \phi_3 sen(2w\lambda) + \phi_4 \cos(2w\lambda) + \ldots$, ou ainda

$$x(\lambda) = c_0 + \sum_{j=1}^{m} \left[c_{2j-1} sen(jwt) + c_{2j} \cos(jwt) \right] \qquad (6)$$

em que $l = 1 + 2m$ é o número total de bases. A constante ω está relacionada ao período *T* pela relação ω = 2π / T , ou seja, as primeiras parcelas de seno e cosseno oscilaram uma vez durante o domínio de $x(\lambda)$, as parcelas associadas a ϕ_3 e ϕ_4 oscilaram duas vezes durante o domínio, e, assim por diante, e *T* pode ser definido como próprio período (amplitude) do espaço da posição (max(λ) – min(λ)), conforme sugere Ramsay, Hooker e Graves (2009) e, além disso, note que para garantir a ortogonalidade, o número total de bases *Fourier l* é sempre ímpar (intercepto e sucessivos pares seno/cosseno), de modo que para cada senoide que entrar no modelo, um cossenoide também terá que fazer parte. Com esta configuração, a estimativa de $x(\lambda_i)$ pode ser alcançada por meio da estimativa dos coeficientes $\mathbf{\Phi} = (\phi_0, \phi_1, ..., \phi_K)^t$ através da matriz de bases *Fourier* **F**, em que:

$$\mathbf{F} = \begin{bmatrix} 1 & sen(w\lambda_1) & \cos(w\lambda_1) & sen(2w\lambda_1) & \cos(2w\lambda_1) & \dots & sen\left(\frac{l-1}{2} w\lambda_1\right) & \cos\left(\frac{l-1}{2} w\lambda_1\right) \\ 1 & sen(w\lambda_2) & \cos(w\lambda_2) & sen(2w\lambda_2) & \cos(2w\lambda_2) & \dots & sen\left(\frac{l-1}{2} w\lambda_2\right) & \cos\left(\frac{l-1}{2} w\lambda_2\right) \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ 1 & sen(w\lambda_n) & \cos(w\lambda_n) & sen(2w\lambda_n) & \cos(2w\lambda_n) & \dots & sen\left(\frac{l-1}{2} w\lambda_n\right) & \cos\left(\frac{l-1}{2} w\lambda_n\right) \end{bmatrix}$$

Logo, para nosso exemplo fictício, temos $\hat{y} \approx \mathbf{F}\hat{\mathbf{\Phi}}$.

Assim, considerando a Série de *Fourier* com ***l*** bases (sempre ímpar), é possível escrever a função *γ* como:

$$\gamma(\lambda) \approx x(\lambda) = \sum_{j=0}^{l-1} \phi_j \mathrm{F}_j(\lambda) = \mathbf{F}\mathbf{\Phi}, \qquad \text{com} \quad \lambda \in [0, L] \tag{7}$$

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

$$\hat{\gamma}(\lambda) \approx \sum_{j=0}^{l-1} \hat{\phi}_j \mathrm{F}_j(\lambda) = \mathbf{F}\hat{\mathbf{\Phi}} \tag{8}$$

Substituindo (8) em (5) e considerando *C* cromossomos, o modelo pode ser reescrito como:

$$y_i = \beta + \sum_{c=1}^{C} \sum_{j=1}^{l-1} Z_{jc}(\lambda)\phi_j(\lambda)\mathrm{F}_j(\lambda) \qquad (9)$$

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

Note que se adotamos *l* bases por cromossomos, teremos com *C* cromossomos *C*l* bases no genoma e, consequentemente, teremos *C*l* parâmetros a serem estimados ao todo. Logo, o modelo básico, em formato matricial, é:

$$\begin{aligned} \mathbf{y}_{n\mathrm{x}1} &= \boldsymbol{\mu}_{n\mathrm{x}1} + \mathbf{Z}_{n\mathrm{x}p}\boldsymbol{\gamma}_{p\mathrm{x}1} + \boldsymbol{\varepsilon}_{n\mathrm{x}1} = \\ &= \boldsymbol{\mu}_{n\mathrm{x}1} + \mathbf{Z}_{n\mathrm{x}p}\mathbf{F}_{p\mathrm{x}b}\boldsymbol{\Phi}_{b\mathrm{x}1} + \boldsymbol{\varepsilon}_{n\mathrm{x}1} \end{aligned} \qquad (10)$$

em que $\boldsymbol{\mu}$ é o vetor da média geral, **Z** é a matriz de estado genotípico dos marcadores (por exemplo, {aa, Aa, AA} = {0,1,2} para SNPs), **F** é a matriz de bases *Fourier* gerada através da informação da posição λ dos marcadores, $\boldsymbol{\Phi}$ é o vetor de coeficientes a serem estimados e $\boldsymbol{\varepsilon}$ é o vetor dos erros.

Dado que as matrizes **Z** e **F** são conhecidas, podemos reescrevê-las por **ZF** = **W** e o modelo a ser trabalhado é:

$$\mathbf{y}_{n\mathrm{x}1} = \boldsymbol{\mu}_{n\mathrm{x}1} + \mathbf{W}_{n\mathrm{x}b}\boldsymbol{\Phi}_{b\mathrm{x}1} + \boldsymbol{\varepsilon}_{n\mathrm{x}1}. \qquad (11)$$

Logo, podemos assumir que $\boldsymbol{\mu} \sim N(\mathbf{W}_{n\mathrm{x}1}\boldsymbol{\Phi}_{n\mathrm{x}b\ b\mathrm{x}1}, \sigma^2)$. Dado que nesse trabalho foi realizado uma busca em grade pelo número de bases que maximiza a acurácia, embora houve uma vasta redução dimensional, gerou um número de bases maior do que o número de indivíduos fenotipados (***l*** >>***n***). Assim, podemos obter as soluções das bases *Fourier* (**F**) pelo método de mínimos quadrados penalizado (*Ridge Regression*). Logo, a solução para o problema de *Ridge Regression* é:

$$\hat{\boldsymbol{\Phi}}^{ridge} = (\mathbf{W}'\mathbf{W} + \alpha\mathbf{I})^{-1}\mathbf{W}\mathbf{y} \quad (12)$$

em que α é um escalar maior ou igual a zero (α é frequentemente chamado de *Penalty*) e $\mathbf{I}_l$ é uma matiz identidade de ordem ***l***. Quando $\alpha \to 0$, o $\hat{\boldsymbol{\Phi}}^{ridge}$ tende para estimador dos mínimos quadrados $\hat{\boldsymbol{\Phi}}^{ols}$, e quando $\alpha \to \infty$, $\hat{\boldsymbol{\Phi}}^{ridge} \to 0$.

A predição do valor genético genômico é dada pelo preditor de mínimos quadrados penalizados $\hat{g} = \mathbf{Z}\hat{\boldsymbol{\gamma}} \approx \mathbf{W}\hat{\boldsymbol{\Phi}}^{ridge}$.

2.4.2 *B-Spline*

Uma *Spline* é uma combinação linear de um conjunto de funções base que são determinadas pelo número e localização de pontos de quebra denominados nós (*knots*), bem como pelo grau da curva a ser ajustada aos dados em cada subintervalo. A quantidade dos nós, 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). Existe uma família de funções *Splines*, e, nesse estudo, abordaremos o sistema de bases *B-Spline*, proposto por De Boor (1978). São mais versáteis que a série de *Fourier* devido sua característica intrínseca de suavização local (nós), contemplando, assim, diferentes estruturas de dados. O termo base se refere a uma transformação realizada na variável dependente (λ) antes da etapa de estimação de parâmetros.

No contexto de seleção genômica serão assumidas as mesmas pressuposições da seção anterior. *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 *d* = *q* – 1. Seja uma sequência não decrescente de *knots* (números reais) tal que $\xi_0 \leq \xi_1 \leq \cdots \leq \xi_{k+1}$, em um intervalo [min(x), max(x)], de modo que haja *k knots* $\{\xi_1, \ldots, \xi_k\}$ internos. A determinação de uma *B-Spline* é baseada numa relação de recorrência proposta

inicialmente por De Boor (1978), em que o *t*-ésimo *B-Spline* de ordem 1 (constante por parte) é:

$$B_{t,1}(x) = \begin{cases} 1, \xi_t \leq x < \xi_{t+1} \\ 0, \ caso\, contrário \end{cases}. \tag{13}$$

As *B-Splines* de ordem superiores podem ser construídas pela seguinte relação de recorrência:

$$B_{t,q}(x) = \theta q B_{t,(q-1)}(x) + [1 - \theta_{(t+1),q}(x)] B_{(t+1),(q-1)}(x). \tag{14}$$

em que

$$\theta_{t,q}(x) = \begin{cases} \dfrac{x - \xi_t}{\xi_{t+q-1} - \xi_t}, \ se \ \ \xi_{t+q-1} - \xi_j \neq 0 \\ 0, \ caso\ contrário \end{cases}. \tag{15}$$

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

$$\gamma(\lambda) = \sum_{t=1}^{k+q} B_{t,q}(\lambda)\phi_t(\lambda) = \mathbf{B\Phi}, \qquad \text{com} \quad \lambda \in [0, L]. \tag{16}$$

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

$$\hat{\gamma}(\lambda) = \sum_{t=1}^{k+q} B_{t,q}(\lambda)\hat{\phi}_t(\lambda) \tag{17}$$

Substituindo (17) em (5) e considerando *C* cromossomos, o modelo pode ser reescrito como:

$$y_i = \beta + \sum_{c=1}^{C} \sum_{t=1}^{k+q} Z_{ic}(\lambda) B_{t,q}(\lambda) \phi_t(\lambda) \tag{18}$$

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

Um polinômio de grau $d = 2$ (ou seja, ordem $q = 3$) foi escolhido para ser ajustado em cada intervalo de *knots*, pois supõe-se 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:

$$\begin{aligned} \mathbf{y}_{n\text{x}1} &= \boldsymbol{\mu}_{n\text{x}1} + \mathbf{Z}_{n\text{x}p} \boldsymbol{\gamma}_{p\text{x}1} + \boldsymbol{\varepsilon}_{n\text{x}1} = \\ &= \boldsymbol{\mu}_{n\text{x}1} + \mathbf{Z}_{n\text{x}p} \mathbf{B}_{p\text{x}b} \boldsymbol{\Phi}_{b\text{x}1} + \boldsymbol{\varepsilon}_{n\text{x}1} \end{aligned} \tag{19}$$

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*, $\boldsymbol{\Phi}$ é o vetor de coeficientes a serem estimados e $\boldsymbol{\varepsilon}$ é o vetor dos erros.

Para exemplificar, considere um cromossomo com *p* SNPs em *n* indivíduos, temos que o modelo *B-Spline* de grau 2 com *k knot*s é descrito da seguinte forma:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} \beta \\ \beta \\ \vdots \\ \beta \\ \beta \end{bmatrix} + \begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ z_{(n-1)1} & z_{(n-1)2} & \cdots & z_{(n-1)p} \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} \begin{bmatrix} B_{1,3}(\lambda_1) & B_{2,3}(\lambda_1) & \cdots & B_{k+q,3}(\lambda_1) \\ B_{1,3}(\lambda_2) & B_{2,3}(\lambda_2) & \cdots & B_{k+q,3}(\lambda_2) \\ \vdots & \vdots & \vdots & \vdots \\ B_{1,3}(\lambda_{p-1}) & B_{2,3}(\lambda_{p-1}) & \cdots & B_{k+q,3}(\lambda_{p-1}) \\ B_{1,3}(\lambda_p) & B_{2,3}(\lambda_p) & \cdots & B_{k+q,3}(\lambda_p) \end{bmatrix} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_{k+q-1} \\ \phi_{k+q} \end{bmatrix} + \boldsymbol{\varepsilon}_{n\text{x}1} = \quad (20)$$

$$\mathbf{y}_{n\text{x}1} = \boldsymbol{\beta}_{n\text{x}1} + \mathbf{Z}_{nxp}\mathbf{B}_{pxb}\boldsymbol{\Phi}_{b\text{x}1} + \boldsymbol{\varepsilon}_{n\text{x}1}$$

em que z_{ij} é o estado do genótipo do *j*-ésimo SNP no *i*-ésimo indivíduo mostrado em (02) e (03), λ_j é a distância do SNP no genoma, dado em pb ou cM, e ϕ a solução polinomial para o número de *knots* específicos. Note que o número de bases a serem estimadas independe do número *p* de marcadores dentro do *knot*. A soluções das bases *B-Spline*, assim como a predição do valor genético genômico ($\hat{g}$) foram realizadas de maneira análoga a seção anterior e serão suprimidas aqui. É importante ressaltar que não utilizamos a teoria de *Splines* penalizados e sim uma penalização de *Ridge Regression* para os dois métodos propostos e, por isso, denominamos de *P-Bspline* e *P-Fourier* respectivamente.

2.5 Implementação da análise

Para os modelos propostos, a matriz **F** de bases *Fourier* foi obtida utilizando a função ***fourier*** disponível pela biblioteca *fda* (RAMSAY; WICKHAM; GRAVES, 2015) e a matriz **B** de base *B-Spline* foi obtida utilizando a função ***bSpline*** disponível pela biblioteca *splines2* (WANG; YAN, 2017), ambas do *software* R (R CORE TEAM, 2018); para a determinação do parâmetro de penalização desses métodos foi utilizada a função ***lm.ridge*** disponível pela biblioteca *MASS* (RIPLEY et al., 2019). Para os modelos concorrentes, os dados foram analisados utilizando o modelo Bayes B, por meio da função ***BGLR*** contida no pacote *BGLR* (PÉREZ; DE LOS CAMPOS, 2014), e os modelos GBLUP e rr-BLUP, com a função ***mixed.solve*** do pacote *rr-BLUP* (ENDELMAN, 2011).

2.6 Acurácia preditiva

Como critérios para avaliar a acurácia preditiva dos métodos, primeiramente, foi realizada uma busca em grade da esquerda para direita a fim de determinar a quantidade ótima de *knots* para o método *P-Bspline* e o número ideal de bases para o método *P-Fourier*. Como critério de informação, utilizou-se o coeficiente de correlação de Pearson (r) entre o valor genético genômico predito pelos métodos e o observado. Esse processo de seleção de *knot* igualmente espaçado via grade foi denominado de algoritmo de busca completa por Ruppert, Wand e Carrol (2003) e Montoya, Ulloa e Miller (2014).

Na análise de dados reais, a quantidade dos *knots* para o método *P-Bspline* e o número de bases para o método *P-Fourier* foram determinados utilizando a validação cruzada 10-*fold*. Este procedimento foi repetido para uma grade de *knots* igualmente espaçados para o método *P-Bspline*, e uma grade de número de bases para o método *P-Fourier*. A fim de determinar o número ótimo de knots e um número ótimo de bases *Fourier*. O objetivo desse procedimento é avaliar situações em que a formação de *knots* naturais, dado pelo desequilíbrio de ligação, não é possível. Dessa forma, trata-se de um procedimento de duas etapas em que dados os parâmetros estimados que otimizam, é realizada uma nova análise para comparação com os métodos concorrentes.

2.7 Escolhendo o parâmetro de suavização

O papel do parâmetro de suavização em análise de dados funcionais (penalização) é controlar a suavidade da curva ajustada. Para calcular o valor ideal de tal parâmetro, o critério de seleção considerado foi o de validação cruzada generalizada (GCV). O método GCV é computacionalmente simples, muito bem usado na literatura em regressão *Splines* (CRAVEN; WAHBA, 1979; CAO et al., 2010) e consiste em selecionar α de modo que minimize

$$GCV(\alpha) = \frac{SSE}{(1 - df_{\alpha} / n)^2} \tag{21}$$

em que df_α são os "graus de liberdade" correspondentes ao parâmetro de suavização α e é definido como traço da matriz chapéu $\mathbf{H}_\alpha$, ou matriz suavizadora no contexto de regressão não paramétrica, em que $\mathbf{H}_\alpha = \mathbf{W}(\mathbf{W}^T\mathbf{W} + \alpha\mathbf{I})^{-1}\mathbf{W}^T$. Assim, $df_\alpha = tr(\mathbf{H}_\alpha)$ e a soma dos quadrados dos resíduos dada por $SSE = \sum_{i=1}^{n}(g_i - \hat{g}_i)^2 = \sum_{i=1}^{n}(y_i - \mathbf{W}\hat{\mathbf{\Phi}}^{ridge})^2$. Note que a matriz chapéu $\mathbf{H}_\alpha$ é quadrada, simétrica e de ordem *n*, sendo uma função de α. Assim, calcula-se o GCV para uma grade de valores de α e escolhe-se o minimizador desse critério sobre a grade, como ótimo.

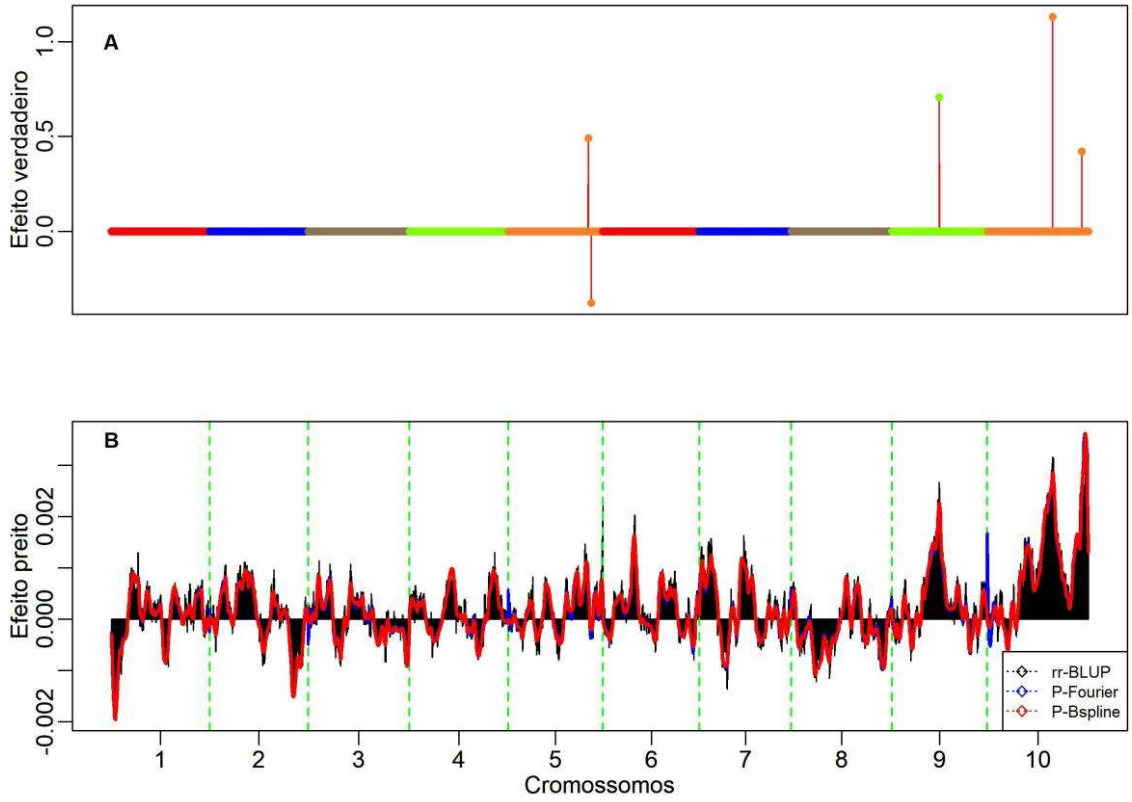
3 RESULTADOS

Nas primeiras três sessões apresenta-se os resultados para diferentes cenários de simulação. Na seção 3.4 apresenta-se resultados para um conjunto de dados reais.

3.1 Cenário simulados I: Modelo oligogênico

Na Figura 1 encontram-se os efeitos dos QTLs simulados (painel A), e os efeitos preditos pelos métodos rr-BLUP (em preto), *P-Fourier* (em azul) e *P-Bspline* (em vermelho) encontram-se no painel B. Para suavizar a curva *P-Fourier* foram adotadas 31 bases por cromossomos, perfazendo, portanto, um total de $l = 310$ bases *Fourier*. Para suavizar a curva *P-Bspline* foram adotados $k = 71$ *knots* por cromossomos; logo, o número de bases *B-Spline* por cromossomo é $k+d+1 = 71+2+1 = 74$, em que d é o grau do polinômio suavizado, perfazendo, portanto, um total de $b = 740$ bases. Nesse contexto, o número de bases reflete o número de parâmetros a serem estimados. O número ideal de *knot*s foi determinado via validação cruzada dentro de cada ponto da grade.

Figura 1 - Efeitos simulados verdadeiros de QTL (painel A) ao longo do genoma para a herdabilidades 0,5 e estimados (painel B) a partir dos métodos, respectivamente, rr-BLUP, *P-Fourier* e *P-Bspline*. Pontos coloridos representam os cinco verdadeiros QTL distribuídos em 12150 SNP ao longo de dez grupos de ligação. Linhas tracejadas em verde separam os cromossomos.



Fonte: Do autor (2021).

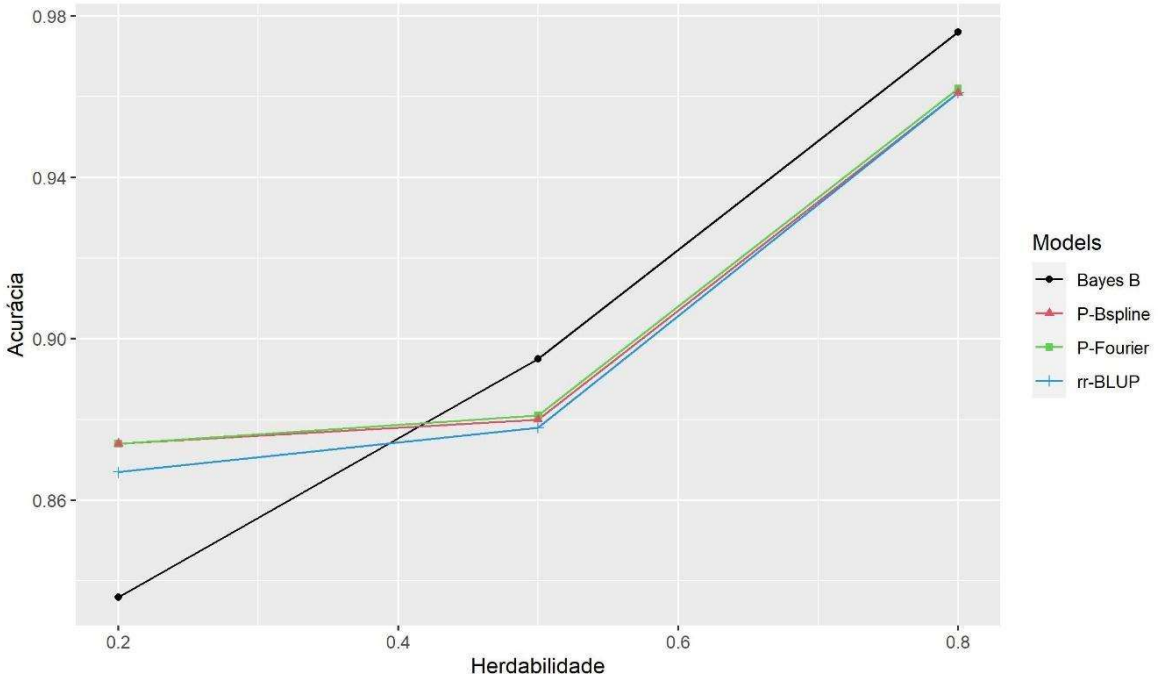
Observa-se que enquanto os efeitos QTL simulados variaram de -0,37719 a 1,13057 (Figura 1A), os efeitos de marcadores estimados variaram de -0.00164 a 0.00323 para rr-BLUP (Figura 1B**,** em preto), de -0,00195 a 0,00361 para *P-Bspline* (Figura 1B**,** em vermelho) e para *P-Fourier* foi de -0,00182 a 0,00362 (Figura 1B**,** em azul). A maioria dos segmentos contendo grandes QTLs foram mapeados por todos os métodos. Os perfis de efeito de SNP dos dois métodos propostos e o rr-BLUP são bastante semelhantes (Figura 1B**)**, com ligeira superioridade em resolução aos métodos *P-Bspline* e *P-Fourier*. Todos os métodos mostram um grande pico no cromossomo 9, dois grandes picos no cromossomo 10, região em que há de fato um QTL simulado. Todavia, um pico expressivo de efeito negativo foi capitado no cromossomo 1 por

todos os métodos, sendo que não há QTL simulado nesse cromossomo, ou seja, um falso positivo.

Para os demais níveis de herdabilidades, os métodos alternativos propostos obtiveram ligeiramente melhor resolução do que o rr-BLUP (resultado não mostrado). Conforme esperado, à medida que se aumentou a herdabilidade, a resolução dos métodos também aumentou.

Na Figura 2 é apresentado o coeficiente de correlação de Pearson (*r*), para diferentes métodos ao variar a herdabilidade, a fim de avaliar a capacidade preditiva dos modelos em estudo. Nota-se o aumento na capacidade preditiva de todos os modelos avaliados quando se aumenta a herdabilidade. Dentre os modelos em estudo, com exceção do Bayes B, que foi o mais acurado nas herdabilidades 0,5 e 0,8, os dois métodos propostos foram mais acurados que os demais.

Figura 2 - Correlação (*r*) dos métodos, respectivamente, Bayes B (em preto), *P-Bspline* (em vermelho), *P-Fourier* (em verde) e rr-BLUP (em azul) para as três herdabilidades.



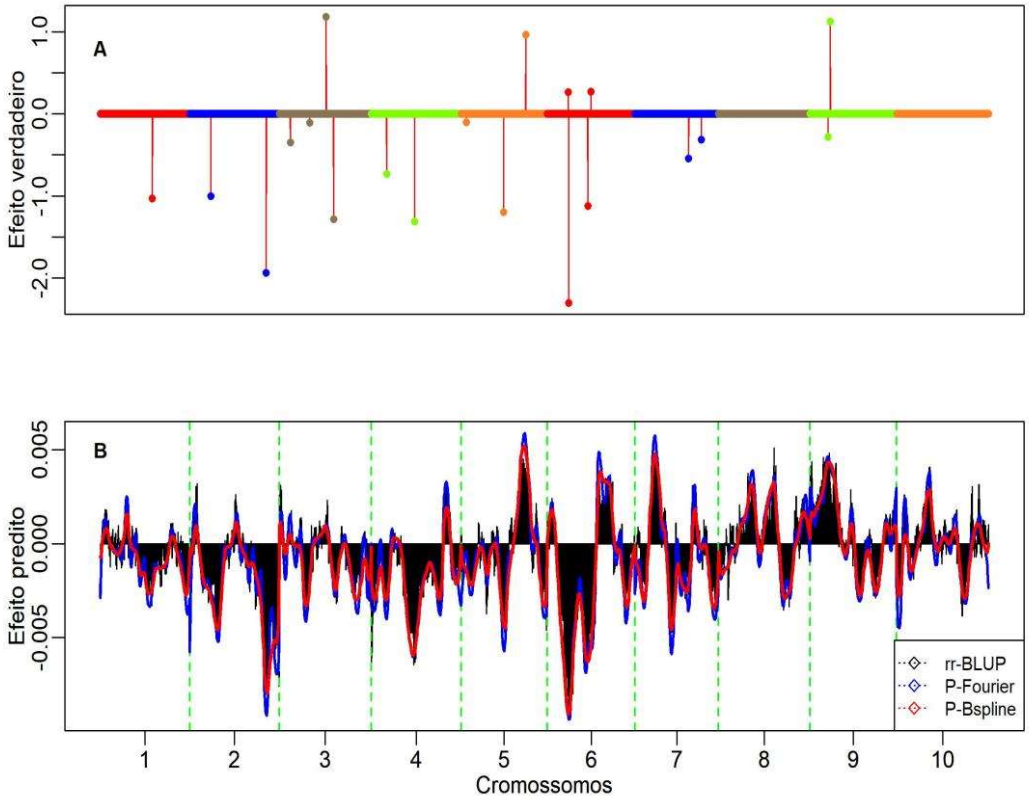
Fonte: Do autor (2021).

3.2 Cenário simulado II: Modelo poligênico I

Os efeitos simulados (verdadeiros) de QTL (Painel A) e os efeitos estimados de marcadores a partir dos métodos propostos e do rr-BLUP (Painel B), para o presente cenário, são apresentados na Figura 3. Para suavizar a curva *P-Fourier*, foram adotadas 23 bases por cromossomos, perfazendo, portanto, um total de $l = 230$ bases *Fourier*. Para ajustar a curva *P-Bspline*, foram adotados $k = 25$ *knots* por cromossomos; logo, o número de bases *B-Spline* por cromossomo é $k+d+1 = 25+2+1 = 28$, perfazendo, portanto, um total de $b = 280$ bases (número de parâmetros a serem estimados). Esse número de *knots* também foi escolhido mediante uma busca em grade de validação cruzada. Nessa arquitetura genética em que há quatro vezes mais QTL simulados do que o cenário anterior, o número de bases *Fourier* e *Bspline* diminuíram em relação a cenário anterior.

Enquanto os efeitos QTL simulados variaram de -2,3031 a 1,1821 (Figura 3A), os efeitos de marcadores estimados variaram de -0,00888 a 0,00511 para rr-BLUP, -0,00908 a 0,00521 para *P-Bspline* e, finalmente, -0,00935 a 0,00588 para *P-Fourier* (Figura 3B). A maioria dos segmentos contendo QTL grande foram mapeados por todos os métodos estudados. Contudo, o perfil de efeito de SNP dos dois métodos propostos (*P-Bspline* e *P-Fourier*) e o rr-BLUP são bastante semelhantes (Figura 3B), com ligeira melhoria em resolução aos métodos propostos.

Figura 3 - Efeitos simulados verdadeiros de QTL (painel A) ao longo do genoma para a herdabilidades 0,5 e estimados (painel B) a partir dos métodos, respectivamente, rr-BLUP, *P-Fourier* e *P-Bspline*. Pontos coloridos representam os 20 verdadeiros QTL distribuídos em 12150 SNP ao longo de dez grupos de ligação. Linhas tracejadas em verde separam os cromossomos.

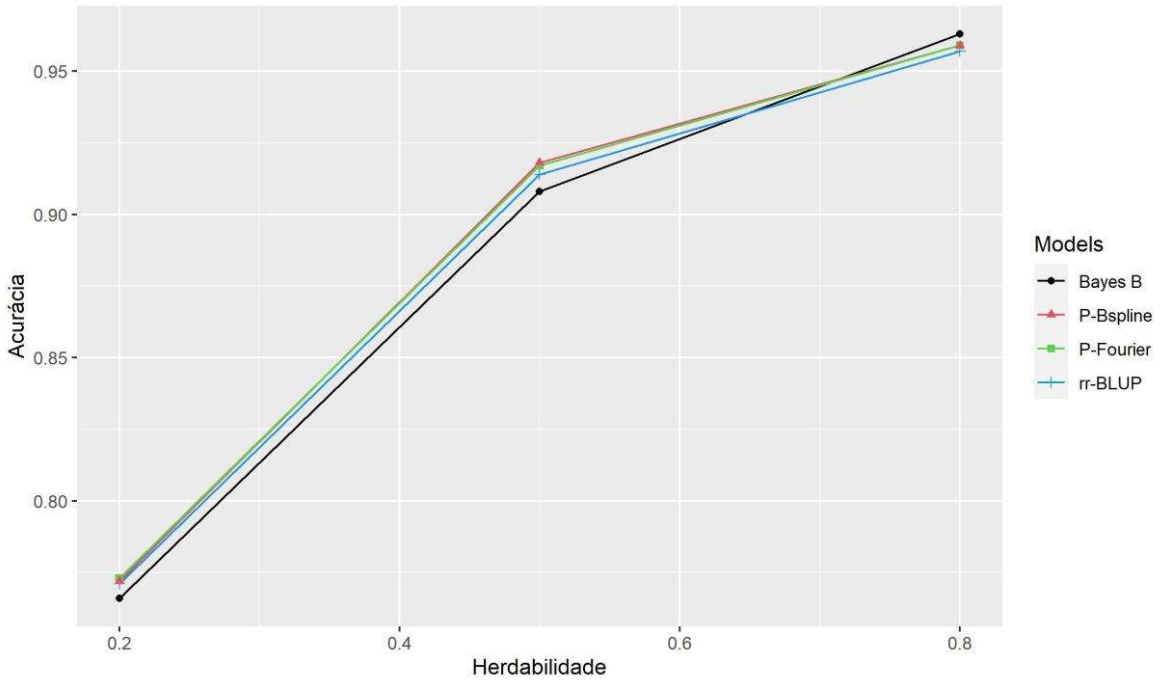


Fonte: Do autor (2021).

Na Figura 4 estão apresentadas as correlações de Pearson (*r*) entre os valores genômicos simulados e preditos pelos diferentes métodos ao variar a herdabilidade, a fim de avaliar suas capacidades preditivas. Mais uma vez os métodos propostos foram mais acurados que os demais (com exceção da herdabilidade 0,8, em que Bayes B foi o mais acurado). Com o aumento da herdabilidade, todos os modelos aumentaram consideravelmente a capacidade preditiva

.

Figura 4 - Correlação (*r*) dos métodos respectivamente, Bayes B (em preto), *P-Bspline* (em vermelho), *P-Fourier* (em verde) e rr-BLUP (em azul) para as três herdabilidades.



Fonte: Do autor (2021).

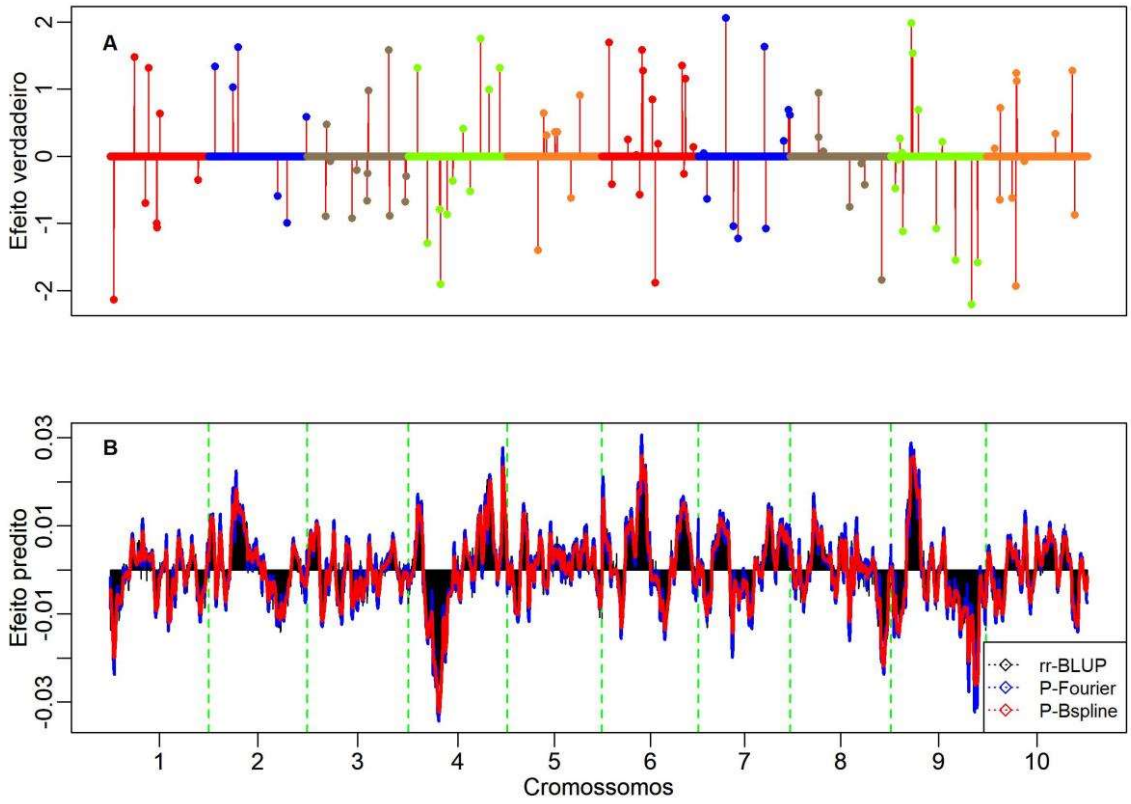
3.3 Cenário simulado III: Modelo poligênico II

A Figura 5 engloba os efeitos dos QTL simulados (painel A) e os perfis de efeitos de SNP preditos por rr-BLUP (em preto), *P-Fourier* (em azul) e *P-Bspline* (em vermelho) encontram-se no painel B. Para suavizar a curva *P-Fourier* foram adotadas 95 bases por cromossomos, perfazendo, portanto, um total de $l = 950$ bases *Fourier*. Para ajustar a curva *P-Bspline*, foram adotados $k = 100$ *knots* por cromossomos; logo, o número de bases *Bspline* por cromossomo é $k+d+1 = 100+2+1 = 103$, perfazendo, portanto, um total de $b = 1030$ bases (número de parâmetros a serem estimados).

Em comparação com os efeitos verdadeiros (painel A), enquanto os efeitos QTL simulados variaram de -2,2075 a 2,0645 (Figura 5A), os efeitos de marcadores estimados variaram de -0,02662 a 0,02317 para rr-BLUP (Figura 5B), de -0,03236 a 0,02599 para *P-*

Bspline e, finalmente, -0,03071 a 0,03048 para *P-Fourier* (Figura 5B). Observa-se que os QTL simulados com grandes efeitos foram mapeados pelos métodos supracitados, isto é, em comparação com os efeitos simulados, os efeitos preditos pelos métodos mostraram padrão semelhante, mas em média foram viciados para baixo. Além disso, os perfis de efeito de SNP dos dois métodos propostos e o rr-BLUP são bastante semelhantes (Figura 5B), com ligeira superioridade em resolução aos métodos *P-Bspline* e *P-Fourier*.

Figura 5 - Efeitos simulados verdadeiros de QTL (painel A) ao longo do genoma para a herdabilidades 0,5 e estimados (painel B) a partir dos métodos, respectivamente, rr-BLUP, *P-Fourier* e *P-Bspline*. Pontos coloridos representam os 100 verdadeiros QTL distribuídos em 12150 SNP ao longo de dez grupos de ligação. Linhas tracejadas em verde separam os cromossomos.

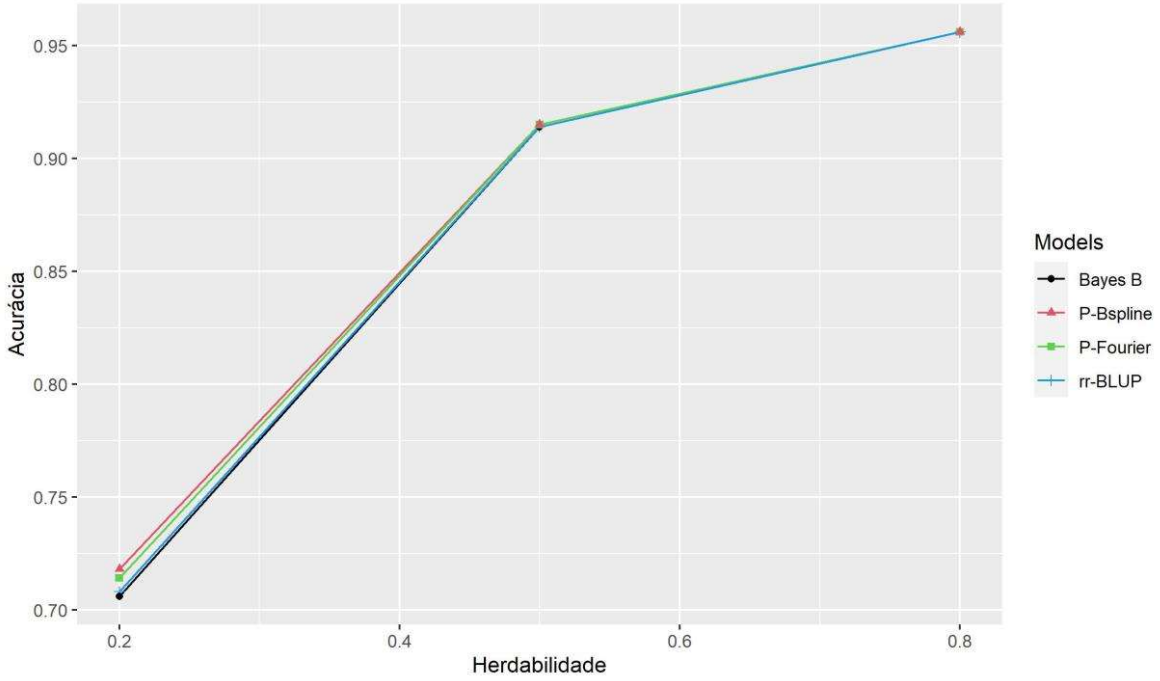


Fonte: Do autor (2021).

Na Figura 6 é apresentado o coeficiente de correlação de Pearson (*r*) para avaliar a capacidade preditiva dos modelos em estudo. No presente cenário, os métodos *P-Bspline* e *P-Fourier* foram mais acurados do que rr-BLUP e Bayes B para herdabilidades 0,2, e igualmente

acurado nas herdabilidades 0,5 e 0,8. Observa-se na Figura 6, o aumento na capacidade preditiva de todos os métodos avaliados quando se aumenta a herdabilidade.

Figura 6 - Correlação (*r*) dos métodos respectivamente, Bayes B (em preto), *P-Bspline* (em vermelho), *P-Fourier* (em verde) e rr-BLUP (em azul) para as três herdabilidades.



Fonte: Do autor (2021).

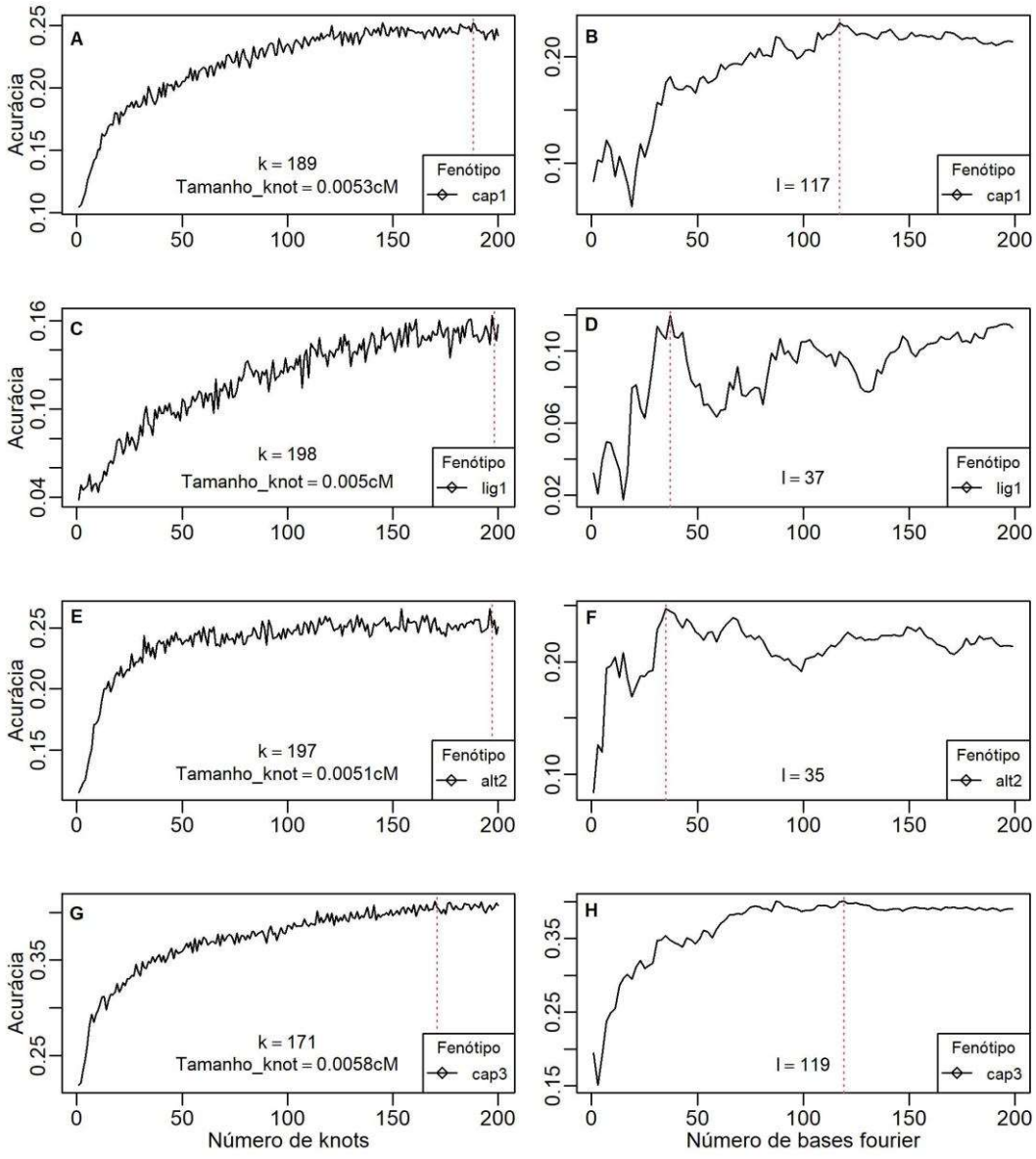
É notável que os dois métodos alternativos propostos, em relação à capacidade preditiva, obtiveram igualdade ou foram superiores aos dois métodos concorrentes (rr-BLUP e Bayes B). Ademais, tais métodos obtiveram resolução igual ou superior ao rr-BLUP em todos os cenários.

3.4 Análise de dados reais

No presente cenário, o número ótimo de *knots* tanto quanto o número ótimo de bases *Fourier* (*l*), foram determinados mediante uma validação cruzada *K-fold* (*ten-fold*). O fator de penalização α, para ambos os métodos, foi determinado via validação cruzada generalizada (GCV). Na Figura 7, no painel de A-H estão representados gráficos com valor de acurácia preditiva obtido por validação cruzada em função do número de *knots* (A, C, E e G) para *P-Bspline* e em função do número de bases *Fourier* (B, C, F e H).

O número ótimo de bases para os dois métodos para cada característica foram: para circunferência à altura do peito na Época 1 (cap1) foi 192, pois como foram $k = 189$ *knots* por cromossomos, o número de bases *B-Spline* por cromossomo é $b = k+d+1 = 189+2+1 = 192$ (essa quantidade de *knots* foi distribuída igualmente espaçada, sendo o tamanho de cada *knot* de 0,0053 cM) e, para essa mesma característica, o número ótimo de bases *Fourier* por cromossomo foi de $l = 117$. Para lignina na Época 1, foram 201 bases *B-Spline* e 37 bases *Fourier* por cromossomo. Para altura de planta na Época 2, foram 200 bases *B-Spline* e 35 bases *Fourier* por cromossomos e, para circunferência à altura do peito na Época 3, foram 174 bases *B-Spline* e 119 bases *Fourier* por cromossomos.

Figura 7 - Acurácias obtidas via validação cruzada (10-*fold*) na busca em grade para número ótimo de *knots* em *Bspline* e número ótimo de bases *Fourier* (*l*) em diferentes características do eucalipto: (A-B) cap1 - circunferência à altura do peito na Época 1; (C-D) lig1 - lignina na Época 1; (E-F) alt2 - altura de planta na Época 2 ;(G-H) cap3 - circunferência à altura do peito na Época 3. As linhas tracejadas vermelhas representam o número de *knots* (painéis A, C, E e G) para *B-Spline* e o número de bases *Fourier* (painéis B, D, F e H) que maximizam a acurácia preditiva.

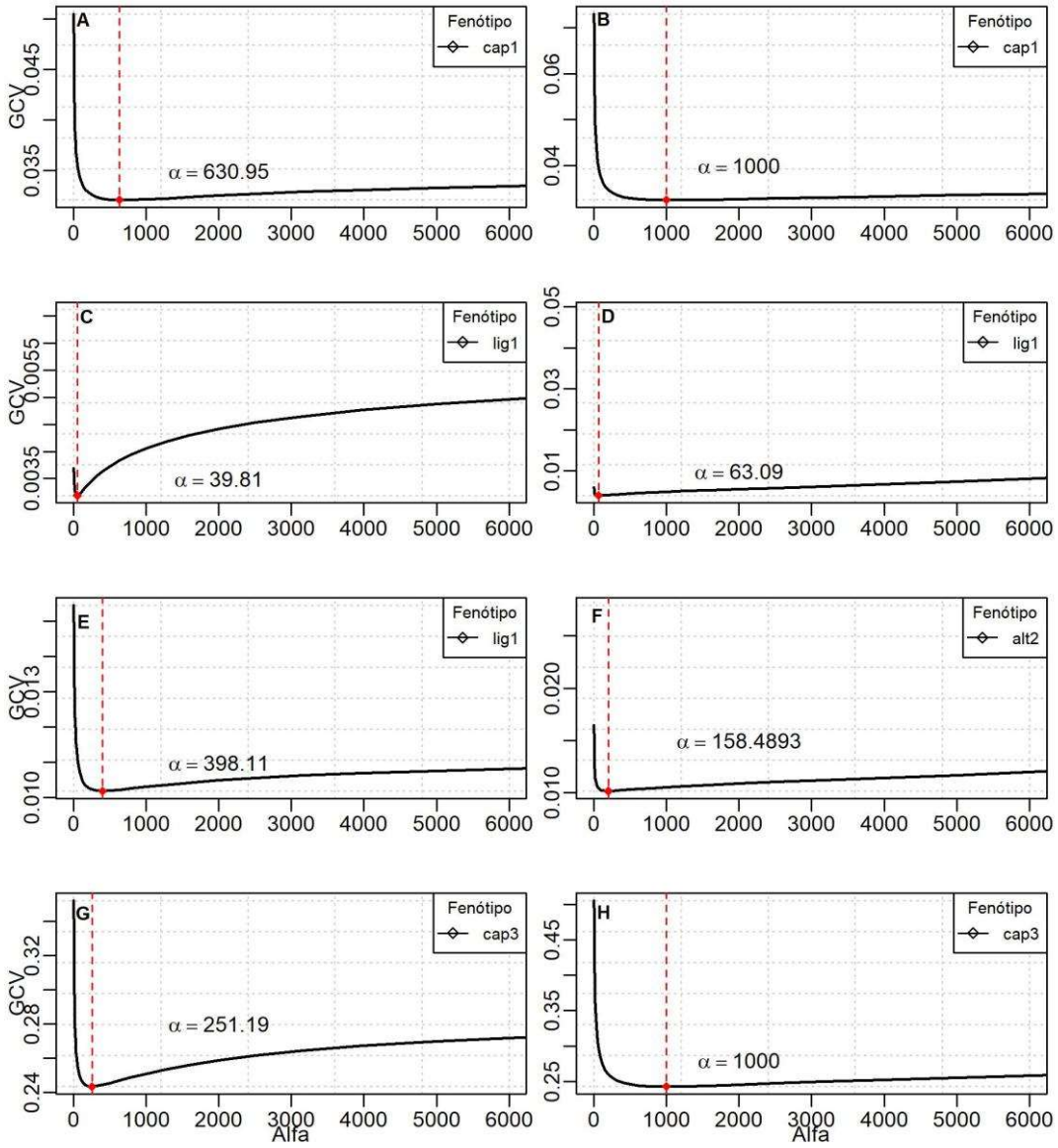


Fonte: Do autor (2021).

Nos *n folds* de validação cruzada, propomos minimizar a média das correlações das observações com observações estimadas. O método GCV consiste em selecionar o parâmetro

suavização α, que minimiza a expressão dada em (21). A Figura 8 mostra os valores de GCV em função de uma grade de candidatos α. Note que α ideal é aquele que minimiza a GCV. Dessa forma, os valores expressos nos painéis de A-H foram utilizados no modelo final para cômputo de tempo de análise em relação aos modelos concorrentes, em que α ideal para *P-Bspline* se encontra nos painéis (A, C, E e G) e para *P-Fourier* nos painéis (B, C, F e H).

Figura 8 -Valores de GCV para selecionar o parâmetro de suavização para os dois métodos *P-Bspline* e *P-Fourier* em diferentes características do eucalipto: (A-B) cap1 - circunferência à altura do peito na Época 1; (C-D) lig1 - lignina na Época 1; (E-F) alt2 - altura de planta na Época 2 ;(G-H) cap3 - circunferência à altura do peito na Época 3. As linhas tracejadas vermelhas representam o valor α que minimiza a GCV.



Fonte: Do autor (2021).

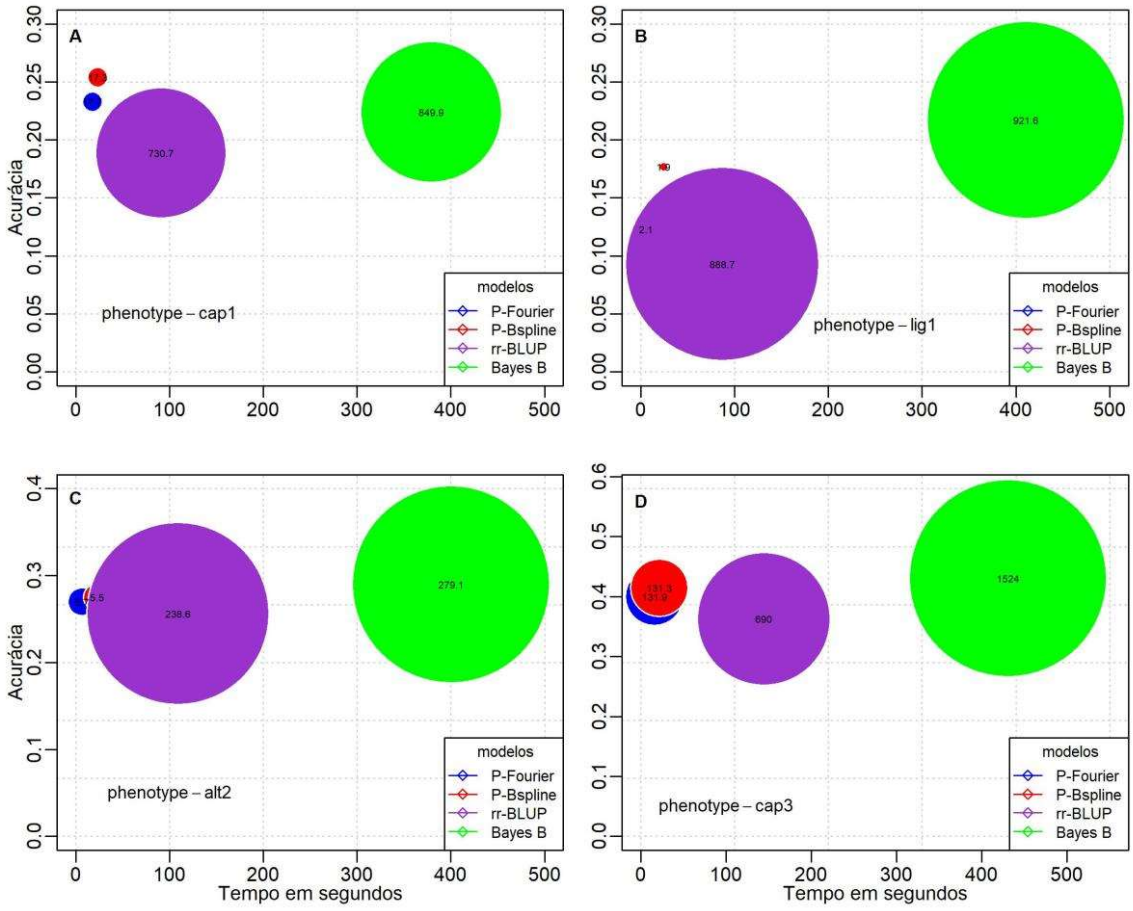
3.4.1 Acurácia *versus* tempo de análise

O tamanho da amostra foi $n = 610$ e o número de marcadores DArT foi $p = 15104$. Na Figura 9 estão apresentadas as acurácias para os quatro fenótipos (do painel de A-D) em relação

ao tempo de análise referentes aos dois modelos concorrentes (rr-BLUP e Bayes B) e os dois modelos propostos. Na figura 9, os diâmetros dos círculos são proporcionais ao desvio quadrático médio (DQM- raiz quadrada de EQM) de cada método. Dessa forma, o centro dos círculos representa a acurácia (correlação de Pearson ***r***, eixos das ordenadas) correspondente ao respectivo tempo de análise (eixo das abscissas) , mas o valor expresso no centro de cada círculo é o valor de DQM ao respectivo método.

Vale ressaltar que os dois métodos propostos além de serem competitivos em termos de acurácia em relação aos modelos concorrentes, pode-se observar pela Figura 9, que tiveram menor custo computacional e, além disso, valores de DQM muito menores do que rr-BLUP e Bayes B.

Figura 9 - Acurácia em relação ao tempo de análise obtida a partir da validação cruzada 10-*fold*, usando os quatro modelos avaliados para os fenótipos (cap1, lig1, alt2 e cap3) (painel A-D), em que significam, respectivamente, a circunferência à altura do peito na Época 1, lignina na Época 1, altura de planta na Época 2 e circunferência à altura do peito na Época 3. Os números no interior de cada circunferência representam o DQM do respectivo método.



Fonte: Do autor (2021).

4 DISCUSSÃO

Com os avanços na tecnologia de coleta e armazenamento de dados, aumentou consideravelmente a presença de dados cujas representações gráficas são curvas, imagens ou formas. Novos tipos de dados exigem, portanto, novas ferramentas analíticas, e a análise funcional de dados é uma área da estatística que estende as metodologias e teorias estatísticas convencionais ao contexto da análise funcional (MONTESINOS-LÓPEZ et al., 2018). Corroborando com essa ideia, Ramsay, Hooker e Graves (2009) afirmam que se faz

necessário uma estratégia para construir funções que funcionem com parâmetros que sejam fáceis de estimar, que possam acomodar praticamente qualquer tipo de curvas e que não use mais parâmetros do que necessário, pois isso aumenta muito o tempo computacional e complica as análises. Assim, segundo Ramsay e Silverman (2002) e Montesinos-López et al. (2018), uma suposição chave na análise de dados funcionais é que é possível aproximar qualquer curva em um espaço menor com uma série de funções base, tomando uma soma de combinações lineares de um número suficientemente grande de funções base.

Hu, Wang e Xu (2012) e Xu (2013) abordam a ideia de que os efeitos de expressão gênica possuem uma estrutura funcional, dado que esses efeitos podem ser modelados em função da posição de cada marcador. É evidente que em modelos de marcas específicas como utilizados no alfabeto bayesiano (GIANOLA et al., 2009), quanto maior o painel de marcadores, maior a demanda computacional para a estimação dos efeitos e maiores os problemas de multicolinearidade. Hu, Wang e Xu (2012) propuseram o modelo genoma contínuo a fim de obter uma técnica de redução da dimensão dos parâmetros: ao invés de buscar uma expressão polinomial para a função sinal de expressão gênica, os autores buscaram uma aproximação da integral por meio da divisão do genoma em *bins* (com o uso dos pontos de interrupção) como representação do modelo infinitesimal.

No presente trabalho, explora-se o uso de métodos de regressão semi-paramétricos via sistemas de bases (*B-Spline* e Série de *Fourier*) para realizar seleção genômica via marcadores SNP e Dart-Seq. Para estabelecimento do número ideal de *knots* (k) para *P-Bspline* e o número de bases (l) para *P-Fourier*, há três artifícios comumente utilizados: o método de seleção fixa (baseado no quantil das observações), o algoritmo Míope (*Myopic Algorithm*) e a Pesquisa completa (*Full-Search Algorithm*) (RUPPERT; WAND; CARROL, 2003). Segundo esses mesmos autores, o algoritmo *Myopic* funciona bem para muitos problemas práticos, mas possui a desvantagem de parar prematuramente e fornecer número ótimo de bases equivocado, razão pela qual não o abordamos no presente estudo. O método de seleção de *knot* fixo baseado no quantil das observações não apresentou bons resultados (resultados não mostrados). Acredita-se que esse fato pode ser explicado por duas razões: resulta em número muito pequeno de *knots* igualmente espaçados e, mais ainda, a característica intrínseca dos dados em relação ao comportamento da função sinal de expressão gênica requer uma quantidade relevante de *knots* para que aumente a chance de haver algum *knot* em regiões críticas dessa função.

O algoritmo de pesquisa completa funcionou bem em todos os exemplos examinados e, por essa razão, utilizamos como padrão para o presente trabalho, em que foi realizado uma busca em grade fina para o estabelecimento do número ideal de *knots* para *P-Bspline* e para número ideal de bases *Fourier* para método *P-Fourier*. Os parâmetros de suavização (α) para ambos os métodos foram escolhidos com base no método de validação cruzada generalizada (GCV).

É evidente que se $k \to p$ (número de *knots* é igual ao número de marcadores) ou, respectivamente, $l \to p$ (número de bases *Fourier* é igual ao número de marcadores), os modelos propostos apresentarão os mesmos problemas que os métodos clássicos de seleção genômica. No entanto, nossos resultados demonstram que o número de bases que maximizam a acurácia é muito menor que o número de marcadores (no geral, o número de bases não ultrapassou 15% do número de marcadores). Note, então, que a escolha dos *knots* é um problema importante quando se trabalha com *B-Splines*. Se muitos *knots* forem selecionados, terá um *overfitting* nos dados. Por outro lado, poucos *knots* fornecem um ajuste inadequado (RAMSAY; SILVERMAN, 2005b). Isto significa que a aplicação de dados funcionais bem-sucedida depende fortemente de alguns parâmetros que o pesquisador precisa definir como, por exemplo, o tipo de função base (*Fourier*, *B-Splines* etc.), o número requerido de funções base, o grau do polinômio para *Spline*, o período (*T*) para *Fourier*, o método de regularização, entre outros (MONTESINOS-LÓPEZ et al., 2018).

Assim, é importante destacar que um método estatístico ideal para a GWS contempla a redução de dimensionalidade, soluciona problemas de multicolinearidade e considera a regularização no processo de estimação. Na prática, escolher o método certo de seleção genômica é um desafio para os pesquisadores dessa área. Embora Meuwissen, Hayes e Goddard (2001) destaca que em estudo de simulação quase sempre o Bayes B é melhor que rr-BLUP, no presente trabalho, só constatamos essa superioridade num cenário em que há poucos genes explicando o caráter (cenário oligogênico) e com alta herdabilidade. Isso sugere que a arquitetura genética subjacente a algumas características está mais próxima do modelo infinitesimal do que o esperado.

O efeito da arquitetura genética sobre os métodos de seleção genômica foi investigado por Daetwyler et al. (2010) e Van Den Berg et al. (2015). Ademais, os dois modelos propostos, além de serem competitivos em todos os cenários abordados, mostraram superioridade aos

métodos tradicionais de SG, principalmente em baixa herdabilidade. Essa particularidade culmina em grande vantagem de nossa abordagem, pois, segundo Su et al. (2017), QTL para características de baixa herdabilidade são muitas vezes difíceis de detectar, mesmo que os traços possam ser altamente hereditários. O presente estudo mostra que Bayes B é mais sensível ao número de QTL subjacente a uma característica do que os métodos P-*Bspline*, *P-Fourier* e rr-BLUP. Portanto, os métodos *P-Bspline* e *P-Fourier* podem ser vantajosos quando aplicados a dados reais onde a arquitetura genética subjacente aos traços de interesse é desconhecida, dado que mesmo em cenários infinitesimais nos quais os métodos propostos possuem acurácia similar ao rr-BLUP, têm-se vantagem substancial em tempo de análise e expressivamente menores valores de DQM, conforme ilustram as Figuras 6 e 9.

Além disso, de acordo com Hu, Wang e Xu (2012), métodos estatísticos isolados podem não ser suficientes para lidar com uma infinidade de marcadores de alta densidade. É importante destacar que os modelos aqui propostos não sofrem desse "mal de alta dimensionalidade", isto é, podemos lidar com um número virtualmente ilimitado de marcadores, dado que $\mathbf{\Phi}$ depende da matriz transformada **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**, como representado na equação (11). Além do que, pode-se utilizar métodos de regularização atualmente disponíveis, por exemplo, o método rr-BLUP para estimar os efeitos do modelo de dimensão reduzida (dimensão do modelo é número de bases e não o painel de SNPs).

Este estudo enfatiza a seleção genômica. Todavia, a partir dos resultados obtidos, percebe-se claramente que esses métodos também poderiam ser utilizados como ferramentas para o mapeamento de QTL e, além disso, os dois métodos propostos obtiveram desempenhos melhores nesse aspecto (veja as Figuras 1, 3 e 5). Para o cenário padrão estudado, os modelos de regressão funcional propostos (*P-Bspline* e *P-Fourier*) mostraram-se muito competitivos em relação aos modelos de regressão convencionais. Além disso, os modelos de regressão funcional têm a vantagem de serem parcimoniosos, pois são necessários menos parâmetros a serem estimados, obtendo acurácias semelhantes aos modelos convencionais.

Diante do exposto, a proposta alternativa atinge um desempenho muito competitivo com métodos tradicionais de seleção genômica e tem vantagem substancial em tempo computacional em relação ao Bayes B e rr-BLUP. Finalmente, a efetividade do método proposto é ilustrada por uma aplicação para um conjunto de dados reais. Espera-se que o uso dessas funções não paramétricas gere uma boa ferramenta alternativa em termos de análise de dados genômicos. Dessa forma, os métodos propostos são promissores para a seleção genômica e merecem maior investigação.

REFERÊNCIAS BIBLIOGRÁFICAS

BALESTRE, M.; VON PINHO, R. G.; SOUZA JUNIOR, C. L.; BUENO FILHO, J. S. S. Bayesian mapping of multiple traits in maize: the importance of pleiotropic effects in studying the inheritance of quantitative traits. **Theoretical and Applied Genetics**, v. 125, n. 3, p. 479-493, 2012. doi: 10.1007/s00122-012-1847-1.

BEISSINGER, T. M.; ROSA, G. J.; KAEPPLER, S. M.; GIANOLA D.; LEON N. Defining window-boundaries for genomic analyses using smoothing spline techniques. **Genet. Sel. Evol.** v.47, n.30, 2015.

BERG, S. V. D.; CALUS M. P. L.; MEUWISSEN, T. H. E.; WIENTJES, Y. C. J. Across population genomic prediction scenarios in which Bayesian variable selection outperforms GBLUP. **BMC Genet.**, v.16, n.146, 2015.

DE BOOR, C. **A pratical guide to spline.** Springer-Verlag, New York. 1978. 392p.

CAO, Y.; LIN, H.; WU, T. Z.; YU, Y. Penalized spline estimation for functional coefficient regression models. **Comput. Stat. Data Anal.**, v.54, p.891–905, 2010.

CHEN, Z.; WANG, B.; DONG, X.; LIU, H.; REN L.; CHEN, J.; HAUCK, A.; SONG, W.; LAI, J. An ultra-high density bin-map for rapid QTL mapping for tassel and ear architecture in a large F2 maize population. **BMC Genomics**, v.15, p.1–10, 2014.

CRAVEN, P.; WAHBA, G. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. **Numerische Mathematik**, v.31,p. 377-403, 1979.

DAETWYLER, H. D.; PONG-WONG, R.; VILLANUEVA, B.; WOOLLIAMS, J. A. He impact of genetic architecture on genome-wide evaluation methods. **Genetics**, v. 185, p. 1021-1031, 2010.

DE LOS CAMPOS G.; SORENSEN D.; GIANOLA D. Genomic Heritability: What Is It? **PLoS Genet.**, v.11, p.1–21, 2015.

EILERS, P. H. C.; MARX B. D. Flexible smoothing with B -splines and penalties. **Stat. Sci.** v.11, p.89–121, 1996.

ELSHIRE, R. J.; GLAUBITZ, J. C.; SUN, Q.; POLAND, J. A.; KAWAMOTO, K.; BUCKLER, E.S.; MITCHEL, S.E. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. **PLoS One**, v.6, p.1–10, 2011.

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

GIANOLA, D.; DE LOS CAMPOS, G.; HILL, W. G.; MANFREDI, E.; FERNANDO, R. Additive genetic variability and the Bayesian alphabet. **Genetics**, v.183, p.347–363, 2009.

GODDARD, M.; HAYES, B. J. Genomic selection. **J Anim Breed Genet**., v.124, p.323–330, 2007.

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

HUANG, X.; FENG, Q.; QIAN, Q.; ZHAO, Q.; WANG, L.; WANG, A.; GUAN, J.; FAN, D.; WENG, Q.; HUANG, T.; DONG, G.; SANG, T.; HAN, B. High-throughput genotyping by whole-genome resequencing. **Genome Res.**, v.19, p.1068–1076, 2009.

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

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

MONTESINOS-LOPEZ, A.; MONTESINOS-LOPEZ, O. A.; DE LOS CAMPOS, G.; CROSSA, J.; BURGUEÑO, J.; LUNA-VAZQUEZ, F. J. Bayesian functional regression as an alternative statistical analysis of high-throughput phenotyping data of modern agriculture. **Plant Methods**, v.14, n.46, p.1-17, 2018.

MONTOYA, E. L.; ULLOA, N.; MILLER, V. A Simulation Study Comparing Knot Selection Methods With Equally Spaced Knots in a Penalized Regression Spline. **Int. J. Stat. Probab.**, v.3, p.96–110, 2014.

MOURA, E. G.; PAMPLONA, A. K. A.; BALESTRE, M. Functional models in genome-wide selection. **Plos One**, v. 14, n. 10, p.1:27, 2019.

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

R CORE TEAM (2021). **R**: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. 2021. Disponível em: <http://www.R-project.org/>.

RAMSAY, J. O.; SILVERMAN, B. W. Applied Functional Data Analysis: Methods and Case Studies. **J. R. Stat. Soc. Ser. A (Statistics Soc.)**, v.167, p.378–379, 2002.

RAMSAY, J. O.; SILVERMAN, B. W. **Functional Data Analysis**. Springer Series in Statistics, 2005a. 429p.

RAMSAY, J. O.; SILVERMAN, B. W. Comparative study of different B-Spline approaches for functional data. **Math. Comput. Model.**, v.58, p.1568–1579, 2005b.

RAMSAY, J.; WICKHAM, H.; GRAVES, S. (2015). **fda: Functional Data Analysis**. R package version 5.1.1. Disponível em: https://cran.r-project.org/web/packages/fda/fda.pdf.

RAMSAY, J.; HOOKER, G.; GRAVES, S. **Functional data analysis with R and MATLAB.** Springer Science & Business Media, New York, 2009. p. 1–19.

RIPLEY, B.; VENABLES, B.; BATES, D.M.; HORNIK, K.; GEBHARDT, A.; FIRTH, D.; RIPLEY, M.B. (2019). **Support Functions and Datasets for Venables and Ripley's MASS**. R package version 7.3.50. Disponível em: https://cran.r-project.org/web/packages/MASS/index.html.

RUPPERT, D.; WAND, M.P.; CARROLL, R. J. **Semiparametric Regression**. Cambridge university press. 2003. 382p.

SU, C.; WANG, W.; GONG, S.; ZUO, J.; LI, S.; XU, S. High Density Linkage Map Construction and Mapping of Yield Trait QTLs in Maize (Zea mays) Using the Genotyping-by-Sequencing (GBS) Technology. **Front. Plant Sci.**, v.8, p.1–14, 2017.

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**, v.5, p.1991–1998, 2015.

WANG, W.; YAN, J. (2017). **Package 'splines2**. R package version 0.4.5. Disponível em: https://cran.r-project.org/web/packages/splines2/index.html.

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

YU, H.; XIE, W.; WANG, J.; XING, Y.; XU, C.; LI, X.; XIAO, J.; ZHANG, Q. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. **PLoS One**, v.6, n.3, 2011.

ARTIGO 2 Saltos reversíveis para a modelagem conjunta da dimensão de modelos pseudo-funcionais em janelas cromossômicas na seleção genômica ampla

RESUMO

Com a genotipagem ampla de espécies e a disponibilidade de marcadores de polimorfismo de nucleotídeo único (SNP), a seleção genômica e suas ferramentas tornaram-se um importante tópico de pesquisa em melhoramento de plantas e animais. Realizar seleção genômica (SG) com grandes volumes de dados, apresenta três principais problemas: alta dimensionalidade, multicolinearidade e potencial custo computacional elevado. Desenvolvemos um método de redução dimensional utilizando um modelo funcional com seleção automática de janelas cromossômicas (*bins*) que permite inferir conjuntamente a dimensão e respectivos conjuntos de parâmetros de modelos de regressão utilizados em GS, combinando informações de SNPs vizinhos. Tal inferência simultânea é feita usando a técnica dos saltos reversíveis na amostragem Monte Carlo via cadeias de Markov (MCMC-RJ). Comparamos o método proposto a dois padrões da literatura (rr-BLUP e Bayes B). Para tanto utilizamos uma população simulada F2 com 300 indivíduos e 12.150 SNP em 10 grupos de ligação, na qual foram gerados fenótipos com duas herdabilidades (0,2 e 0,5). Um painel de SNP real de humanos com 9.307 marcadores do projeto HapMap (genoma humano) também foi empregado para simular genótipos e GS apenas como referência de população de cruzamentos complexos. A técnica proposta apresentou capacidade preditiva semelhante aos métodos padrão. O grande diferencial dessa técnica foi na alta resolução de mapear regiões causais, obtendo resultados muitos melhores nesse quesito, o que sugere que pode ser empregada no mapeamento de QTL.

Palavras-chave: Inferência Bayesiana, Modelos Funcionais, Seleção Genômica, Simulação

1 INTRODUÇÃO

A seleção genômica (SG) proposta inicialmente por Meuwissen, Hayes e Goddard (2001), diz respeito à predição de valores genotípicos usando marcadores que saturam o genoma. O resultado estatístico foi simplificar os aspectos genéticos dos modelos de regressão prévios de mapeamento de QTLs e seleção assistida por marcadores. Devido à revolução tecnológica que gerou a disponibilidade de milhares de marcadores, os métodos de SG passaram a enfrentar dois problemas teóricos de modelos de regressão: alta dimensionalidade e multicolinearidade. Isto tem levado ao desenvolvimento de técnicas estatísticas que permitam ajustar modelos lineares saturados. Em geral, os métodos propostos se agrupam em duas estratégias básicas: seleção de variáveis e modelos de encolhimento (PÉREZ; DE LOS CAMPOS, 2014; DE LOS CAMPOS et al., 2015; MOURA; PAMPLONA; BALESTRE, 2019).

Embora a seleção de variáveis seja uma técnica importante para a redução da dimensão dos modelos, avaliar todos os possíveis submodelos é impraticável. Isto faz com que seja difícil estabelecer o critério adequado para a seleção de variáveis. Por outro lado, os estimadores de encolhimento (bayesianos) são em geral, computacionalmente intensivos e não têm sido amplamente aplicados à análise de regressão. A alta dimensão dos modelos continua sendo um problema desafiador e a melhoria dos métodos estatísticos é crucial para a realização do potencial de marcadores genéticos cada vez mais densos (XU, 2007; HU; WANG; XU, 2012; FAN et al., 2016; WANG et al., 2018b).

A técnica da análise de dados funcionais (FDA) tem sido empregada recentemente e com maior sucesso na análise de associação. Essa abordagem parte do pressuposto que, o genoma de um indivíduo pode ser pensado como uma função das posições cromossômicas e assim trazer informações sobre desequilíbrios de ligação (*Linkage disequilibrium* - LD) entre regiões genômicas e QTLs (FAN et al., 2013, 2016). Em outras palavras, a segregação particular dos marcadores genéticos de um indivíduo numa região é tratada como imagens de uma função da posição cromossômica, em vez de observações discretas. Contudo, dada a descontinuidade natural do genoma entre os cromossomos, para o ajuste destas curvas, têm sido empregadas funções contínuas por partes, por exemplo, funções *Splines*. Uma curva *Spline* é uma sequência de curvas polinomiais por partes que se conectam suavemente nos pontos de quebra (nós) para formar uma única curva (RAMSAY; DALZELL, 1991).

Hu, Wang e Xu (2012) propuseram um novo método de seleção genômica, denominado genoma contínuo. A ideia desse método é dividir o genoma em blocos de alto LD denominado de *bins*, ideia similar aos intervalos entre os nós em uma *Spline*. Sob essa ótica, um LD alto requer um pequeno número de *bins* para capturar todas as informações genômicas, enquanto o LD baixo precisa de um grande número de *bins* para capturar a mesma quantidade de informações. Janelas genômicas assim definidas, foram denominados de *bins* naturais por (HU; WANG; XU, 2012; AN et al., 2020; XU, 2013b).

Um *bin* pode conter diversos marcadores, produzindo, portanto, um modelo de redução dimensional. Embora o método desenvolvido por esses autores apresentasse maior acurácia do que modelos tradicionais de SG, blocos com alto desequilíbrio de ligação podem não ser diretamente aplicáveis em algumas situações, uma vez que o número de *bins* depende, entre outros fatores, do tamanho da amostra. Contudo, o autor introduz o conceito de *bins* artificiais, definidos a *priori* pelo pesquisador usando critérios genéticos (como por exemplo anotações genômicas de outros estudos) sem necessidade de estabelecimento prévio de blocos de alto LD (*bins* naturais), tornando a análise muito mais simples.

Como visto acima, é possível combinar vários SNP (*Single Nucleotide Polymorphism*) no genoma em número gerenciável de janelas genômicas, denominadas de *bins* (HU; WANG; XU, 2012; XU, 2013b; MOURA; PAMPLONA; BALESTRE, 2019). Técnicas semelhantes, de dividir os marcadores em grupos, foram propostas por outras equipes de pesquisadores com nomes diferentes em contextos diferentes, por exemplo, Fan et al. (2014) realizaram um estudo sobre a construção de haplótipos no genoma bovino para detectar genes que afetam características relacionadas à qualidade da carne. Outra técnica similar denominada cBLUP (do Inglês *compact BLUP*) foi proposta por Wang et al. (2018) para aproximar o parentesco (ou a identidade em estado) entre os indivíduos em GWAS (*Genome-Wide Association Studies*). Os resultados desses autores, mostram que associação baseada em grupos de SNP, podem ser mais eficientes do que a análise SNP única, com custo computacional menor.

Na definição clássica de *bins*, é utilizado o efeito médio dos *bins* como medida de informação. Isto é, os *bins* são tratados como novos marcadores "sintéticos" e devido isso, assume-se a suposições que todos os marcadores na mesma janela genômica devem ter efeitos homogêneos. Isso significa que o tamanho das janelas não pode ser muito grande. Se os *bins* forem grandes é necessário, outra suposição para contrabalançar, ou seja, todos os marcadores

no *bin* devem terem efeitos na mesma direção (HU; WANG; XU, 2012; XU 2013b; MOURA; PAMPLONA; BALESTRE, 2019; AN et al., 2020).

Para contornar este problema, Moura, Pamplona e Balestre (2019) propuseram ponderar os efeitos dos marcadores dentro de cada janela genômica pela frequência relativa de visitas ao marcador em um processo estocástico de amostragem da distribuição *a posteriori* via MCMC (*Markov Chain Monte Carlo*). A ponderação evita a necessidade do cancelamento de efeitos direcionais opostos, tornando a estratégia de dividir o genoma em blocos mais direta. Contudo, embora fossem encontrados resultados satisfatórios, na prática, não se sabe a quantidade ou tamanho ideal dos *bins*, que é um parâmetro muito influente no sucesso da técnica. Uma abordagem bastante promissora para este tipo de problema parece ser a adoção do método de amostragem MCMC com saltos reversíveis (MCMC-RJ, do Inglês *Reversible Jump*), proposto por Green (1995) no contexto seleção de modelos de regressão e posteriormente experimentado em modelos de QTL por diversos autores (SILLANPÄÄ; ARJAS, 1998; YI; XU, 2000; WAAGEPETERSEN; SORENSEN, 2001; YI; XU; ALLISON, 2003; YI, 2004; LIU et al. 2007) .

Esta é a motivação do presente trabalho, em que se propõe um método que possa simultaneamente inferir sobre a dimensão dos *bins* e os efeitos de suas marcas combinando SNPs vizinhos. Dado que o número de *bins* é substancialmente menor que o painel de SNP, vários métodos podem ser empregados para estimar os efeitos dos marcadores contidos nos *bins* para uma característica de interesse. Nesse estudo, utilizou-se o método de encolhimento proposto por Xu (2003) por ser eficiente e de fácil implementação. No presente estudo o número (K) de *bins* é tratado como uma variável aleatória. A inferência simultânea sobre a dimensão do modelo é feita usando o método MCMC-RJ.

2 MATERIAL E MÉTODOS

2.1 Dados simulados

No primeiro cenário, foram simulados dez grupos de ligação, cada um com tamanho de 1,20 cM e distância média de 0,001 cM entre marcadores adjacentes. A função Haldane foi usada como função de mapeamento e passeio aleatório como método meiose para construir um painel com 300 indivíduos e 12150 SNPs. A população gerada foi a F2 de um cruzamento biparental usando o software QGenes (JOEHANES; NELSON, 2008). Dentre os SNPs

simulados, dez foram assumidos como QTL para representar o cenário oligogênico e seus efeitos amostrados de uma distribuição normal. Para o cenário poligênico, 60 marcadores foram assumidos como QTL e seus efeitos também foram amostrados de uma distribuição normal. Os resíduos foram amostrados para cada SNP de distribuições normais com variâncias calibradas para as duas herdabilidades simuladas (0,2 e 0,5).

2.2 Dados reais

Neste cenário, uma estrutura natural de genótipos com desequilíbrio de ligação em um painel de SNP com 9.307 marcadores do projeto *HapMap* (genoma humano) foi usada para simular os fenótipos. O painel está distribuído nos vinte e dois cromossomos autossômicos com posição de SNP conhecida (medida em kb). Informações sobre duas populações diferentes estão disponíveis em https://cran.r-project.org/web/packages/SNPassoc/SNPassoc.pdf.

Os genomas referem-se a duas populações humanas, uma europeia (UE) e outra Yorubá (YRI) e estão descritos em González et al. (2014). As anotações genômicas (nomes de SNPs, cromossomos e posição genética) também estão disponíveis. Uma análise inicial corrigiu o painel preservando a frequência alélica mínima (MAF >5%) e imputando valores perdidos com a função *A.Mat(.)* da biblioteca *rr-BLUP* (ENDELMAN, 2011). O painel resultante apresenta 7.574 SNPs em 120 indivíduos. Foram simulados os fenótipos com as mesmas configurações descritas na seção 2.1.

2.3 Modelo funcional

Neste trabalho, será empregado o modelo funcional bayesiano desenvolvido por Moura, Pamplona e Balestre (2019), alterado para a seleção automática de *bins*, por meio do método MCMC com Saltos Reversíveis (MCMC-SR). Dada à alta densidade de marcadores, utilizamos a ideia de dividir o genoma em janelas (análogo ao método *bin* artificial de Xu 2013b) como estratégia de redução dimensional. Com marcadores genéticos que cobrem o genoma inteiro, as posições dos marcadores são tão próximas que podem ser consideradas como um conjunto saturado de observações no domínio de uma função contínua (no intervalo do cromossomo), isto é, cada ponto (posição) que representa uma região candidata pode ser suavizada por uma relação funcional.

Dessa forma, o modelo funcional genômico supõe que a função sinal de um gene no genoma pode ser descrito por uma função unidimensional (MOURA; PAMPLONA; BALESTRE, 2019). Por outro lado, por não ser possível verificar todas as posições no genoma, por se tratar de uma variável contínua, utiliza-se as posições dos marcadores (λ) como grade fina de realizações de uma função desconhecida $f(\lambda)$. Em outras palavras, tem-se que $f(\lambda)=\gamma(\lambda)$, em que γ é a representação funcional do efeito de expressão do gene em dada posição. O objetivo final é estimar γ e, a partir dessa estimativa, predizer o valor genético genômico de novos indivíduos.

Como mostrado em Hu, Wang e Xu (2012), Xu (2013b) e Moura, Pamplona e Balestre (2019), a integral $\int Z(\lambda)\hat{\gamma}(\lambda)d\lambda$ retorna o valor genético genômico para o *i*-ésimo indivíduo, ao passo que nos demais modelos de seleção genômica, este é recomposto pela soma das combinações lineares entre os genótipos dos SNP e seus efeitos aditivos, na suposição de que cada SNP é potencialmente um QTL. Em modelos com alta dimensão pode-se considerar a igualdade $\lim_{p\to\infty}\left(\sum_{i=1}^{p} z_{ij}\hat{\gamma}_i\right)=\int z(\lambda)\ \hat{\gamma}(\lambda)d\lambda$, sendo "$p$" o número de SNP considerados. A analogia com modelos infinitesimais é direta e o modelo linear pseudo-funcional (HU; WANG; XU, 2012) para uma característica quantitativa pode ser definido como

$$\mathbf{y}_i=\boldsymbol{\mu}+\sum_{t=1}^{C}\int_{0}^{L_t} Z_{it}(\lambda)\gamma(\lambda)d\lambda+\boldsymbol{\varepsilon}_i,\quad \forall i=1,...,n \tag{1}$$

No modelo acima adaptado por Moura, Pamplona e Balestre (2019), a soma descreve a descontinuidade da função ao longo dos cromossomos e L_t é o tamanho do *t*-ésimo cromossomo, com $t=1,\cdots,C$. Sob esse modelo, μ é o intercepto comum a todos os indivíduos da população, $\gamma(\lambda)$ é efeito aditivo do marcador na posição λ (expresso como uma função

desconhecida), $\mathbf{Z}(\lambda)$ o estado genotípico contínuo que só é conhecido nas posições dos marcadores. Assim, Z exerce a função de ponderar o quanto cada ponto γ contribui para a integral em (1) de acordo com o estado genotípico (2, 1 e 0) correspondendo aos três genótipos possíveis para o *i*-ésimo indivíduo na posição λ, como *AA*, *Aa* e *aa*. Por fim, $\boldsymbol{\varepsilon}_i$ é o erro de medida para o indivíduo *i*, sendo $\varepsilon_i \sim N(0, \sigma^2)$.

A função $\gamma(\lambda)$ é desconhecida e a integral em (1) não é explícita. Existem algumas possibilidades de aproximação numérica sendo adotada a estratégia de divisão dos cromossomos em *bins*. A função contínua desconhecida $f(\lambda) = \gamma$ foi estimada de forma empírica utilizando algoritmo Metropolis-Hasting (METROPOLIS et al., 1953; HASTINGS, 1970). Diferentemente da abordagem original em que λ_k é a posição do ponto médio do *k*-ésimo *bin* no genoma e $Z_j(\lambda_k)$ é o valor médio de todos os marcadores do *k*-ésimo *bin*, neste trabalho propomos utilizar λ_k como sendo a posição do marcador amostrado no $bin\left[\lambda_{j(\min)}, \lambda_{j(\max)}\right]$. A novidade é utilizar $Z_j(\lambda_k)$ para representar o estado genotípico no marcador amostrado dentro do *k*-ésimo *bin*, em que $j = 1, \cdots, p_k$ e $k = 1, \cdots, K$, sendo p_k o número de marcadores dentro do *k*-ésimo *bin*, K é número de *bins*. Notar que $\gamma_j(\lambda_k)$ passa a ser, simplesmente, o efeito do *j*-ésimo marcador no *k*-ésimo *bin*. Assim, o total de marcadores em todo o genoma é dado por $p = \sum_{k=1}^{K} p_k$. Por simplicidade, doravante, adota-se Z_k em substituição $Z_j(\lambda_k)$ e γ_k ao invés de $\gamma_j(\lambda_k)$.

Dado um valor candidato amostrado dentro do *k*-ésimo *bin*, o modelo pseudo-funcional a ser considerado durante a amostragem da distribuição conjunta *a posteriori* é descrito praticamente como um modelo linear dado por:

$$\mathbf{y}_i = \boldsymbol{\mu} + \sum_{k=1}^{K} Z_{ik}\gamma_k + \boldsymbol{\varepsilon}_i, \quad \forall i = 1,...,n \tag{2}$$

Aqui, o número de *bins* é desconhecido e foi estimado conjuntamente.

2.3.1 Distribuições *a priori*

As variáveis observáveis são os valores fenotípicos (***y***) e os genótipos dos marcadores ($\mathbf{Z}$), a posição λ coincide com a posição dos marcadores (λ_k) em cada janela genômica e, portanto, será atribuído uma *priori* para esse parâmetro. As variáveis não observáveis são os coeficientes de regressão ($\boldsymbol{\mu}$, γ), o números de *bins* (*K*) e as variâncias residual e dos efeitos aditivos (σ_e^2 , σ_γ^2). Estabelecemos distribuições *a priori* não informativas (de Jeffreys) para a variância residual e para a média geral, dadas por:

$$p(\boldsymbol{\mu}) \propto 1 \quad \text{e} \qquad p\left(\sigma_e^2\right) \propto \frac{1}{\sigma_e^2} \tag{3}$$

As distribuições conjuntas *a priori* para o efeito de marcador e sua variância seguiram a conjugação normal-inversa-qui-quadrada:

$$p\left(\gamma_k \mid \sigma_{\gamma_k}^2\right) \propto N\left(0, \sigma_{\gamma_k}^2\right), \qquad p\left(\sigma_{\gamma_k}^2\right) \propto \chi_{esc}^{-2}\left(\upsilon, S^2\right) \tag{4}$$

Quanto à distribuição *a priori* para a posição λ_k dentro do *k*-ésimo *bin*, considerando Δ_k o número de marcadores dentro do *bin*, assumiu-se que a posição é uniformemente distribuída nos *bins*, isto é, $p(\lambda_k) = \frac{1}{\Delta_k}$.

A *priori* conjunta para as variáveis não observáveis é dada por:

$$p\left(K, \lambda_k, \theta^{(K)}\right) \propto p(K)\, p(\boldsymbol{\mu})\, p\left(\sigma_e^2\right) \prod_{k=1}^{K} p\left(\boldsymbol{\gamma}_k \mid \lambda_k, \sigma_{\gamma_k}^2\right) p\left(\boldsymbol{\sigma}_{\boldsymbol{\gamma}_k}^{\mathbf{2}}\right) \tag{5}$$

em que, $\theta^{(K)} = \{\boldsymbol{\mu}, \sigma_e^2, \gamma_k, \sigma_{\gamma_k}^2\}$ é um vetor de parâmetros do modelo.

2.3.2 Verossimilhança conjunta do fenótipo e da posição dos marcadores

A probabilidade conjunta para a observação fenotípica e a posição do marcador pode ser descrito como:

$$p\left(\mathbf{y}, \lambda_k \mid K, \theta^{(K)}\right) = p\left(\mathbf{y} \mid K, \theta^{(K)}\right) p\left(\lambda_k \mid K, \theta^{(K)}, y\right) \tag{6}$$

A verossimilhança para os dados fenotípicos pode ser descrita por:

$$\begin{aligned} p\left(\mathbf{y} \mid K, \theta^{(K)}\right) &= \prod_{i=1}^{n} p\left(\mathrm{y}_i \mid \theta^{(K)}\right) \\ &\propto \left(\sigma_e^2\right)^{-n/2} exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{i=1}^{n} \left(y_i - \mu - \sum_{t=1}^{C} \int_0^{L_t} Z_{kit} \gamma_k d\lambda \right)^2 \right\} \end{aligned} \tag{7}$$

que pode ser aproximada por:

$$p\left(\mathbf{y} \mid K, \theta^{(K)}\right) \propto \left(\sigma_e^2\right)^{-n/2} exp\left\{ -\frac{1}{2\sigma_e^2} \sum_{i=1}^{n} \left(y_i - \mu - Z_{k(i)} P_{\lambda_k} \gamma_k \right)^2 \right\} \tag{8}$$

sendo P_{λ_k} a função peso determinada implicitamente pela frequência com que λ_k é visitado ao longo das cadeias que aproximam a distribuição *a posteriori* conjunta.

A distribuição *a priori* para o número de *bins* por cromossomo é *Poisson-truncada* (limite mínimo de um *bin* por cromossomo e máximo pré-definido como o número de marcadores, ou seja, $N_{bin\max} = \mathrm{K}$ e $N_{cro} \leq \phi \leq K$), com média ϕ:

$$p(K/\phi) \propto \frac{\phi^K e^{-\phi}}{K!} \tag{9}$$

a distribuição *a priori* para ϕ utilizada foi a *gama* (τ=1, υ =1), onde τ é o parâmetro de forma e υ é o parâmetro de escala.

2.3.3 Distribuição posteriori conjunta

A distribuição *a posteriori* conjunta pode ser descrita por:

$$p\left(\theta^{(K)},\mathrm{K},\phi \mid \mathbf{y},\lambda_k\right) \propto p\left(\mathbf{y} \mid K,\theta^{(K)}\right) p\left(\lambda_k \mid K,\theta^{(K)},y\right) p\left(\lambda_k,K,\theta^{(K)}\right) p(\mathrm{K}/\phi) p(\phi) \tag{10}$$

2.3.4 Processo de amostragem via MCMC com *Reversible Jump*

Esse método consiste na execução das etapas de Metropolis-Hastings que aceitam ou rejeitam movimentos diferentes, como "nascimento" ou "morte" de um *bin*. Essas etapas permitem transições do modelo atual para modelos de dimensões superiores ou inferiores (ZUANETTI; MILAN, 2016). No contexto aqui, não há necessidade de mudança de variável, pois o determinante do Jacobiano da transformação é igual a um, dado que, é o determinante de uma matriz identidade, semelhante às propostas de (SILLANPÄÄ; ARJAS, 1998; YI; XU, 2000; WAAGEPETERSEN; SORENSEN, 2001; BALESTRE et al., 2012).

Devido à variabilidade da dimensionalidade do nosso problema, deve-se projetar tipos de movimento entre os subespaços *K*. Para esse problema, as transições possíveis são:

(I) a adição de um *bin* (etapa do nascimento),

(II) a exclusão de um *bin* (uma etapa da morte) e

(III) manter o número atual de *bins*.

Nas etapas (I) e (II), altera-se a dimensão do modelo.

As probabilidades a *priori* para os três tipos de movimentos adicionar, excluir ou manter o número atual de *bins* são respectivamente, p_a para o movimento que eleva o modelo do estado *K* para *K*+1, p_d para o movimento *K* para *K*–1 e p_0 que não altera a dimensão do modelo, em que, $p_0 + p_a + p_d = 1$ para todos os *k*. Nesse contexto, os detalhes do algoritmo condicional e RJMCMC completo, mais a regra de decisão para incluir ou excluir um *bin* no modelo serão apresentados a seguir.

2.3.5 Adição de *bin* (etapa de nascimento)

Nesta etapa é sorteado um único *bin* entre os existentes e uma única marca contida nesse *bin*, que será transformada em limite de novos dois *bins* e será a marca representante do novo *bin* na iteração.

Assim, a regra de decisão para adicionar um novo *bin* ao modelo é dada por $min[1, \alpha(\mathrm{K}, \mathrm{K}+1)]$, em que

$$\alpha\left(\mathrm{K}, \mathrm{K}+1\right) = \frac{\prod_{i=1}^{n} p(y|K+1)}{\prod_{i=1}^{n} p(y|K)} \frac{p(\mathrm{K}+1|\phi)}{p(\mathrm{K}|\phi)} \frac{\xi(\mathrm{K}+1, \mathrm{K})}{\xi(\mathrm{K}, \mathrm{K}+1)} \tag{11}$$

em que, *K* é o número de *bins* no modelo atual, $K+1$ é o modelo candidato à nascimento de um *bin* e, $p(\mathrm{K}|\phi)$ e $p(\mathrm{K}+1|\phi)$ são as probabilidade *a priori* dos modelos com dimensões *K* e

K+1, como visto na equação (7), é dada por uma distribuição de Poisson truncada. $p(y|\mathrm{K})$ e $p(y|K+1)$ são as funções de probabilidade dos dados mediante aos parâmetros dos modelos *K* e *K*+1, respectivamente. As distribuições de Hastings $\xi(\mathrm{K},\mathrm{K}+1)$ e $\xi(\mathrm{K}+1,\mathrm{K})$ propostas, são necessárias para permitir a reversibilidade durante o processo MCMC, em que $\xi(\mathrm{K},\mathrm{K}+1)=\mathrm{p}_a$ e seu reverso $\xi(\mathrm{K}+1,\mathrm{K})=p_d$ (Xu 2013a). Note que, $\xi(\mathrm{K},\mathrm{K}+1)=\mathrm{p}_a$ é a probabilidade proposta para inclusão de um *bin* e $\xi(\mathrm{K}+1,\mathrm{K})=p_d$ é probabilidade proposta de deleção.

Assim, a expressão (11) pode ser simplificada da seguinte maneira:

$$\alpha\left(\mathrm{K},\mathrm{K}+1\right)=\frac{\prod_{i=1}^{n}p(y|K+1)}{\prod_{i=1}^{n}p(y|K)}\,\frac{\phi}{K+1}\,\frac{p_d}{p_a} \tag{12}$$

em que ϕ é a média da Poisson e $p\left(K\,|\,\phi\right)\propto\frac{\phi^k e^{-\phi}}{\mathrm{K!}}$.

Se a proposta for aceita, isto é, se o $min[1,\alpha(K,K+1)]$ for maior que uma variável aleatória amostrada de uma distribuição uniforme [0,1], um *bin* é incluído no modelo. Caso contrário, a cadeia permanecerá no modelo atual, isto é, *K* não será alterado.

2.3.6 Deleção de *bin* (etapa de morte)

Na etapa de deleção é sorteado um único *bin* dentre os existentes a ser excluído. Em seguida é calculada a probabilidade de aceitação do estado (deleção), dada pelo $min[1,\alpha(K,K-1)]$, sendo

$$\alpha\left(\mathrm{K},\mathrm{K}-1\right)=\frac{\prod_{i=1}^{n}p(y|K)}{\prod_{i=1}^{n}p(y|K-1)}\,\frac{p(\mathrm{K}-1|\phi)}{p(\mathrm{K}|\phi)}\,\frac{\xi(K-1,K)}{\xi(K,K-1)} \tag{13}$$

Assim, a expressão (13) pode ser simplificada e o fator de Bayes $min[1,\alpha(K-1,K)]$ para testar a morte do *k*-ésimo *bin*, é dado por:

$$\alpha\left(K-1,K\right)=\frac{\prod_{i=1}^{n}p(y|K)}{\prod_{i=1}^{n}p(y|K-1)}\,\frac{K}{\phi}\,\frac{p_a}{p_d} \tag{14}$$

sendo $p(y|K-1)$ a distribuição condicional dos dados para o modelo com dimensão reduzida, $\xi(K-1,K)$ e $\xi(K,K-1)$ são as propostas de Hastings que garantem reversibilidade do modelo durante o processo MCMC. Se a proposta for aceita, isto é, se o $min[1,\alpha(K-1,K)]$ for maior que uma variável aleatória amostrada de uma distribuição uniforme [0, 1], excluiremos um *bin* do modelo. Caso contrário, a dimensão do modelo atual é mantida.

A atualização dos parâmetros, com exceção de *K,* segue modelos usuais de MCMC padrão. O resultado do RJ-MCMC é uma cadeia da *posteriori* conjunta que é formada com frequências de ocorrência variável de *bins* e marcadores. Isto permite a inferência marginal em parâmetros como o número de *bins*, posição e efeitos de marcadores (QTLs) e valores genéticos individuais para a SG.

2.3.7 Monte Carlo Cadeias de Markov para modelos funcionais genômicos

Nesta subseção, são apresentados apenas os passos referentes ao algoritmo Gibbs e Metropolis-Hasting para todos os parâmetros do modelo, com exceção de *K*. Ressalta-se que o processo de amostragem iterativa é uma combinação dos algoritmos de Gibbs, Metropolis-

Hasting e MCM-RJ. Utilizou-se o algoritmo Monte Carlo Cadeias de Markov (MCMC), via amostragem de Gibbs para os parâmetros do modelo e Metropolis-Hasting para a integração numérica de $p\left(\lambda_k \mid K, \theta^{(K)}, y\right)$. Para isso, dividiu-se o genoma em *bins* e dentro de cada *bin* uma posição λ_k foi sorteada de modo que a dimensão do modelo fosse restrita ao número de *K* *bins* durante o processo MCMC, cujos passos são descritos a seguir.

1) *Inicialização:* Os parâmetros μ e σ_e^2 são inicializados com a média e a variância dos dados fenotípicos, respectivamente; o vetor γ é b_j inicializado com o valor zero e dimensão *K*, onde *K* é o número de *bins* do modelo atual. A matriz do estado genotípico Z_k de dimensão (*n* x *K*) foi amostrada da matriz completa de marcadores **Z** de dimensão (*n* x *p*), em que $K \leq p$.

E o índice λ_k correspondente à posição inicial do marcador amostrado no *k*-ésimo *bin*. Assim, Z_k foi inicialmente amostrada com base na posição média λ_k do *k*-ésimo *bin*. As variâncias dos efeitos de cada marcador $\sigma^2_{\gamma_k}$ foram assumidas inicialmente como 0,5.

$$I^{(0)} = \left[\mu^{(0)}, \gamma_1^{(0)}, \ldots, \gamma_K^{(0)}, \sigma_e^{2^{(0)}}, \sigma_{\gamma_1}^{2^{(0)}}, \ldots, \sigma_{\gamma_K}^{2^{(0)}}, Z_K\right] \tag{15}$$

2) Neste passo atualiza-se μ utilizando a seguinte condicional b_0:

$$\mu \mid \ldots \sim N\left[\sum_{i=1}^{n}\left(y_i - \sum_{k=N_{cro}}^{K} Z_{k(i)}\gamma_k\right)\Big/ n\ ,\ \frac{\sigma_e^2}{n}\right] \tag{16}$$

A partir dessa condicional, um novo μ é amostrado. O μ amostrado é denotado por $\mu^{(t)}$ e substituirá $\mu^{(0)}$ em todos os processos de amostragem subsequentes.

3) A distribuição *a posteriori* referente aos efeitos dos marcadores (γ_{k^t}) dada a posição λ_k são amostradas da *t*-ésima iteração é dada pela seguinte distribuição:

$$\gamma_{k^t} | \ldots N\left(\left(\sum_{i=1}^{n} Z^2_{k^t(i)} + \frac{\sigma_e^2}{\sigma^2_{\gamma_{k^t}}}\right)^{-1} \sum_{i=1}^{n} Z_{k^t_{(i)}}\left(y_i - \mu - \sum_{k'=1}^{K-1} Z_{k'^{(t)}_{(i)}} \gamma_{k^t}\right), \left(\sum_{i=1}^{n} Z^2_{k^t_{(i)}} + \frac{\sigma_e^2}{\sigma^2_{\gamma_{k^t}}}\right)^{-1} \sigma_e^2\right) \quad (17)$$

O γ_k recém-amostrado é denotado $\gamma^{(t)}{}_k$ e substituirá $\gamma^{(0)}{}_k$ em todos os processos de amostragem subsequentes.

4) Como mencionado anteriormente, não existe uma expressão analítica para $p\left(\lambda_k \mid K, \theta^{(K)}, y\right)$, mas o algoritmo Metropolis-Hasting (METROPOLIS et al., 1953; HASTINGS, 1970), pode ser utilizado dado que não exige que o parâmetro tenha uma função de probabilidade fechada. Para isso, faz-se uso de uma função geradora de candidatos que podem ser aceitos com α_k de probabilidade. Foi utilizada aqui a distribuição uniforme para λ_k, amostrada em cada *bin*, em todos os cromossomos, no intervalo $[\,max\left(LI_k, \lambda_k - c\right); min\left(LS_k, \lambda_k + c\right)\,]min(0{,}2; \lambda_j + c)$, sendo *c* uma constante discreta positiva que define o caminhamento (salto) dentro do *k*-ésimo *bin*, normalmente fixado um valor de 10% do número de posições alocadas dentro de cada *bin*. A densidade geradora de candidatos, é representada por $u\left(\lambda_k^*, \lambda_k^{(t)}\right) u(\lambda_j^{(k+1)}, \lambda_j^{(k)}) u(\lambda^*, \lambda)$, para o *k*-ésimo *bin* e a nova posição λ_k^* é aceita na *t*-ésima iteração com $min\left(1, \alpha_k\right)$ de probabilidade. Assim, se α_k for aceito, uma nova posição é estabelecida λ_k^* e o estado do marcador Z_{k^*} é amostrado do painel completo. A regra de decisão para mudança da posição do marcador dentro do *bin* é dada por:

$$\alpha_k = \frac{p\left(\lambda_k^* \mid K, \theta^{(K)}, y\right) u\left(\lambda_k^*, \lambda_k^{(t)}\right)}{p\left(\lambda_k^{(t)} \mid K, \theta^{(K)}, y\right) u\left(\lambda_k^{(t)}, \lambda_k^*\right)} \quad (18)$$

em que

$$p\left(\lambda^{*} \mid K, \theta^{(k)}, y\right) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i - \mu - \sum_{k=N_{cro}}^{K-1} Z_{k(i)}\gamma_k - Z_{k^*}\gamma_{k^*}\right)^2\right\} \tag{19}$$

e

$$p\left(\lambda_k^{(t)} \mid K, \theta^{(K)}, y\right) \propto \exp\left\{-\frac{1}{2\sigma_e^2}\sum_{i=1}^{n}\left(y_i - \mu - \sum_{k=N_{cro}}^{K} Z_{k(i)}\gamma_k\right)^2\right\} \tag{20}$$

Na maioria das situações, o algoritmo Metropolis é suficiente, mas se o λ_k^* é amostrado perto dos limites (superior ou inferior) de uma janela genômica, o ajuste ou correção de Hastings é necessário para garantir que o algoritmo não fique preso a uma posição fixa e, além disso, garantir que todas as posições visitadas pelo algoritmo no *k*-ésimo *bin*, devem serem amostradas dentro dos limites desse *bin*.

A correção de Hasting foi dada por:

$$u\left(\lambda_k^{(t)}, \lambda_k^{*}\right) = \begin{cases} \dfrac{1}{2c}, & \text{se } \lambda_k^{(t)} + c \le LS_k \text{ e } \lambda_k^{(t)} - c \ge LI_k \\ \dfrac{1}{c + \lambda_k^{(t)} - LI_k}, & \text{se } \lambda_k^{(t)} + c < LS_k \text{ e } \lambda_k^{(t)} - c < LI_k \\ \dfrac{1}{c + LS_k - \lambda_k^{(t)}}, & \text{se } \lambda_k^{(t)} + c > LS_k \text{ e } \lambda_k^{(t)} - c > LI_k \end{cases}$$

$$u\left(\lambda_k^{*}, \lambda_k^{(t)}\right) = \begin{cases} \dfrac{1}{2c}, & \text{se } \lambda_k^{*} + c \le LS_k \text{ e } \lambda_k^{*} - c \ge LI_k \\ \dfrac{1}{c + \lambda_k^{*} - LI_j}, & \text{se } \lambda_k^{*} + c < LS_k \text{ e } \lambda_k^{*} - c < LI_k \\ \dfrac{1}{c + LS_k - \lambda_k^{*}}, & \text{se } \lambda_k^{*} + c > LS_k \text{ e } \lambda_k^{*} - c > LI_k \end{cases} \tag{21}$$

5) A distribuição condicional completa para a variância residual após aceitar a *j*-ésima no posição no genoma, contida no *k*-ésimo *bin* é dada por:

$$\sigma_e^2 | \ldots \sim \chi_{esc}^{-2}\left(n, \sum_{i=1}^{n} \left(y_i - \mu - \sum_{k=1}^{K} Z_{k(i)}\gamma_{k(i)}\right)^2\right) \quad (22)$$

6) Finalmente, a variância específica do marcador $\sigma^2_{\gamma_k}$, é amostrada utilizando uma distribuição qui-quadrada invertida-escalada:

$$\sigma^2_{\gamma_k} | \ldots \sim \chi^{-2}_{esc}(\upsilon + 1, \gamma_k^2 + S^2) \quad (23)$$

Isto conclui a atualização dos parâmetros do modelo *K* dimensional na *t*-ésima iteração. Testada a troca de dimensão de modelo, repete-se a sequência, até a convergência da cadeia para uma distribuição estacionária. Na cadeia final, adotou-se. Na cadeia final, adotou-se $f(\lambda_k | K, \theta^{(K)}, y) = P_{\lambda_k}$ como a frequência de visitas realizadas na posição λ_k dentro de um *bin* específico.

Contudo, a integral $\int_0^L Z_{ki}\gamma_k d\lambda_k$ para recompor o valor genético genômico não é conhecida dado que γ_k não tem uma forma analítica fechada. Hu, Wang e Xu (2012) utilizaram o efeito médio dos *bin* como a esperança de $p\left(\lambda_k | K, \theta^{(K)}, y\right)$. Nesse estudo, relaxa-se a suposição do efeito médio e, como para cada λ_k pode-se atribuir a frequência do número de vezes que o modelo visitou o marcador correspondente, é possível aproximar, $\int_0^L Z_i(\lambda)\gamma(\lambda)d\lambda$ por $\mathbf{Z}_i P_\lambda \bar{\hat{\gamma}}$. Note que, γ_k é o efeito do *j*-ésimo marcador no *bin k*. Logo, o efeito total de todos os marcadores do *bin k* pode ser representado por $\gamma'_k = \sum_{j=1}^{p_k} \gamma_k$ e , assim, γ é um vetor $p \times 1$ de

efeitos de todos *bins* e, é dado por $\gamma = \sum_{k=1}^{K} \gamma'_k$. Contudo, como utilizou-se os *bins* apenas como estratégia de integração numérica, γ ode ser estimado pela média *a posteriori* das cadeias dos efeitos dos marcadores. Isso é obtido atribuindo efeito nulo para marcas não visitadas em uma interação. Tomando $\hat{\gamma}$ como a média da cadeia de Markov após a convergência, temos $\hat{\gamma} = \frac{\sum_{l=1}^{N} \hat{\gamma}_l}{N} = \frac{\tau \sum_{l=1}^{n} \hat{\gamma}_l + \tau(N-\tau) \times 0}{\tau N} = \frac{\tau \bar{\hat{\gamma}}}{N} = \hat{P}_\lambda \bar{\hat{\gamma}}$ em que N é o tamanho total da cadeia de Markov, τ é o número de vezes que a marca foi selecionada durante o processo MCMC , $\hat{P}_\lambda = \frac{\tau}{N}$ é a frequência de visitas dos marcadores no genoma e $\bar{\hat{\gamma}}$ é a média dos efeitos dos marcadores que foram visitados. Assim, a predição do valor genético genômico final foi dado por $Z\hat{\gamma} = \hat{g}$.

2.4 Implementação da análise

O algoritmo para o amostrador de Gibbs e Metrópolis-Hastings foi implementado utilizando-se o *software* R (R CORE TEAM, 2016). Considerou-se, nas análises dos dados simulados, um número fixo de 10000 iterações. Para o ajuste do modelo aos dados simulados, foram descartadas as 2000 primeiras iterações para retirar o efeito dos valores iniciais e realizados saltos a cada duas iterações para reduzir a correlação nas amostras armazenadas, de modo que o número total de amostras mantidas foi de 2000 observações para análise pós-MCMC.

2.5 Acurácia preditiva

Além do modelo proposto, os dados simulados foram analisados utilizando os modelos Bayes B, por meio da função *BGLR* contida no pacote *BGLR* (PÉREZ; DE LOS CAMPOS,

2014) e o modelo rr-BLUP com a função *mixed.solve* (ENDELMAN, 2011) do pacote *rr-BLUP*, sendo que ambos são bibliotecas do *software* R (R CORE TEAM, 2016).

Para avaliar a capacidade preditiva dos modelos, utilizou-se o erro quadrático médio (EQM), interpretável como quanto menor melhor é o modelo, definido por:

$$\text{EQM} = \frac{1}{n}\sum_{i=1}^{n}\left(g_i - \hat{g}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{\mu} - Z\hat{\gamma}\right)^2 \qquad (24)$$

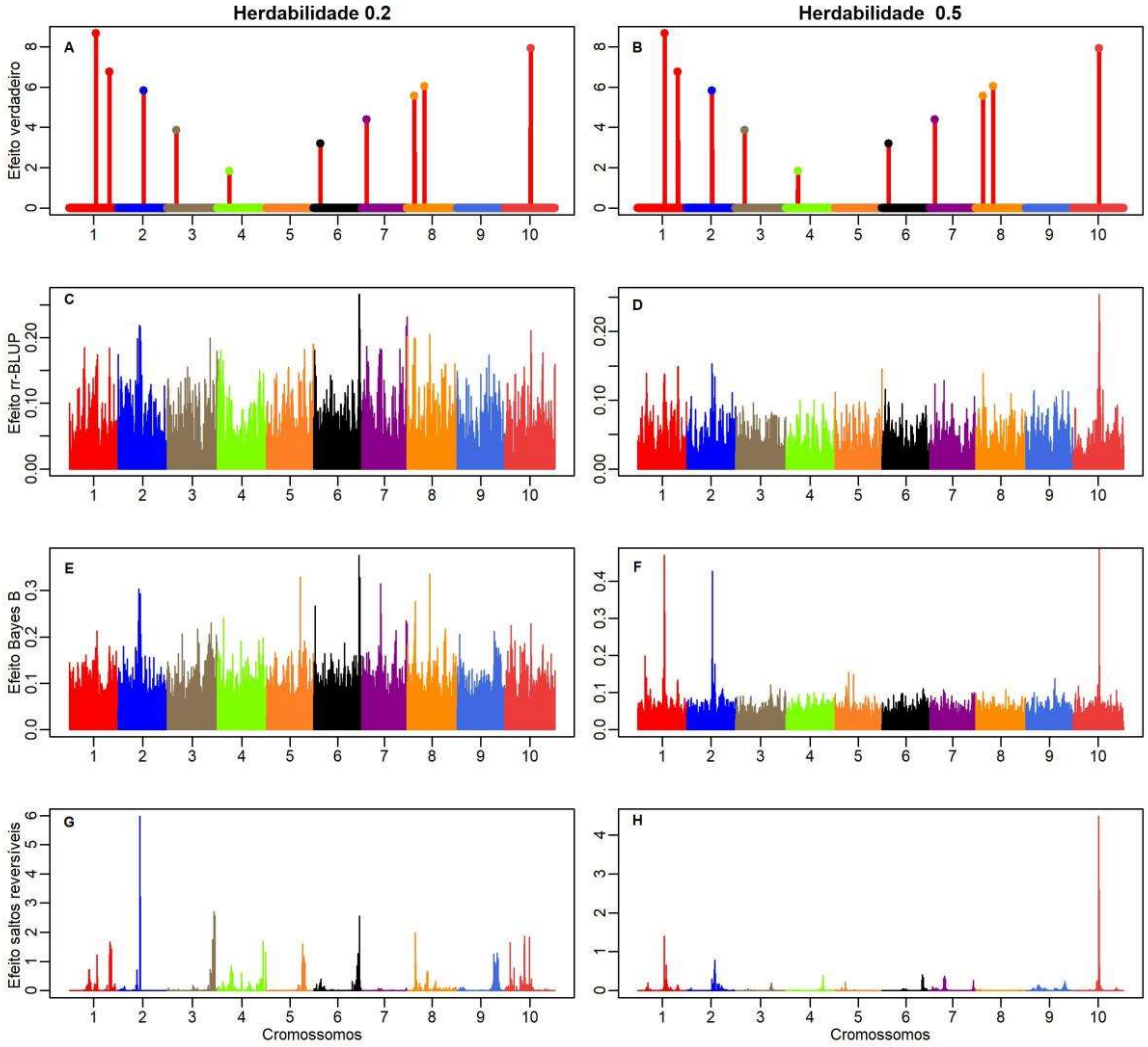
Estimou-se também o coeficiente de determinação R^2 da regressão linear entre valor genético predito $\hat{g}_i$ e verdadeiro (g).

3 RESULTADOS

3.1 Cenários simulado I: Modelo oligogênico

Na Figura 1 encontram-se os efeitos absolutos dos QTL simulados (painéis **A** e **B**) e os efeitos absolutos estimados de marcador a partir dos métodos rr-BLUP (painéis **C** e **D**), Bayes B (painéis **E** e **F**) e o modelo proposto via *Reversible Jump* (painéis **G** e **H**) para as herdabilidades 0,2 e 0,5. Nota-se que a maioria dos grandes QTL simulados foram mapeados pelos três métodos. No entanto, a resolução do método proposto foi superior aos demais, isto é, os efeitos que foram preditos pelo método *bin* proposto apresentaram estimativas mais próximas dos valores paramétricos.

Figura 1 - Efeitos simulados verdadeiros de QTL ao longo do genoma para as herdabilidades 0,2 e 0,5 (painéis A e B) e estimados a partir dos métodos, respectivamente, rr-BLUP (painéis C e D), Bayes B (painéis E e F) e o método proposto via *Reversible Jump* (painéis G e H). Pontos coloridos representam os 10 verdadeiros QTL distribuídos em 12150 SNP ao longo de dez grupos de ligação.

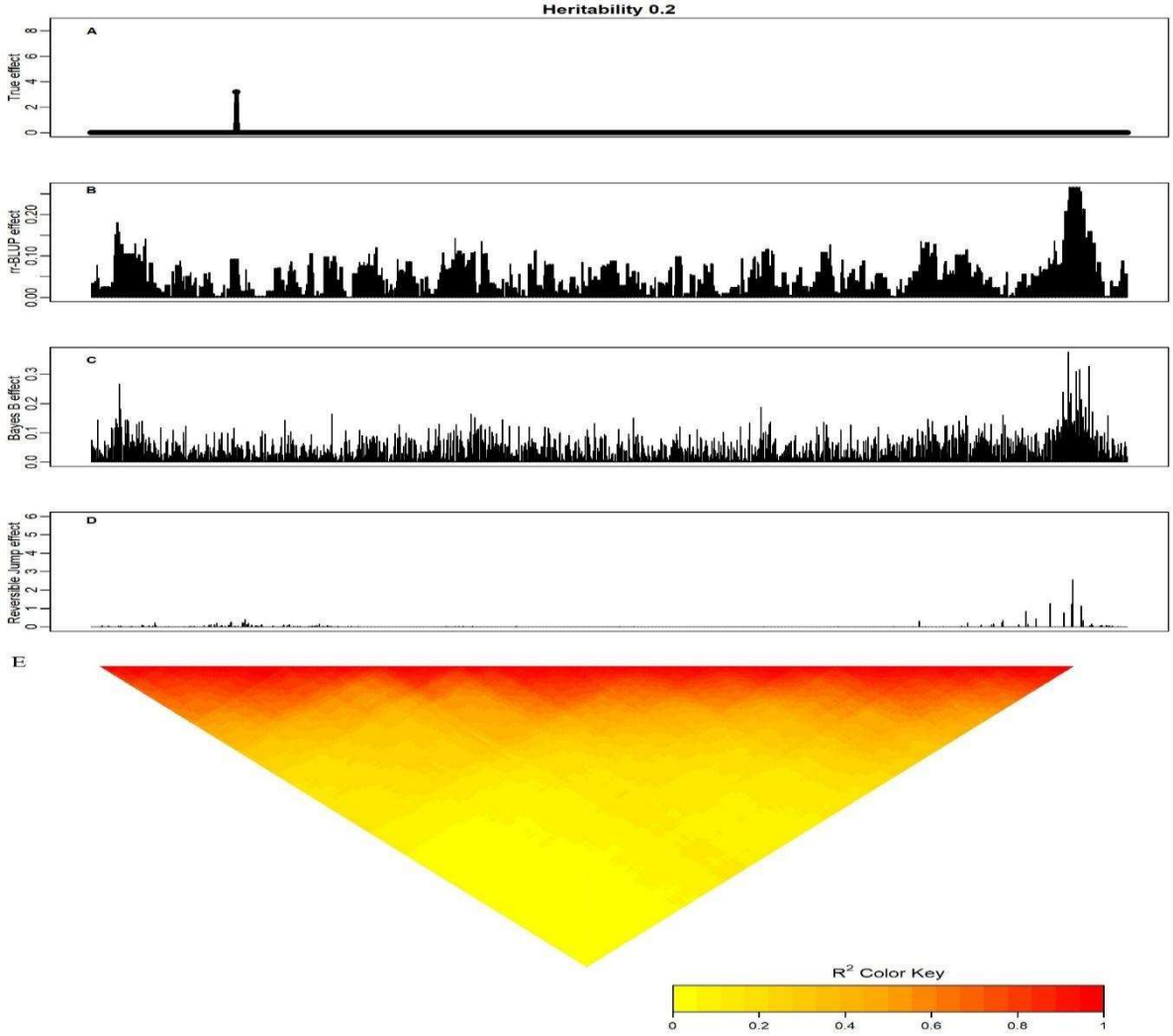


Fonte: Do autor (2021).

É importante destacar que diferentemente dos modelos tradicionais de SG, o método *bin* apresentou poucos sinais falso-positivos. Dessa forma, além da metodologia proposta apresentar uma melhor resolução, mostra sinais onde realmente existem. No entanto, todos os métodos apresentaram um falso positivo no final do cromossomo 6, pois o único QTL simulado nesse cromossomo é no início. Assim sendo, este fato pode ter ocorrido devido ao alto desequilíbrio de ligação entre as marcas contidas nesse grupo de ligação.

Para desafiar esta hipótese foi realizado um gráfico de calor (*LDheatmap*) cujos resultados são mostrados na Figura 2. O mapa de calor agrupa os marcadores com alto LD, em que quanto mais vermelhas as células dos marcadores, maior o desequilíbrio de ligação (no mesmo cluster) (MOURA; PAMPLONA; BALESTRE, 2019). Dessa forma, é possível observar que não há ponto de quebra no cromossomo 6, o que leva a pensar que o cromossomo 6 pode ser considerado como *bin* natural (XU, 2013b). No entanto, apenas a ligação entre as marcas no cromossomo não basta para explicar o efeito, que pode estar associado a um desequilíbrio complexo entre a segregação desta marca e uma combinação de outras ligadas a efeitos de QTL.

Figura 2 - Efeitos simulados verdadeiros de QTL ao longo do cromossomo 6 para a herdabilidade 0,2 (painel A) e estimados a partir dos métodos, respectivamente, rr-BLUP (painel B), Bayes B (painel C) e método proposto via *Reversible Jump* (painel D). O painel E mostra o mapa de calor (*LDheatmap*) relacionado ao padrão de desequilíbrio para o cromossomo 6. Quanto mais vermelho, maior será o desequilíbrio de ligação.



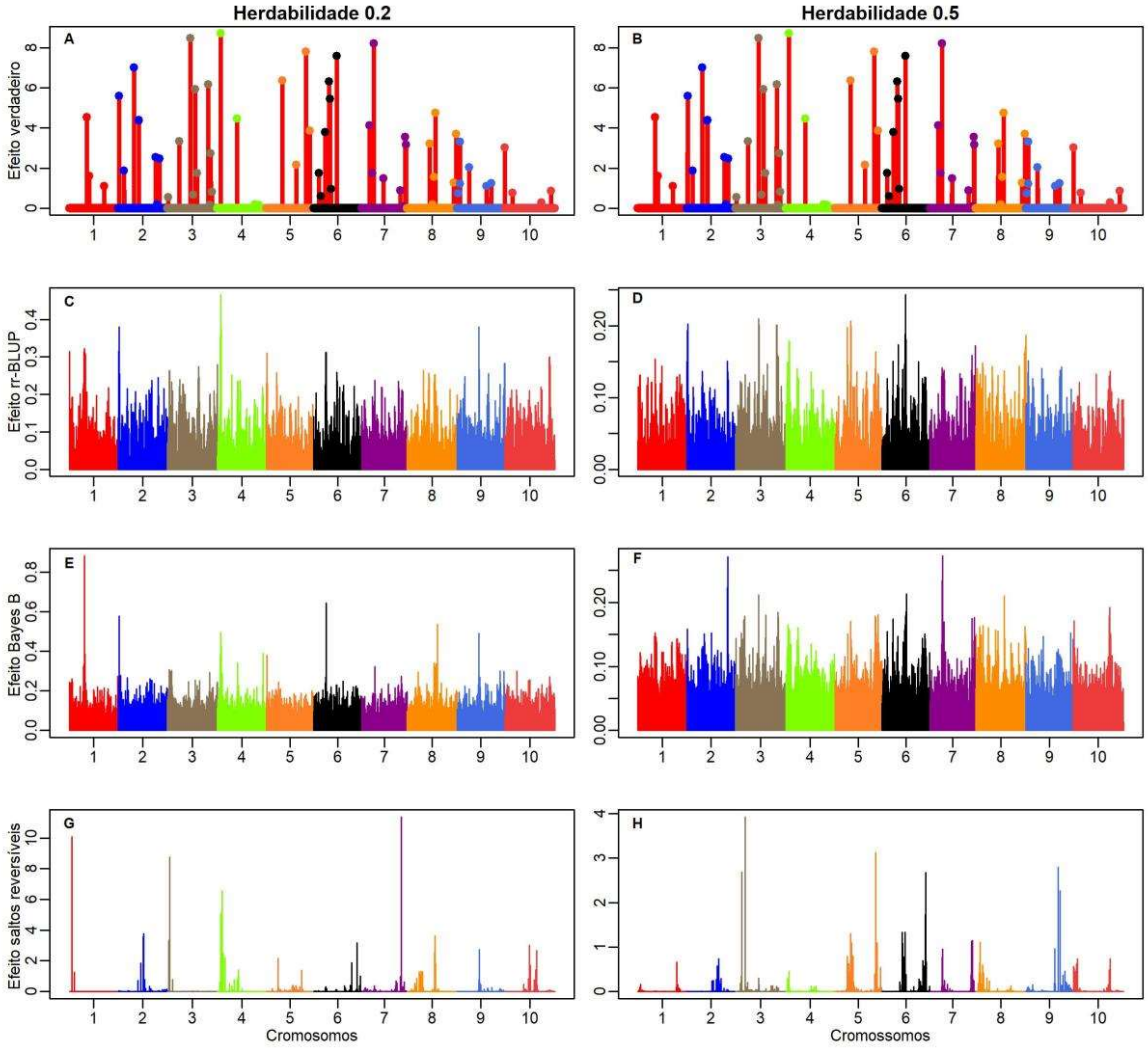
Fonte: Do autor (2021).

3.2 Cenário simulado II: Modelo poligênico

A Figura **3** inclui os efeitos absolutos dos QTL simulados (painéis **A** e **B**) e os perfis absolutos de efeitos de SNP estimados a partir dos métodos rr-BLUP (painéis **C** e **D)**, Bayes B (painéis **E** e **F)** e o modelo proposto via *Reversible Jump* (painéis **G** e **H)** para as duas herdabilidades 0,2 e 0,5. Em geral, os perfis de efeito SNP dos dois métodos concorrentes são semelhantes e apresentam muito ruídos. Em particular, o método proposto apresentou mais

sinais parecidos com os sinais reais, inclusive com magnitude de efeito o que evidência melhor resolução.

Figura 3 - Efeitos simulados verdadeiros de QTL ao longo do genoma para as herdabilidades 0,2 e 0,5 (painéis A e B) e estimados a partir dos métodos, respectivamente, rr-BLUP (painéis C e D), Bayes B (painéis E e F) e o método proposto via *Reversible Jump* (painéis G e H). Pontos coloridos representam os 60 verdadeiros QTL distribuídos em 12150 SNP ao longo de dez grupos de ligação.



Fonte: Do autor (2021).

A Tabela 1 mostra os erros quadráticos médios (EQM) e os coeficientes de determinação (R^2) entre os valores verdadeiros simulados (VGG) e os preditos pelos métodos avaliados, para as duas herdabilidades, nos cenários oligogênico (10 QTL) e poligênico (60 QTL).

Tabela 1 - Coeficiente de determinação (R^2) e Erro Quadrático Médio (EQM) entre os valores verdadeiros e preditos, usando os diferentes métodos nos cenários estudados.

População	**Modelos**	**Herdabilidades**			
		0,2		**0,5**	
		EQM	**R^2**	**EQM**	**R^2**
F2 (Oligogênico)	**rr-BLUP**	12,70	0,290	11,69	0,399
	Bayes B	12,62	0,298	11,51	0,417
	Reversible Jump	**12,37**	**0,326**	**11,04**	**0,463**
F2 (Poligênico)	**rr-BLUP**	**17,71**	**0,218**	**15,61**	**0,392**
	Bayes B	17,77	0,213	15,69	0,386
	Reversible Jump	17,92	0,199	15,96	0,365
Humanas	**rr-BLUP**	12,14	0,260	10,27	0,47
	Bayes B	12,17	0,255	9,288	0,566
	Reversible Jump	**12,07**	**0,268**	**9,151**	**0,579**

Fonte: Do autor (2021).

Em termos de acurácia, para população F_2 no cenário oligogênico, o método proposto obteve melhores resultados que rr-BLUP e Bayes B, nas duas herdabilidades. No entanto, para o cenário poligênico, rr-BLUP e Bayes B foram equivalentes e superaram o método proposto (apresentam valores de R^2 maiores). Todavia, adotando-se o EQM ao invés do R^2, como medida de acurácia seletiva, observa-se que os três métodos estudados foram praticamente equivalentes nos dois cenários de estudo simulados para esta população. Conforme o esperado, ao aumentar a herdabilidade, de 0,2 para 0,5, as precisões de todos os métodos aumentaram.

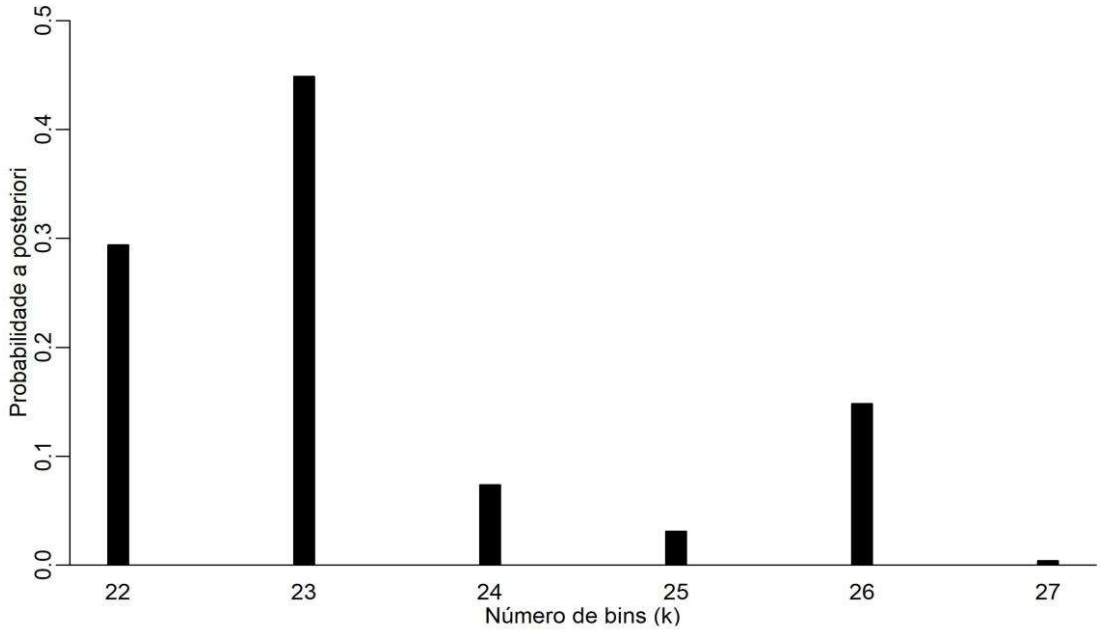
No conjunto de dados de humanos, observou-se que o método proposto foi o mais acurado nas duas herdabilidades. Entre os métodos de regressão clássicos usados neste estudo, apresentaram desempenho muito diferentes de acordo com o cenário de herdabilidade, ou seja, rr-BLUP teve melhor desempenho para o nível de herdabilidade 0,2, e Bayes B para o nível de herdabilidade 0,5. Em geral, o método *bin* via *Reversible Jump* obteve um desempenho muito bom nesta estrutura LD e com modelo mais parcimonioso, dado que foram necessários poucos *bins* para o método ter boa capacidade preditiva, conforme ilustra a Figura 4.

Em geral, o método proposto tem maior similaridade final com o modelo simulado, em especial para situações oligogênicas. Isto o tornaria mais útil para estudos de associação, embora este não seja o objetivo do trabalho. Os métodos consagrados rr-BLUP e Bayes B estabelecem

estimativas pouco realistas, mas usam toda a informação de marcadores e, portanto, têm sido usados para a SG com as devidas precauções de se fazer validação cruzada, dado o maior risco de *overfitting*.

Na Figura 4 são fornecidas as frequências *a posteriori* para dimensão de modelo, isto é, para número de *bins*, em que as maiores frequências se encontram entre os modelos com 22 e 23 *bins*. É importante ressaltar que para esse conjunto de dados (humanos) há 22 cromossomos (quantidade mínima possível de *bin*) e o modelo mais visitado foi de 23 *bins*. Para população F_2, a dimensão mais visitada ficou perto do dobro do número de cromossomos (resultado não mostrado).

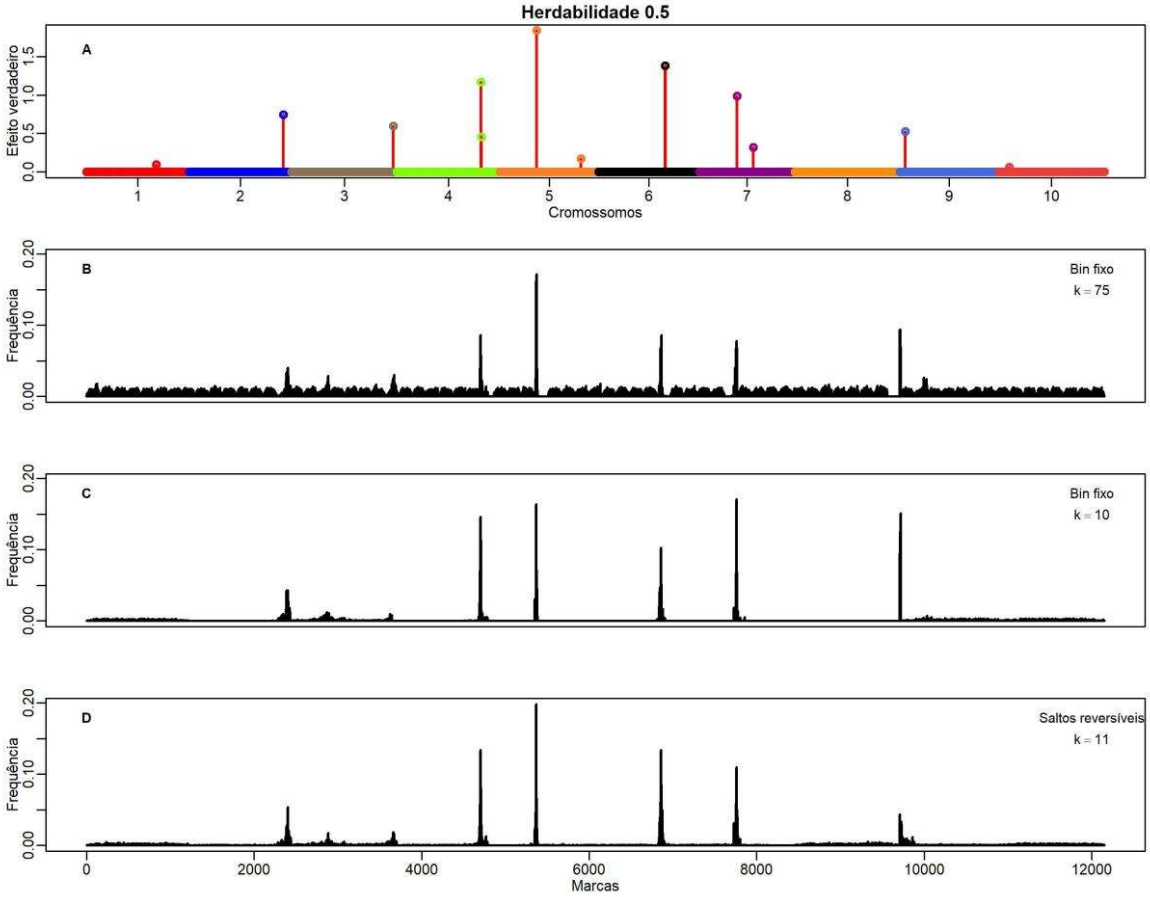
Figura 4 - Distribuição posterior do número de *bins* do Método *Reversible Jump* para dados de humanos.



Fonte: Do autor (2021).

Conforme mostrado em Moura, Pamplona e Balestre (2019), o número ou tamanho dos *bins* que dividem o genoma podem influenciar o resultado das análises, embora os perfis genômicos entre os métodos não sejam substancialmente divergentes. A Figura 5 mostra os efeitos absolutos de QTL simulado (painel **A**), o perfil de varredura em todo o genoma simulado (com as posições QTL) para método *bin fixo* (k=75, painel **B**), *bin fixo* (k=10, painel **C**) e *bin Reversible Jump* (k=11, painel **D**) , para o nível de herdabilidade de 0,5.

Figura 5 - Efeitos simulados e frequência relativa. O painel A mostra os efeitos simulados absolutos. O painel B mostra a frequência relativa do modelo *Bin* fixo (*K*=75), painel C mostra a frequência relativa do modelo *Bin* fixo (*K* =10) e no painel D mostra a frequência relativa do modelo *Bin* (*Reversible Jump*) para o nível de herdabilidade de 0,5.



Fonte: Do autor (2021).

Dada Figura 5, observa-se que o algoritmo percorre o genoma de forma aleatória e com peso constante até identificar o sinal de um possível QTL e gerar um pico agudo. Além disso, são observados pontos de alta frequência próximos aos picos principais. Observa-se também que ao alterar manualmente a quantidade de *bins* de 75 para 10, o perfil genômico não parece mudar muito e é parecido com perfil encontrado via *Reversible Jump*. Isso acontece por causa que os *bins* são adotados apenas para otimizar a pesquisa estocástica. No entanto, como já mencionado acima, o tamanho, consequentemente a quantidade de *bins* e a localização, determinam o sucesso dessa técnica.

4 DISCUSSÃO

Os avanços nas tecnologias de sequenciamento de última geração possibilitam tratar de modelos que eram apenas teorizados por Meuwissen, Hayes e Goddard (2001) e expandir suas possibilidades. O grande volume de dados de sequenciamento a alta densidade de marcadores e potencialmente com alto LD em populações sequenciadas também indica que a enorme maioria dos marcadores disponíveis são redundantes ou não fornecem informação sobre a segregação de QTLs (XU, 2013b; BANSAL; BOUCHER, 2019; MOURA; PAMPLONA; BALESTRE, 2019). Para fins de SG diversas estratégias têm sido utilizadas para tratar da alta dimensionalidade, da multicolinearidade e dos potencialmente elevados custos computacionais (PÉREZ; DE LOS CAMPOS, 2014; AN et al., 2020). O objetivo, no entanto, parece mais simples do que associar QTLs a regiões genômicas e preditores inesperados baseados em segregações não causais podem, por mero acaso, descrever muito bem os caracteres de interesse na população. Por outro lado, se há qualquer interesse em inferir causalidade, estratégias baseadas em dividir o genoma em *bins* parecem ser mais eficientes.

Su et al. (2017) define um *bin* como um bloco de desequilíbrio de ligação de (LD) em que todos os SNPs segregam de forma idêntica. Hu, Wang e Xu (2012) propuseram o modelo de genoma contínuo, que usou a informação média desses blocos em alto LD ao invés de marcador específico. *Bins* assim definidos, foram denominados por (XU, 2013b) de *bins* naturais. Tais *bins* na verdade são semelhantes a haplótipos. O autor propôs uma nova estratégia: montar *bins* artificiais, com pontos de quebra de interesse do investigador. Este método torna as análises mais simples, dado que a busca de *bins* naturais não é trivial e depende, dentre outros fatores, do tamanho da amostra e do número de marcadores.

O uso de *bins* artificiais leva a pequena perda de capacidade preditiva ao utilizar em relação a *bins* naturais ou outros métodos consagrados (RR-BLUP e Bayes B). Um problema passou a ser determinar (estimar) a quantidade ótima de *bins*, para que a perda de acurácia seja mínima (XU, 2013b). Moura, Pamplona e Balestre (2019) propuseram uma nova estratégia de *bins* artificiais. A proposta desses autores foi relaxar a suposição de efeito médio de *bin* e assumir os marcadores candidatos como uma variável aleatória discreta cuja função sinal genômica é desconhecida. Dessa forma, os blocos artificiais de SNP foram utilizados por esses autores apenas como estratégia de integração numérica e redução dimensional, motivo pelo

qual, caracterizaram seu modelo de pseudo-funcional. Embora obtivessem alta acurácia seletiva em diversos cenários de estudo simulado e dados reais, os autores não estudaram variações no número e tamanho de *bins*, embora tenham encontrado evidências de que estes parâmetros de ajuste determinam o sucesso das análises.

No presente estudo, desenvolveu-se uma estratégia semelhante às propostas anteriores de obtenção de *bins* artificiais. Nessa abordagem, adotou-se o número K de *bins* como uma variável aleatória e aproximou-se sua distribuição conjunta com os demais parâmetros do modelo via usando MCMC com RJ. Espera-se que as cadeias resultantes permitam inferência marginal sobre a dimensão do modelo e que se descreva melhor o padrão LD dos dados e sua associação com os QTLs. O algoritmo Reversible Jump implementado parece ser uma excelente ferramenta para integrar esse processo em uma única análise.

Um método similar denominado cBLUP (BLUP compacto) foi proposto por Wang et al. (2018) para derivar o parentesco entre os indivíduos em GWAS (*Genome-Wide Association Studies*) utilizado a estratégia de *bins*. De acordo com os resultados desses autores, em análises de dados simulados e reais esse método demonstrou flexibilidade para avaliar uma variedade de características. Para características com baixa herdabilidade, o cBLUP superou os métodos gBLUP (*Genomic best linear unbiased prediction*) e *Bayesian* LASSO. Contudo, para otimizar o tamanho e o número de *bin*, eles utilizaram o método de máxima verossimilhança ao invés de buscar inferir sobre a dimensão do modelo conjuntamente.

No presente estudo, as melhores dimensões de modelos encontradas para população F2 resultou em torno do dobro do número de cromossomo, nos dois cenários de herdabilidades. Para a população de humanos, a dimensão mais frequente na *posteriori* foi de apenas 23 *bins*, ou seja, um *bin* a mais do que o número mínimo estabelecido, que é a quantidade de cromossomos. Ou seja, não se sabe a *priori* de quão grande pode ser o tamanho e a quantidade de *bins* que apresentará bons resultados, embora o pesquisador saiba que o número, a localização e o tamanho desses blocos influencia a análise em termos de velocidade e poder de resolução (MOURA; PAMPLONA; BALESTRE, 2019). Uma maneira alternativa, seria realizar a análise em etapas, inicialmente condicionando a números pré-fixados de *bins* e realizar análise do perfil das verossimilhanças marginais das dimensões de modelo. Na Figura 5, apresenta-se os resultados de uma análise com K =75 (painel **A**) em que se verifica que há vários *bins* desnecessários que apenas acarretaram a demora na convergência das cadeias da *posteriori*

(maior tempo para concluir a análise). Visualmente se constata que K =10 (painel **B**) levaria a uma análise satisfatória. A nova análise com essa configuração melhorou muito a capacidade preditiva, além disso, obteve resultados parecidos aos obtidos via *Reversible Jump*.

Contudo, embora uma análise em duas etapas possa ser eficiente, não se justifica dado que o tempo gasto nas duas etapas é maior do que realizar a amostragem conjunta via *Reversible Jump*. Além disso, dividir o genoma em *bins* em tamanhos uniformes pode provocar quebra de regiões causais em janelas adjacentes, proporcionando perda de resolução de mapear tais regiões genômicas (BEISSINGER et al., 2015). Assim sendo, a busca automática via MCMC-RJ tende a ser mais eficiente que janelas artificiais fixas, pois, é permitido que a dimensão do modelo mude durante o processo de amostragem MCMC. Ou seja, a quantidade (K) de janelas genômicas é tratado como um problema de seleção de modelos e é inferido pela distribuição posterior de K. Assim, espera-se que marcas que tenham informações redundantes permaneçam numa mesma janela genômica. Esse processo evita criar um número grande de janelas desnecessárias que apenas super-parametrizam o modelo sem nenhum ganho de informação.

A metodologia proposta, funcionou bem para dados simulados de uma população F2 e para uma base de dados reais de humanos, mostrando-se competitiva em relação aos modelos de regressão convencional em relação a capacidade preditiva. O grande diferencial dessa técnica foi na alta resolução de mapear regiões causais, obtendo resultados muitos melhores nesse quesito (veja as Figuras 1 e 3). Esses resultados sugerem que embora o foco do estudo seja seleção genômica, poderia ser empregado no mapeamento de QTL. Além disso, como o número de *bins* pode ser muito menor que a quantidade SNP, necessita, portanto, de menor custo computacional. No entanto, não queremos afirmar que o algoritmo proposto é absolutamente ótimo. Podem ser feitas melhorias em alguns aspectos, por exemplo, quando um *bin* é excluído, todos os parâmetros associados a este *bin* desaparecem. Se um *bin* recém-adicionado ocupar exatamente o mesmo local que um *bin* excluído anteriormente, os parâmetros associados são amostrados novamente a partir da distribuição *a prior*. Além disso, o ganho em tempo computacional, não advém do algoritmo de salto reversível em si, que por sua vez, é computacionalmente intensivo, mas sim da técnica de redução dimensional que é dividir o genoma em *bins* artificiais.

REFERÊNCIAS BIBLIOGRÁFICAS

AN, B.; GAO, X.; CHANG, T.; XIA J., WANG, X.; MIAO, J.; XU, L.; ZHANG, L.; CHEN, Y.; LI, J.XU, S.; GAO, H. Genome-wide association studies using binned genotypes. **Heredity**, v. 124, p.288–298, 2020.

BALESTRE, M.; VON PINHO, R. G.; SOUZA JUNIOR, C. L.; BUENO FILHO, J. S. S. Bayesian mapping of multiple traits in maize: the importance of pleiotropic effects in studying the inheritance of quantitative traits. **Theoretical and Applied Genetics**, v. 125, n. 3, p. 479-493, 2012. doi: 10.1007/s00122-012-1847-1.

BANSAL, V.; BOUCHER, C. Sequencing Technologies and Analyses: Where Have We Been and Where Are We Going? **iScience**, v.18, p.37–41, 2019.

BEISSINGER, T. M.; ROSA, G. J.; KAEPPLER, S. M.; GIANOLA D.; LEON N. Defining window-boundaries for genomic analyses using smoothing spline techniques. **Genet. Sel. Evol.** v.47, n.30, 2015.

DE LOS CAMPOS G.; SORENSEN D.; GIANOLA D. Genomic Heritability: What Is It? **PLoS Genet.**, v.11, p.1–21, 2015.

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

FAN, R.; WANG, Y.; MILLS, J. L.; WILSON, A. F.; BAILEY-WILSON, J. E.; XIONG, M. Functional linear models for association analysis of quantitative traits. **Genet. Epidemiol.**, v.37, p.726–742, 2013.

FAN, H.; WU, Y.; FAN, H.; WANG, Y.; ZHANG, L.; GAO, X.; CHEN, Y.; LI, J.; REN, H.; GAO, H. Genome-Wide Association Studies Using Haplotypes and Individual SNPs in Simmental Cattle. **Plos One**, v.9, n.10, 2014. doi: https://doi.org/10.1371/ journal.pone.01 09330

FAN, R.; WANG, Y.; YAN, Q.; DING, Y.; WEEKS, D. E.; LU, Z.; REN, H.;COOK, R.J.;XIONG, M.; SWAROOP, A.; CHEW, E.Y.;CHEN, W. Gene-Based Association Analysis for Censored Traits Via Fixed Effect Functional Regressions. **Genet. Epidemiol.** v.40, p.133–143, 2016.

GONZÁLEZ, J. R.; ARMENGOL, L.; GUINÓ, E.; SOLÉ, X.; MORENO, V. (2014) **SNPassoc: SNPs-based whole genome association studies**, 2014. R package version 2.0.11 Disponível em: https://cran.r-project.org/web/packages/SNPassoc/SNPassoc.pdf

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

HASTINGS, B. Y. W. K. Monte Carlo sampling methods using Markov chains and their

applications. **Biometrika**, v.57, p.97–109, 1970.

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

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

LIU, J.; LIU, Y.; LIU, X.; DENG, H. W. Bayesian mapping of quantitative trait loci for multiple complex traits with the use of variance components. **Am. J. Hum. Genet.**, v.81, p.304–320, 2007.

METROPOLIS, N.; ROSENBLUTH, A. W.; ROSENBLUTH, M. N.; TELLER, A. H.; TELLER, E. Equation of state calculations by fast computing machines. **J. Chem. Phys.**, v.21, p.1087–1092, 1953.

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

MOURA, E. G.; PAMPLONA, A. K. A.; BALESTRE, M. Functional models in genome-wide selection. **Plos One**, v. 14, n. 10, p.1:27, 2019.

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

R CORE TEAM (2016). **R:** a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2016. Disponível em: . Acesso em: 1 jun. 2016.

RAMSAY, J. O.; DALZELL, C. J. Some Tools for Functional Data Analysis. **J. R. Stat. Soc.** v.53, p.539–572, 1991.

SILLANPÄÄ, M. J.; ARJAS, E. Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data. **Genetics**, v.148, p.1373–1388, 1998.

SU, C.; WANG, W.; GONG, S.; ZUO, J.; LI, S.; XU, S. High Density Linkage Map Construction and Mapping of Yield Trait QTLs in Maize (Zea mays) Using the Genotyping-by-Sequencing (GBS) Technology. **Front. Plant Sci.**, v.8, p.1–14, 2017.

WAAGEPETERSEN, R.; SORENSEN, D. A tutorial on reversible jump MCMC with a view toward applications in QTL-mapping. **Int. Stat. Rev.**, v.69, p.49–61, 2001.

WANG, J.; ZHOU, Z.; ZHANG, Z.; LI, H.; LIU, D.; ZHANG, Q.; BRADBURY, P.J.; BUCKLER, E.S.; ZHANG, Z. Expanding the BLUP alphabet for genomic prediction adaptable to the genetic architectures of complex traits. **Heredity**, v.121, p.648–662, 2018a. https://doi.org/10.1038/s41437-018-0075-0

WANG, X.; XU, Y.; HU, Z.; XU, C. Genomic selection methods for crop improvement: Current status and prospects. **Crop J.**, v.6, p.330–340, 2018b.

XU, S. Estimating polygenic effects using markers of the entire genome. **Genetics** v.163, p.789–801, 2003.

XU, S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. **Biometrics**, v.63, p.513–521, 2007.

XU, S. **Principles of Statistical Genomics.** Springer New York, New York, NY, 2013a. 428p.

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

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

YI, N.; XU, S.; ALLISON, D. B. Bayesian Model Choice and Search Strategies for Mapping Interacting Quantitative Trait Loci. **Genetics**, v.165, p.867–883, 2003.

YI, N. 2004 A unified Markov chain Monte Carlo framework for mapping multiple quantitative trait loci. **Genetics**, v.167, p.967–975, 2004.

ZUANETTI, D. A.; MILAN, L. A. Data-driven reversible jump for QTL mapping. **Genetics**, v. 202, p.25–36, 2016.

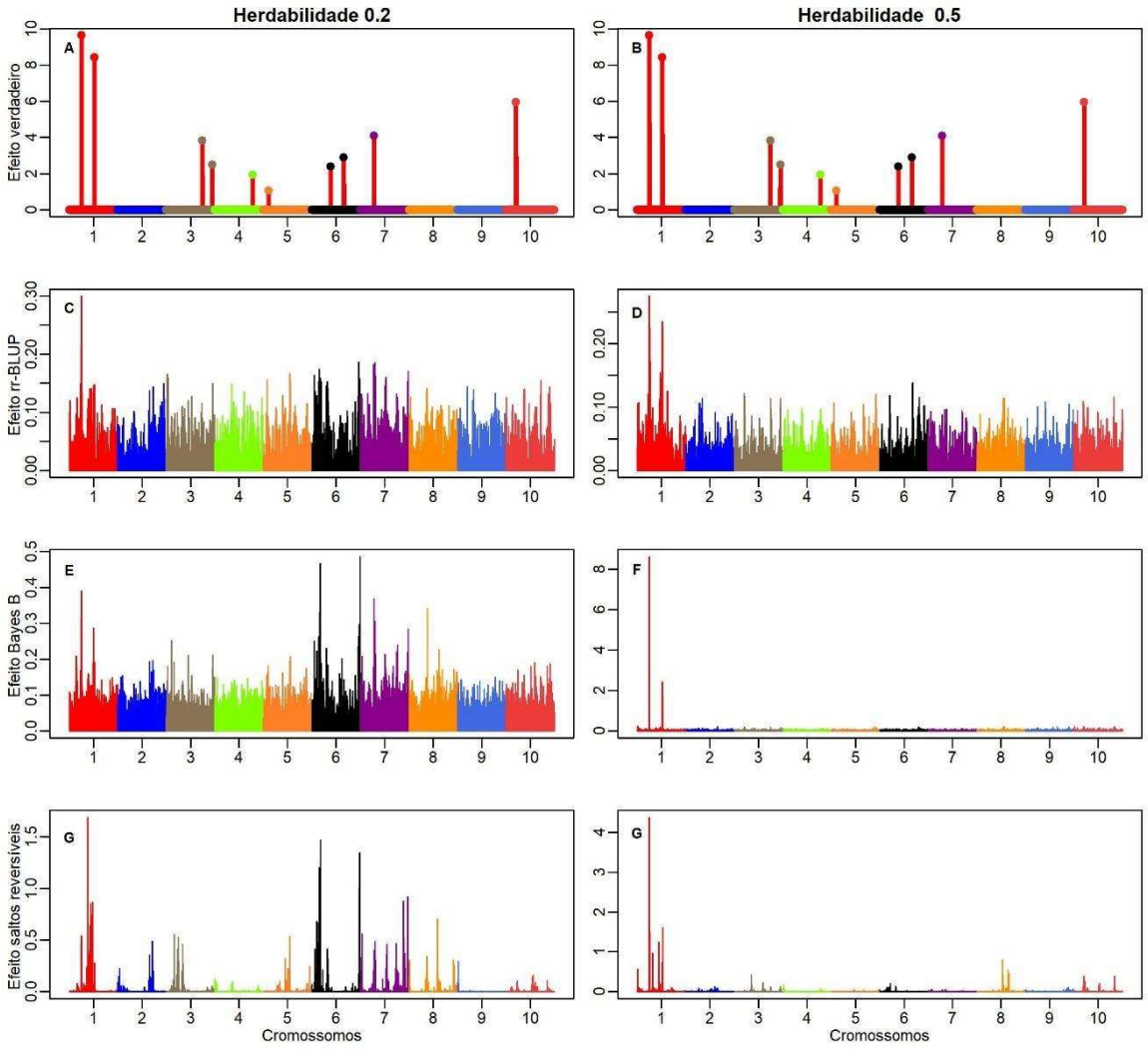
4 CONSIDERAÇÕES FINAIS

No presente trabalho, foram propostas duas formas de regressão funcional para a seleção genômica (GWS). As técnicas de representação do genoma em funções contínuas mostraram-se competitivas às técnicas tradicionais da GWS em termos de predição e têm menor custo computacional. Além disso, foi proposta uma abordagem funcional bayesiana para seleção genômica utilizando a busca estocástica de regiões causais por meio da divisão do genoma em *bins*. Mais ainda, estabelecemos uma estratégia de seleção automática de *bins* via MCMC com Saltos Reversíveis (MCMC-SR). Podemos concluir que modelos de regressão funcional são opções viáveis para a seleção genômica e merecem maior investigação.

APÊNDICE

Apêndice A. A seguir é apresentado um gráfico que ilustram os efeitos absolutos verdadeiros dos QTLs e os efeitos absolutos preditos de uma população F_{10}, num cenário oligogênico. Também é exibida uma tabela que representa a acurácia seletiva de todos os métodos apresentado neste estudo.

Figura A - Efeitos simulados verdadeiros de QTL ao longo do genoma para as herdabilidades 0,2 e 0,5 (painéis A e B) e estimados a partir dos métodos, respectivamente, rr-BLUP (painéis C e D), Bayes B (painéis E e F) e o método proposto via *Reversible Jump* (painéis G e H). Pontos coloridos representam os 10 verdadeiros QTL distribuídos em 10020 SNP ao longo de dez grupos de ligação.



Fonte: Do autor (2021).

Observa-se que os QTL que apresentaram grandes efeitos foram mapeados pelos três métodos. No entanto, em geral, os três métodos apresentaram forte efeito de encolhimento. Todavia, o método proposto, apresentou melhor resolução na distribuição genômica dos efeitos, apresentando menos ruído.

Tabela A - Coeficiente de determinação (R^2) e Erro Quadrático Médio (EQM) entre os valores verdadeiros e preditos, usando os diferentes métodos num cenário lisogênico para uma população F_{10}.

População	**Modelos**	**Herdabilidades**			
		0,2		**0,5**	
		EQM	**R^2**	**EQM**	**R^2**
F_{10} (Oligogênico)	**rr-BLUP**	**16,21**	**0,16,4**	13,77	0,397
	Bayes B	16,25	0,162	13,83	0,392
	Reversible Jump	16,40	0,146	**12,76**	**0,482**

Fonte: Do autor (2021).

Na população F_{10}, observou-se que os modelos de regressão utilizados neste estudo tiveram um desempenho muito diferente de acordo com o cenário de herdabilidade, ou seja, o rr-BLUP teve melhor desempenho para o nível de herdabilidade 0,2, já para herdabilidade 0,5 o método proposto foram substancialmente mais acurado que os demais.