ALINE MARÇAL ROSSINOL

ASSOCIAÇÃO GENÔMICA VIA REGIÕES CROMOSSÔMICAS SOB A ABORDAGEM BAYESIANA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientadora: Camila Ferreira Azevedo
Coorientadores: Ana Carolina C. Nascimento
Moysés Nascimento

VIÇOSA – MINAS GERAIS
2023

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T
R835a
2023

Rossinol, Aline Marçal, 1997-
Associação genômica via regiões cromossômicas sob a abordagem bayesiana / Aline Marçal Rossinol. – Viçosa, MG, 2023.
1 dissertação eletrônica (51 f.): il.

Orientador: Camila Ferreira Azevedo.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.150
Modo de acesso: World Wide Web.

1. Mapeamento cromossômico. 2. Estudo de associação genômica ampla. 3. Teoria bayesiana de decisão estatística. 4. Probabilidades. I. Azevedo, Camila Ferreira, 1988-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 572.8633

Bibliotecário(a) responsável: Alice Regina Pinto Pires CRB-6/2523

ALINE MARÇAL ROSSINOL

ASSOCIAÇÃO GENÔMICA VIA REGIÕES CROMOSSÔMICAS SOB A ABORDAGEM BAYESIANA

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 16 de fevereiro de 2023

Assentimento:

Documento assinado digitalmente
gov.br
ALINE MARCAL ROSSINOL
Data: 30/03/2023 09:44:32-0300
Verifique em https://validar.iti.gov.br

_____

Aline Marçal Rossinol
Autora

Documento assinado digitalmente
gov.br
CAMILA FERREIRA AZEVEDO
Data: 30/03/2023 10:06:28-0300
Verifique em https://validar.iti.gov.br

_____

Camila Ferreira Azevedo
Orientadora

Aos meus Pais, Hélio e Neuza, por sempre acreditarem em mim.

AGRADECIMENTOS

Agradeço a Deus pelo Dom da vida, por iluminar meus caminhos e me dar forças para não desistir perante os desafios.

Aos meus pais, Hélio e Neuza, pelo apoio incondicional e por todos os valores que me ensinaram, os quais levarei para toda vida. A minha irmã, Amanda, pelo apoio e por estar sempre ao meu lado, independentemente da situação.

Aos meus tios, tias, primos e primas pelo carinho e pela torcida, por serem meu apoio, mesmo a quilômetros de distância.

Ao Caio, por toda paciência, carinho e amor que teve comigo durante a trajetória, por ser meu braço direito nessa caminhada.

Aos amigos de longa data do Ensino Fundamental e Médio, em especial a Beatriz, Bruno e Mylena por toda a paciência e carinho que têm comigo.

Aos amigos que a Matemática me trouxe de presente, Daniela, Fernanda, Mauricio, Lucas B, Patrícia, Lucas C, Cesar, Luciano, Vinicius, Nilson e Davi, por tornarem a caminhada até o mestrado mais leve e por me aguentarem nos dias mais difíceis dessa nova caminhada, sempre com palavras de conforto e incentivo.

As minhas madrinhas Isabela e Diane, por todo apoio e incentivo para que eu pudesse chegar até aqui e a Joana Mariani por acreditar e contribuir para o desenvolvimento dessa jornada.

Aos funcionários do Departamento de Estatística e aos docentes do Programa de Pós-graduação em Estatística Aplicada e Biometria pela disponibilidade e os ensinamentos.

Aos integrantes dos Laboratório de Inteligência Computacional e Aprendizado Estatístico companheirismo, amizade, apoio e pelo aprendizado.

A minha orientadora, pela paciência, pelo incentivo, pelos ensinamentos, conselhos, críticas e sugestões que foram essenciais para o meu crescimento.

Aos meu coorientadores, por todo o ensinamento e contribuições durante essa etapa.

A banca composta pelos Professores Doutores Leísa Pires Lima e Moysés Nascimento por todas as contribuições para a pesquisa.

Aos que, de alguma forma, direta ou indiretamente, ajudaram: Meu muito obrigada! Vocês fazem parte de mais essa vitória!

O presente trabalho foi realizado com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e com o auxílio do Cluster da UFV (Universidade Federal de Viçosa).

"Todos os nossos sonhos podem se tornar realidade
se tivermos coragem para persegui-los"
(Walt Disney)

RESUMO

ROSSINOL, Aline Marçal, M.Sc., Universidade Federal de Viçosa, fevereiro de 2023. **Associação Genômica Via Regiões Cromossômicas Sob A Abordagem Bayesiana**. Orientadora: Camila Ferreira Azevedo. Coorientadores: Ana Carolina Campana Nascimento e Moysés Nascimento.

Com os avanços na biotecnologia se tornou possível novas descobertas na área da biologia molecular, o que favorece cada vez mais os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS). A GWAS utiliza marcadores moleculares, principalmente, os SNPs (*Single Nucleotide Polymorphism*), tendo como objetivo identificar as variantes causais no genoma e investigar as regiões do cromossomo em que estas variantes se encontram. Um dos principais métodos estatísticos em GWAS é o método via regressão em marcas únicas, que visa estudar a associação entre o fenótipo e um único marcador. No entanto, esse método apresenta problemas estatísticos, como, por exemplo, necessidade de grandes amostras e alta taxa de falsos positivos. Atualmente, os métodos utilizando grupos de marcadores vêm ganhando cada vez mais destaque, devido ao fato de que, os marcadores moleculares podem estar em alto desequilíbrio de ligação (*Linkage Disequilibrium* – LD) entre si e, com isso, influenciar conjuntamente o fenótipo. Um desses métodos é o Mapeamento de Herdabilidade Regional (*Regional Heritability Mapping* - RHM). Atualmente, os estudos de associação sob abordagem Bayesianas e utilizando grupos de marcadores, ou regiões genômicas, vêm ganhando cada vez mais destaque. Isto ocorre devido ao fato de que nesses métodos é possível estimar simultaneamente os efeitos dos marcadores ao invés de um único marcador, o que reduz a taxa de falsos positivos. A literatura ainda não apresenta nenhuma proposta sobre o método de RHM sob uma abordagem bayesiana e também sobre a estimação simultânea dos efeitos das regiões em um único modelo. Desta forma, no capítulo 1 desta dissertação é realizada uma revisão de literatura sobre as metodologias estatísticas utilizadas. O capítulo 2 visa comparar a eficiência de se estimar o efeito de todas as regiões genômicas simultaneamente através de um modelo bayesiano em relação ao procedimento de se estimar o efeito de cada região por vez através de dados simulados e depois para elucidar a utilização deste modelo nos programas de melhoramento, as estimações usando uma única região e todas as regiões simultaneamente também foram performadas em dados reais de arroz *Oryza sativa*. Esse estudo utilizou dados simulados através do pacote AlphaSimR e dados de arroz provenientes do *Rice Diversity Project*. O tamanho das regiões foi determinado como sendo a distância na qual o LD é metade

do seu valor máximo e, para verificar se as regiões eram associadas as características fenotípicas, foi utilizada a Probabilidade *a Posteriori* da Associação da Janela (*Window Posterior Probability of Association* - WPPA). Para os dados simulados, a eficiência da estimação simultânea dos efeitos das regiões genômicas utilizando a estimação bayesiana, apresentou resultados superiores. Nos dados de arroz, a estimação simultânea detectou uma quantidade superior de regiões já relatadas na literatura em detrimento a estimação única, além de apresentar novas regiões genômicas que podem ser estudadas em analises pós-GWAS. Essa é uma metodologia que apresenta potencial para aplicação, descoberta e investigação de novas regiões genômicas associadas a características fenotípicas.

Palavras-chave: Mapeamento de Herdabilidade Regional. Poder de detecção. Probabilidade *a posteriori*

ABSTRACT

ROSSINOL, Aline Marçal, M.Sc., Universidade Federal de Viçosa, February 2023. **Genome-Wide Association Studies via Bayesian Analysis of Chromosomal Regions.** Adviser: Camila Ferreira Azevedo. Co-advisers: Ana Carolina Campana Nascimento and Moysés Nascimento.

With advances in biotechnology, it has become possible to make discoveries in molecular biology, which increasingly favor Genome-Wide Association Studies (GWAS). GWAS uses molecular markers, mainly Single Nucleotide Polymorphisms (SNPs), to identify causal variants in the genome and investigate the chromosome regions where these variants are located. One of the main statistical methods in GWAS is the single-marker regression method, which aims to study the association between the phenotype and a single marker. However, this method has statistical problems, such as the need for large samples and a high rate of false positives. Currently, methods using groups of markers are gaining more prominence because molecular markers may be in high linkage disequilibrium (LD) with each other and thereby jointly influence the phenotype. One of these methods is Regional Heritability Mapping (RHM). Currently, Bayesian approaches using groups of markers or genomic regions are gaining more and more attention. This increase is because these methods make it possible to estimate the effects of markers simultaneously instead of just one marker, which reduces the false positive rate. The literature still needs to present a proposal for the RHM method under a Bayesian approach, nor for the simultaneous estimation of the effects of regions in a single model. Thus, in chapter 1 of this dissertation, a literature review is conducted on the statistical methodologies used. Chapter 2 compares the efficiency of estimating the effect of all genomic regions simultaneously using a Bayesian model versus the procedure of estimating the effect of each region one at a time using simulated data. Estimations using a single region and all regions simultaneously were also performed on real rice data from the Rice Diversity Project To elucidate the use of this model in breeding programs. This study used simulated data through the AlphaSimR package and rice data. The size of the regions was determined as the distance at which LD is half of its maximum value, and to verify whether the regions were associated with phenotypic characteristics, the Window Posterior Probability of Association (WPPA) was used. For simulated data, the efficiency of simultaneously estimating the effects of genomic regions using Bayesian estimation showed superior results. In rice data, simultaneous estimation detected a higher number of regions already reported in the literature compared to single estimation, as well as presenting new genomic regions that can be studied in post-GWAS

analyses. This methodology has the potential for application, discovery, and investigation of new genomic regions associated with phenotypic traits.

Keywords: Regional Heritability Mapping. Detection Power. Posterior Probability

SUMÁRIO

1. INTRODUÇÃO GERAL

A biotecnologia vem a cada dia mais colaborando e realizando novas descobertas na área da biologia molecular. Devido a isso, é possível que sejam utilizadas informações que venham diretamente do DNA, através de marcadores moleculares. Esses permitem que se tenha amplas informações sobre o genoma dos indivíduos, tornando-se possível os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS) (MEUWISSEN et al. 2001; UFFELMANN et al., 2021). Estes estudos buscam identificar as variantes causais no genoma que afetam uma determinada característica por meio dos marcadores moleculares, principalmente, o SNP (*Single Nucleotide Polymorphism*). Um dos primeiros métodos estatísticos aplicados a GWAS foi o método via regressão em marcas únicas, que visa estudar a associação entre o fenótipo e um único marcador (ZINGLER et al., 2008). No entanto, esse método apresenta problemas estatísticos, como, por exemplo, a necessidade de grandes amostras, alta taxa de falsos positivos e baixo poder de detecção (FERNANDO et al., 2004).

Com isso, estudos que consideram grupos de marcadores, as denominadas regiões genômicas, visando buscar associações entre elas e o fenótipo, vêm ganhando cada vez mais destaque na GWAS (FERNANDO et al., 2017). O mapeamento de herdabilidade regional (*Regional Heritability Mapping* - RHM), proposto por Nagamine et al. (2012), é uma abordagem que utiliza modelos mistos e que é baseada em regiões. Como essa abordagem utiliza regiões, isso faz com que se aumente o desequilíbrio de ligação (*Linkage Disequilibrium* - LD) entre as regiões e os QTL (*Quantitative Trait Loci*) reduzindo as taxas de falsos positivos e aumentando o poder de detecção de QTLs (USAI et al., 2014). No entanto, nesta metodologia, estima-se o efeito de uma região do genoma por vez. Já os métodos bayesianos propostos inicialmente para a predição genômica, como o BayesDπ, também podem ser utilizados para analisar a associação entre as regiões genômicas e o fenótipo, como apresentado por Fernando e Garrick (2013). Nesta abordagem, os efeitos individuais de cada marcador são estimados simultaneamente, e posteriormente, em posse das médias *a posteriori* dos efeitos dos marcadores se constrói e avalia-se as regiões como apresentado por Lima et al. (2022).

Considerando o melhor de nosso conhecimento, até o momento, não há relatos na literatura de estudos de GWAS de uma proposta similar ao RHM, mas que se estimam os efeitos de todas as regiões diretamente e simultaneamente no modelo. Procedendo com a análise de forma simultânea, os efeitos de todas as regiões teriam somente uma única fonte de erro, o que pode permitir o aumento do poder de detecção de regiões associadas. No entanto, a realização da estimação simultânea considerando modelos lineares mistos, sob a abordagem REML/BLUP

(*Restricted Maximum Likelihood/Best Unbiased Linear Prediction*) com matrizes de covariância estruturadas, podem sofrer de problemas de convergência quando os modelos são super-parametrizados, causando sérios problemas nas inferências dos parâmetros (BATES et al., 2015). Já os métodos bayesianos parecem não sofrer com problemas de super-parametrização, pois utilizam informações das distribuições *a priori* além das informações dos dados amostrais (GAMERMAN e LOPES, 2006).

Neste contexto, o capítulo 1 desta dissertação consiste em uma revisão de literatura, em que é apresentado a definição e a importância dos estudos de GWAS. Neste capítulo são apresentados detalhes teóricos da análise via marcas únicas, dos estudos utilizando regiões genômicas e do mapeamento de herdabilidades regionais. São apresentadas vantagens teóricas de se trabalhar com regiões genômicas utilizando metodologias bayesianas, em que essas podem ser aplicadas utilizando o método BayesDπ e o método RHM sob o enfoque bayesiano. Também é apresentada uma descrição do critério utilizado nos estudos para encontrar as regiões associadas.

O capítulo 2 compara a eficiência de se estimar o efeito de todas as regiões genômicas simultaneamente através de um modelo bayesiano em relação ao procedimento de se estimar o efeito de cada região por vez por meio de dados simulados e depois para elucidar a utilização deste modelo nos programas de melhoramento, as estimações única e simultânea foram aplicadas a dados reais de arroz *Oryza sativa*.

2 REVISÃO DE LITERATURA

2.1 Definição e Importância

A biotecnologia vem a cada dia mais facilitando e proporcionando novas descobertas aos estudos na área da biologia molecular. Devido a isso, atualmente é possível que sejam utilizadas informações que venham diretamente do DNA, através de marcadores moleculares. Esses permitem que se tenha ampla informação sobre o genoma dos indivíduos e são economicamente mais viáveis. Com isso, se tornou possível estudos que relacionam fenótipos e genótipos, favorecendo estudos de seleção e associação genômica ampla (*Genomic Wide Selection* – GWS e *Genome Wide Association Studies* – GWAS, respectivamente) (MEUWISSEN et al. 2001; UFFELMANN et al., 2021).

Os marcadores moleculares que veem sendo utilizados em estudos de GWS e de GWAS são do tipo SNP (*Single Nucleotide Polymorphism*), pois estes são a forma mais abundante de variação do DNA no genoma, são codominantes e sua genotipagem pode ser realizada em larga escala. A utilização destes marcadores permitiu a obtenção de informações sobre a variabilidade genética, a identificação e a localização de genes específicos associados a características de interesse dos programas de melhoramento (RESENDE et al., 2012).

Devido a elevada densidade destes marcadores, a probabilidade de que um loco de característica quantitativa (*Quantitative trait locus* - QTL) esteja em desequilíbrio de ligação (*Linkage Disequilibrium* – LD) com pelo menos um marcador é muito alta. Esses marcadores em LD com os QTLs explicarão quase a totalidade da variação genética de um caráter quantitativo, podendo ser em grandes ou pequenos efeitos. Portanto, sob a suposição de existência do LD entre marcadores e QTL, a associação entre QTL e valor genético pode ser acessada indiretamente por meio da associação de marcadores moleculares e valores fenotípicos (RESENDE et al., 2012). Essa associação favorece os estudos de GWAS, que buscam a associação entre o QTL e os valores genéticos dos indivíduos referentes a uma determinada característica.

Os estudos de GWAS foram inicialmente desenvolvidos e validados para estudos epidemiológicos em humanos (MCCARTHY et al., 2008), mas já vem sendo consolidados e utilizados no melhoramento vegetal e animal (ARORA et al. 2019; QUERO et al. 2018; MÜLLER et al. 2018; ZHANG J. et al. 2016; CHENG et al.,2022). Várias abordagens estatísticas vêm sendo utilizadas, como as análises via marcas únicas, análises via modelos lineares mistos, modelos de haplótipos, modelos mistos baseados em genealogia e modelos de seleção de marcadores via abordagens bayesianas (FERNANDO e GARRICK, 2013; ZHOU et al., 2014; WU et al., 2014; GUO et al., 2016; BENNEWITZ et al., 2017; FERNANDO et al.,

2017; BRAZ et al., 2019; CHENG et al.,2022). Um dos primeiros e principais métodos estatísticos aplicados a GWAS é o método via regressão em marcas únicas, que visa estudar a associação entre o fenótipo e um único marcador a cada vez (ZINGLER et al., 2008). No tópico seguinte, será apresentada a descrição metodológica do método de análise via marcas únicas.

2.2. Análise via Marcas Únicas

O método estatístico com teoria mais simples aplicado a GWAS é o método via marcas únicas cujo modelo utilizado é o de regressão linear simples:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{W_i} m_i + \boldsymbol{e}$$

em que:

y é o vetor de informações fenotípicas;

$\boldsymbol{\beta}$ é vetor de efeitos fixos, como a média e os efeitos de estrutura de populações e $\boldsymbol{X}$ é a matriz de incidência dos efeitos fixos;

m_i é o efeito do i-ésimo marcador (considerado fixo) sendo $i = 1,2,\dots,n$ e n o número de marcadores;

$\boldsymbol{W_i}$ é o vetor de incidência para o i-ésimo marcador (composto por 0, 1 e 2);

$\boldsymbol{e}$ é o vetor de erros aleatórios com $\boldsymbol{e} \sim N(0, I\sigma_e^2)$ sendo σ_e^2 a variância residual.

Como se trata de um modelo de regressão linear simples, a média geral e o efeito do i-ésimo marcador são estimados através do método dos mínimos quadrados ordinários, utilizando as equações a seguir:

$$\begin{bmatrix} \boldsymbol{X^T X} & \boldsymbol{X^T W}_i \\ \boldsymbol{W}_i^T \boldsymbol{X} & \boldsymbol{W}_i^T \boldsymbol{W}_i \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{m_i} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X^T y} \\ \boldsymbol{W}_i \boldsymbol{y} \end{bmatrix}.$$

Após a estimação do efeito do i-ésimo marcador ($\widehat{m}_i$), é necessária a realização de um teste de hipótese para verificar se de fato aquele efeito do marcador sobre o fenótipo de interesse é significativo. O teste utilizado para este propósito é o teste *t-student* (RESENDE et al., 2012), cuja a hipótese de nulidade (H_0) é definida como "o i-ésimo marcador não apresenta qualquer efeito sobre o fenótipo ($H_0: m_i = 0$)" e a hipótese alternativa (H_a) é definida como "o i-ésimo marcador tem efeito sobre o fenótipo ($H_0: m_i \neq 0$)", ou seja, o i-ésimo marcador e o QTL encontram-se em LD para aquele determinado fenótipo.

O procedimento de estimação e de teste de hipótese é repito por n vezes, ou seja, repetido para cada um dos marcadores contidos no banco de dados genotípicos. Desta forma, é necessária a realização de múltiplos testes, o que ocasiona nesta análise estatística um alto índice de falsos positivos, ou seja, considerar que o efeito do marcador sobre o fenótipo é

significativo quando não é (RESENDE et al., 2012, SAHANA et al., 2010). Na prática isso significa que se concluirá que um determinado marcador afeta um determinado fenótipo, quando isso não ocorre. Uma forma de contornar esse problema estatístico foi apresentada por Fernando et al. (2004), esta consiste em controlar o número de falsos positivos em relação ao número total de resultados positivos por meio da taxa de descobertas falsas (*False Discovery Rate* - FDR). Para considerar a FDR no teste de significância é realizada uma correção no *p-value* associado ao teste, denominado de *q-value* (STOREY e TIBSHIRANI, 2003).

Os marcadores moleculares dentro do genoma podem estar altamente em LD e influenciar conjuntamente o fenótipo. Desta forma, um marcador sozinho pode não apresentar uma forte associação com o fenótipo de interesse, o que faz com que o efeito deste seja difícil de ser identificado no fenótipo, ou seja, o método via marcas únicas pode apresentar baixo poder de detecção. Assim, estudos como o de Fernando et al. (2017), Li et al. (2021) e Lima et al. (2022) recomendam a utilização de grupos de marcadores, as denominadas regiões genômicas, que possibilitam com que se capture maior parte da variabilidade genética e aumente o poder de detecção de QTL. Para isso, os tópicos a seguir contemplam maneiras de se construir regiões genômicas e diferentes abordagens estatísticas para a análise destas regiões.

2.3. Estudos de Regiões Cromossômicas

Os estudos de GWAS baseados em regiões genômicas foram propostos para superar o problema de que um marcador sozinho pode apresentar uma baixa associação com o fenótipo de interesse e, consequentemente, a análise apresentar baixo poder de detecção (LI et al., 2021). Desde então, muitas abordagens estatísticas começaram a serem desenvolvidas para a realização de inferências baseadas em regiões genômicas, podendo essas serem sobrepostas ou não. As metodologias podem utilizar a inferência clássica, como a metodologia de modelos mistos (NAGAMINE et al., 2012; SUELA et al., 2022; RESENDE, 2017) e a inferência Bayesiana (HAYES et al., 2010; FAN et al., 2011; LEGARRA et al., 2018).

Algumas estratégias apresentadas pela literatura para a definição de regiões são: i) Após a construção de um gráfico de decaimento de LD entre marcadores (r^2, que representa estimativas pareadas do LD) e o ajuste de uma curva que represente a relação entre o LD e a distância dos marcadores no genoma, uma alternativa seria definir o tamanho da região como sendo a distância que fornece o LD estimado entre marcadores igual a $0{,}20$. Kim et al. (2007), Lam et al. (2010) e Vos et al. (2017) consideram que abaixo de $0{,}20$, não há desequilíbrio de ligação entre os marcadores, ou seja, os marcadores se encontram em equilíbrio de ligação; ii)

Também após a construção de um gráfico de decaimento de LD, o tamanho da região seria definido como a distância na qual o LD estimado é metade do seu valor máximo conforme avaliado por Suela et al. (2022); ou iii) cada região deve possuir a mesma quantidade de marcadores, como proposto por Li et al. (2021), logo se teria regiões genômicas de diferentes tamanhos.

Após a descrição da construção destas regiões genômicas, é importante descrever como avaliar estas regiões quanto a associação das mesmas com os fenótipos de interesse. Desta forma, a seguir, serão apresentadas uma metodologia baseada em modelos mistos e duas metodologias baseadas em inferência bayesiana.

2.4. Mapeamento de herdabilidades regionais (*Regional Heritability Mapping* - RHM)

O método de mapeamento de herdabilidades regionais (*Regional Heritability Mapping* - RHM) foi proposto por Nagamine et al. (2012) e é baseado na teoria de modelos mistos. É uma das principais metodologias estatísticas aplicadas a GWAS que utilizam em seus procedimentos as regiões genômicas. O modelo da RHM visa a estimação do efeito de uma única região no fenótipo e é definido por meio de:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_k r_k} + \boldsymbol{e}$$

em que:

y é o vetor de informações fenotípicas;

$\boldsymbol{\beta}$ é vetor de efeitos fixos, como a média e os efeitos de estrutura de populações e $\boldsymbol{X}$ é a matriz de incidência dos efeitos fixos;

$\boldsymbol{r_k}$ é o vetor de efeitos genômicos aditivos de indivíduos referente a k-ésima região do genoma com matriz de incidência $\boldsymbol{Z_k}$, sendo $\boldsymbol{r_k} \sim N\left(0, \boldsymbol{G_{r_k}}\sigma^2_{r_k}\right)$ em que $\boldsymbol{G_{r_k}}$ é a matriz de parentesco genômico e $\sigma^2_{r_k}$ é a variância genética referentes a k-ésima região ($k = 1,2, \dots, K$ e K é o número de regiões genômicas formadas);

$\boldsymbol{e}$ é o vetor de erros aleatórios com $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma^2_e)$ sendo σ^2_e a variância residual.

A matriz $\boldsymbol{G_{r_k}}$ utiliza a matriz de incidência dos marcadores pertencentes apenas a k-ésima região, e é dada por:

$$\boldsymbol{G_{r_k}} = \frac{\boldsymbol{W_{r_k} W'_{r_k}}}{\sum_{i=1}^{n_k} 2p_i(1-p_i)}$$

em que:

$\boldsymbol{W_{r_k}}$ é a matriz de incidência dos marcadores pertencentes a k-ésima região;

p_i é a frequência alélica associada ao i-ésimo SNP;

n_k é o número de SNPs pertencentes à k-ésima região.

Para a estimação dos efeitos genéticos da k-ésima região do genoma sobre o fenótipo são utilizadas as equações de modelos mistos propostas por Henderson (1973) e dadas por:

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z_k} \\ \boldsymbol{Z'_kX} & \boldsymbol{Z'_kZ_k + G^{-1}_{r_k}} \frac{\sigma^2_e}{\sigma^2_{r_k}} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\boldsymbol{r_k}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'_ky} \end{bmatrix},$$

em que os componentes de variância, $\sigma^2_{r_k}$ e σ^2_e, são estimados via método de máxima verossimilhança restrita (*Restricted maximum likelihood* – REML - PATTERSON e THOMPSON, 1971).

O modelo acima considera a presença do efeito da k-ésima região $(\boldsymbol{r_k})$, dessa forma é definido como sendo o modelo completo. Já o modelo sem a presença dos efeitos genéticos da k-ésima região $(\boldsymbol{r_k})$, é considerado como modelo restrito. Para avaliar a significância dos efeitos da região k-ésima sobre o fenótipo, os modelos, reduzido e completo, são comparados por meio do teste da razão de verossimilhança (TRV), em que se considera o logaritmo da função de verossimilhança $(ln\, L)$, conforme mostrado abaixo:

$$TRV = 2l\, n\left(\frac{L_1}{L_0}\right)$$

em que L_1 é a função de verossimilhança referente ao modelo completo e L_0 é a função de verossimilhança referente ao modelo restrito. A hipótese de nulidade (H_0) do TRV é definida como "os modelos não diferirem entre si", ou seja, isso indica que a k-ésima região não apresenta efeitos genéticos sobre o fenótipo. Já a hipótese alternativa (H_a) é definida como "os modelos diferem entre si", ou seja, a k-ésima região apresenta efeitos significativos sobre o fenótipo. Como são determinadas K regiões no genoma, os procedimentos de estimação e de teste de hipótese devem ser realizados K vezes.

Dessa forma, assim como ocorre na análise via marcas únicas apresentada no tópico **2.2,** os procedimentos de teste usados no RHM também sofrem com a ocorrência de múltiplos teste, e assim, elevando a taxa de falsos positivos. Uma forma de contornar esse problema é a mesma já apresentada anteriormente, o *q-value*.

2.5. Métodos Bayesianos

As abordagens bayesianas vêm ganhando cada vez mais destaque em GWAS, devido ao fato de que nesses métodos é possível estimar simultaneamente os efeitos dos marcadores, o que reduz a taxa de falsos positivos (LI et al., 2021). Os métodos bayesianos propostos para

a predição genômica, como os métodos Bayes A, Bayes B, BayesCπ e BayesDπ, também podem ser utilizados para analisar a associação entre as regiões genômicas e o fenótipo, como apresentado por Fernando e Garrick (2013). Os métodos bayesianos têm a vantagem de quantificar a incerteza de cada parâmetro, quando assumimos uma distribuição de probabilidade *a priori* para os mesmos e combinam-se a elas na inferência, as informações provenientes dos dados (ZHAO et al., 2019). Os efeitos individuais de cada marcador são estimados simultaneamente, e posteriormente, em posse das médias, medianas e/ou modas *a posteriori* dos efeitos dos marcadores se constrói e avalia-se as regiões como apresentado por Lima et al. (2022). A utilização de abordagens bayesianas é uma estratégia promissora para a GWAS quando combinada com inferência baseada em regiões, como apresentado por Chunyu et al. (2017).

Desta forma, são apresentadas a seguir descrições metodológicas básicas sobre o principal método bayesiano proposto para a GWS e aplicado a GWAS, chamado método BayesDπ.

2.5.1 Método BayesD$\boldsymbol{\pi}$

Considere o seguinte modelo linear proposto por Meuwissen et al. (2001):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Wm} + \boldsymbol{e}$$

em que:

y é o vetor de informações fenotípicas;

$\boldsymbol{\beta}$ é vetor de efeitos fixos, como a média e os efeitos de estrutura de populações e $\boldsymbol{X}$ é a matriz de incidência dos efeitos fixos;

$\boldsymbol{m}$ é o vetor de efeitos genéticos aditivos dos marcadores;

$\boldsymbol{W}$ é a matriz de incidência de todos os marcadores moleculares;

$\boldsymbol{e}$ é o vetor de erros aleatórios com $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$ sendo σ_e^2 a variância residual.

Com base no modelo linear apresentado acima, Habier et al. (2011) propuseram o método Bayes Dπ. A distribuição dos dados e as distribuições *a priori* referentes a este método são definidas, respectivamente, como:

$$y \sim Normal(\boldsymbol{X\beta} + \boldsymbol{Wm}, I\sigma_e^2)$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{I}10^8)$$

$$m_j|\pi \sim (1-\pi)N(0, \sigma_{m_j}^2) + \pi N(0, \sigma_{m_j}^2 = 0);$$

$$\sigma_{m_j}^2|\pi \sim \chi^{-2}(\upsilon, S_m^2),$$

$$\pi \sim U(0,1)$$

$$S_m^2 \sim Gama(\alpha, \beta)$$
$$\sigma_e^2 \sim \chi^{-2}(s_e, df_e)$$

em que $j = 1, \dots, n$, n é o número de marcadores, s_e, df_e, α, υ e β são os hiperparâmetros das distribuições *a priori* e devem ser conhecidos, π é a probabilidade de inclusão de variáveis. O BayesDπ visa limitar a influência do parâmetro de escala (S_m^2) no fator de *shrinkage* para os efeitos dos marcadores com intuito de minimizar os vieses das estimativas e tratar π como um parâmetro desconhecido do modelo, assumindo para ele uma distribuição de probabilidade (AZEVEDO et al., 2015).

O vetor de parâmetros a ser estimado é dado por $\boldsymbol{\theta} = \left[\boldsymbol{\beta}, m_1, \dots, m_n, \sigma_{m_1}^2, \dots, \sigma_{m_n}^2, \pi, S_m^2, \sigma_e^2\right]'$. O processo de estimação bayesiana será definido no tópico **2.5.3**, considerando o vetor $\boldsymbol{\theta}$. Após a estimação dos efeitos de marcadores, é possível construir as regiões e avaliá-las conforme será apresentado no tópico **2.5.4**.

2.5.2 Mapeamento de herdabilidades regionais (*Regional Heritability Mapping* – RHM) bayesiano

Assim como o RHM sob a abordagem de modelos mistos, pode-se considerar o efeito da região genômica diretamente no modelo e a estimação deste efeito ser realizada sobre o enfoque bayesiano, diferentemente do que ocorre com o método BayesDπ, que considera os efeitos de marcadores no modelo estatístico e depois constrói-se as regiões. Para tanto, considere o seguinte modelo:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z_k r_k} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$ é o vetor de informações fenotípicas;

$\boldsymbol{\beta}$ é vetor de efeitos fixos, como a média e os efeitos de estrutura de populações e $\boldsymbol{X}$ é a matriz de incidência dos efeitos fixos;

$\boldsymbol{r_k}$ é o vetor de efeitos genéticos aleatórios aditivos dos indivíduos da k-ésima região ($k = 1, 2, \dots, K$);

$\boldsymbol{Z_k}$ é a matriz de incidência que relaciona o efeito da k-ésima região ao fenótipo;

$\boldsymbol{e}$ é o vetor de erros aleatórios do modelo com $e \sim N(0, \boldsymbol{I}\sigma_e^2)$, sendo σ_e^2 a variância residual.

A distribuição dos dados e as distribuições *a priori* referentes ao modelo acima são definidas, respectivamente, como:

$$\boldsymbol{y} \sim N(\boldsymbol{X\beta} + \boldsymbol{Z_k r_k}, I\sigma_e^2)$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{I}10^8)$$

$$\boldsymbol{r_k} \sim N(0, \boldsymbol{G_{rk}}\sigma^2_{rk})$$
$$\sigma^2_{r_k} \sim \chi^{-2}(s_{r_k}, df_{r_k})$$
$$\sigma^2_e \sim \chi^{-2}(s_e, df_e)$$

em que $\sigma^2_{r_k}$ é a variância genética associada a k-ésima região; $s_{r_k}, df_{r_k}, s_e, df_e$ são denominados hiperparâmetros sendo s o parâmetro de escala e df os graus de liberdade.

A matriz $\boldsymbol{G_{r_k}}$ utiliza a matriz de incidência dos marcadores pertencentes a k-ésima região, e é dada por:

$$\boldsymbol{G_{r_k}} = \frac{\boldsymbol{W_{r_k}}\boldsymbol{W'_{r_k}}}{\sum_{i=1}^{n_k} 2p_i(1-p_i)}$$

em que:

$\boldsymbol{W_{r_k}}$ é a matriz de incidência de marcadores pertencentes a k-ésima região;

p_i é a frequência alélica associada ao i-ésimo SNP;

n_k é o número de SNPs pertencentes à k-ésima região.

O vetor de parâmetros a ser estimado é $\boldsymbol{\theta} = [\boldsymbol{\beta}, \boldsymbol{r_k}, \sigma^2_{r_k}, \sigma^2_e]'$. O processo de estimação bayesiana será definido no tópico **2.5.3**, considerando o vetor $\boldsymbol{\theta}$. Após a estimação do componente de variância associado a k-ésima região, é possível avaliá-la conforme será apresentado no tópico **2.5.4**.

2.5.3 Estimação Bayesiana

A distribuição a *posteriori* conjunta dos parâmetros é proporcional a multiplicação da função de verossimilhança e das distribuições *a priori* de cada parâmetro. No entanto, a distribuição de probabilidade desejada para a realização de inferências é a distribuição *a posteriori* marginal de cada parâmetro, porém, esta distribuição é de difícil ou impossível obtenção analítica, principalmente, no contexto de alta dimensionalidade (RESENDE et al., 2012). A obtenção de amostras das distribuições *a posteriori* marginais é realizada por meio dos algoritmos MCMC (*Markov Chain Monte Carlo*). Esse método consiste na técnica de gerar valores de uma distribuição de probabilidade por meio de um processo de Cadeias de *Markov*. Os valores das distribuições *a posteriori* marginais são gerados indiretamente por meio de uma classe de distribuição denominada Distribuição Condicional Completa *a posteriori* (D.C.C.P.) e por meio da teoria de Cadeias de *Markov* é possível mostrar que, após atingir a condição de equilíbrio, gerar valores da D.C.C.P. equivale a gerar valores da distribuição *a posteriori* marginal do parâmetro. Quando a distribuição assumida para os dados e as distribuições *a priori* assumidas para os parâmetros são ditas conjugadas, temos que as respectivas D.C.C.P.'s são

distribuições de probabilidade conhecidas e a da mesma família de distribuições da distribuição *a priori,* como é o caso do método apresentado no tópico **2.5.2**. Desta forma, pode-se gerar valores diretamente da D.C.C.P. e, consequentemente, utilizar o algoritmo MCMC denominado *Gibbs Sampler* (GEMAN e GEMAN, 1984) que será descrito a seguir.

Considere p parâmetros. Assim, o algoritmo geral de amostragem de *Gibbs* é dado por:

Passo 1: Inicie o contador de iterações em $j = 0$ e defina o número J de iterações;

Passo 2: Especifique valores iniciais para cada um dos parâmetros respeitando seu respectivo espaço paramétrico, formando o vetor de parâmetros de valores na iteração 0, ou seja, $\boldsymbol{\theta}^{(0)} = (\theta_1^{(0)}\ \theta_2^{(0)}\ \ldots\ \theta_p^{(0)})$;

Passo 3: Avance j fazendo $j = j + 1$ e obtenha $\theta^{(j)}$ a partir do vetor de parâmetros $\theta^{(j-1)}$ por geração sucessiva de valores, como:

$$\theta_1^{(j)} \sim \pi(\theta_1 |\ \theta_2^{(j-1)}\ \theta_3^{(j-1)}\ \ldots\ \theta_p^{(j-1)})$$
$$\theta_2^{(j)} \sim \pi(\theta_2 |\ \theta_1^{(j)}\ \theta_3^{(j-1)}\ \ldots\ \theta_p^{(j-1)})$$
$$\theta_3^{(j)} \sim \pi(\theta_3 |\ \theta_1^{(j)}\ \theta_2^{(j)}\ \ldots\ \theta_p^{(j-1)})$$
$$\vdots$$
$$\theta_p^{(j)} \sim \pi(\theta_p |\ \theta_1^{(j)}\ \theta_2^{(j)}\ \ldots\ \theta_{p-1}^{(j)})$$

em que $\pi(\theta_1 |\ \theta_2^{(j-1)}\ \theta_3^{(j-1)}\ \ldots\ \theta_p^{(j-1)})$ é a D.C.C.P. de θ_1, $\pi(\theta_2 |\ \theta_1^{(j)}\ \theta_3^{(j-1)}\ \ldots\ \theta_p^{(j-1)})$ é a D.C.C.P. de θ_2 e assim, sucessivamente. A cada iteração tem-se um vetor de parâmetros dado por $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}\ \theta_2^{(j)}\ \ldots\ \theta_p^{(j)})$.

Passo 4: Se $j < J$, volte para o Passo 3. Se $j = J$, encerrar.

Em casos em que a D.C.C.P. é uma distribuição de probabilidade desconhecida, utiliza-se o algoritmo MCMC denominado *Metropolis Hastings* (GELMAN et al., 2014), que é o caso do método BayesDπ. Pois, nestes casos, há um impedimento quanto a geração de valores aleatórios destas distribuições. Dessa forma, a ideia deste algoritmo é utilizar distribuições de probabilidade candidatas (sendo estas distribuições de probabilidades conhecidas), e aceitar ou não os valores gerados. Esta aceitação é feita mediante a um critério probabilístico. Este é o caso do método BayesDπ apresentado no tópico **2.5.1**.

Após a definição do número de iterações (J), duas quantidades devem ser definidas *burn-in* (B) e *thin* (T). O *burn-in* consiste em descartar as primeiras B iterações da cadeia, sendo que após esse descarte tem-se a distribuição supostamente em equilíbrio. E como o algoritmo MCMC é um processo *markoviano*, a dependência entre as amostras diminui com o aumento

da distância entre as iterações, dessa forma o *thin* consiste em descartar iterações entre cada *T* amostras a serem salvas, com intuito de obter independência entre elas. Ao final, a cadeia teria um total de $(J - B)/T$ amostras. O diagnóstico de convergência ou de equilíbrio pode ser avaliado por meio de critérios propostos por Geweke (1992), Heidelberger e Welch (1983), Raftery e Lewis (1992) e Gelman e Rubin (1992). Quando se tem a convergência do algoritmo, pode-se garantir que o valor gerado da D.C.C.P. de um determinado parâmetro equivale a gerar valores da sua distribuição *a posteriori* marginal. A partir disso, toda a informação necessária para as inferências a respeito dos parâmetros pode ser obtida por meio desta distribuição, como a média *a posteriori*.

Depois de realizadas as inferências, é necessária a utilização de critérios de seleção das regiões genômicas utilizando os efeitos de marcadores estimados via métodos bayesianos (BayesDπ, por exemplo) ou os componentes de variância associados a cada região estimados (método RHM bayesiano, por exemplo). Lima et al. (2022) avaliaram vários critérios de seleção de regiões genômicas e o que apresentou resultados superiores em termos de baixas taxas de falsos positivos e alto poder de detecção foi a medida denominada de probabilidade a *posteriori* da associação da região genômica (*Window Posterior Probability of Association* - WPPA) e é esta que será apresentada a seguir.

2.5.4 Seleções pela probabilidade a *posteriori* da associação da região genômica – (*Window Posterior Probability of Association*- WPPA)

A medida WPPA é utilizada para verificar se a k-ésima região é associada ou não à um fenótipo de interesse. Esse valor é obtido com base na proporção da variância genética que é explicada pelos marcadores de cada região genômica. No RHM bayesiano, a variância genética associada a k-ésima região ($\sigma^2_{r_k}$) é diretamente estimada. No entanto, nos métodos bayesianos baseados em efeitos de marcadores, a variância genética associada a k-ésima região pode ser estimada através da seguinte expressão:

$$\hat{\sigma}^2_{r_k} = \sum_{i=1}^{n_k} 2p_i(1 - p_i)\hat{m}_i^2$$

em que $\hat{m}_i$ é a média *a posteriori* do efeito do i-ésimo SNP pertencente a k-ésima região estimado pelo método bayesiano, p_i é a frequência alélica associada ao i-ésimo SNP e n_k é o número de SNPs pertencentes à k-ésima região.

A proporção da variância genética que é explicada pelos marcadores na k-ésima região, denotada por q_k, pode ser definida de duas formas como apresentado por:

$$q_k = \frac{\sigma^2_{r_k}}{E(\sigma^2_{r_k})} \quad ou \quad q_k = \frac{\sigma^2_{r_k}}{\frac{\sigma^2_g}{K}}$$

em que $E(\sigma^2_{r_k}) = \sum_{j \epsilon k} 2p_j \, q_j \, E(m^2_j)$, $E(m^2_j) = \frac{\sigma^2_g}{n\bar{H}}$, σ^2_g é a média *a posteriori* da variância genética total, n é o número de SNPs, $\bar{H}$ é a média de $2p(1-p)$ e K é o número de regiões genômicas construídas.

Em situações em que q_k> 1, existe a presença de uma mutação causativa dentro da k-ésima região, uma vez que esta apresenta variância maior do que a média das variâncias de todas as regiões ou do valor esperado da variância genética para aquela região e, quando isso ocorrer, conclui-se que a região avaliada é associada ao fenótipo de interesse (PETERS et al., 2012; BENNEWITZ et al., 2017). A medida WPPA é obtida pela razão entre o número de iterações em que q_k é maior que 1 e o número total de iterações. Quando $WPPA > limiar$, então a região será dita associada. O limiar é definido pelo pesquisador, no entanto, existem propostas como as de Fernando e Garrick (2013) e Fernando et al. (2017), que definem o limiar como sendo igual a 0,95.

2.6. REFERÊCIAS BIBLIOGRAFICAS

ARORA, S. et al. Genome-wide association mapping of grain micronutrients concentration in *Aegilops tauschii*. **Frontiers in plant science**, 2019

AZEVEDO, C.F., de Resende, M.D.V., e Silva, F.F. *et al.* Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC Genet 16**, 105 (2015).

BENNEWITZ, J., Edel, C., Fries, R., Meuwissen, T.H., Wellmann, R. (2017). Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. **Genetics Selection Evolution**, *49*, 7.

BRAZ, C. U., Taylor, J. F., Bresolin, T., Espigolan, R., Feitosa, F. L. B., Carvalheiro, R., Baldi, F., Albuquerque, L. G., Oliveira, H. N. (2019). Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. **BMC genetics**, *20(1)*, 8.

CHENG, J. et al. Genome-wide association study of disease resilience traits from a natural polymicrobial disease challenge model in pigs identifies the importance of the major histocompatibility complex region. **G3**, v. 12, n. 3, p. jkab441, 2022.

CHUNYU C., Juan P S., Robert J T., Genome-Wide Association Analyses Based on Broadly Different Specifications for Prior Distributions, Genomic Windows, and Estimation Methods, **Genetics** , Volume 206, Issue 4, 1 August 2017, Pages 1791–1806.

FAN, B.; Onteru, S.K.; Du, Z.Q.; Garrick, D.J.; Stalder, K.J.; Rothschild, M.F. 2011. Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. **PloS one** 6(2): e14726.

FERNANDO, R. et al. Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. **Journal of Agricultural, Biological and Environmental Statistics**, v. 22, n. 2, p. 172-193, 2017

FERNANDO, R. L. et al. Controlling the proportion of false positives in multiple dependent tests. **Genetics**, v. 166, n. 1, p. 611-619, 2004.

FERNANDO, R. L., & Garrick, D. (2013). **Bayesian methods applied to GWAS**. In C. Gondro et al. (Eds), Genome-wide association studies and genomic prediction (p. 237-274). **Humana Press**, Totowa, NJ.

GELMAN, A., Hwang, J. & Vehtari, A. Understanding predictive information criteria for Bayesian models. **Stat Comput 24,** 997–1016 (2014). https://doi.org/10.1007/s11222-013-9416-2

GELMAN, A.; Rubin, D. B. Inference from iterative simulation using multiple sequences. **Statistical Science**, Hayward, v.7, n.4, p.457-472, 1992.

GEMAN S., Geman D., 1984. Stochastic relaxation, gibbs distributions and the Bayesian restoration of images. **IEEE Trans. Pattern Anal. Mach. Intell**. **6**(6): 721–741

GEWEKE, J., (1992) Evaluating the accurary of sampling-based approaches to the calculation of posterori moments. Bayesian Statistics 4 (eds. Bernardo, J.M; Berger, J.O; Dawid, A.P.; Smith, A.F.M.) **New York: Oxford University Press**, p.625-631

GUO, X. et al. Genome-wide association analyses using a Bayesian approach for litter size and piglet mortality in Danish Landrace and Yorkshire pigs. **BMC genomics**, v. 17, n. 1, p. 468, 2016.

HABIER, D. et al. Extension of the Bayesian alphabet for genomic selection. **BMC Bioinformatics**, v. 12, n. 1, p. 186, 2011.

HAYES, Ben J. et al. Genetic architecture of complex traits and accuracy of genomic prediction: coat colour, milk-fat percentage, and type in Holstein cattle as contrasting model traits. **PLoS genetics**, v. 6, n. 9, p. e1001139, 2010

HEIDELBERG, P.; Welch, P. Simulation run lenght control in the presence of an initial transient. **Operations Research**, Baltimore, v.31, n.6, p.1109-1114, 1983.

HENDERSON, C. R. Sire evaluation and genetic trends. **Journal of Animal Science**, v. 1973, n. Symposium, p. 10-41, 1973.

KIM S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D and Nordborg M (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. **Nature Genetics** 39, 1151-1155.

LAM HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, Weiming H, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SSM and Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. **Nature Genetics** 42:1053-1059

LEGARRA, A.; RICARD, A.; VARONA, L. GWAS by GBLUP: single and multimarker EMMAX and Bayes factors, with an example in detection of a major gene for horse gait. **G3: Genes, Genomes, Genetics**, v. 8, n. 7, p. 2301-2308, 2018.

LI, J., Wang, Z., Fernando, R. *et al.* Tests of association based on genomic windows can lead to spurious associations when using genotype panels with heterogeneous SNP densities. **Genet Sel Evol 53,** 45 (2021).

LIMA, L.P; et al. Evaluation of bayesian methods of genomic association via chromosomic regions using simulated data. **Scientia Agricola**, v. 79, n. 3, 2022.

MCCARTHY, M. I. et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. **Nature reviews genetics**, v. 9, n. 5, p. 356-369, 2008.

MEUWISSEN, T. et al. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819–29, 2001.

MÜLLER, B. S. et al. Independent and Joint-GWAS for growth traits in Eucalyptus by assembling genome-wide data for 3373 individuals across four breeding populations. **New Phytologist**, 2019.

NAGAMINE, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, Hicks, A. A., Pramstaller, P. P., Hastie, N., Wright, A. F., Haley, C. S. (2012). Localising loci underlying complex trait variation using regional genomic relationship mapping. **PLoS One**, 7:e46501.

PATTERSON, H. D.; THOMPSON, R. Recovery of inter-block information when block sizes are unequal. **Biometrika**, v. 58, n. 3, p. 545-554, 1971.

PETERS, S. O. et al. Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. **Journal of animal science**, v. 90, n. 10, p. 3398-3409, 2012.

QUERO, G. et al. Genome-wide association study using historical breeding populations discovers genomic regions involved in high-quality rice. **The plant genome**, 2018.

RAFTERY, A. L.; Lewis, S. Comment: one long run with diagnostics: implementation strategies for Markov chain Monte Carlo. **Statistical Science**, Hayward, v.7, n.4, p.493- 497, 1992.

RESENDE, M.D.V.; Silva, F.F.; Lopes, P.S.; Azevedo, C.F. Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial. **Viçosa: Universidade Federal de Viçosa/Departamento de Estatística**. 2012. 291 p. http://www.det.ufv.br/ppestbio/corpo_docente.php

RESENDE, R. T. Regional Heritability Mapping and GWAS for molecular breeding in eucalyptus hybrids. 2017.

SAHANA, G., Guldbrandtsen, B., Janss, L., & Lund, M. S. (2010). Comparison of association mapping methods in a complex pedigreed population. **Genetic Epidemiology***, 34(5), 455–462.* doi:10.1002/gepi.20499

STOREY, J. D.; TIBSHIRANI, R. Statistical significance for genomewide studies. **PNAS** 100:9440-9445, 2003.

SUELA, M.M., Azevedo, C.F., Nascimento, M., Nascimento, A.C.C. and de Resende, M.D.V. (2022), Regional Heritability Mapping and Genome-Wide Association Identify Loci For Rice Traits. **Crop Sci**.. Accepted Author Manuscript.

UFFELMANN, E., Huang, Q.Q., Munung, N.S. et al. Genome-wide association studies. **Nat Rev Methods Primers** 1, 59 (2021).

VISSCHER, Peter M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. **American Journal of Human Genetics**, v. 101, n. 1, p. 5–22, 2017. Disponível em: <http://dx.doi.org/10.1016/j.ajhg.2017.06.005>.

VOS PG, Paulo MJ, Voorrips RE, Visser RG, van Eck HJ and van Eeuwijk FA (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. **Theoretical and Applied Genetics** 130(1), 123-135.

WU, Y. et al. Genome-wide association studies using haplotypes and individual SNPs in Simmental cattle. **PloS one**, v. 9, n. 10, p. e109330, 2014.

ZHANG J. et al. Genome wide association study, genomic prediction and marker assisted selection for seed weight in soybean (*Glycine max*). **Theor Appl Genet**, 2016.

ZHAO, Y. et al. Structured Genome-Wide Association Studies with Bayesian Hierarchical Variable Selection. **Genetics**, v. 212, n. 2, p. 397-415, 2019

ZHOU, X.; STEPHENS, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. **Nature methods**, v. 11, n. 4, p. 407, 2014.

ZIEGLER, A.; König, I. R.; Thompson, J. R. Biostatistical aspects of genome-wide association studies. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, v. 50, n. 1, p. 8-28, 2008.

3-ASSOCIAÇÃO GENÔMICA VIA REGIÕES CROMOSSÔMICAS SOB A ABORDAGEM BAYESIANA

RESUMO

A biotecnologia vem proporcionando novas descobertas na área da biologia molecular, o que favorece cada vez mais os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS). A GWAS busca identificar e investigar as regiões do cromossomo em que as variantes significativas se encontram. Atualmente, estudos utilizando grupos de marcadores vêm ganhando cada vez mais destaque na GWAS, devido ao fato de que, os marcadores moleculares podem estar em alto desequilíbrio de ligação (*Linkage Disequilibrium* – LD) entre si e, com isso, influenciar conjuntamente o fenótipo. As abordagens bayesianas também, pois, permitem quantificar a incerteza de cada parâmetro, quando se é assumido uma distribuição de probabilidade *a priori* para os mesmos e combinam-se a elas, na inferência, as informações provenientes dos dados. Os objetivos deste estudo foram: (i) comparar a eficiência de se estimar o efeito de todas as regiões genômicas simultaneamente através de um modelo bayesiano em relação ao procedimento de se estimar o efeito de cada região por vez (ii) elucidar a utilização deste modelo nos programas de melhoramento, aplicando em dados reais de arroz *Oryza sativa*. Esse estudo utilizou dados simulados através do pacote *AlphaSimR* e dados de arroz provenientes do *Rice Diversity Project.* O tamanho das regiões foi determinado como sendo a distância na qual o LD é metade do seu valor máximo e, para verificar se as regiões eram de fato associadas as características fenotípicas, foi utilizada a Probabilidade *a Posteriori* da Associação da Janela (*Window Posterior Probability of Association* - WPPA). Para os dados simulados, a eficiência da estimação simultânea dos efeitos das regiões genômicas utilizando a estimação bayesiana, apresentou resultados satisfatórios. Nos dados de arroz, a estimação simultânea detectou uma quantidade superior de regiões já relatadas na literatura em detrimento a estimação única, além de apresentar novas regiões genômicas que podem ser estudadas em analises pós-GWAS. Essa é uma metodologia que apresenta potencial para aplicação, descoberta e investigação de novas regiões genômicas associadas a características fenotípicas.

Palavras-chave: Marcadores Moleculares, Poder de detecção, Falsos Positivos, Simulação, Desequilíbrio de Ligação

1- Introdução

A biotecnologia vem a cada dia mais facilitando e proporcionando novas descobertas aos estudos na área da biologia molecular. Devido a isso, é possível que sejam utilizadas informações que venham diretamente do DNA, através de marcadores moleculares. Esses permitem que se tenha amplas informações sobre o genoma dos indivíduos, tornando-se possível os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS) (UFFELMANN et al., 2021). Estes estudos buscam identificar as variantes causais no genoma que afetam uma determinada característica por meio dos marcadores moleculares, principalmente, o SNP (*Single Nucleotide Polymorphism*). Um dos primeiros métodos estatísticos aplicados a GWAS foi o método via regressão em marcas únicas, que visa estudar a associação entre o fenótipo e um único marcador (ZINGLER et al., 2008). No entanto, esse método apresenta problemas estatísticos, como, por exemplo, a necessidade de grandes amostras, alta taxa de falsos positivos e baixo poder de detecção (FERNANDO et al., 2004).

Com isso, estudos que consideram grupos de marcadores, as denominadas regiões genômicas, visando buscar associações entre elas e o fenótipo, vêm ganhando cada vez mais destaque na GWAS (FERNANDO et al., 2017). O mapeamento de herdabilidade regional (*Regional Heritability Mapping* - RHM), proposto por Nagamine et al. (2012), é uma abordagem que utiliza modelos mistos e que é baseada em regiões e, com isso, tende-se a aumentar o desequilíbrio de ligação (*Linkage Disequilibrium* - LD) entre as regiões e os QTL (*Quantitative Trait Loci*) reduzindo as taxas de falsos positivos e aumentando o poder de detecção de QTLs (USAI et al., 2014). No entanto, nesta metodologia, estima-se o efeito de uma região do genoma por vez. Já os métodos bayesianos propostos inicialmente para a predição genômica, como o BayesDπ, também podem ser utilizados para analisar a associação entre as regiões genômicas e o fenótipo, como apresentado por Fernando e Garrick (2013). Nesta abordagem, os efeitos individuais de cada marcador são estimados simultaneamente, e posteriormente, em posse das médias *a posteriori* dos efeitos dos marcadores se constrói e avalia-se as regiões como apresentado por Lima et al. (2022).

Considerando o melhor de nosso conhecimento, até o momento, não há relatos na literatura de estudos de GWAS de uma proposta similar ao RHM, mas que se estimam os efeitos de todas as regiões diretamente e simultaneamente no modelo. Procedendo com a análise de forma simultânea, os efeitos de todas as regiões teriam somente uma única fonte de erro, o que pode permitir o aumento do poder de detecção de regiões associadas. No entanto, a realização da estimação simultânea considerando modelos lineares mistos, sob a abordagem REML/BLUP com matrizes de covariância estruturadas, podem sofrer de problemas de convergência quando

os modelos são super-parametrizados, causando sérios problemas nas inferências dos parâmetros (BATES et al., 2015). Já os métodos bayesianos parecem não sofrer com problemas de super-parametrização, pois utilizam informações das distribuições *a priori* além das informações dos dados amostrais (GAMERMAN e LOPES, 2006).

Nesse contexto, o objetivo deste estudo é: (i) comparar a eficiência de se estimar o efeito de todas as regiões genômicas simultaneamente através de um modelo bayesiano em relação ao procedimento de se estimar o efeito de cada região por vez por meio de dados simulados e (ii) elucidar a utilização deste modelo nos programas de melhoramento, aplicá-los a dados reais de arroz *Oryza sativa*.

2- Materiais e Métodos

2.1- Dados Simulados

Os conjuntos de dados genotípicos e fenotípicos de uma população F2 foram simulados utilizando o *software* R versão 4.1.2. (R CORE TEAM, 2021) por meio do pacote *AlphaSimR* versão 1.0.4 (GAYNOR et al., 2021). As características simuladas têm alta ($h^2 = 0{,}50$), moderada ($h^2 = 0{,}30$) e baixa ($h^2 = 0{,}10$) herdabilidades com, respectivamente, três arquiteturas genéticas com 3, 10 e 100 QTL. Os indivíduos foram gerados com genomas diploides com comprimento cromossômico de 12 Morgans, supondo que os cromossomos fossem do mesmo tamanho, os marcadores foram distribuídos de forma aleatória entre os 12 cromossomos. Cada um dos cromossomos possui 250 SNPs, totalizando 3.000 SNPs. Desta forma, tem-se 3 cenários distintos, cada qual com 10 repetições e com 1.000 indivíduos cada. Foi obtido para cada um dos cenários a proporção da variação dos locus de características quantitativas (QTL) explicadas pelos SNPs (r^2_{mq}), em que este é calculado através da seguinte expressão (GODDARD et al. 2011):

$$r^2_{mq} = \frac{n}{n + n_{Qtl}}$$

em que n é o número de SPNs e n_{Qtl} é o número de QTL. A Tabela 1 apresenta os cenários utilizados para a simulação dos dados e a proporção da variação genética explicada pelos SNPs.

Tabela 1. Descrição dos cenários com a proporção da variação dos locus de características quantitativas (QTL) explicadas pelos SNPs (r^2_{mq}), arquitetura genética, número de QTL e herdabilidade aditiva (h^2_a).

Cenarios	r^2_{mq}	Arquitetura Genetica	Numero de QTL	h^2_a
Cenário 1	0.9990	3 QTL em 12 cromossomos*	3	0.50
Cenário 2	0.9967	1 QTL em cada um dos 10 cromossomos*	10	0.30
Cenário 3	0.9677	10 QTL em cada um dos 10 cromossomos*	100	0.10

* Os dois últimos cromossomos não possuem QTL.

2.2- Banco de dados de arroz (*Oryza sativa*)

O conjunto de dados genotípicos e fenotípicos de arroz (*Oryza sativa)*, utilizados para o estudo foram descritos em Ammiraju et al. (2006) e Zhao et al. (2011) e estão disponíveis em http://www.ricediversity.org/data/sets/44kgwas/. Foram selecionadas 11 características fenotípicas em 413 acessos de arroz, genotipados para 44.100 marcadores SNPs. Foi realizado o controle de qualidade para os marcadores genotípicos, $call\ rate < 70\%$ e baixa frequência do alelo mais raro (*Minor Allele Frequency* - MAF) < 1%, após a realização desses, os marcadores que não atendiam os limites foram eliminados, restando 36.901 marcadores.

As características fenotípicas avaliadas foram: i) comprimento da folha bandeira (CFB); ii) largura da folha bandeira (LFB); iii) número de panículas por planta (NPP); iv) número de ramos da panícula primária (NRPP); v) altura da planta (AP); vi) comprimento da panícula (CP); vii) Flores por panícula (FLP); viii) Resistência à Brusone (RB); ix) fertilidade da panícula (FP); x) teor de proteína (TP); xi) número de sementes por panícula (NSPP). Os fenótipos foram pré-corrigidos para estrutura de população por meio de quatro componentes principais (Zhao et al., 2011).

2.3- Regiões Genômicas

Neste estudo, construiu-se regiões genômicas de tamanhos determinados com base no desequilíbrio de ligação médio entre os marcadores. Após a construção do gráfico de

decaimento de LD, os tamanhos das regiões foram definidos, como sendo a distância na qual o LD é metade do seu valor máximo. Essa medida também foi utilizada por Kim et al. (2007), Suela et al. (2022) e Vos et al. (2017). O número de regiões e marcadores em cada região com base na análise da metade do decaimento máximo do LD para todos os cenários simulados são apresentados na Tabela 2. Para os dados reais o valor que forneceu a metade do valor máximo de LD foi 0,21 Mb conforme também foi apresentado por Suela et al. (2022).

Tabela 2. Estatísticas relacionadas aos dados genotípicos para os cenários: média e erro Padrão da distância entre marcadores dados em Morgans (Dist), máximo e respectivo erro padrão de desequilíbrio de ligação por cromossomo (maxLD), distância na qual temos metade do LD máximo (Dist. Metade LDmax), número de regiões (NR) e número de marcadores (N) por região e respectivos erros padrão.

Cenário	Dist.	maxLD	Dist. Metade LDmax	NR	N
1	0.003 ± 0.001	0.41 ± 0.01	0.04 ± 0.01	308.10 ± 1.15	9.74 ± 1.09
2	0.004 ± 0.001	0.41 ± 0.01	0.04 ± 0.01	304.50 ± 1.49	9.85 ± 1.10
3	0.002 ± 0.001	0.40 ± 0.01	0.04 ± 0.01	302.80 ± 0.47	9.91 ± 1.14

2.4- Modelo sob o enfoque bayesiano

O presente estudo visa propor o seguinte modelo para estimar os efeitos das regiões genômicas, por simultaneamente:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 r_1} + \boldsymbol{Z_2 r_2} + \cdots + \boldsymbol{Z_K r_K} + \boldsymbol{e},$$

em que **y** é o vetor de informações fenotípicas ($N \times 1$ sendo N o número de indivíduos); 1 é um vetor com a mesma dimensão de **y** e valores iguais a 1; $\boldsymbol{\mu}$ se refere a média geral; $\boldsymbol{r_k}$ é o vetor de efeitos genéticos aleatórios aditivos dos indivíduos para a k-ésima região ($k = 1, 2, \ldots, K$ em que K é o número total de regiões por cromossomo); $\boldsymbol{Z_k}$ é a matriz de incidência que relaciona o efeito da k-ésima região ao fenótipo; $\boldsymbol{e}$ é o vetor de erros aleatórios do modelo com $e \sim N(0, \boldsymbol{I}\sigma_e^2)$, sendo σ_e^2 a variância residual.

A distribuição dos dados e as distribuições *a priori* referentes ao modelo acima são definidas, respectivamente, como:

$$\boldsymbol{y} \sim N(\mathbf{1}\mu + \boldsymbol{Z_1 r_1} + \boldsymbol{Z_2 r_2} + \cdots + \boldsymbol{Z_K r_k}, \boldsymbol{I}\sigma_e^2)$$

$$\mu \sim N(0, 10^8)$$

$$\boldsymbol{r_1} \sim N(0, \boldsymbol{G_{r_1}}\sigma_{r_1}^2)$$

...

$$r_K \sim N(0, \boldsymbol{G}_{\boldsymbol{r_K}} \sigma^2_{r_K})$$
$$\sigma^2_{r_1} \sim \chi^{-2}(s_{r_1}, df_{r_1})$$
$$\ldots$$
$$\sigma^2_{r_K} \sim \chi^{-2}(s_{r_K}, df_{r_K})$$
$$\sigma^2_e \sim \chi^{-2}(s_e, df_e)$$

em que $\sigma^2_{r_k}$ é a variância associada a k-ésima região ($k = 1, 2, \ldots, K$); $s_{r_k}, df_{r_k}, s_e, df_e$ são denominados hiperparâmetros sendo s o parâmetro de escala e df os graus de liberdade.

A matriz de parentesco aditiva da k-ésima região é dada pela seguinte expressão:

$$\boldsymbol{G}_{\boldsymbol{r_k}} = \frac{W_{r_k} W'_{r_k}}{\sum_{i \in rk} 2p_i(1 - p_i)}$$

em que W_{r_k} é a matriz de incidência de marcadores da k-ésima região; p_i é a frequência alélica associada ao i-ésimo SNP ($i = 1,2, \ldots, n_k$ em que n_k é o número de SNPs da k-ésima região) pertencente à k-ésima região.

A obtenção das distribuições a *posteriori* marginais é realizada por meio dos algoritmos MCMC (*Markov Chain Monte Carlo*). Esse método consiste na técnica de gerar valores de uma distribuição de probabilidade por meio de um processo de Cadeias de *Markov*. Os valores das distribuições *a posteriori* marginais são gerados indiretamente por meio de uma classe de distribuição denominada Distribuição Condicional Completa *a posteriori* (D.C.C.P.) e por meio da teoria de Cadeias de *Markov* é possível mostrar que, após atingir a condição de equilíbrio, gerar valores de uma D.C.C.P. equivale a gerar valores da distribuição a *posteriori* marginal do parâmetro. Como a distribuição assumida para os dados e as distribuições *a priori* assumidas para os parâmetros foram as distribuições Normal e Qui-quadrado invertida escalada, os quais são conjugadas, temos que a D.C.C.P. é uma distribuição de probabilidade conhecida. Desta forma, pode-se gerar valores diretamente da D.C.C.P. e, consequentemente, utilizar o algoritmo MCMC denominado *Gibbs Sampler* (GEMAN & GEMAN, 1984).

Para as análises foram utilizados um número de interações de 300.000, um *burn-in* de 20.000 e um *thin* de 10. O diagnóstico de convergência foi avaliado por meio do critério proposto por Geweke (1992).

2.5-Seleção pela Probabilidade *a Posteriori* da Associação da Janela – (*Window Posterior Probability of Association* - WPPA)

Para verificar se a k-ésima região é de fato associada ou não à característica de interesse foi utilizada a proporção da variância genética explicada (PVGE), e definida como,

$$PVGE_k = \frac{\sigma^2_{r_k}}{\frac{\sum_{i=1}^{K} \sigma^2_{r_i}}{K}}$$

em que $\sigma^2_{r_i}$ é a média *a posteriori* da variância genética da i-ésima região ($i = 1,2, \dots, k \dots, K$) e K é o número de regiões por cromossomo. Em situações em que $PVGE_k$> 1, existe a presença de uma mutação causativa dentro da k-ésima região, uma vez que esta apresenta variância maior do que a média das variâncias de todas as regiões e, quando isso ocorrer, temos que a região avaliada é considerada associada à característica de interesse (PETERS et al., 2012; BENNEWITZ et al., 2017). A medida WPPA (*Window Posterior Probability of Association*) foi obtida pela razão entre o número de iterações em que q_k é maior que 1 e o número total de iterações. Quando $WPPA > limiar$, então a região será dita associada.

O limiar ótimo, para o WPPA, foi o ponto que minimizou a distância euclidiana entre a curva ROC (*Receiver Operating Characteristic* - proposta por Metz (1978)) construída em cada cenário e o ponto ideal para a GWAS ($x = 0, y = 1$), ou seja, taxa de falso-positivo igual a 0 e poder de detecção igual a 1 (Figura Suplementar 1). Este critério de escolha de determinação do ponto ótimo da curva ROC foi descrito por Perkins e Schisterman (2006).

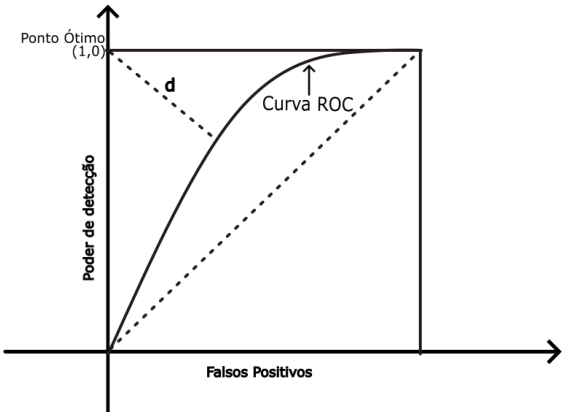

Figura Suplementar 1: O limiar ótimo é mostrado como o ponto que minimizou a distância euclidiana, d, entre a curva ROC e o ponto ideal (0,1) onde temos uma taxa de falsos positivos igual a 0% e um poder de detecção de 100%.

2.6- Comparação de Metodologias

Com o intuito de comparar a eficiência de se estimar o efeito das regiões simultaneamente e estimar o efeito de uma região por vez, serão calculadas as seguintes medidas:

i) O poder de detecção que se refere à capacidade de detectar um QTL na população, quando ele realmente existe (RESENDE et al., 2012), ou seja, declarar que o efeito

de uma região é associado ao fenótipo quando de fato a região está em LD com o QTL. Este é definido pela razão do número de regiões consideradas associadas e que afetam a característica pelo número total de regiões que afetam a característica.

ii) A taxa de falsos positivos que consiste em declarar que uma região é associada, quando na verdade a região não está em LD com o QTL. É possível calculá-la pela razão entre o número de regiões ditas associadas e que não afetam a característica e o número de regiões que não afetam a característica.

iii) A porcentagem da variância genética capturada foi obtida pela razão entre a soma das médias *a posteriori* das variâncias genéticas das regiões consideradas associadas e a média *a posteriori* da variância genética total (LIMA et al., 2022).

iv) Área abaixo da curva obtida entre as taxas de falsos positivos e o poder de detecção (curva ROC).

v) A porcentagem de regiões detectadas como associadas nos cromossomos 11 e 12 (sem QTLs) foi obtida como sendo a razão entre o número de regiões ditas associadas pelo modelo e o número total de regiões nestes cromossomos.

Dessa forma, o procedimento que fornece o maior poder de detecção, menor taxa de falsos positivos, maior porcentagem da variância genética capturada, maior área abaixo da curva e menor número de regiões detectadas como associadas nos cromossomos sem QTL será considerado mais eficiente.

Para elucidar a aplicação dos modelos propostos em programas de melhoramento, foram utilizados dados reais de arroz (*Oryza sativa).* Foi utilizado um limiar de 0,95 para verificar se de fato as regiões eram associadas ou não a fenótipo de interesse. A escolha desse limiar se deve ao fato de várias pesquisas como Fernando e Garrick (2013) e Fernando et al. (2017), já utilizarem esse valor. A localização (cromossomo e posição) no genoma dos SNPs e as regiões de pico foram comparadas com os QTL relatados anteriormente, utilizando as bases de dados Q-TARO (http://qtaro.abr.affrc.go.jp/qtab/table) e *Gramene* QTL (https://archive.gramene.org/qtl/).

2.7- Recursos Computacionais

Todas as rotinas computacionais utilizaram o *software* R versão 4.1.2. (R CORE TEAM, 2021). Para a simulação dos dados foi utilizado o pacote *AlphaSimR* versão 1.0.4 (GAYNOR et al., 2021), o cálculo do LD foi realizado pelo pacote *sommer* (COVARRUBIAS-PAZARAN, 2018) usando a função LD.decay e a rotina computacional implementada no GenomicLand

(AZEVEDO et al., 2019), para as análises bayesianas e de convergência foram utilizados, respectivamente, os pacotes os pacotes BGLR versão 1.0.9 (DE LOS CAMPOS E PÉREZ., 2021) e coda versão 0.19-4 (PLUMMER et al., 2006).

3- Resultados e Discussões

Os resultados da taxa de falsos positivos, poder de detecção, área sob a curva ROC, porcentagem de regiões ditas associadas nos cromossomos 11 e 12 e porcentagem da variância genética capturada pelo modelo sob o enfoque bayesiano, considerando a estimação do efeito de uma única região e a estimação simultânea dos efeitos das regiões, utilizando limiar ótimo e o limiar de 0,95 são apresentados na Tabela 3.

Para o cenário com características controladas por apenas 3 QTLs de efeitos maiores (Cenário 1), foi considerado o limiar ótimo, para detectar as regiões associadas utilizando o critério do WPPA, de 0,55 para estimação simultânea e de 0,67 para a estimação única. A estimação única obteve uma taxa de falsos positivos (0,11) menor que a estimação simultânea (0,18), mas apresentou menor poder de detecção. No estudo de Schmid e Bennewitz (2017), os autores afirmam que a diminuição do número de falsos positivos em GWAS pode comprometer o poder de detecção, o que corrobora com os resultados encontrados para estimação única neste cenário.

Os resultados apresentados pela estimação simultânea, quando considerado o limiar ótimo, foram superiores para a taxa de falsos positivo (0,18), poder de detecção (0,73), a porcentagem de regiões associadas nos cromossomos 11 e 12 (18%), o percentual da variância genética capturada (89,86) e a área abaixo a curva ROC (0,72). Ou seja, apesar de apresentar uma taxa maior de falsos positivos e possuir uma maior da porcentagem de regiões associadas nos cromossomos 11 e 12, a estimação simultânea consegue detectar um maior número de regiões associadas.

O cenário com características controladas por 10 QTLs (Cenário 2), obteve o valor médio de limiar ótimo muito similar para ambas as abordagens, sendo 0,40 para estimação simultânea e 0,41 para a estimação única. Pode-se observar que a estimação simultânea obteve os resultados de poder de detecção, área sob a curva ROC e porcentagem da variância genética capturada superiores (Tabela 3), para a taxa de falso positivo e porcentagem de regiões associadas nos cromossomos 11 e 12 os resultados foram inferiores quando comparados a estimação única, como desejado. Lima et al. (2022) mostraram que o critério WPPA em detrimento a outras abordagens para detectar regiões associadas, consegue uma solução desejável, que é reduzir os falsos positivos, sem que isso comprometa tanto o poder de detecção

da análise. A estimação simultânea utilizando o WPPA mostrou que é possível reduzir a taxa de falso positivo sem que isso afetasse o poder de detecção. Já a estimação única utilizando o critério WPPA, apresentou uma redução no poder de detecção, devido à redução da taxa de falsos positivos. Para o cenário com características controladas por muitos marcadores de pouco efeito (Cenário 3), foi considerado o limiar ótimo de 0,35 para estimação simultânea e 0,37 para a estimação única. O resultado da estimação única foi similar ao resultado apresentado pela estimação simultânea.

Analisando a diferença entre os valores de poder encontrados para os cenários 1,2 e 3, estes estão de acordo com o que é descrito no estudo de Shin e Lee (2015), que concluíram que arquitetura genética e a herdabilidade afetam o poder. No estudo realizado pelos autores, foi observado que o poder aumentou muito quando se avaliou características oligogênicas com uma maior herdabilidade. Podemos observar que o mesmo acontece quando consideramos um limiar de 0,95 como proposto por Fernando e Garrick (2013) e Fernando et al. (2017). O cenário 1 que possui característica oligogênica, com 3QTLs de grande efeito, foi o único cenário que apresentou um poder de detecção diferente de zero em ambas as estimações. A simultânea apresentou resultados superiores na porcentagem da variância genética explicada (70,72), além de possuir uma taxa de falsos positivos igual a zero.

Para elucidar a aplicação do modelo proposto, este foi aplicado em um banco de dados de arroz (*Oryza sativa)*, para 11 características fenotípicas, utilizando a estimação única e a estimação simultânea dos efeitos das regiões genômicas. Os valores de WPPA juntamente com os cromossomos e a posição inicial das regiões foram plotados no gráfico *manhattan* e são apresentados na Figura 1. Para todas a 11 características fenotípicas, a estimação simultânea apresentou mais regiões associada, foram consideradas como associadas regiões que obtiveram um valor de WPPA igual ou superior a 0,95, como proposto por Fernando e Garrick (2013) e Fernando et al. (2017). A estimação simultânea detectou regiões associadas a todas as 11 caraterísticas fenotípicas (Tabela 4). Já a estimação única apresentou regiões associadas somente para o comprimento da folha da bandeira, altura da plana, fertilidade do pináculo, teor de proteína e resistência a Brusone (Tabela 5).

A estimação simultânea encontrou, em uma grande maioria, as mesmas regiões associadas, no cromossomo 3 (posição inicial – posição final: 36827238 - 37243733 e 37037934 - 37243733), no cromossomo 8 (28349924 - 28467527) e no cromossomo 11 (30754510 - 30751187 e 30544386- 30751187) para as características fenotípicas em estudo

(Tabela 4). Isso sugere que um único gene controle diversas características fenotípicas. Nos estudos de Suela et al. (2022), os métodos bayesianos aplicados ao mesmo banco de dado de arroz, detectaram regiões pertencentes ao mesmo cromossomo para todas as características, mas as regiões encontradas por eles diferem das regiões encontradas na estimação simultânea deste estudo.

Tabela 3. Médias e erros-padrão para taxa de falso positivo (FP), poder de detecção (PD), área sob a curva ROC (Área), porcentagem de regiões associadas nos cromossomos 11 e 12 (N (%)) e porcentagem da variância genética capturada (PE) para os cenários avaliados com estimação simultânea e única, para o limiar ótimo e o limiar de 0,95.

Limiar	Cenário	Estimação	FP	Poder	Área	N (%)	PE
Ótimo	1	Simultânea	0,18 ± 0,07	0,73 ±0,08	0,72 ±0,05	18 ± 7	89,86 ± 2,93
		Única	0,11 ± 0,05	0,62 ± 0,09	0,64 ± 0,05	9 ± 4	77,34 ± 5,36
	2	Simultânea	0,26 ±0,03	0,74 ±0,05	0,76 ± 0,03	21 ± 4	81,14 ±3,67
		Única	0,34 ± 0,04	0,65 ± 0,03	0,61 ± 0,04	28 ± 4	70,09 ± 2,85
	3	Simultânea	0,51 ± 0,03	0,52 ± 0,03	0,48 ± 0,01	56 ± 5	55,92 ± 2,81
		Única	0,48 ± 0,03	0,42 ± 0,03	0,43 ± 0,01	54 ± 3	48,42 ± 3,32
0,95	1	Simultânea	0,00 ± 0,00	0,38 ±0,05	0,72 ±0,05	0,00 ± 0,00	70,72 ± 8,10
		Única	0,01 ± 0,01	0,30 ±0,06	0,64 ±0,05	0,00 ± 0,00	46,55 ± 10,29
	2	Simultânea	0,01 ± 0,01	0,00 ± 0,00	0,75 ±0,03	0,00 ± 0,00	0,00 ± 0,00
		Única	0,00 ± 0,00	0,00 ± 0,00	0,61 ±0,03	0,00 ± 0,00	0,00 ± 0,00
	3	Simultânea	0,00 ± 0,00	0,00 ± 0,00	0,47 ±0,01	0,00 ± 0,00	0,00 ± 0,00
		Única	0,00 ± 0,00	0,00 ± 0,00	0,43 ±0,006	0,00 ± 0,00	0,00 ± 0,00

Para a altura de planta, a estimação simultânea encontrou no cromossomo 8 uma região associada e está localizada na região onde Huang et al. (1996) encontraram um QTL (5421297 – 25592993). Como os efeitos foram estimados simultaneamente, ao se ter um forte desequilíbrio de ligação entre regiões vizinhas, o método pode ter capturado regiões próximas as regiões identificadas pela literatura. Dessa forma, para a altura de planta foi encontrado no cromossomo 3 a região (36827238 – 37035869) que é próxima ao QTL (32945649 – 36396286) identificado no cromossomo 3 por Li et al. (2003). A estimação única, considerando altura de plantas, encontrou o QTL (ph1.2 – 34937981– 41541798) que foi identificado por Thomson et at. (2003).

Considerando a característica de número de panículas por planta, a estimação simultânea encontrou regiões no cromossomo 8 que são próximas as regiões encontradas por Ishimaru et al. (2001) na posição 26138197–28248441 e Lanceras et.al. (2004) na posição 22471837 – 27735542 (Tabela 6). Pela estimação única não foi possível encontrar nenhuma região. Para comprimento de panícula, a estimação simultânea encontrou regiões próximas, no cromossomo 8 e no cromossomo 3, as encontradas pela literatura por Kobayadhi et al. (2003) na posição 22885196 – 27831422 e Hittalmani et al. (2003) na posição 36395749 – 36396286 (Tabela 6). Já a estimação única não encontrou nenhuma região. Para as características de largura de comprimento da folha da bandeira, flores por panículas, largura para folha de bandeira, número de sementes por panícula, nenhuma das regiões encontradas foram relacionadas a nenhum banco de genes anotados nos bancos de dados. Com a estimação simultânea foi possível capturar mais regiões associadas as características fenotípicas no banco de dados de arroz em detrimento a estimação única.

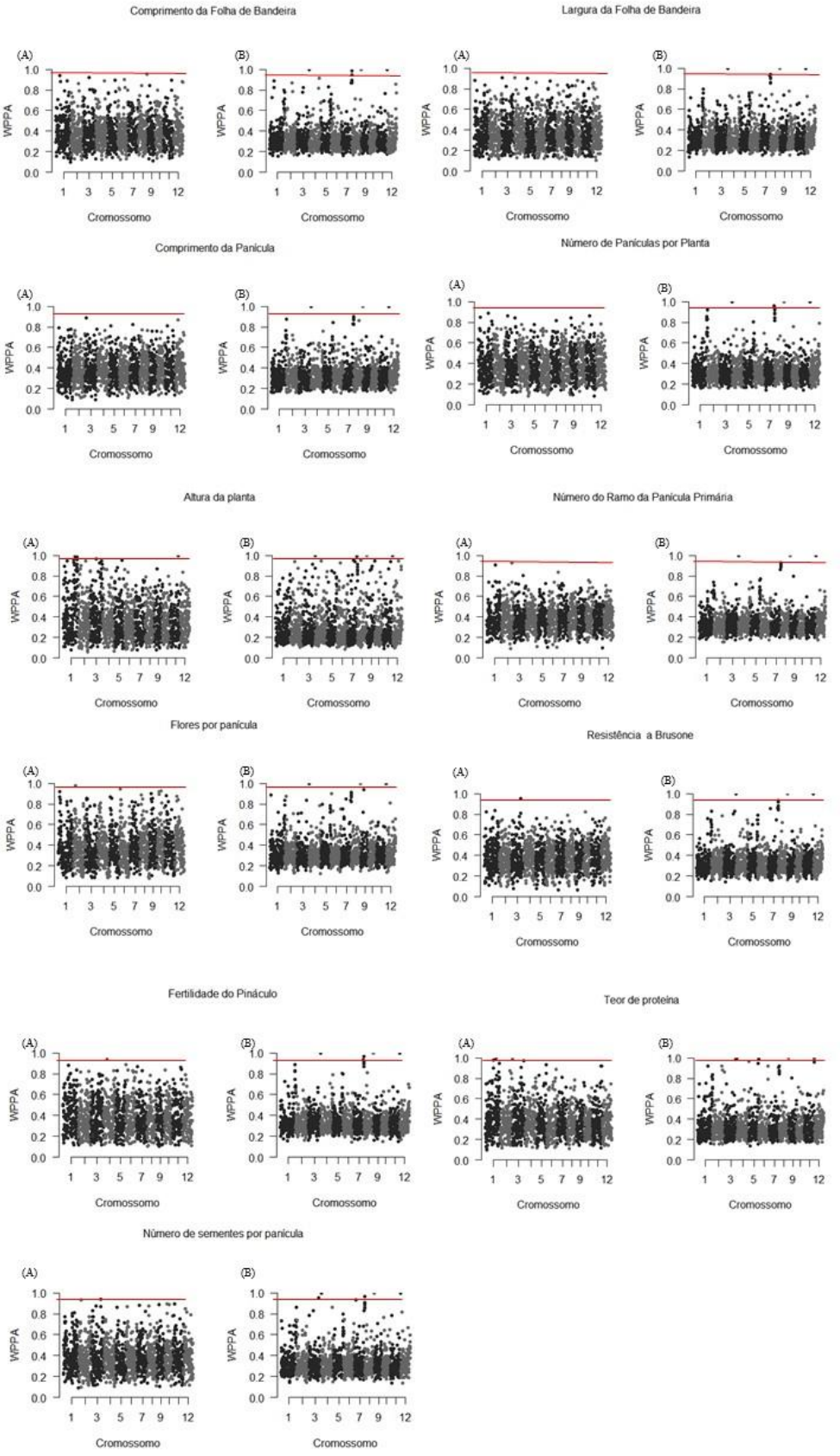

Figura 1. Gráficos de Manhattan para características do arroz considerando a estimação única (A) e a estimação simultânea (B).

Tabela 4. Regiões associadas identificadas através da estimação simultânea, identificando o cromossomo (Chr.), a posição inicial da região, posição final da região e probabilidade de associação *a posteriori* da janela (*Window Posterior Probability of Association* -WPPA).

Característica fenotípica	Chr.	Posição inicial (bp)	Posição final (bp)	*WPPA*
CBF	3	37037934	37243733	1.00
	3	36827238	37035869	0.99
	7	29512969	29716783	0.98
	8	28349924	28467527	1.0
	11	30754510	30790496	0.99
	11	30544386	30751187	0.99
LFB	3	36827238	37035869	1.0
	3	37037934	37243733	0.99
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0
NPP	3	36827238	37035869	1.0
	3	37037934	37243733	0.99
	7	28835594	29040548	0.96
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0
NRPP	3	36827238	37035869	1.0
	3	37037934	37243733	1.0
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0
AP	3	36827238	37035869	0.99
	3	37037934	37243733	0.99
	7	28835594	29040548	0.99
	8	28349924	28467527	1.0
	8	10505299	10708432	0.95
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0
CP	3	37037934	37035869	1.0
	3	36827238	37243733	0.99
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0
NSPP	3	37037934	37035869	1.0
	3	36827238	37243733	0.99
	7	29946177	30152555	0.96
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0
FLP	3	36827238	37243733	1.0
	3	37037934	37035869	0.99
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	1.0

FP	3	36827238	37243733	1.0
	3	37037934	37035869	0.99
	7	29512969	29716783	0.96
	8	28349924	28467527	1.0
	11	30544386	30751187	1.0
	11	30754510	30790496	0.96
TP	3	37037934	37035869	0.99
	3	36827238	37243733	1.0
	3	32856764	33047141	0.98
	4	34943601	35148884	0.96
	5	29211489	29418755	0.99
	8	28349924	28467527	1.0
	11	30754510	30751187	0.99
	11	30544386	30751187	0.99
RB	3	37037934	37035869	1.0
	3	36827238	37243733	0.99
	8	28349924	28467527	1.0
	11	30754510	30751187	0.99
	11	30544386	30751187	1.0

CFB-Comprimento da folha bandeira; LFB- largura da folha bandeira; NPP- número de panículas por planta; NRPP - número de ramos da panícula primária; AP- altura da planta; CP- comprimento da panícula; FLP- Flores por panícula; RB- Resistência à Brusone; FP-fertilidade da panícula; TP- teor de proteína; NSPP- número de sementes por panícula.

Tabela 5. Regiões associadas identificadas através da estimação única, identificando o cromossomo (Chr.), a posição inicial da região, posição final da região e probabilidade de associação *a posteriori* da janela (*Window Posterior Probability of Association* -WPPA).

Característica fenotípica	Chr.	Posição inicial (bp)	Posição final (bp)	*WPPA*
CBF	8	25718163	25922432	0.95
AP	1	39775856	39984340	0.99
	1	36240171	36446723	0.98
	1	41346576	41555667	0.98
	1	37123586	37316370	0.96
	1	42201610	42409233	0.95
	1	31925204	32129017	0.95
	3	17048990	17238494	0.96
	5	22110402	22308407	0.95
	11	23987074	24176941	0.99
FP	2	6586892	6781371	0.98
TP	1	31706523	31900761	0.99
	1	24430588	24613272	0.97
	2	36365085	36570079	0.99
	3	32856764	33047141	0.97
RB	3	28368015	28577096	0.95

CFB-Comprimento da folha bandeira; AP- altura da planta; FP- Floretes por panícula; RB-Resistência à Brusone; TP- teor de proteína.

Tabela 6. Lista de regiões de loci de características quantitativas (cromossomo, posição inicial da região e posição final da região) associadas ao comprimento da panícula (CP), resistência a brusone (RB), altura de planta (AP) e número de panículas por planta (NPP) relatados no banco de dados Q-TARO e *Gramene* QTL.

Caraterística Fenotípica	Cromossomo	Posição inicial	Posição Final	Referência
CP	8	22885196	27831422	Kobayashi et al. (2003)
	3	34925198	35443016	Cui et al. (2002)
	3	36395749	36396286	Hittalmani et al. (2003)
RB	8	4105519	17438003	Wang et al. (1994)
	11	13671613	28412347	Tabien et al. (2002)
AP	8	5421297	25592993	Huang et al. (1996)
	11	28409788	28412347	Li et al. (2003)
	3	34230739	34231482	Macmillan et al. (2006)
	3	32945649	36396286	Li et al. (2003)
	1	34937981	41541798	Thomson et al. (2003)
NPP	8	26138197	28248441	Ishimaru et al. (2001)
	8	22471837	27735542	Lanceras et al. (2004)

4- Conclusão

Para os dados simulados, a eficiência da estimação simultânea dos efeitos das regiões genômicas utilizando a estimação bayesiana, apresentou resultados satisfatórios, uma vez que apresentou resultados superiores ou semelhantes a estimação única. Nos dados de arroz, a estimação simultânea detectou uma quantidade superior de regiões já relatadas na literatura em detrimento a estimação única. Foram também encontradas novas regiões que podem ser potencialmente estudas em análises pós-GWAS. Essa é uma metodologia que apresenta potencial para aplicação, descoberta e investigação de novas regiões genômicas associadas a características fenotípicas.

Referências Bibliográficas

AMMIRAJU, J. S., Luo, M., Goicoechea, J. L., Wang, W., Kudrna, D., Mueller, C., Talag, J., Kim, H., Sisneros, N. B., Blackmon, B., Fang, E., Tomkins, J. B., Brar, D., MacKill, D., McCouch, S., Kurata, N., Lambert, G., Galbraith, D. W., Arumuganathan, K., Rao, K., … Wing, R. A. (2006) The Oryza Bacterial Artificial Chromosome library resource: Construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. **Genome Research**, *16*, 140–147.

AZEVEDO, C. F. et al. *GenomicLand*: Software for genome-wide association studies and genomic prediction. **Acta Scientiarum. Agronomy**, v. 41, 2019.

BATES, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. https://doi.org/10.48550/arXiv.1506.04967

BENNEWITZ, J., Edel, C., Fries, R., Meuwissen, T.H., Wellmann, R. (2017). Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. **Genetics Selection Evolution**, *49*, 7.

COVARRUBIAS-PAZARAN G (2018) Software update: Moving the R package sommer to multivariate mixed models for genome-assisted prediction.

CUI-K-H, Peng-S-B, Xing-Y-Z, Yu-S-B, Xu-C-G, "Genetic analysis of the panicle traits related to yield sink size of rice", **Acta genetica Sinica**, 2002, vol. 29, pp. 144-152

DE LOS CAMPOS, G; Pérez Rodrigues, P. BGLR: Bayesian Generalized Linear Regression. R package version 1.0.9**.** Disponível em: < https://cran.r-project.org/web/packages/BGLR/BGLR.pdf>. Acesso em: 8 Jan. 2022.

FERNANDO, R. et al. Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. **Journal of Agricultural, Biological and Environmental Statistics**, v. 22, n. 2, p. 172-193, 2017.

Fernando, R. L., & Garrick, D. (2013). **Bayesian methods applied to GWAS**. In C. Gondro et al. (Eds), Genome-wide association studies and genomic prediction (p. 237-274). **Humana Press**, Totowa, NJ

FERNANDO, R. L., Nettleton, D., Southey, B. R., Dekkers, J. C. M., Rothschild, M. F., Soller, M. (2004). Controlling the proportion of false positives in multiple dependent tests. **Genetics**, 166(1), 611-9.

GAMERMAN, D., & Lopes, H. F. (2006). Markov chain Monte Carlo: stochastic simulation for Bayesian inference. **CRC press**.

GAYNOR, R. Chris, Gregor Gorjanc, and John M. Hickey. 2021. AlphaSimR: an R package for breeding program simulations**. G3 Gene|Genomes|Genetics** 11(2):jkaa017. https://doi.org/10.1093/g3journal/jkaa017.

GEMAN S., Geman D., 1984. Stochastic relaxation, gibbs distributions and the Bayesian restoration of images. **IEEE Trans. Pattern Anal. Mach. Intell**. **6**(6): 721–741

GEMAN S., Geman D., 1984. Stochastic relaxation, gibbs distributions and the Bayesian restoration of images. **IEEE Trans. Pattern Anal. Mach. Intell**. **6**(6): 721–741

GEWEKE, J., (1992) Evaluating the accurary of sampling-based approaches to the calculation of posterori moments. Bayesian Statistics 4 (eds. Bernardo, J.M; Berger, J.O; Dawid, A.P.; Smith, A.F.M.) **New York : Oxford University Press**, p.625-631

GODDARD, M.E.; Hayes, B.J.; Meuwissen, T.H. 2011. Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of animal breeding and genetics** 128(6): 409-421.

HITTALMANI-S, Huang-N, Courtois-B, Venuprasad-R, Shashidhar-H-E, Zhuang-J-Y, Zheng-K-L, Liu-G-F, Wang-G-C, Sidhu-J-S, Srivantaneeyakul-S, Singh-V-P, Bagali-P-G, Prasanna-H-C, McLaren-G, Khush-G-S, "Identification of QTL for growth- and grain yield-related traits in rice across nine locations of Asia", **Theoretical and applied genetics**, 2003, vol. 107, pp. 679-690

HUANG-N, Courtois-B, Khush-G-S, Lin-H-X, Wang-G-L, Wu-P, Zheng-K-G, "Association of quantitative trait loci for plant height with major dwarfing genes in rice", **Heredity**, 1996, vol. 77, pp. 130-137

ISHIMARU-K, Yano-M, Aoki-N, Ono-K, Hirose-T, Lin-S-Y, Monna-L, Sasaki-T, Ohsugi-R, "Toward the mapping of physiological and agronomic characters on a rice function map: QTL analysis and comparison between QTLs and expressed sequence tags", **Theoretical and applied genetics**, 2001, vol. 102- (6/7-)., pp. 793-800

KIM S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, Ecker JR, Weigel D andNordborg M (2007) Recombination and linkage disequilibrium in Arabidopsis thaliana. **Nature Genetics** 39, 1151-1155

KOBAYASHI-S, Fukuta-Y, Sato-T, Osaki-M, Khush-G-S, "Molecular marker dissection of rice (Oryza sativa L.) plant architecture under temperate and tropical climates", **Theoretical and applied genetics**, 2003, vol. 107, pp. 1350-1356

LANCERAS-J-C, Pantuwan-G, Jongdee-B, Toojinda-T, "Quantitative trait Loci associated with drought tolerance at reproductive stage in rice", **Plant physiology**, 2004, vol. 135, pp. 384-399

LIMA, L.P; et al. Evaluation of bayesian methods of genomic association via chromosomic regions using simulated data. **Scientia Agricola**, v. 79, n. 3, 2022.

LI-Z-K, Yu-S-B, Lafitte-H-R, Huang-N, Courtois-B, Hittalmani-S, Vijayakumar-C-H, Liu-G-F, Wang-G-C, Shashidhar-H-E, Zhuang-J-Y, Zheng-K-L, Singh-V-P, Sidhu-J-S, Srivantaneeyakul-S, Khush-G-S, "QTL x environment interactions in rice. I. heading date and plant height", **Theoretical and applied genetics**, 2003, vol. 108, pp. 141-153

MACMILLAN-K, Emrich-K, Piepho-H-P, Mullins-C-E, Price-A-H, "Assessing the importance of genotype x environment interaction for root traits in rice using a mapping population II: conventional QTL analysis", **Theor Appl Genet**, 2006, vol. 113, pp. 953-964

METZ, C. E. Basic principles of ROC analysis. In: **Seminars in nuclear medicine**. WB Saunders, p. 283-298, 1978.

NAGAMINE, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, Hicks, A. A., Pramstaller, P. P., Hastie, N., Wright, A. F., Haley, C. S. (2012). Localising loci underlying complex trait variation using regional genomic relationship mapping. **PLoS One**, 7:e46501.

NAGAMINE, Y., Pong-Wong, R., Navarro, P., Vitart, V., Hayward, C., Rudan, I., Campbell, H., Wilson, J., Wild, Hicks, A. A., Pramstaller, P. P., Hastie, N., Wright, A. F., Haley, C. S. (2012). Localising loci underlying complex trait variation using regional genomic relationship mapping. **PLoS One**, 7:e46501.

PETERS, S. O. et al. Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. **Journal of animal science**, v. 90, n. 10, p. 3398-3409, 2012.

PLUMMER, M. et al. CODA: convergence diagnosis and output analysis for MCMC. **R news**, v. 6, n. 1, p. 7-11, 2006.

RESENDE, M.D.V.; Silva, F.F.; Lopes, P.S.; Azevedo, C.F. Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial. **Viçosa: Universidade Federal de Viçosa/Departamento de Estatística**. 2012. 291 p. http://www.det.ufv.br/ppestbio/corpo_docente.php

Schmid, M.; Bennewitz, J. 2017. Invited review: Genome-wide association analysis for quantitative traits in livestock–a selective review of statistical models and experimental designs. **Archiv fuer Tierzucht** 60(3): 335-346.

SHIN, J.; LEE, C. Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. **Genomics**, v. 105, n. 1, p. 1-4, 2015.

SUELA, M.M., Azevedo, C.F., Nascimento, M., Nascimento, A.C.C. and de Resende, M.D.V. (2022), Regional Heritability Mapping and Genome-Wide Association Identify Loci For Rice Traits. **Crop Sci**.. Accepted Author Manuscript.

TABIEN-E, Li-Z, Paterson-H, Marchetti-A, Stansel-W, Pinson-M, "Mapping QTLs for field resistance to the rice blast pathogen and evaluating their individual and combined utility in improved varieties", **Theoretical and applied genetics**, 2002, vol. 105, pp. 313-324

THOMSON-M-J, Tai-T-H, McClung-A-M, Lai-X-H, Hinga-M-E, Lobos-K-B, Xu-Y, Martinez-C-P, McCouch-S-R, "Mapping quantitative trait loci for yield, yield components and

morphological traits in an advanced backcross population between Oryza rufipogon and the Oryza sativa cultivar Jefferson", **Theoretical and applied genetics**, 2003, vol. 107, pp. 479-493

UFFELMANN, E., Huang, Q.Q., Munung, N.S. et al. Genome-wide association studies. **Nat Rev Methods Primers** 1, 59 (2021).

USAI, M. G., Gaspa, G., Macciotta, N. P., Carta, A., Casu, S. (2014). XVIth QTLMAS: simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation. **BMC Proceedings**, 8, 1–9.

VOS PG, Paulo MJ, Voorrips RE, Visser RG, van Eck HJ and van Eeuwijk FA (2017) Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. **Theoretical and Applied Genetics** 130(1), 123-135.

WANG-G-L, Mackill-D-J, Bonman-J-M, McCouch-S-R, Champoux-M-C, Nelson-R-J, "RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistance rice cultivar", **Genetics**, 1994, vol. 136, pp. 1421-1434

Zhao, K., Tung, C. W., Eizenga, G. C., Wright, M. H., Ali, M. L., Price, A. H., Norton, G. J., Islam, M. R., Reynolds, A., Mezey, J., McClung, A. M., Bustamante, C. D., McCouch, S. R. (2011). Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nature Communication*, *2*, 467.

ZIEGLER, A.; König, I. R.; Thompson, J. R. Biostatistical aspects of genome-wide association studies. **Biometrical Journal: Journal of Mathematical Methods in Biosciences**, v. 50, n. 1, p. 8-28, 2008.