VINICIUS SILVA BEGNAMI

SELEÇÃO DE MARCADORES UTILIZANDO PROBABILIDADE *A POSTERIORI* DE INCLUSÃO NO MODELO PARA PREDIÇÃO GENÔMICA

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

VIÇOSA - MINAS GERAIS
2023

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

T

B471s 2023

Begnami, Vinicius Silva, 1997-

Seleção de marcadores utilizando probabilidade a *posteriori* de inclusão no modelo para predição genômica / Vinicius Silva Begnami. – Viçosa, MG, 2023.

1 dissertação eletrônica (56 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.2023.607

Modo de acesso: World Wide Web.

1. Bioestatítica. 2. Genômica. 3. Marcadores genéticos. 4. Melhoramento genético. 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. 570.15195

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

VINICIUS SILVA BEGNAMI

SELEÇÃO DE MARCADORES UTILIZANDO PROBABILIDADE *A POSTERIORI* DE INCLUSÃO NO MODELO PARA PREDIÇÃO GENÔMICA

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: 18 de julho de 2023.

Assentimento:

gov.br
Documento assinado digitalmente
VINICIUS SILVA BEGNAMI
Data: 02/10/2023 16:55:36-0300
Verifique em https://validar.iti.gov.br

Vinicius Silva Begnami
Autor

gov.br
Documento assinado digitalmente
CAMILA FERREIRA AZEVEDO
Data: 03/10/2023 09:24:16-0300
Verifique em https://validar.iti.gov.br

Camila Ferreira Azevedo
Orientadora

À minha família.

AGRADECIMENTOS

Agradeço primeiramente a Deus por ter me guiado e abençoado ao longo de minha vida. Permitiu e deu forças para chegar até aqui.

A minha mãe Maria do Carmo, ao meu irmão Marcelo e a todos os familiares que sempre estiveram ao meu lado nos momentos difíceis e enfermos.

A minha noiva e futura esposa Daniele que acompanhou de perto toda a luta e os desafios enfrentados nesta passagem pela UFV. Ela foi (é/será) minha motivação diária. Sempre esteve (estará) ao meu lado me dando apoio, carinho, atenção e cuidados.

Aos amigos da zona rural (vulgo roça) passando pelo ensino fundamental e médio que sempre me proporcionaram momentos de alegrias e risadas.

Aos professores do ciclo básico, fundamental e médio que sempre acreditaram e investiram em mim. Meu mais profundo agradecimentos aos professores Tininha (*in memoriam*), Sueli Antonucci (“prezinho”), Ricardo Miranda, Mirian Becari (matemática), João Renato (ciências), Manoel, Eder (geografia), Elizete (inglês), Vanda (religião), Elenice Capobiango, Solange (português), Virgínia Samôr, Ricardo Fiorilo (biologia), Alessandra (Física), Luciano Lino (química), Silvio Mota, Michel Alex (filosofia), Zulmira e José Geraldo (diretores).

Aos amigos da graduação na UFV, meu agradecimento pela amizade, conselhos, orações e pelos bons momentos que passamos juntos. Meu obrigado ao Fagner Darlan, Cesar Augusto, Marco Luís, Cláudio Henrique, Gabriel Morkazel (Zé), Bianca Assis, Juliana Eliete, Joel Teixeira, Marco Aurélio, Michel Rena, Adalto Luís, Wilker Teodoro, Jean Carlos, Daniel Mendes, Alvim Lucas, Emanuel Ferrari, Alípio, Augusto Goretti, Gabriel Júnior e Ronan Capobiango.

A coordenadora do programa de tutoria da UFV, Daniela Rodrigues que me acolheu e cuidou de mim como um filho, muito obrigado.

Aos amigos/irmãos do alojamento 1911 que me proporcionaram um dos melhores períodos que vivi até hoje. Foram várias histórias e momentos épicos que me fizeram não duvidar de nada nessa vida. Me ensinaram a ser organizado, paciente, viver em harmonia mesmo estando com pessoas de diferentes culturas e mentalidades, ter jogo de cintura, a ser mais responsável, que a vida é difícil e nada vai cair do céu e, por fim, me ensinaram a me esforçar para que o meu melhor fosse feito. Sem sobra de dúvidas, o alojamento me tornou uma pessoa melhor e mais

preparada para viver em sociedade. Fica meus agradecimentos aos amigos que fizeram parte destes momentos: Lucas (Mocotó), Victor (Mouse), Michael (Jackson), Patrick (Queijo), Joel (Aderbal), Igor (Avatar), Laio (Gepeto), Kayto (Katiuço), Alex (Galfo), Cesár (Berg), Joel (Sadboy), João (Patrimônio), Natanael (Hamut) e Adriano (Pogba).

Aos amigos da pós-graduação que me apoiaram e me ajudaram tanto que falta palavras para expressar minha gratidão. Agradeço muito a todos vocês: Aline Marçal, Luciano Gonçalves, Samantha Gouvêa, Noé Eiterer, Matheus Massariol, Thais Hashimoto, Wagner Barbosa, Gabriela França e Mayara Rodrigues.

Aos professores da graduação em matemática, em especial ao Rogério Picanço, Marli Duffles, Rejane Faria e Laís Moreira.

Aos professores e funcionários do Programa de Pós-graduação em Estatística Aplicada e Biometria pelos ensinamentos e conselhos.

Aos integrantes dos Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE) e Laboratório de Análises e Pesquisas em Estatísticas (LAPEA) pelo companheirismo, apoio e aprendizado.

A minha orientadora Professora Doutora Camila Ferreira Azevedo, agradeço pela paciência, pelos conselhos, pelos ensinamentos, pelas críticas e pelas oportunidades. Serei eternamente grato por tudo que fez e faz por mim.

Aos meus coorientadores Professor Doutor Moysés Nascimento e Professora Doutora Ana Carolina Campana Nascimento pelos ensinamentos e por sempre estarem dispostos a me ajudar quando precisava.

A banca composta pelas Professoras Doutoras Leísa Pires Lima e Laís Mayara Azevedo Barroso, agradeço pelas contribuições para esta pesquisa.

A Universidade Federal de Viçosa, pela oportunidade de realizar minha graduação e pós-graduação.

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

Aos que, de alguma forma, contribuíram e me ajudaram. Muito obrigado.

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001.

“Talvez não tenha conseguido fazer o melhor, mas lutei para que o melhor fosse feito. Não sou o que deveria ser, mas Graças a Deus, não sou o que era antes”.

(Marthin Luther King)

RESUMO

BEGNAMI, Vinicius S., M.Sc., Universidade Federal de Viçosa, julho de 2023. **Seleção de Marcadores Utilizando Probabilidade *a Posteriori* de Inclusão no Modelo de Predição Genômica**. Orientadora: Camila Ferreira Azevedo.

Com o aumento constante da população mundial, a demanda por alimentos está crescendo diariamente, embora as áreas agricultáveis estejam chegando ao seu limite territorial. Uma solução para enfrentar esse desafio é a aplicação do melhoramento genético, que ganha cada vez mais destaque devido à sua capacidade de aumentar a produtividade e melhorar a qualidade dos alimentos em uma área de cultivo limitada. Com os avanços na genética molecular, é possível obter informações genéticas diretamente do DNA por meio de marcadores moleculares, especialmente os SNP (*Single Nucleotide Polimorphism*), que têm sido utilizados em estudos de Seleção Genômica Ampla (GWS, *Genome Wide Selection*). A GWS busca estimar os valores genéticos genômicos (GEBV, *Genomic Estimated Breeding Value*) dos indivíduos com base em informações genotípicas. No entanto, ao ajustar o modelo de predição, a alta dimensionalidade e multicolinearidade representam desafios, uma vez que o número de marcadores é muito superior ao número de indivíduos avaliados. Como nem todos os marcadores do genoma influenciam uma característica fenotípica específica, é comum realizar uma seleção prévia desses marcadores. Neste contexto, este estudo propõe a seleção os marcadores mais relevantes para a predição genômica com base em sua probabilidade de inclusão. Para atingir esse objetivo, a dissertação foi dividida em dois capítulos. O Capítulo 1 consiste em uma revisão de literatura sobre as metodologias estatísticas que serão aplicadas no próximo capítulo. O Capítulo 2 tem como principal objetivo a seleção dos marcadores mais relevantes a partir de um conjunto de dados reais originários do arroz *Oryza Sativa*. Este conjunto de dados contém 413 acessos genotipados para 44.100 marcadores do tipo SNP. A seleção dos marcadores é realizada com base na probabilidade a posteriori de inclusão, com cálculos apoiados na matriz de efeitos dos marcadores moleculares, estimados por meio do método BayesDπ, e no número total de iterações salvas. Após a seleção dos marcadores mais relevantes, eles são agrupados em conjuntos de 2.000, 4.000, 6.000, ..., até 36.901, de acordo com sua importância. Em seguida, cada grupo tem

seu efeito estimado pelo método BayesA, e a capacidade preditiva do modelo de predição é calculada. Essa métrica é comparada com a capacidade preditiva dos modelos de predição ajustados pelos métodos bayesianos BayesA e BayesDπ, quando aplicados separadamente e sem a prévia seleção dos marcadores. Os resultados obtidos indicam que a seleção de marcadores mais relevantes para a predição genômica se mostra eficaz, com alta capacidade preditiva em comparação aos métodos BayesA e BayesDπ quando usados isoladamente e sem a prévia seleção. Além disso, a probabilidade *a posteriori* de inclusão também demonstrou ser eficaz na compreensão da arquitetura genética da característica em estudo. Assim, a seleção de marcadores contribui para a redução da alta dimensionalidade, o aumento da capacidade preditiva do modelo de predição genômica e a redução do esforço computacional, abordando problemas recorrentes na seleção genômica.

Palavras-chave: Marcadores Moleculares. Arroz. Seleção Genômica. Genética. Melhoramento Genético.

ABSTRACT

BEGNAMI, Vinicius S., M.Sc., Universidade Federal de Vicosa, July, 2023. **Marker Selection Using Posterior Probability of Inclusion in the Genomic Prediction Model**. Adviser: Camila Ferreira Azevedo.

With the growing global population, the demand for food is increasing every day, even as arable land areas approach their territorial limits. One solution to address this challenge is the practice of genetic improvement, which is gaining increasing prominence due to its ability to enhance productivity and improve the quality of food within the confines of existing cultivation areas. With advances in molecular genetics, it has become possible to obtain genetic information directly from DNA through molecular markers, particularly Single Nucleotide Polymorphism (SNP), which have been used in Genome-Wide Selection (GWS) studies. GWS aims to estimate genomic breeding values (GEBV) of individuals under study based on genotypic information. However, when adjusting the prediction equation, high dimensionality and multicollinearity pose challenges, as the number of markers is much larger than the number of evaluated individuals. Since not all markers in the genome influence a specific phenotypic trait, it is common practice to conduct a prior selection of these markers. In this context, this study proposes to select the most important markers for genomic prediction based on their inclusion probability. To achieve this, the dissertation is divided into two chapters. Chapter 1 consists of a literature review on the statistical methodologies to be applied in the following chapter. Chapter 2 aims to select the most important markers from a real dataset derived from Oryza Sativa rice, containing 413 genotyped accessions with 44,100 SNP markers, using their posterior inclusion probability. The calculation of this probability is supported by the marker molecular effects matrix, estimated through the BayesDπ method, and the total number of saved iterations. After the selection of the most important markers, they are grouped into sets of 2,000, 4,000, 6,000, ..., up to 36,901 markers, according to their importance. Subsequently, each group has its effect estimated by the BayesA method, and the predictive ability of the prediction model is calculated. This metric is compared to the predictive ability of prediction models adjusted by the Bayesian methods, BayesA and BayesDπ separately, without prior marker selection. The results obtained indicate that the selection of the most important markers for genomic

prediction has proven to be efficient, as its exhibits high predictive ability compared to the BayesA and BayesDπ methods when used in isolation and without prior selection. Furthermore, the posterior inclusion probability has also proven effective in understanding the genetic architecture of the trait under study. Thus, marker selection contributes to the reduction of high dimensionality, an increase in the predictive ability of the genomic prediction model, and a reduction in computational effort, addressing recurring issues in genomic selection.

Keywords: Molecular Markers. Rice. Genomic Selection. Genetics. Genetic Breeding.

SUMÁRIO

INTRODUÇÃO GERAL 12
CAPÍTULO 1 15
REVISÃO DE LITERATURA 15
1. Seleção Genômica 15
2. Inferência Bayesiana 17
3. Regressão Linear Múltipla 18
4. Algoritmo MCMC 20
4.1 Algoritmo *Metropolis-Hasting* 21
5. Métodos Bayesianos Aplicados a Predição Genômica 23
5.1 BayesA 23
5.2 BayesDπ 25
6. Validação cruzada K-*fold* 26
7. Métricas de comparação 27
7.1 Herdabilidade 28
7.2 Viés da Predição 28
7.3 Capacidade Preditiva 29
8. Referências 29
CAPÍTULO 2 33
SELEÇÃO DE MARCADORES UTILIZANDO PROBABILIDADE *A POSTERIORI* DE INCLUSÃO NO MODELO PARA PREDIÇÃO GENÔMICA EM DADOS DE ARROZ 33
RESUMO 33
1. Introdução 35
2. Materiais e métodos 37
2.1 Dados reais 37
2.2 Modelo de Regressão Linear Múltipla Sob Enfoque Bayesiano 38
2.2.1 Método BayesA 39
2.2.2 Método BayesDπ 40
2.3 Seleção de marcadores 41
2.4 Comparação de metodologias 42
2.5 Recursos computacionais 42
3. Resultados e discussões 43
4. Conclusão 53
5. Referências 53

INTRODUÇÃO GERAL

Atualmente, a população mundial é estimada em aproximadamente 8 bilhões de habitantes e, há uma perspectiva de que, em 2059, o número de habitantes atinja 10 bilhões (Nações Unidas, 2022). Com isso, surge a necessidade de produzir mais alimentos em uma quantidade limitada de área agricultável. Nesse cenário, o melhoramento genético vem se tornando uma importante ferramenta para a sobrevivência humana (BORÉM, MIRANDA e NETO, 2021). Diante deste propósito, o avanço da genética molecular tornou possível utilizar informações vindas diretamente do DNA, por meio de marcadores moleculares, tornando o processo de seleção de indivíduos geneticamente superiores mais rápido, eficiente e menos oneroso (RESENDE, 2008).

Dentre os diversos marcadores, destaca-se o marcador denominado Polimorfismo de Nucleotídeo Único (SNP, *Single Nucleotide Polimorphism*) que consiste em uma alteração de um único nucleotídeo na molécula de DNA. Esses marcadores são extremamente abundantes no genoma, possuem facilidade em genotipagem em larga escala e baixo custo em comparação aos demais (RESENDE et al., 2014).

No início do século XXI, Meuwissen et al. (2001) idealizaram a Seleção Genômica Ampla (GWS, *Genomic Wide Selection*), que visa obter os valores genéticos genômicos estimados (GEBV, *Genomic Estimated Breeding Value*) dos indivíduos com base apenas nas informações genotípicas. Para isso, é necessário estimar o efeito genético de um grande número de marcadores utilizando dados fenotípicos e genotípicos de uma população de treinamento, afim de capturar os efeitos de todos os locos de um caráter quantitativo. No entanto, é necessário a existência de desequilíbrio de ligação (LD, *Linkage Disequilibrium*) entre os marcadores e o QTL (*Quantitative Trait Loci*), uma vez que somente tais marcadores contribuirão para a determinação dos fenótipos e na explicação da variabilidade genética (HAYES et al., 2009; CAVALCANTI et al., 2012).

Após estimar os efeitos genéticos dos marcadores, eles são testados em uma população de validação, em seguida, os marcadores podem ou não serem selecionados para explicação de grande parte da variância genética do caráter em estudo. Com isso, os efeitos genéticos estimados dos marcadores (selecionados ou

não) são somados e usados para predição de valores genéticos genômicos de indivíduos genotipados candidatos a seleção (RESENDE et al., 2010). A predição genômica tem a vantagem de reduzir o custo e o ciclo para o desenvolvimento de uma nova variedade nos programas de melhoramento. Por exemplo, na cultura do milho, a redução de custos pode chegar em até 50% (CROSSA et al., 2017).

As primeiras aplicações da GWS foram para o melhoramento animal e mais tarde no melhoramento de planta (WANG et al., 2018). A GWS utiliza um grande número de marcadores distribuídos em todo o genoma. No entanto, ainda há menos indivíduos fenotipados e genotipados do que o número de marcadores disponíveis (LONG et al., 2010). Essa alta dimensionalidade faz com que os graus de liberdade sejam insuficientes para estimar o efeito de todos os marcadores simultaneamente por meio dos métodos dos mínimos quadrados (NEVES et al., 2012). Caso o efeito de cada marcador fosse estimado isoladamente, e posteriormente a predição fosse realizada, estaria associada a uma baixa capacidade preditiva (DESTA et al., 2014).

Os métodos propostos para a predição genômica são capazes de solucionar tais desafios estatísticos, como a alta dimensionalidade e a consequente multicolinearidade entre os marcadores. Dentre eles, os métodos baseados em modelos mistos e os bayesianos, como o BayesA proposto por Meuwissen et al. (2001), incluem todos os marcadores moleculares no modelo de predição genômica em todas as iterações do algoritmo MCMC (*Markov Chain Monte Carlo*), o que resulta uma complexidade desnecessária para as características fenotípicas, uma vez que nem todos os marcadores estão em LD com os QTL associados a elas. Assim, torna-se útil o uso de métodos que considerem a seleção de variáveis tanto para o contexto biológico quanto para o contexto estatístico.

Neste sentido, o método BayesDπ, proposto por Habier et al. (2011), difere do BayesA ao selecionar, a cada iteração do MCMC, uma proporção π de marcadores para o ajuste do modelo e, consequentemente anular o efeito dos (1-π) marcadores restantes. Este método pressupõe que os π efeitos não-nulos assumem uma distribuição normal com média 0 e variância específica para cada marcador, além do parâmetro π assumir uma distribuição de probabilidade, podendo ser a uniforme ou beta devido ao seu espaço paramétrico (RESENDE et al., 2014).

No capítulo 1, foi realizada uma revisão de literatura sobre os aspectos da predição genômica, inferência bayesiana e dos modelos bayesianos BayesA e

BayesDπ. No capítulo 2, inspirado no algoritmo *Reversible Jump Markov Chain Monte Carlo* (RJMCMC) (GREEN, 1995) que visa ajustar um modelo com diferentes números de marcadores a cada iteração do algoritmo MCMC e calcular a probabilidade *a posteriori* de inclusão de cada marcador no contexto da predição genômica, o estudo apresentado teve como objetivo selecionar os marcadores mais importantes para um conjunto de características fenotípicas por meio da probabilidade *a posteriori* de inclusão de cada marcador. Tal probabilidade foi obtida pela razão entre o número de marcadores que tiveram efeito diferente de zero e o número de iterações salvas. Os efeitos dos marcadores foram estimados pelo método BayesDπ. Após a seleção, os marcadores foram ordenados em grupos de 2.000, 4.000, 6.000, ..., até 36.901 no qual cada grupo teve seus efeitos reestimados via BayesA. Os resultados foram comparados, em termos de capacidade preditiva, com os obtidos pelos métodos bayesianos, BayesA e BayesDπ, separadamente e sem a prévia seleção. Para elucidar este estudo, foi utilizado o conjunto de dados público de arroz *Oryza Sativa*, que possui 11 características fenotípicas, 44.100 marcadores do tipo SNP e 413 indivíduos.

CAPÍTULO 1
REVISÃO DE LITERATURA

1. Seleção Genômica

O melhoramento genético tem como objetivo selecionar indivíduos geneticamente superiores para características desejáveis, geralmente visando maior produtividade e qualidade. No passado, o melhoramento genético era baseado apenas em características fenotípicas, mas com o avanço da ciência e da biotecnologia, tornou-se possível também selecionar indivíduos com base em informações obtidas diretamente do DNA por meio de marcadores moleculares, principalmente do tipo SNP (*Single Nucleotide Polimorphism*). Esses marcadores são preferíveis por serem abundantes no genoma e apresentarem facilidade no processo de genotipagem em larga escala (RESENDE et al., 2014).

No final do século XX, Lander e Thompson (1990) propuseram um dos primeiros métodos a utilizar marcadores moleculares no melhoramento genético, denominado de Seleção Assistida por Marcador (MAS, *Marker Assisted Selection*). A MAS tem como objetivo correlacionar informações genéticas de um grupo de marcadores moleculares com os genes de interesse, levando em consideração o desequilíbrio de ligação (LD, *Linkage Disequilibrium*) entre o marcador e os diferentes genes envolvidos no controle da característica desejável. Além disso, é necessário o conhecimento prévio da associação entre o marcador e o QTL (*Quantitative Trait Loci*) (PEIXOTO et al., 2022).

Com o decorrer dos estudos, a MAS mostrou-se ser mais eficiente para características oligogênicas com alta herdabilidade e/ou monogênicas. No entanto, as características de interesse agronômico, em geral, são poligênicas e rastrear um pequeno número de genes para explicar uma pequena porção da variação genética torna a implementação da MAS inviável (ALKIMIM et al., 2020; BERNARDO, 2008). Apesar das limitações, a MAS trouxe avanços nos programas de melhoramento como agilidade no processo de desenvolvimento de novas variedades e baixo custo financeiro e operacional (ALKIMIM et al., 2017).

O avanço da genética molecular vem contribuindo de forma significativa para o melhoramento genético animal e vegetal. A informações vindas diretamente do

DNA permiti alta eficiência seletiva, rapidez na obtenção de ganhos genéticos, redução no intervalo de gerações e redução nos custos a longo prazo quando comparado com a seleção baseada em dados fenotípicos (RESENDE et al., 2010).

Dessa forma, Meuwissen et al. (2001) propuseram a Seleção Genômica Ampla (GWS, *Genomic Wide Selection*), na qual utiliza informações genotípicas para selecionar indivíduos geneticamente superiores baseado nos valores genéticos genômicos estimados (GEBV, *Genomic Estimated Breeding Value*) dos indivíduos a serem selecionados. As informações genotípicas utilizadas na GWS são os efeitos genéticos de um alto número de marcadores moleculares, preditos de forma simultânea e distribuídos no genoma buscando capturar os efeitos de todos os locos, desde que pelo menos um marcador esteja em LD com os QTL (PEIXOTO et al., 2022).

A GWS foi pouco difundida na época devido à necessidade de um grande número de marcadores, bem como ao alto custo de genotipagem dos mesmos, disponíveis naquele momento (GODDARD e HAYES, 2007). Porém, com o desenvolvimento de novas tecnologias foi possível desenvolver marcadores moleculares, principalmente do tipo SNP, e desde então a GWS vem sendo empregada em grande escala, apresentando vantagens como a não exigência do conhecimento prévio do mapeamento dos QTL, alta eficiência seletiva, velocidade na obtenção de ganhos genéticos e alta acurácia seletiva (RESENDE et al., 2008).

Os marcadores do tipo SNP têm a vantagem de serem abundantes no genoma, apresentarem facilidade na genotipagem de larga escala e baixo custo. Devido ao grande número desses marcadores, a probabilidade de se encontrar pelo menos um marcador em LD com um QTL é muito alta, seja esse marcador de grande ou pequeno efeito, e estes explicarão quase a totalidade da variação genética de uma característica quantitativa (RESENDE et al., 2014).

Devido à grande quantidade de marcadores do tipo SNP, ao ajustar um modelo de predição, nos deparamos com um problema matemático em que o número de marcadores (p) é maior que o número de indivíduos genotipados e fenotipados (n). Além disso, os marcadores estão altamente correlacionados, como resultado da superparametrização do modelo (JANNINK et al., 2010). A alta dimensionalidade, devido ao grande número de marcadores, faz com que os graus de liberdade sejam insuficientes para estimar todos os efeitos dos marcadores

simultaneamente por meio dos Métodos dos Mínimos Quadrados, ou seja, assumindo os marcadores como efeitos fixos (NEVES et al., 2012).

Na hipótese de estimar o efeito de cada marcador separadamente e, posteriormente, realizar a predição, teríamos uma baixa capacidade preditiva (DESTA et al., 2014). Já a multicolinearidade conduz à instabilidade das estimativas dos efeitos dos marcadores sobre os fenótipos por meio dos Mínimos Quadrados Ordinários, caso fosse possível essa estimação (PEIXOTO et al., 2022).

2. Inferência Bayesiana

Segundo Casella e Berger (2021), a inferência estatística tem como objetivo estudar uma população por meio de informações oriundas de uma amostra. Dessa forma, ela visa inferir a respeito de uma ou mais características da população, o que denominamos de parâmetros. Além disso, podemos elucidá-la, principalmente, por meio da inferência clássica ou da inferência bayesiana (FARAHANI et al., 2014).

Na inferência clássica, baseada no método da máxima verossimilhança, o parâmetro θ de interesse é um escalar desconhecido, porém fixo (BOLSTAD e CURRAN, 2016). Enquanto no modelo bayesiano, o parâmetro θ é desconhecido e é sempre tratado como uma variável aleatória, pois os procedimentos bayesianos consideram tudo que é desconhecido como incerto e, portanto, toda a incerteza deve ser quantificada em termos de probabilidade (MURTEIRA, 1990). Sendo assim, surge o conceito de distribuição *a priori* que consiste na distribuição de probabilidade assumida para o parâmetro θ, independente dos dados amostrais (PAULINO et al., 2018).

Ainda de acordo com Paulino et al. (2018), podemos dizer que o conhecimento que dispomos a respeito do vetor de parâmetros $\boldsymbol{\theta}$, antes da observação dos dados, é contabilizado probabilisticamente por meio da distribuição de probabilidade *a priori* denotada por $p(\boldsymbol{\theta})$. Porém, é possível aumentar o conhecimento de $\boldsymbol{\theta}$ utilizando, além da distribuição *a priori*, as informações dos dados amostrais $(\boldsymbol{Y})$ e que estejam relacionadas a esses parâmetros, a essa relação dá-se o nome de função de verossimilhança e é denotada por $p(y|\boldsymbol{\theta})$**.**

Espera-se que com as informações a respeito dos dados, aumente o conhecimento a respeito de $\boldsymbol{\theta}$, e este conhecimento acumulado é representado pela

distribuição *a posteriori* denotada por $p(\boldsymbol{\theta}|y)$. Dessa forma, a distribuição *a posteriori* pode ser interpretada como sendo a distribuição de probabilidade do vetor de parâmetro $\boldsymbol{\theta}$ após a observação dos dados experimentais. Tal distribuição é utilizada para inferência bayesiana uma vez que detém todo conhecimento a respeito do vetor de parâmetros $\boldsymbol{\theta}$ (RESENDE, 2000). Gamerman e Lopes (2006) relatam que a relação entre a distribuição *a posteriori*, a distribuição *a priori* e a função de verossimilhança dar-se pelo Teorema de Bayes:

$$p(\boldsymbol{\theta}|y) = \frac{p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(y)}$$

em que $p(y)$ é a função de verossimilhança marginal que independe de $\boldsymbol{\theta}$ e tem como função normalizar $p(\boldsymbol{\theta}|y)$. Comumente, para simplificação dos cálculos analíticos, o Teorema de Bayes é definido em termos de proporcionalidade como segue:

$$p(\boldsymbol{\theta}|y) \propto p(y|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Uma dificuldade enfrentada pela inferência bayesiana ocorre em casos de inferência multiparamétrica onde é necessário a resolução de integrais complexas e/ou multidimensionais (SORENSEN e GIANOLA, 2002). Nestes casos, é necessário o uso de métodos de simulação de dados para fazer aproximações, e um dos métodos mais utilizados são os algoritmos MCMC (*Markov Chain Monte Carlo*) que serão detalhados posteriormente.

3. Regressão Linear Múltipla

A regressão linear permite determinar a relação funcional entre uma variável aleatória dependente (Y) e uma ou mais variáveis independentes de efeito fixo (X) com base nas estimativas dos parâmetros. O principal objetivo da regressão é estimar os parâmetros do modelo proposto, sendo comumente usado o método de estimação de Mínimos Quadrados Ordinários (CECON et al., 2012). O modelo estatístico da regressão linear múltipla com k variáveis independentes é dado por:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i \ , \qquad i = 1, \cdots, n. \tag{1}$$

em que

- n é o número de observações;

- X_{ij} é o i-ésimo valor observado da j-ésima variável explicativa ou independente (X) ($j = 1, \dots, k$). Neste estudo, X_{ij} representa a incidência do j-ésimo marcador (codificado como 0, 1 ou 2 de acordo com a dosage alélica) para o i-ésimo indivíduo;
- Y_i é o i-ésimo valor observado da variável dependente ou resposta (Y). Neste estudo, Y_i representa o valor fenotípico do i-ésimo indivíduo;
- β_0 é a constante da regressão. Neste estudo, β_0 é a media geral da característica fenotípica;
- β_j é o j-ésimo coeficiente de regressão associado a variável X_j. Neste estudo, β_j é o efeito do j-ésimo marcador no fenótipo;
- ϵ_i é o i-ésimo erro aleatório, assumindo-se que $\epsilon_i \sim N(1, \sigma_e^2)$ onde σ_e^2 representa a variância residual.

Uma alternativa para obter a estimativa dessa relação funcional é a Inferência Bayesiana, que difere das abordagens comumente utilizadas como os métodos de mínimos quadrados ordinários e da máxima verossimilhança (FARAHANI et al., 2014).

A Inferência Bayesiana tem como característica principal tratar o vetor de parâmetros desconhecidos como uma variável aleatória, permitindo que qualquer informação sobre esses parâmetros seja representada por meio de um modelo probabilístico. Dessa forma, toda informação é descrita como uma distribuição de probabilidade denominada distribuição *a priori* (RESENDE et al., 2014).

Neste estudo, a distribuição dos dados e as distribuições *a priori* dos parâmetros do modelo (1) são definidas, respectivamente, como sendo:

$$y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}, \sigma^2)$$

$$\beta_0 \sim N(1, \sigma_{\beta_0}^2),\ \beta_1 \sim N(1, \sigma_{\beta_1}^2),\ \beta_2 \sim N(1, \sigma_{\beta_2}^2),\ \cdots,\ \beta_k \sim N(1, \sigma_{\beta_k}^2)$$

$$\sigma^2 \sim GI(a, b)$$

$$\sigma_{\beta_1}^2 \sim GI(a_1, b_1),\ \sigma_{\beta_2}^2 \sim GI(a_2, b_2),\ \cdots,\ \sigma_{\beta_k}^2 \sim GI(a_k, b_k)$$

em que, $\sigma_{\beta_0}^2 = 10^8$, a, b, a_j e b_j $(j = 1, \dots, k)$ são hiperparâmetros, definidos geralmente com base em revisões de literatura ou pela experiência do pesquisador. Os termos $\sigma_{\beta_1}^2, \sigma_{\beta_2}^2, \cdots, \sigma_{\beta_k}^2$ representam os componentes de variância associados a

cada marcador. A distribuição $N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}, \sigma^2)$ representa uma distribuição normal com média $\mu = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}$ e variância σ^2 e $GI(a, b)$ é a distribuição gama inversa com hiperparâmetros a e b.

Sob o Teorema de Bayes, a distribuição conjunta *a posteriori* de todos os parâmetros desconhecidos é proporcional ao produto da função de verossimilhança e das distribuições *a priori* dos parâmetros. A formulação geral deste teorema é dada por:

$$p\left(\beta_0, \beta_1, \beta_2, \cdots, \beta_k, \sigma^2, \sigma^2_{\beta_1}, \cdots, \sigma^2_{\beta_k} \middle| y\right)$$

$$\propto p\left(y \middle| \beta_0, \beta_1, \beta_2, \cdots, \beta_k, \sigma^2, \sigma^2_{\beta_1}, \cdots, \sigma^2_{\beta_k}\right) p(\beta_0) p(\sigma^2) \prod_{i=1}^{k} p(\beta_i)\, p(\sigma^2_{\beta_i})$$

A inferência realizada sobre o vetor de parâmetros desconhecidos do modelo, ou seja, $\boldsymbol{\theta}' = \left[\beta_0, \beta_1, \beta_2, \cdots, \beta_k, \sigma^2, \sigma^2_{\beta_1}, \cdots, \sigma^2_{\beta_k}\right]'$, se baseia nas distribuições marginais *a posteriori* de cada parâmetro, as quais são obtidas por meio de algoritmos MCMC.

4. Algoritmo MCMC

Os algoritmos MCMC são utilizados para a geração de amostras sob a abordagem bayesiana. Eles combinam a simulação de Monte Carlo e a teoria de cadeias de *Markov* (PAULINO et al., 2018).

As simulações de Monte Carlo são baseadas em amostragens aleatórias de uma variável aleatória com uma certa distribuição de probabilidade conhecida. Isso é feito para obter uma quantidade que, analiticamente, era de difícil ou impossível obtenção, mas que pode ser estimada numericamente (BOLSTAD e CURRAN, 2016).

Por outro lado, as Cadeias de *Markov* são um tipo especial de processo estocástico no qual a ocorrência de um evento no tempo/iteração t depende apenas da ocorrência do mesmo evento no tempo/iteração $t - 1$ (GAMERMAN e LOPES, 2006).

Conforme Paulino et al. (2018) explicam, os algoritmos MCMC consistem em gerar valores de uma distribuição de probabilidade sob a perspectiva de Cadeias de *Markov*. As amostras aleatórias das distribuições marginais *a posteriori* são geradas

indiretamente a partir de uma classe de distribuições denominadas Distribuições Condicionais Completas *a posteriori* (DCCP).

Quando o algoritmo MCMC é executado por um número suficientemente grande de iterações, a cadeia de valores atinge a condição de equilíbrio. É possível demonstrar, por meio da teoria de cadeias de Markov, que essa cadeia representa uma amostra da distribuição de probabilidade marginal *a posteriori* (BOLSTAD e CURRAN, 2016). A DCCP também orienta a escolha do algoritmo MCMC a ser utilizado para a gerar esses valores, que pode ser o *Gibbs Sampler* ou *Metropolis-Hastings* e seus refinamentos, dependendo da forma da distribuição de probabilidade (GAMERMAN e LOPES, 2006).

Para determinar a DCCP de um parâmetro θ_i, basta assumir que os demais parâmetros θ_{-i} são constantes na distribuição *a posteriori* conjunta. Assim, as DCCP dos parâmetros $\theta_1, \theta_2, \cdots, \theta_p$ são, respectivamente, proporcionais ao produto da função de verossimilhança e da distribuição a priori de cada parâmetro:

$$p(\theta_1|y, \theta_2, \cdots, \theta_p) \propto p(y|\theta_1, \theta_2, \cdots, \theta_p)p(\theta_1)$$
$$p(\theta_2|y, \theta_1, \theta_3 \cdots, \theta_p) \propto p(y|\theta_1, \theta_2, \cdots, \theta_p)p(\theta_2)$$
$$\vdots$$
$$p(\theta_p|y, \theta_1, \cdots, \theta_{p-1}) \propto p(y|\theta_1, \theta_2, \cdots, \theta_p)p(\theta_p)$$

Quando a DCCP de um parâmetro é uma distribuição de probabilidade conhecida, o algoritmo *Gibbs Sampler* é utilizado. No entanto, se não for o caso, recorre-se ao algoritmo *Metropolis-Hastings,* que se baseia na geração de valores a partir de uma distribuição de probabilidade candidata conhecida, $q(\cdot \mid \cdot)$, relacionada à DCCP $\pi(\cdot)$ desse parâmetro, a fim de calcular a probabilidade de aceitação de cada valor amostrado para pertencer a cadeia de valores de cada parâmetro (PAULINO et al., 2018).

4.1 Algoritmo *Metropolis-Hasting*

Considere p parâmetros de um modelo com as DCCP desconhecidas. Gamerman e Lopes (2006) descrevem o algoritmo *Metropolis-Hasting* em alguns passos, conforme a seguir:

1) Inicie o contador de iterações $t = 0$ e determine o número total de iterações T do algoritmo;

2) Defina valores iniciais para o vetor de parâmetros dado pelo vetor $\boldsymbol{\theta}^{(0)} = \left(\beta_0^{(0)}, \beta_1^{(0)}, \dots, \beta_p^{(0)}, \sigma_{\beta 1}^{2(0)}, \dots, \sigma_{\beta p}^{2(0)}, \sigma^{2(0)}\right)$, respeitando o espaço paramétrico de cada um dos elementos de $\boldsymbol{\theta}$;

3) Faça $t = t + 1$ e gere um valor para $\theta_i^{(t)}$ de uma distribuição candidata conhecida denotada por $q(\theta_i \,|\theta_{-i})$;

4) A probabilidade de aceitação deste valor é dada por:

$$\alpha_{\mathrm{i}}^{(\mathrm{t})} = \min\left\{1, \frac{\pi\left(\theta_i^{(t)}\middle|y, \theta_{k<i}^{(t)}, \theta_{k>i}^{(t-1)}\right)}{\pi\left(\theta_i^{(t-1)}\middle|y, \theta_{k<i}^{(t)}, \theta_{k>i}^{(t-1)}\right)} \frac{q\left(\theta_i^{(t-1)}\middle|\theta_i^{(t)}, \theta_{k<i}^{(t)}, \theta_{k>i}^{(t-1)}\right)}{q\left(\theta_i^{(t)}\middle|\theta_i^{(t-1)}, \theta_{k<i}^{(t)}, \theta_{k>i}^{(t-1)}\right)}\right\}$$

5) Gerar um valor $u_i \sim U(1,1)$;

6) Comparar u_i com $\alpha_i^{(t)}$:

i. Se $u_i \leq \alpha_i^{(t)}$, então aceita $\theta_i^{(t)}$

ii. Se $u_i > \alpha_i^{(t)}$, rejeita $\theta_i^{(t)}$ e faça $\theta_i^{(t)} = \theta_i^{(t-1)}$

7) Voltar ao passo 4 e calcular a probabilidade de aceitação para cada parâmetro;

8) Se $t < T$ voltar ao passo 3. Caso $t = T$ encerrar o algoritmo.

É importante notar que embora os valores $\theta_i^{(t)}$ sejam gerados de forma independente, a cadeia resultante não será independente e identicamente distribuída, uma vez que a probabilidade de aceitação ainda depende do valor na iteração anterior, $\theta_i^{(t-1)}$. Além disso, pode-se tomar a distribuição candidata $q(\cdot\,|\,\cdot)$ como sendo a própria DCCP, caso a DCCP seja conhecida. Dessa forma, a probabilidade de aceitação será sempre igual a um ($\alpha_{\mathrm{i}}^{(\mathrm{t})} = 1$) e com isso todos os valores serão aceitos, definindo assim o algoritmo de *Gibbs Sampler* (PAULINO et al., 2018).

O MCMC, em sua formulação original, quando aplicado a modelos de regressão como definido anteriormente em (1), ajusta o mesmo modelo em todas as iterações, uma vez que o algoritmo gera valores para todos os parâmetros do

modelo e dificilmente, quando se trata de variável aleatória contínua, estes assumirão valor igual a zero nas iterações (RESENDE et al., 2014).

Após a realização do algoritmo MCMC, a convergência das cadeias de cada parâmetro deve ser avaliada e esta avaliação pode ser feita por meio de critérios como o de Geweke (1992) e de Raftery e Lewis (1992). Além disso, quantidades como *thin* e *burn-in* também devem ser definidas.

5. Métodos Bayesianos Aplicados a Predição Genômica

Visando solucionar os desafios estatísticos presentes na seleção genômica, Meuwissen et al. (2001) propuseram métodos bayesianos, como por exemplo o BayesA, que impõe diferentes encolhimentos (*shrinkage*) aos efeitos de cada um dos marcadores moleculares. No entanto, desde então, surgiram novas metodologias com melhorias, dentre elas o método Bayes$D\pi$ (HABIER et al., 2011) que além de contemplar um encolhimento heterogêneo também seleciona marcadores.

Os métodos bayesianos propostos para a predição genômica são eficazes para situações de alta dimensionalidade. Além disso, esses métodos possuem uma caracterização própria para lidar com incertezas e, quando combinados com os algoritmos MCMC possuem vastas aplicações (GIANOLA et al., 2009).

5.1 BayesA

Proposto por Meuwissen et al. (2001) para a predição genômica, o método BayesA visa estimar os parâmetros do modelo (1). Para isso, assume que o efeito de cada marcador (ou seja, os coeficientes de regressão do modelo) segue uma distribuição normal com média zero e variância específica. Dessa forma, é gerado um encolhimento heterogêneo nos efeitos dos marcadores. Sendo assim, podemos definir as distribuições *a priori* do BayesA como sendo:

$$\beta_j \sim N\left(0, \sigma^2_{\beta_j}\right)$$

$$\sigma^2_{\beta_j} \sim GI\left(a_j, b_j\right)$$

em que

- β_j é o j-ésimo coeficiente de regressão associado à j-ésima variável explicativa X_j, representando o efeito do j-ésimo marcador no fenótipo;
- $\sigma^2_{\beta_j}$ é o componente de variância associada ao efeito do j-ésimo marcador;
- a_j e b_j são os hiperparâmetros;
- $N\left(0, \sigma^2_{\beta_j}\right)$ é a distribuição normal com média zero e variância $\sigma^2_{\beta_j}$;
- $GI(a_j, b_j)$ é a distribuição gama inversa com hiperparâmetros a_j e b_j.

Vale ressaltar que a escolha da distribuição gama inversa, assim como a qui-quadrado invertida, para os componentes de variância é matematicamente preferível, pois é uma distribuição *a priori* conjugada com a distribuição dos dados. Dessa forma, a DCCP encontrada para a variância também será uma distribuição gama inversa (ou qui-quadrado invertida), tendo apenas seus parâmetros atualizados (RESENDE et al., 2011). O mesmo é válido para a distribuição dos coeficientes de regressão, ou seja, distribuição *a priori* é conjugada com a distribuição dos dados. Assim a DCCP obtida também será uma distribuição normal na qual os parâmetros serão atualizados.

Pelo fato das DCCP encontradas serem conhecidas pode-se usar o algoritmo *Gibbs Sampler* para gerar valores das distribuições marginais *a posteriori* de cada parâmetro. Ao final do processo e análise de convergência, é possível obter as estimativas do efeito de cada marcador por meio das médias, medianas ou modas *a posteriori* (RESENDE, 2000).

No método BayesA, ao assumir a média igual a zero para cada marcador na distribuição normal, faz com que os coeficientes de regressão estimados (efeitos de marcadores) sejam valores próximos de zero, mas nunca zero. As estimativas dos coeficientes da regressão são ditadas pelas distribuições *a priori* dos efeitos dos marcadores e, consequentemente, a precisão com que esses valores são estimados difere (PEIXOTO et al., 2022).

5.2 BayesDπ

No método BayesA, os hiperparâmetros, (a_j e b_j) da gama inversa visto anteriormente, atribuídos a distribuição *a priori* têm muita influência sobre o encolhimento nos efeitos de cada marcador, impossibilitando o aprendizado bayesiano e fazendo com que as estimativas se tornam viesadas (AZEVEDO et al., 2015; HABIER et al., 2011).

Visando contornar esse problema e incorporar a seleção de marcadores, Habier et al. (2011) propuseram o método BayesDπ para ajustar o modelo (1). Este método apresenta semelhança ao BayesA para uma porção π de marcadores, ou seja, se assume a eles uma distribuição normal com média zero e variância específica para cada marcador, enquanto a outra porção (1- π) de marcadores tem seus efeitos anulados. Dessa forma, há uma redução do número de efeitos de marcadores a serem estimados nas iterações, o que acarreta uma maior precisão no ajuste do modelo (RESENDE et al., 2011).

Outro ponto importante é que além dos parâmetros β_j e $\sigma^2_{\beta_j}$, no BayesDπ, também se tem a quantidade π tratada como uma variável aleatória, em que se assume que π segue uma distribuição uniforme com hiperparâmetros 0 e 1 (ou uma distribuição Beta). Dessa forma, as distribuições *a priori* são definidas como:

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

$$\beta_j \sim \pi N\left(0, \sigma^2_{\beta_j}\right) + (1-\pi)N(0,0)$$

$$\sigma^2_{\beta_j} \sim GI\left(a_j, b_j\right)$$

em que

- $j = 1, 2, 3, \cdots, k.$
- β_j é o j-ésimo coeficiente de regressão associado à j-ésima variável explicativa X_j, no qual representará o efeito do j-ésimo marcador no fenótipo;
- $\sigma^2_{\beta_j}$ é o componente de variância associada ao efeito do j-ésimo marcador;
- a_j e b_j são os hiperparâmetros;

- $N\left(0, \sigma^2_{\beta_j}\right)$ é a distribuição normal com media zero e variância $\sigma^2_{\beta_j}$;
- $N(0,0)$ é uma distribuição degenerada centrada em zero;
- $GI\left(a_j, b_j\right)$ é a distribuição gama inversa com hiperparâmetros a_j e b_j.

Como a distribuição *a priori* dos coeficientes da regressão não é conjugada com a distribuição dos dados, o uso do algoritmo *Metropolis-Hasting* se faz necessário, uma vez que a DCCP não possui forma conhecida. No entanto, a distribuição da variância de cada marcador é conjugada com a distribuição dos dados e com isso a DCCP é conhecida, o que possibilita a aplicação do algoritmo *Gibbs Sampler* (HABIER et al., 2011).

Como o parâmetro π é uma quantidade variável ao longo das iterações do MCMC, é possível a estimação de um modelo contendo diferentes números de marcadores a cada iteração. Ao final do algoritmo também é possível obter a probabilidade de inclusão de cada uma das variáveis explicativas/marcadores no modelo. Assim, esta probabilidade p_j se torna um indicativo da importância da j-ésima variável explicativa X para a predição de Y.

Esta probabilidade p_j, que chamaremos de probabilidade *a posteriori* de inclusão do j-ésimo marcador no modelo, é dada pela razão entre o número de iterações em que o efeito deste marcador foi não-nulo e o número total de iterações. Neste estudo, a probabilidade p_j é um indicativo da importância do j-ésimo marcador SNP na predição do fenótipo de interesse.

6. Validação cruzada K-*fold*

De acordo com Resende et al. (2014), na predição genômica no contexto estatístico podemos definir dois tipos de subpopulações considerando a população em estudo: população de treinamento (também denominada como população de descoberta, estimação ou referência) e a população de validação.

A população de treinamento é constituída por um grande número de marcadores moleculares, avaliados em um número moderado de indivíduos nos quais os fenótipos são conhecidos para as características de interesse. Essa

população é utilizada para ajustar o modelo, ou seja, para obter as estimativas dos parâmetros do modelo, incluindo o efeito de cada marcador sobre uma característica específica. Portanto, os indivíduos da população de treinamento devem ter os genótipos e fenótipos conhecidos (RESENDE, 2008).

Por outro lado, a população de validação é composta por um número menor de indivíduos em comparação com a população de treinamento. Esses indivíduos também são necessitam das informações de marcadores e para fenótipos de interesse. Nesta população, o modelo é testado para verificar a acurácia e/ou a capacidade preditiva do modelo de predição. Para isso, os valores genéticos genômicos são preditos com base nos efeitos estimados dos marcadores na população de treinamento, e em seguida, são submetidos à análise de correlação (ou outra medida avaliativa) com os valores fenotípicos observados. Na população de validação também é essencial ter conhecimento dos genotípicos e fenotípicos (RESENDE, 2008).

No entanto, segundo James et al. (2013), este processo de validação apresenta dois problemas, a saber: (i) as estimativas de validação da taxa de erro podem variar dependendo de quais observações são incluídas nos conjuntos de treinamento e de validação; (ii) se apenas um subconjunto de observações é usado para ajustar o modelo, a taxa de erro do conjunto de validação tende a superestimar a taxa de erro do modelo em todo o conjunto de dados.

Para contornar esses problemas, pode-se utilizar o processo de validação cruzada k-*fold*. Esse procedimento consiste em dividir aleatoriamente o conjunto de N observações em k grupos de tamanho r, de modo que $N = rk$. O primeiro grupo é definido como a população de validação, enquanto os outros $k - 1$ grupos são usados para ajustar o modelo (população de treinamento). O processo é repetido k vezes (ciclos) com grupos diferentes atuando como a população de validação em cada ciclo. Em cada ciclo, é possível calcular medidas de avaliação e comparação da predição para a população de validação, e, ao final, calcula-se a média dos valores obtidos (JAMES et al., 2013).

7. Métricas de comparação

7.1 Herdabilidade

A herdabilidade é um parâmetro amplamente utilizado no campo do melhoramento genético, pois permite aos melhoristas estimar o ganho de seleção antes mesmo de realizar a seleção. Esse parâmetro indica a proporção da variabilidade fenotípica que pode ser atribuída à variabilidade genotípica (RAMALHO et al., 2012). Em geral, o componente de valor aditivo é o mais relevante nos programas de melhoramento, pois representa a parte herdável da variação genética. Portanto, a herdabilidade no sentido restrito é definida como a proporção da variância fenotípica que se deve aos efeitos aditivos dos genes, conforme a seguinte equação:

$$h^2 = \frac{\sigma^2}{\sigma_F^2} = \frac{\sigma_A^2}{\sigma_A^2 + \sigma_e^2}$$

em que σ_A^2 é a variância genotípica dos efeitos aditivos, σ_F^2 é a variância fenotípica, que pode ser decomposta em σ_A^2 e σ_e^2, sendo que σ_e^2 é a variância residual. Portanto, a herdabilidade no sentido restrito considera apenas a porção herdável da variância genética (JUNG et al., 2008; VIANA, 2002). Ela expressa o quanto uma característica específica pode ser transmitida de uma geração para a próxima.

7.2 Viés da Predição

O coeficiente da regressão de um modelo linear pode ser utilizado como uma medida de viés da predição para cada método proposto, desde que esse coeficiente represente a relação entre o valor genômico estimado e o valor fenotípico. Essa medida de viés é definida como:

$$Viés = 1 - \hat{\beta}$$

Dessa forma, quando o coeficiente estimado for igual a 1 ($\hat{\beta} = 1$), significa que os GEBV preditos não são viesados. Se for maior que um ($\hat{\beta} > 1$), indica que os GEBV preditos estão subestimados, enquanto se for menor que 1 ($\hat{\beta} < 1$), implica que os GEBVs preditos estão superestimados (RESENDE JR et al., 2012; RESENDE et al., 2014).

7.3 Capacidade Preditiva

Ao ajustar um modelo de regressão, é fundamental avaliar sua capacidade de predição, especialmente no contexto da predição genômica, onde o objetivo é prever os valores genômicos de indivíduos para os quais os fenótipos não foram coletados. Essa capacidade é conhecida como capacidade preditiva e pode ser quantificada calculando a correlação entre os fenótipos (y) e os valores genéticos genômicos estimados $(\hat{y} = G\hat{E}BV)$, ou seja:

$$r_{\hat{y}y} = cor(\hat{y}, y) = cor(G\hat{E}BV, y)$$

8. Referências

ALKIMIM, E. R. et al. Selective efficiency of genome-wide selection in Coffea canephora breeding. **Tree Genetics & Genomes**, v. 16, n. 3, p. 1-11, 2020.

ALKIMIM, E. R. et al. Marker-assisted selection provides arabica coffee with genes from other Coffea species targeting on multiple resistance to rust and coffee berry disease. **Molecular Breeding**, v. 37, n. 1, p. 1-10, 2017.

AZEVEDO, C. F. et al. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC genetics**, v. 16, n. 1, p. 1-13, 2015.

AUTON, A.; MCVEAN, G. Recombination rate estimation in the presence of hotspots. **Genome research**, v. 17, n. 8, p. 1219-1227, 2007.

BERNARDO, R. Molecular markers and selection for complex traits in plants: learning from the last 20 years. **Crop science**, v. 48, n. 5, p. 1649-1664, 2008.

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. 8ª edição. São Paulo: Oficina de Textos, 2021.

BOLSTAD, W. M.; CURRAN, J. M. **Introduction to Bayesian statistics**. John Wiley & Sons, 2016.

CASELLA, George; BERGER, Roger L. **Statistical inference**. Cengage Learning, 2021.

CROSSA, J. et al. Genomic selection in plant breeding: methods, models, and perspectives. **Trends in plant science**, v. 22, n. 11, p. 961-975, 2017.

CECON, P. R. et al. **Métodos estatísticos (Série Didática)**. Viçosa, MG: Editora da UFV, 2012.

CAVALCANTI, J. J. V. et al. Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. **Revista Brasileira de Fruticultura**, v. 34, p. 840-846, 2012.

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

FARAHANI, Z. S. M.; KHORRAM, E. Bayesian statistical inference for weighted exponential distribution. **Communications in Statistics-Simulation and Computation**, v. 43, n. 6, p. 1362-1384, 2014.

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

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

GAMERMAN, D.; LOPES, H. F. **Markov chain Monte Carlo: stochastic simulation for Bayesian inference**. CRC press, 2006.

GEWEKE, J. 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, 1992.

GREEN, P. J. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. **Biometrika**, v. 82, n. 4, p. 711-732, 1995.

HABIER, D. et al. **Extension of the bayesian alphabet for genomic selection**. BMC Bioinformatics, 12:186. 2011.

HAYES, B. J. et al. Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of dairy science**, v. 92, n. 2, p. 433-443, 2009.

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

JAMES, G. et al. **An introduction to statistical learning**. New York: springer, 2013.

JUNG, M. S. et al. Herdabilidade e ganho genético em caracteres do fruto do maracujazeiro-doce. **Revista Brasileira de Fruticultura**, v. 30, p. 209-214, 2008.

LANDE, R.; THOMPSON, R. Efficiency of marker-assisted selection in the improvement of quantitative traits. **Genetics**, v. 124, n. 3, p. 743-756, 1990.

LONG, N. et al. Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genetics research**, v. 92, n. 3, p. 209-225, 2010.

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

MURTEIRA, B. J. F. **Probabilidade e estatística**: inferência estatística. 2.ed. Lisboa: McGraw-Hill, 1990. v.2.

NEVES, H. H. R.; CARVALHEIRO, R.; QUEIROZ, S. A. A comparison of statistical methods for genomic selection in a mice population. **BMC genetics**, v. 13, n. 1, p. 1-17, 2012.

Nações Unidas, Departamento de Assuntos Econômicos e Sociais, Divisão de População (2022). Perspectivas da População Mundial 2022: Nota de lançamento.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B.; SILVA, G. L. **Estatística Bayesiana**. Lisboa, Portugal: Fundação Calouste Gulbenkian, 2018.

PEIXOTO, L. A.; BHERING, L. L.; CRUZ, C. D. **Seleção Genômica Aplicada ao Melhoramento Genético**. Viçosa, MG: Editora UFV, 2022.

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.

RAMALHO, M. A. P. et al. **Genética na Agropecuária**. 5º ed. Lavras, MG: Editora UFLA, 2012.

RESENDE, M. D. V. **Inferência Bayesiana e simulação estocástica (amostragem de Gibbs) na estimação de componentes de variância e de valores genéticos em plantas perenes**. 2000.

RESENDE, M. D. V. et al. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa florestal brasileira**, n. 56, p. 63-63, 2008.

RESENDE, M. D. V. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Embrapa Florestas, 2008.

RESENDE, M. D. V. et al. Computação da seleção genômica ampla (GWS). Colombo, Embrapa Florestas. 79p. **Série Documentos**, v. 210, 2010.

RESENDE, M. D. V. et al. **Métodos estatísticos na seleção genômica ampla**. Colombo: Embrapa Florestas, 2011.

RESENDE JR, M. F. R. et al. Accuracy of genomic selection methods in a standard data set of loblolly pine (Pinus taeda L.). **Genetics**, v. 190, n. 4, p. 1503-1510, 2012.

RESENDE, M. D. V. de; SILVA, F. F. e; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa: UFV, 2014.

SILVA, J. P.; LEANDRO, R. A. Uma abordagem bayesiana para o mapeamento de QTLS utilizando o método MCMC com saltos reversíveis. **Ciência e Agrotecnologia**, v. 33, p. 1061-1070, 2009.

SILVA, C. P. et al. Use of the reversible jump Markov chain Monte Carlo algorithm to select multiplicative terms in the AMMI-Bayesian model. **Plos one**, v. 18, n. 1, p. e0279537, 2023.

SORENSEN, D.; GIANOLA, D. **Likelihood, Bayesian and MCMC methods in quantitative genetics**. 2002.

VIANA, J. M. S. Heritability at family mean level. **Revista Árvore**, v. 26, p. 271-278, 2002.

WANG, X. et al. Genomic selection methods for crop improvement: Current status and prospects. **The Crop Journal**, v. 6, n. 4, p. 330-340, 2018.

CAPÍTULO 2
SELEÇÃO DE MARCADORES UTILIZANDO PROBABILIDADE *A POSTERIORI* DE INCLUSÃO NO MODELO PARA PREDIÇÃO GENÔMICA EM DADOS DE ARROZ

RESUMO

Para atender à crescente demanda por alimento nos próximos anos, os métodos estatísticos aplicados ao melhoramento genético desempenham um papel fundamental. Dentre as diversas culturas que alimentam a população mundial, destaca-se o arroz, devido à sua alta produção e consumo global. Com os avanços da genética molecular, têm sido desenvolvidos métodos de melhoramento que visam selecionar indivíduos geneticamente superiores de forma mais rápida, eficiente e econômica. Um desses métodos é a seleção genômica ampla (GWS, *Genomic Wide Selection*), que permite prever o valor genético genômico dos indivíduos com base em informações genotípicas. Geralmente, as informações genotípicas utilizadas na GWS são obtidas a partir de marcadores moleculares, principalmente do tipo SNP (*Single Nucleotide Polymorphism*), desde que estejam em desequilíbrio de ligação (LD, *Linkage Disequilibrium*) com os QTL (*Quantitative Trait Loci*). Para aplicar a GWS, é necessário estimar o efeito genético dos marcadores com base em uma população de treinamento. No entanto, o alto número de marcadores impõe desafios, como a alta dimensionalidade e a multicolinearidade. Para contornar esses desafios, têm sido propostos métodos, como o BayesA, que promove um encolhimento heterogêneo para cada marcador. No entanto, existem marcadores que não estão em LD com os QTL, e dessa forma, métodos que contemplam a seleção de variáveis, como o BayesDπ, surgem como alternativas que tornam o processo de estimação dos efeitos dos marcadores menos complexo. Outra abordagem para a redução do número de marcadores é a seleção dos mesmos com base na probabilidade *a posteriori* de inclusão (PIB). Para calcular essa probabilidade é necessário estimar os efeitos dos marcadores pelo método BayesDπ, pois permite ajustar diferentes números de marcadores a cada iteração do algoritmo MCMC (*Markov Chain Monte Carlo*). Com isso, a PIB é calculada pela razão entre o número de marcadores com efeito diferente de zero e o número total

de iterações salvas. Essa probabilidade torna-se um indicativo de importância do marcador para característica fenotípica, permitindo selecionar os marcadores mais relevantes. Neste estudo, foram selecionados os marcadores mais importantes para 11 características do arroz *Oryza Sativa*, utilizando um conjunto de dados públicos disponível em https://ricediversity.org/data/. Após a estimação das probabilidades de inclusão *a posteriori*, os marcadores foram organizados em grupos contendo 2.000, 4.000, 6.000, ..., e 36.901 marcadores. Após, cada grupo teve seus efeitos reestimados pelo método BayesA. Os resultados demonstraram uma maior capacidade preditiva, maior herdabilidade e menor viés em comparação com os métodos BayesA e BayesDπ sem a seleção prévia dos marcadores.

Palavras-chave: Marcadores Moleculares. Arroz. Seleção Genômica. Genética. Melhoramento Genético.

1. Introdução

A população mundial está crescendo em ritmo mais lento desde os anos de 1950, no entanto, espera-se que o número de habitantes em 2030 alcance aproximadamente 8,5 bilhões de pessoas, e que na década de 2080 atinja o ápice populacional de 10,4 bilhões de habitantes (Nações Unidas, 2022). Dessa forma, torna-se necessário aumentar a produção de alimentos, embora as áreas agricultáveis sejam limitadas. Uma solução para essa limitação é o melhoramento genético, pois permite aumentar a produtividade e qualidade dos alimentos considerando as mesmas áreas de cultivo (BORÉM et al., 2021).

Entre os cultivos, o arroz *Oryza sativa* se destaca como um dos cereais mais produzidos e consumidos no mundo, sendo a principal fonte de alimentação para mais da metade da população mundial. Além disso, o arroz é uma fonte de energia, devido à concentração de amido, e fornece proteínas, vitaminas, minerais e possui baixo teor de lipídios (WALTER et al., 2008). Segundo Childs e LeBeau (USDA, 2023), a produção mundial de arroz na safra 2022/2023 foi estimada em 503 milhões de toneladas e no Brasil a produção foi de 10,169,3 mil toneladas (CONAB, 2023).

Devido à sua importância, é necessário o desenvolvimento de novas linhagens visando maior produtividade e qualidade em relação às variedades já existentes. De acordo com o CONAB (2023), a quantidade de arroz produzida na safra 2022/2023 caiu cerca de 5,7% em relação à safra anterior. Para o desenvolvimento de novas linhagens utilizando métodos convencionais é estimado um tempo médio de 10 anos (SPINDEL et al., 2015).

Uma possibilidade para acelerar o processo de melhoramento é a aplicação da Seleção Genômica Ampla (GWS, *Genomic Wide Selection*), desenvolvida por Meuwissen et al. (2001), no qual utiliza informações diretamente do DNA para obter os valores genéticos genômicos estimados (GEBV, *Genomic Estimated Breeding Value*) dos indivíduos. Essas informações do DNA são os efeitos dos marcadores moleculares, que são estimados a partir de uma população de treinamento, na qual há um grande número de indivíduos genotipados e fenotipados para as caraterísticas de interesse. No entanto, para o uso da GWS, é necessário que haja desequilíbrio de ligação (LD, *Linkage Disequilibrium*) entre o marcador e o QTL (*Quantitative Trait Loci*).

Os marcadores moleculares são preferíveis os do tipo SNP (*Single Nucleotide Polimorphism*) por serem abundantes no genoma, o que aumenta a probabilidade de encontrar pelo menos um marcador em LD com algum QTL, independentemente do marcador ter pequeno ou grande efeito. Além disso, os marcadores SNP possuem facilidade e baixo custo de genotipagem (RESENDE et al., 2014).

No entanto, a grande quantidade de marcadores faz com que se tenha problemas de alta dimensionalidade no ajuste do modelo de predição dos efeitos, além da possibilidade de os marcadores estarem correlacionados (JANNINK et al., 2010). Isso torna inviável o uso de métodos de estimação como o Métodos dos Mínimos Quadrados (MMQ), pois há uma insuficiência de graus de liberdade caso estime todos os efeitos dos marcadores simultaneamente. Por outro lado, caso se estime o efeito de cada marcador separadamente e, em seguida, a predição, dos valores genéticos, a capacidade preditiva associada a eles seria baixa (NEVES et al., 2012; DESTA et al., 2014).

Ao longo dos anos, foram propostos diversos métodos para predição genômica, dentre eles os métodos bayesianos, como o BayesA proposto por Meuwissen et al. (2001), que inclui o efeito de todos os marcadores moleculares no modelo, atribuindo a cada marcador variância específica. Hayes et al. (2009) estimaram os efeitos marcadores moleculares do tipo SNP, pelo método BayesA, para predizer o GEBV de touros de uma população de referência multirracial. Enquanto Haile et al. (2020) aplicaram o BayesA para avaliar a precisão de modelos de seleção genômica com múltiplas e únicas características de lentilhas, bem como modelos que incluíram marcadores significativos atestados por estudos de associação genômica e modelos que inclui a interação genótipo e ambiente.

No entanto, o BayesA conduz uma complexidade desnecessária visto que nem todos os marcadores estão em LD com os QTL (PEIXOTO et al., 2022). Dessa forma, se faz necessário métodos que contemplam a seleção de variáveis, como o BayesDπ, proposto por Habier et al. (2011), que seleciona, a cada iteração, uma quantidade aleatória de π marcadores para o ajuste do modelo e anula o efeito dos $1-\pi$ marcadores restantes. Para os π marcadores utilizados no ajuste, assume-se uma distribuição normal com média zero e variância específica para cada marcador.

Lopez et al. (2019) aplicaram o método BayesDπ para estudar a arquitetura genética e a influência das características, produção de fotossíntese e eficiência do uso da água, no rendimento dos grãos de soja.

Por meio do ajuste de uma quantidade variável π de marcadores não nulos a cada iteração, com base em uma distribuição de probabilidade específica, torna-se possível calcular uma probabilidade *a posteriori* de inclusão (PIP) de cada marcador no modelo. Assim, utilizando a PIP obtida para cada marcador, é possível selecionar os marcadores mais relevantes para o modelo, empregando o método BayesDπ.

Posteriormente, pode-se avaliar a capacidade preditiva, herdabilidade viés preditivo desses marcadores utilizando o método BayesA. Dessa forma, é viável selecionar os marcadores que contribuem para a variância genética e para a capacidade preditiva do modelo de predição. Outros autores, como Bukszar et al. (2009), Guan e Stephens (2011), Yang et al. (2015), Wang et al. (2020), vêm abordando também o termo PIP, mas em contextos de associação genômica ou utilizando outros métodos bayesianos de predição.

Neste estudo, o método combinado, que consiste em utilizar os métodos BayesDπ para a seleção de marcadores, baseados em suas PIB's e, em seguida, o BayesA para a avaliação da capacidade preditiva, herdabilidade e viés preditivo, foi aplicado a um conjunto de características fenotípicas. Os resultados obtidos com esse método combinado foram comparados, em termos de capacidade preditiva, herdabilidade e viés preditivo com os resultados dos métodos bayesianos, BayesA e BayesDπ, utilizados separadamente. Para esse propósito, foram utilizados dados públicos provenientes de arroz *Oryza Sativa*, que contêm 11 características de interesse e 44.100 marcadores do tipo SNP de 413 variedades de arroz.

2. Materiais e métodos

2.1 Dados reais

O conjunto de dados utilizado neste estudo é composto por informações fenotípicas e genotípicas do arroz *Oryza sativa*, fazendo parte de dois projetos, o Projeto *Oryza* SNP e o Projeto OMAP (AMMIRAJU et al., 2006; ZHAO et al., 2011). Os dados estão disponíveis no seguinte *link* https://ricediversity.org/data/.

Esses dados abrangem 413 variedades de arroz provenientes de 82 países e foram genotipados para 44.100 marcadores do tipo SNPs. O controle de qualidade foi realizado considerando um *call rate* menor que 70% e baixa frequência do alelo mais raro menor que 1% a fim de remover marcadores não informativos e sem relevância genética para a população. Ao final desse processo de filtragem, restaram, 36.901 marcadores. As variedades foram fenotipadas para 11 características, são elas: 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).

O experimento de plantio e o cultivo do arroz foi conduzido na cidade de Arkansas, Estados Unidos, durante o período de maio a outono nos anos de 2006 e 2007. Para o experimento foi utilizado duas repetições por ano em um delineamento em blocos completos casualizados. Cada parcela consistia em fileiras de 5 metros de comprimento, com um espaçamento de 25 cm entre plantas e 50 cm entre as fileiras (ZHAO et al., 2011).

2.2 Modelo de Regressão Linear Múltipla Sob Enfoque Bayesiano

Os métodos bayesianos propostos para predição genômica são baseados no modelo estatístico da regressão linear múltipla com k variáveis independentes, que é expresso da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i \ , \qquad i = 1, \cdots, n. \tag{2}$$

em que

- n é o número de observações;
- X_{ij} é o i-ésimo valor observado da j-ésima variável explicativa ou independente (X) ($j = 1, \ldots, k$). Neste estudo, X_{ij} representa a incidência do j-

ésimo marcador (codificado como 0, 1 ou 2 de acordo com a dosagem alélica) para o i-ésimo indivíduo;

- Y_i é o i-ésimo valor observado da variável dependente ou resposta (Y). Neste estudo, Y_i representa o valor fenotípico do i-ésimo indivíduo;
- β_0 é a constante da regressão. Neste estudo, β_0 é a média de característica fenotípica;
- β_j é o j-ésimo coeficiente de regressão associado a variável X_j. Neste estudo, β_j é o efeito do j-ésimo marcador no fenótipo;
- ϵ_i é o i-ésimo erro aleatório onde assume-se que $\epsilon_i \sim N(1, \sigma^2)$ e σ^2 representa a variância residual.

Na Inferência Bayesiana, o vetor de parâmetros desconhecidos é tratado o como uma variável aleatória, e qualquer informação sobre esses parâmetros pode ser representada por um modelo probabilístico. No modelo (2), a distribuição dos dados e a distribuição *a priori* do componente de variância residual foram definidas, respectivamente, como sendo:

$$y_i \sim N(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik}, \sigma^2)$$
$$\sigma^2 \sim GI(a, b)$$

em que GI é a distribuição gama inversa e $a\ e\ b$ são seus hiperparâmetros, definidos com base em Azevedo et al. (2022). Os métodos Bayesianos aplicados à GWS utilizam diferentes distribuições *a priori* para os coeficientes de regressão, sendo estas distribuições o que os diferencia.

2.2.1 Método BayesA

O método BayesA, proposto por Meuwissen et al. (2001), assume que o efeito de cada marcador segue uma distribuição normal com média zero e variância específica. Com isso, é gerado um encolhimento heterogêneo nos efeitos dos marcadores, fazendo com que seus valores se aproximem de zero, mas nunca sejam zero. As distribuições *a priori* do BayesA podem ser definidas como:

$$\beta_j \sim N\left(0, \sigma^2_{\beta_j}\right)$$
$$\sigma^2_{\beta_j} \sim GI\left(a_j, b_j\right)$$

em que;

- β_j é o efeito do j-ésimo marcador no fenótipo;
- $\sigma^2_{\beta_j}$ é o componente de variância associada ao efeito do j-ésimo marcador;
- a_j e b_j são os hiperparâmetros da distribuição gama inversa, definidos com base em Azevedo et al. (2022) ;
- $N\left(0, \sigma^2_{\beta_j}\right)$ é a distribuição normal com media zero e variância $\sigma^2_{\beta_j}$.

A escolha da distribuição gama inversa para os componentes de variância é preferível devido ao fato de ser uma distribuição *a priori* conjugada com a distribuição dos dados. Isso significa que, quando se utiliza a distribuição gama inversa como distribuição *a priori*, a distribuição condicional completa *a posteriori* (DCCP) encontrada para a variância também será uma distribuição gama inversa, e seus parâmetros atualizados. Portanto, pode-se empregar o algoritmo *Gibbs Sampler* para gerar valores das distribuições marginais *a posteriori* de cada parâmetro, o que permite obter as estimativas do efeito de cada marcador por meio das médias, medianas ou modas *a posteriori*.

2.2.2 Método BayesDπ

O método BayesA, embora útil, não incorpora a seleção de marcadores, o que significa que ele inclui no modelo marcadores que não possuem efeito genético ou não estejam em LD com algum QTL (PEIXOTO et al., 2022). Para contornar esse problema e permitir a seleção de marcadores, Habier et al. (2011) propuseram o método denominado BayesDπ.

No método BayesDπ, assume-se que uma fração π de marcadores tem efeitos seguindo uma distribuição normal com média zero e variância específica, smilar ao método BayesA. No entanto, os demais marcadores (1- π) têm seus efeitos anulados, ou seja, não contribuem para o modelo. Isso resulta em uma redução no número de efeitos de marcadores a serem estimados nas iterações, o que aumenta a precisão no ajuste do modelo de predição.

Para a fração π de marcadores, assume-se uma distribuição uniforme com hiperparâmetros 0 e 1. As distribuições *a priori* são definidas da seguinte forma:

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

$$\beta_j \sim \pi N\left(0, \sigma^2_{\beta_j}\right) + (1-\pi)N(0,0)$$

$$\sigma^2_{\beta_j} \sim GI\left(a_j, b_j\right)$$

em que;

- β_j é o efeito do j-ésimo marcador no fenótipo;
- $\sigma^2_{\beta_j}$ é o componente de variância associada ao efeito do j-ésimo marcador;
- a_j e b_j são os hiperparâmetros da distribuição gama inversa, definidos com base em Azevedo et al. (2022) ;
- $N\left(0, \sigma^2_{\beta_j}\right)$ é a distribuição normal com media zero e variância $\sigma^2_{\beta_j}$;
- $N(0,0)$ é uma distribuição degenerada centrada em zero.

Neste método, utiliza-se o algoritmo *Metropolis-Hasting*, uma vez que não se conhece a forma da DCCP. No entanto, a distribuição da variância de cada marcador é conjugada com a distribuição dos dados, o que significa que a DCCP é conhecida, possibilitando a aplicação do algoritmo *Gibbs Sampler* (HABIER et al., 2011).

2.3 Seleção de marcadores

Ao final do método BayesDπ, utilizando as cadeias dos efeitos dos marcadores, é possível estimar a probabilidade *a posteriori* de inclusão de cada marcador. Essa probabilidade pode ser calculada pela razão entre o número de iterações em que o marcador teve seu efeito não nulo e o número total de iterações. Após o cálculo de todas a probabilidades *a posteriori* de inclusão, elas foram ordenadas em ordem decrescente. Em seguida, grupos de marcadores SNP com as maiores probabilidades foram selecionados para a avaliação da capacidade preditiva, herdabilidade e viés de predição utilizando o método BayesA. Esses

grupos foram formados por N marcadores, variando de 2.000 até 36.901, com um aumento de de 2.000 em 2.000 marcadores a cada vez.

2.4 Comparação de metodologias

Os métodos BayesA e BayesDπ foram submetidos ao procedimento de validação cruzada *k-fold*, com $k = 10$, para avaliar as seguintes medidas de eficiência: i) Capacidade preditiva, que é a correlação entre o valor predito pelos métodos e o fenótipo; ii) Viés de predição, calculado como 1 menos o coeficiente da regressão entre o valor predito pelos métodos e o fenótipo, juntamente com o intervalo de confiança a 95% do coeficiente de regressão.

Além disso, a herdabilidade para cada um dos métodos também foi calculada como a média *a posteriori* dos seguintes valores: $h^{2(t)} = \frac{\sigma_A^{2(t)}}{\sigma_A^{2(t)} + \sigma_e^{2(t)}}$, em que $h^{2(t)}$ é a herdabilidade na t-ésima iteração, $\sigma_A^{2(t)}$ é a variância genética aditiva na t-ésima iteração e $\sigma_e^{2(t)}$ é a variância residual na t-ésima iteração.

Para avaliar os resultados apresentados pela seleção de marcadores, foi feita uma comparação entre a combinação dos métodos BayesA e BayesDπ com os próprios métodos BayesA e BayesDπ, sem a seleção de marcadores. Para isso, os dados com os grupos de marcadores ajustados no método BayesA, também foram submetidos ao procedimento de validação cruzada *k-fold*, com $k = 10$, e suas capacidades preditivas, herdabilidades e viés de predição foram comparadas com as dos métodos BayesA e BayesDπ.

2.5 Recursos computacionais

Todas as rotinas computacionais foram desenvolvidas no *software* R versão 4.0.2 (R CODE TEAM, 2022). Para a análises bayesianas e de convergência, foram utilizadosos pacotes BLGR (PÉREZ e DE LOS CAMPOS, 2014) e coda (PLUMMER et al., 2006), respectivamente.

2. Resultados e discussões

A Tabela 1 apresenta os resultados da média e do desvio-padrão da capacidade preditiva, viés de predição e os intervalos de confiança dos coeficientes de regressão, bem como a média *a posteriori* da herdabilidade, para as 11 características de arroz estudadas por meio dos métodos BayesA e BayesDπ, de forma isolada. Além disso, também são fornecidas as médias *a posteriori* da quantidade π de marcadores para cada característica proveniente do método BayesDπ.

Tabela 1. Média ($CP_{Média}$) e desvio-padrão (CP_{SD}) da capacidade preditiva, viés de predição e os intervalos de confiança (IC) dos coeficientes de regressão e a média *a posteriori* das herdabilidades (h^2) das 11 características fenotípicas de arroz estimados pelos métodos BayesA e BayesDπ e a média *a posteriori* da probabilidade $\boldsymbol{\pi}$ associada ao método BayesDπ.

Métodos	Característica fenotípica	$\boldsymbol{\pi}$	$CP_{Média}$	CP_{SD}	Viés	IC	h^2
BayesA	CFB		0,47	0,08	-0,44	(1,38 ; 1,50)	0,58
	LFB		0,72	0,06	-0,22	(1,17 ; 1,27)	0,68
	NPP		0,79	0,05	-0,13	(1,09 ; 1,18)	0,75
	AP		0,73	0,10	-0,18	(1,15 ; 1,22)	0,76
	CP		0,64	0,11	-0,26	(1,21 ; 1,31)	0,68
	NRPP		0,62	0,10	-0,33	(1,26 ; 1,40)	0,57
	NSPP		0,56	0,11	-0,44	(1,37 ; 1,51)	0,55
	FPP		0,66	0,07	-0,28	(1,24 ; 1,33)	0,68
	FP		0,51	0,12	-0,38	(1,31 ; 1,44)	0,59
	RB		0,65	0,08	-0,26	(1,21 ; 1,31)	0,66
	TP		0,46	0,08	-0,50	(1,42 ; 1,58)	0,51
BayesDπ	CFB	0,41	0,47	0,08	-0,43	(1,37 ; 1,49)	0,59
	LFB	0,30	0,72	0,056	-0,21	(1,16 ; 1,26)	0,69
	NPP	0,45	0,79	0,05	-0,13	(1,09 ; 1,18)	0,75
	AP	0,48	0,73	0,10	-0,18	(1,14 ; 1,22)	0,77
	CP	0,46	0,64	0,11	-0,26	(1,21; 1,30)	0,68
	NRPP	0,46	0,62	0,10	-0,33	(1,26 ; 1,40)	0,57
	NSPP	0,47	0,56	0,11	-0,42	(1,36 ; 1,49)	0,57
	FPP	0,41	0,66	0,07	-0,27	(1,24 ; 1,33)	0,69
	FP	0,33	0,51	0,12	-0,37	(1,31 ; 1,42)	0,61
	RB	0,42	0,65	0,08	-0,26	(1,20 ; 1,31)	0,67
	TP	0,44	0,46	0,08	-0,50	(1,42 ; 1,58)	0,51

Comprimento da folha da bandeira (CFB); Largura da folha da bandeira (LFB); Número de panículas por planta (NPP); Altura da planta (AP); Comprimento da panícula (CP); Número do ramo primário da panícula (NRPP); Número de sementes por panícula (NSPP); Flores por panículas (FPP); Fertilidade da panícula (FP); Resistência à brusone (RB); Teor de proteína (TP).

Os resultados indicam que, para todas as características, os métodos BayesDπ e BayesA apresentaram herdabilidades e capacidades preditivas semelhantes. Essas herdabilidades estimadas também se mostraram comparáveis

às encontradas por Guo et al. (2014), que utilizaram o método G-BLUP para ajustar o modelo de predição no mesmo conjunto de dados.

De acordo com Jilo et al. (2018), os valores das herdabilidades variando de 0,30 a 0,60 são consideradas de magnitude moderada, enquanto herdabilidades superiores à 0,60 são de alta magnitude. Portanto, as herdabilidades relatadas neste estudo são de moderada a alta magnitude, variando entre 0,51 e 0,76 estimadas pelo método BayesA e de 0,51 a 0,77 estimadas pelo método BayesDπ.

Em comparação com o estudo de Wang et al. (2017), que aplicou o G-BLUP em dados de arroz da Universidade de Wuhan, observamos que tanto a capacidade preditiva quanto a herdabilidade da característica altura da planta (AP) se mostraram superiores nos resultados de Wang et al. (2017) em comparação aos métodos bayesianos neste estudo. No entanto, o comprimento da panícula (CP) apresentou maior capacidade preditiva, embora com uma herdabilidade ligeiramente menor.

No estudo de Azevedo et al. (2015), que utilizou dados simulados para representar características oligogênicas e poligênicas e para comparar o desempenho de alguns métodos de predição genômica, incluindo G-BLUP, Lasso e Regressão Bayesiana Ridge, os métodos bayesianos neste estudo também apresentaram capacidades preditivas similares às apresentadas na Tabela 1.

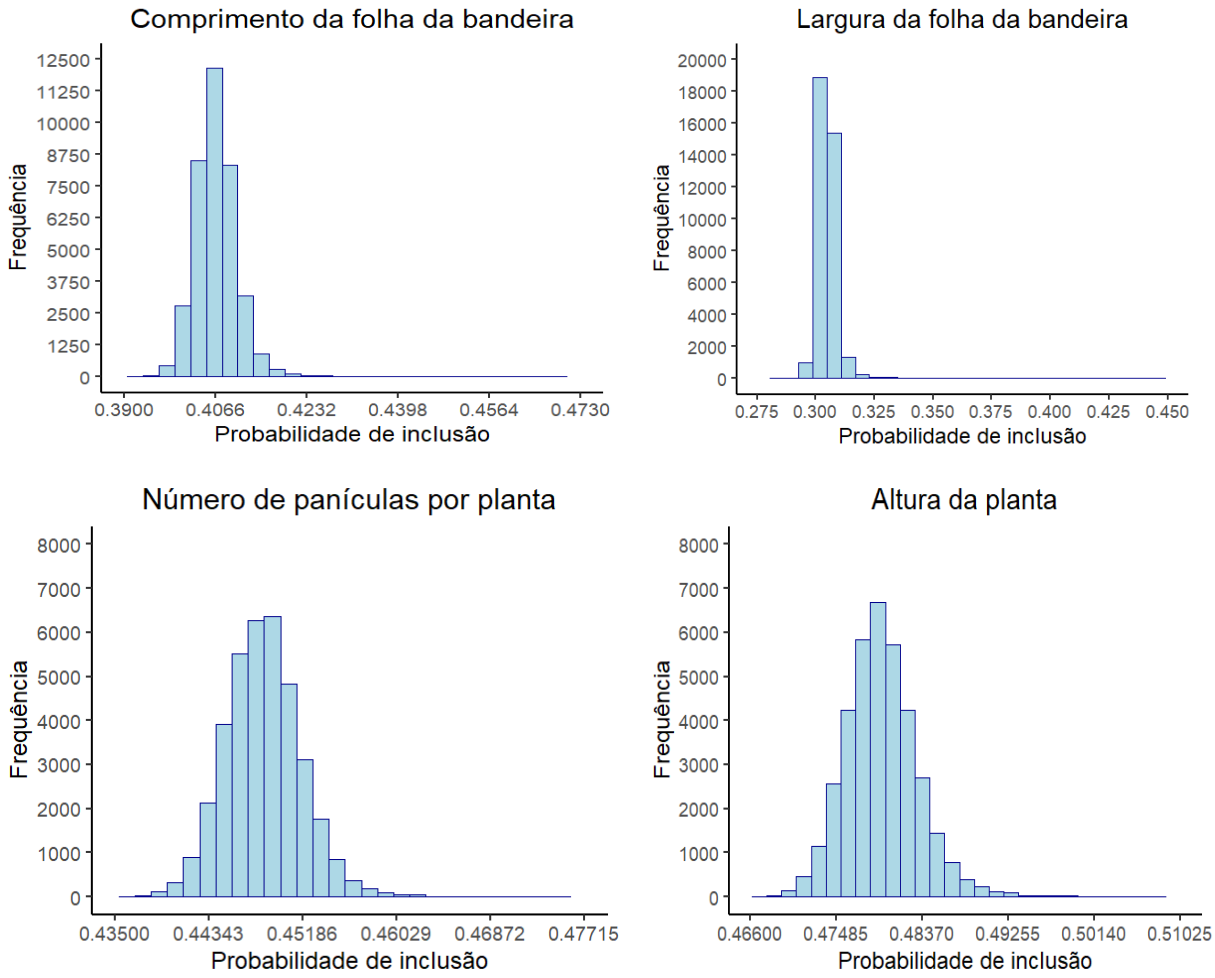
Este estudo revelou que ambos os métodos bayesianos, BayesA e BayesDπ, tendem a subestimar os valores genéticos genômicos preditos. Esses resultados são condizentes com os resultados de Suela et al. (2019) que utilizaram o mesmo conjunto de dados deste estudo para comparar a eficiência dos métodos de predição BayesCπ e BLASSO para nove características de arroz e também observaram uma subestimação dos GEBV, com exceção da característica teor de proteína, que não apresentou viés.

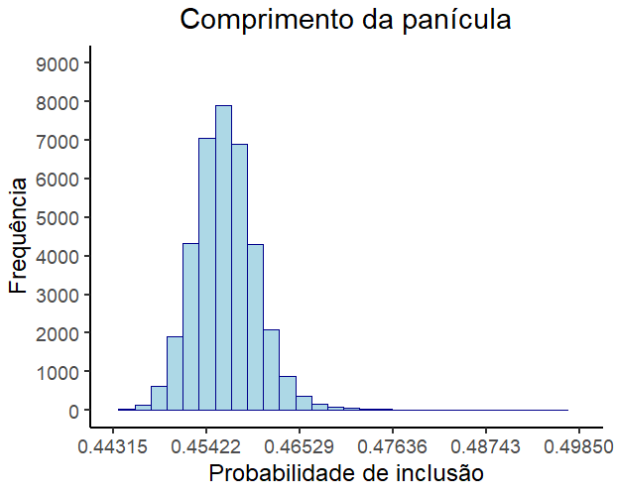
A Figura 1 apresenta os histogramas das probabilidades *a posteriori* de inclusão dos marcadores no modelo de predição genômica para cada característica. Essas probabilidades permitem classificar os marcadores por ordem de importância na predição, identificando aqueles com maior probabilidade de inclusão. Essa informação é útil para selecionar os marcadores que mais influenciam uma determinada característica.

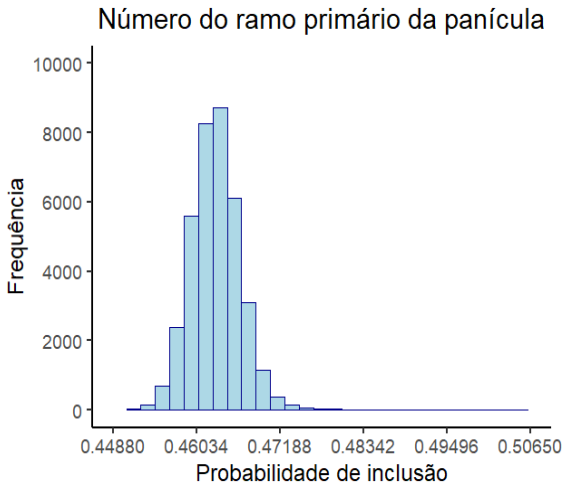
A característica resistência a brusone exibe uma maior variação nos valores de PIP, com a probabilidade variando entre aproximadamente 0,40 e 0,70. Observa-

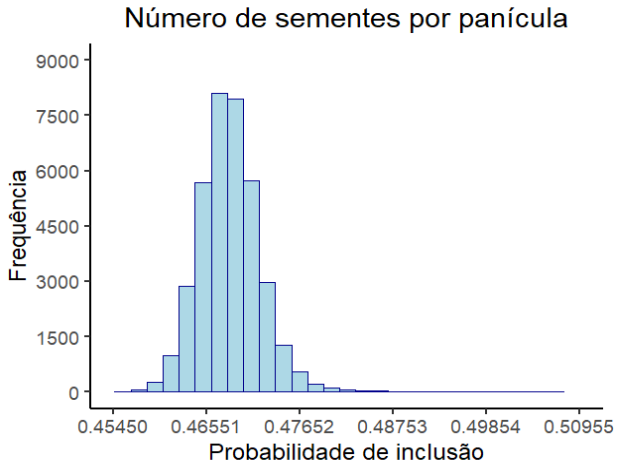
se também a existência de poucos marcadores com alta PIP, sugerindo que essa característica pode ser de natureza oligogênica, o que está em acordo com a literatura, conforme relatado por diversos autores como Hasan et al. (2015) e Wisser et al. (2005). Por outro lado, as demais características apresentam uma menor variação na probabilidade *a posteriori* de inclusão, indicando que podem ser de natureza mais poligênica, com muitos marcadores apresentando valores de PIP próximos.

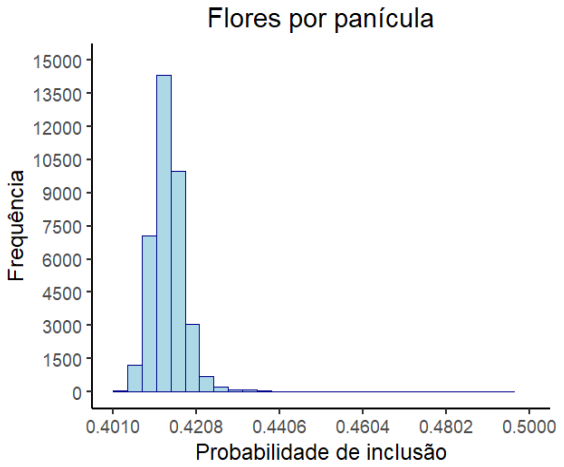
A distribuição das frequências da probabilidade de inclusão dos marcadores pode auxiliar os melhoristas a compreender a arquitetura genética das características de interesse e, consequentemente, pode ser aplicada em estudos de Associação Genômica Ampla (GWAS, Genome Wide Association), conforme realizado por Guan e Stephens (2011) ao aplicar a Regressão Bayesiana de Seleção Variável aos estudos de GWAS.

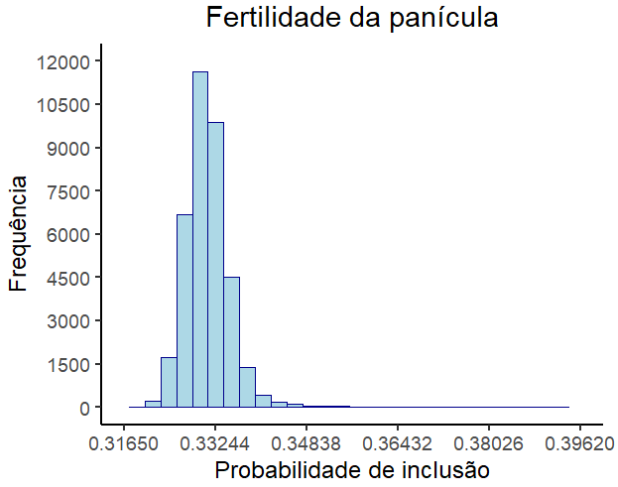


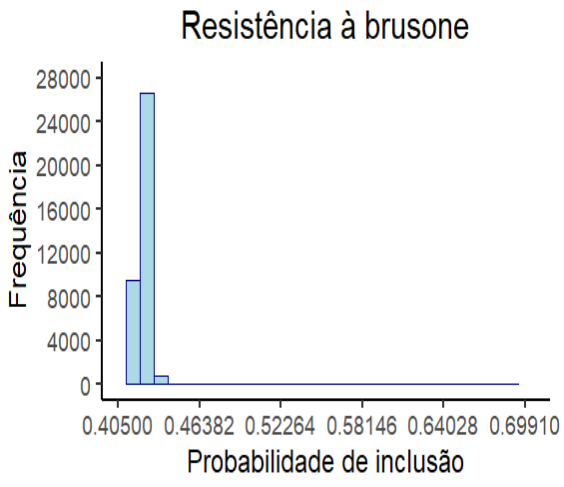

Comprimento da panícula
Frequência
9000
8000
7000
6000
5000
4000
3000
2000
1000
0
0.44315
0.45422
0.46529
0.47636
0.48743
0.49850
Probabilidade de inclusão


Número do ramo primário da panícula
Frequência
10000
8000
6000
4000
2000
0
0.44880
0.46034
0.47188
0.48342
0.49496
0.50650
Probabilidade de inclusão


Número de sementes por panícula
Frequência
9000
7500
6000
4500
3000
1500
0
0.45450
0.46551
0.47652
0.48753
0.49854
0.50955
Probabilidade de inclusão


Flores por panícula
Frequência
15000
13500
12000
10500
9000
7500
6000
4500
3000
1500
0
0.4010
0.4208
0.4406
0.4604
0.4802
0.5000
Probabilidade de inclusão


Fertilidade da panícula
Frequência
12000
10500
9000
7500
6000
4500
3000
1500
0
0.31650
0.33244
0.34838
0.36432
0.38026
0.39620
Probabilidade de inclusão


Resistência à brusone
Frequência
28000
24000
20000
16000
12000
8000
4000
0
0.40500
0.46382
0.52264
0.58146
0.64028
0.69910
Probabilidade de inclusão

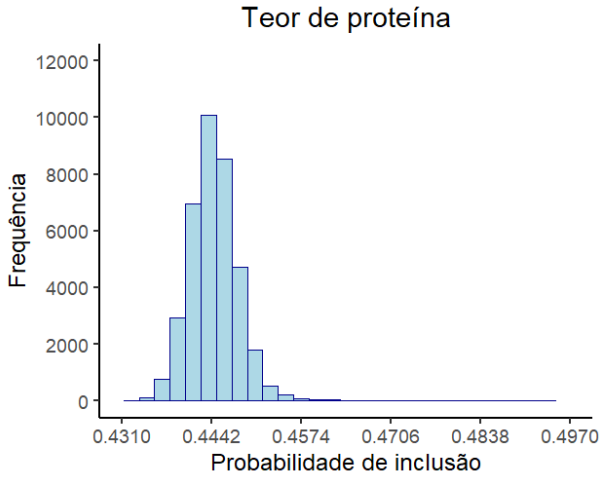
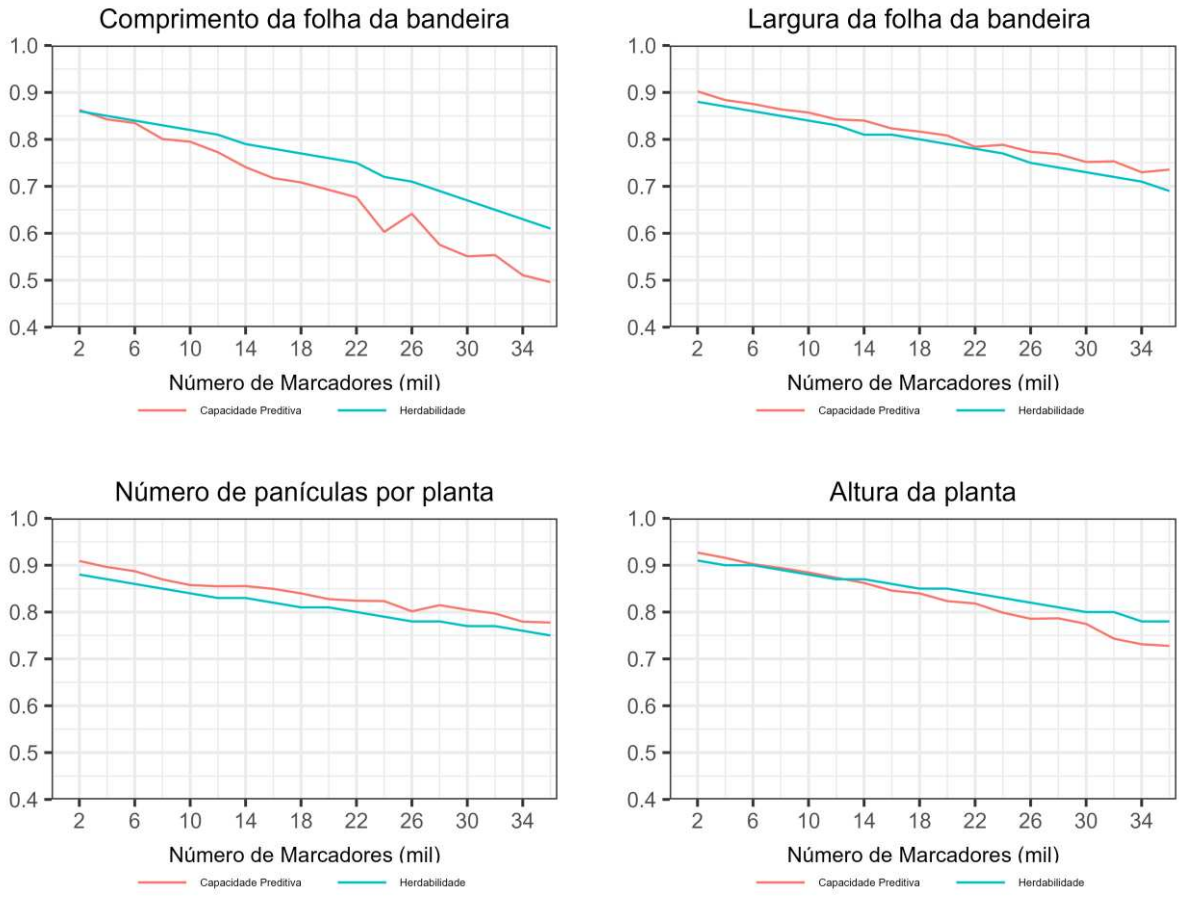
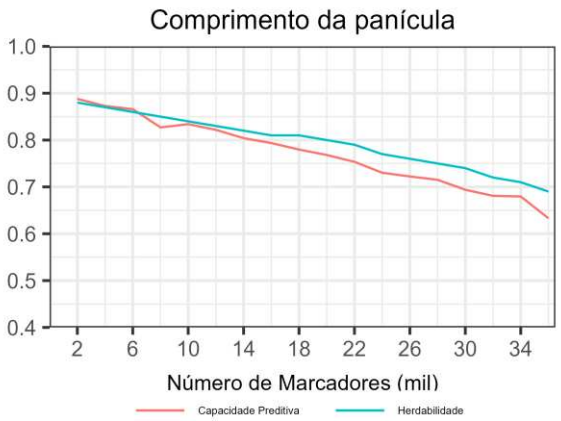
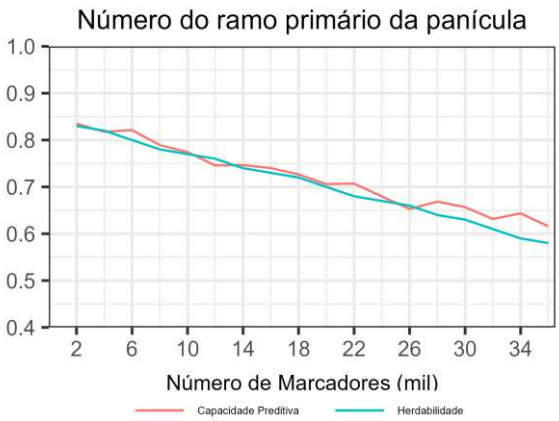


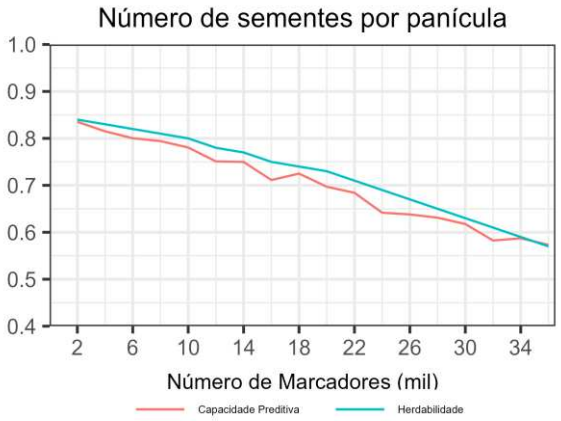
Figura 1. Histograma contendo a probabilidade *a posteriori* de inclusão dos marcadores moleculares para cada uma das características de arroz.

Na Figura 2 é apresentada a média *a posteriori* da herdabilidade e a capacidade preditiva do modelo de predição, ajustado pelo método BayesA, utilizando os grupos de marcadores com as maiores probabilidades *a posteriori* de inclusão no modelo.

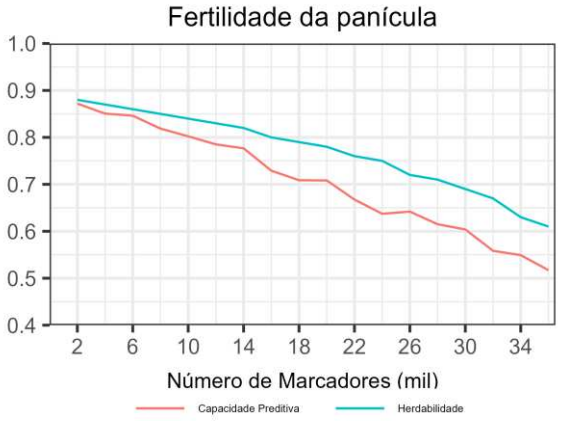


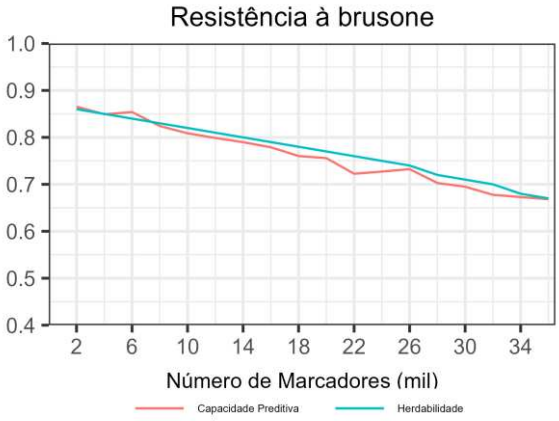

Comprimento da panícula
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2 6 10 14 18 22 26 30 34
Número de Marcadores (mil)
Capacidade Preditiva
Herdabilidade


Número do ramo primário da panícula
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2 6 10 14 18 22 26 30 34
Número de Marcadores (mil)
Capacidade Preditiva
Herdabilidade


Número de sementes por panícula
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2 6 10 14 18 22 26 30 34
Número de Marcadores (mil)
Capacidade Preditiva
Herdabilidade


Flores por panículas
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2 6 10 14 18 22 26 30 34
Número de Marcadores (mil)
Capacidade Preditiva
Herdabilidade


Fertilidade da panícula
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2 6 10 14 18 22 26 30 34
Número de Marcadores (mil)
Capacidade Preditiva
Herdabilidade


Resistência à brusone
1.0
0.9
0.8
0.7
0.6
0.5
0.4
2 6 10 14 18 22 26 30 34
Número de Marcadores (mil)
Capacidade Preditiva
Herdabilidade

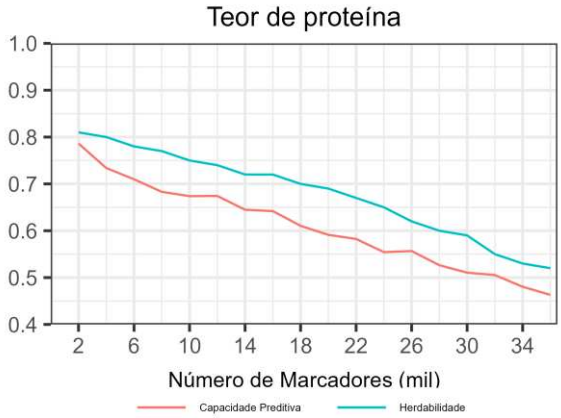


Figura 2. Capacidade preditiva do modelo de predição do método BayesA utilizando grupos de marcadores com as maiores probabilidades *a posteriori* de inclusão no modelo e a média *a posteriori* da herdabilidade.

Já na Tabela 2, são apresentados o viés de predição da regressão, juntamente com o intervalo de confiança do coeficiente de regressão, para as 11 características fenotípicas de arroz estudadas.

Tabela 2. Viés de predição e o intervalo de confiança (IC) dos coeficientes de regressão das 11 características fenotípicas de arroz estimados pelo método BayesA utilizando os grupos de marcadores com as maiores probabilidades *a posteriori* de inclusão no modelo.

Número de Marcadores	Características fenotípicas											
	CFB		LFB		NPP		AP		CP		NRPP	
	Viés	IC	Viés	IC	Viés	IC	Viés	IC	Viés	IC	Viés	IC
2000	-0,10	(1,08 ; 1,13)	-0,08	(1,05 ; 1,10)	-0,07	(1,04 ; 1.09)	-0,06	(1,04 ; 1.08)	-0,09	(1,06 ; 1,11)	-0,12	(1,09 ; 1,15)
4000	-0,12	(1,09 ; 1,14)	-0,09	(1,06 ; 1,12)	-0,08	(1,05 ; 1,10)	-0,07	(1,05 ; 1.09)	-0,10	(1,07 ; 1,12)	-0,13	(1,10 ; 1,16)
6000	-0,13	(1,11 ; 1,16)	-0,10	(1,07 ; 1,13)	-0,08	(1,06 ; 1,11)	-0,08	(1,06 ; 1,10)	-0,11	(1,08 ; 1,13)	-0,15	(1,12 ; 1,19)
8000	-0,14	(1,12 ; 1,17)	-0,11	(1,08 ; 1,14)	-0,09	(1,06 ; 1,12)	-0,08	(1,06 ; 1,10)	-0,11	(1,09 ; 1,14)	-0,16	(1,13 ; 1,20)
10000	-0,16	(1,13 ; 1,18)	-0,12	(1,09 ; 1,14)	-0,09	(1,06 ; 1,13)	-0,09	(1,07 ; 1,11)	-0,12	(1,10 ; 1,15)	-0,18	(1,14 ; 1,21)
12000	-0,17	(1,14 ; 1,20)	-0,12	(1,09 ; 1,15)	-0,10	(1,07 ; 1,13)	-0,10	(1,07 ; 1,12)	-0,13	(1,10 ; 1,16)	-0,19	(1,15 ; 1,23)
14000	-0,18	(1,15 ; 1,21)	-0,13	(1,10 ; 1,16)	-0,10	(1,07 ; 1,14)	-0,10	(1,08 ; 1,13)	-0,14	(1,11 ; 1,17)	-0,20	(1,16 ; 1,24)
16000	-0,19	(1,16 ; 1,22)	-0,14	(1,10 ; 1,17)	-0,11	(1,07 ; 1,14)	-0,11	(1,08 ; 1,13)	-0,15	(1,12 ; 1,18)	-0,21	(1,17 ; 1,25)
18000	-0,21	(1,18 ; 1,24)	-0,14	(1,11 ; 1,18)	-0,11	(1,07 ; 1,14)	-0,11	(1,09 ; 1,14)	-0,15	(1,12 ; 1,18)	-0,22	(1,18 ; 1,26)
20000	-0,22	(1,19 ; 1,26)	-0,15	(1,12 ; 1,19)	-0,11	(1,08 ; 1,15)	-0,12	(1,09 ; 1,14)	-0,16	(1,13 ; 1,19)	-0,23	(1,19 ; 1,28)
22000	-0,23	(1,20 ; 1,27)	-0,16	(1,12 ; 1,19)	-0,11	(1,08 ; 1,15)	-0,12	(1,10 ; 1,15)	-0,17	(1,14 ; 1,20)	-0,25	(1,20 ; 1,30)
24000	-0,26	(1,22 ; 1,30)	-0,17	(1,13 ; 1,21)	-0,12	(1,08 ; 1,16)	-0,13	(1,10 ; 1,16)	-0,18	(1,15 ; 1,22)	-0,26	(1,21 ; 1,31)
26000	-0,28	(1,24 ; 1,32)	-0,18	(1,14 ; 1,22)	-0,12	(1,08 ; 1,16)	-0,14	(1,11 ; 1,17)	-0,19	(1,16 ; 1,23)	-0,27	(1,22 ; 1,32)
28000	-0,30	(1,25 ; 1,34)	-0,18	(1,14 ; 1,23)	-0,12	(1,08 ; 1,17)	-0,14	(1,11 ; 1,17)	-0,20	(1,16 ; 1,24)	-0,28	(1,23 ; 1,34)
30000	-0,33	(1,28 ; 1,37)	-0,19	(1,15 ; 1,24)	-0,13	(1,08 ; 1,17)	-0,15	(1,12 ; 1,18)	-0,21	(1,17 ; 1,25)	-0,29	(1,23 ; 1,35)
32000	-0,35	(1,30 ; 1,40)	-0,20	(1,15 ; 1,25)	-0,13	(1,09 ; 1,17)	-0,16	(1,12 ; 1,19)	-0,22	(1,18 ; 1,27)	-0,31	(1,24 ; 1,37)
34000	-0,38	(1,33 ; 1,43)	-0,21	(1,16 ; 1,25)	-0,13	(1,09 ; 1,18)	-0,17	(1,13 ; 1,20)	-0,24	(1,19 ; 1,28)	-0,32	(1,25 ; 1,38)
36000	-0,40	(1,35 ; 1,46)	-0,21	(1,16 ; 1,27)	-0,13	(1,09 ; 1,18)	-0,17	(1,14 ; 1,21)	-0,25	(1,20 ; 1,30)	-0,32	(1,25 ; 1,39)

Número de Marcadores	Características fenotípicas									
	NSPP		FPP		FP		RB		TP	
	Viés	IC	Viés	IC	Viés	IC	Viés	IC	Viés	IC
2000	-0,12	(1,09 ; 1,15)	-0,08	(1,06 ; 1.09)	-0,09	(1,07 ; 1,11)	-0,09	(1,07 ; 1,12)	-0,14	(1,11 ; 1,17)
4000	-0,13	(1,11 ; 1,16)	-0,08	(1,07 ; 1,10)	-0,10	(1,08 ; 1,12)	-0,10	(1,08 ; 1,13)	-0,16	(1,13 ; 1,19)
6000	-0,15	(1,12 ; 1,17)	-0,09	(1,07 ; 1,11)	-0,11	(1,09 ; 1,13)	-0,11	(1,09 ; 1,14)	-0,18	(1,15 ; 1,21)
8000	-0,16	(1,13 ; 1,19)	-0,10	(1,08 ; 1,12)	-0,12	(1,10 ; 1,15)	-0,12	(1,10 ; 1,15)	-0,20	(1,16 ; 1,23)
10000	-0,17	(1,14 ; 1,20)	-0,11	(1,09 ; 1,13)	-0,14	(1,11 ; 1,16)	-0,13	(1,10 ; 1,16)	-0,21	(1,18 ; 1,25)
12000	-0,19	(1,16 ; 1,22)	-0,12	(1,10 ; 1,14)	-0,14	(1,12 ; 1,17)	-0,14	(1,11 ; 1,17)	-0,23	(1,19 ; 1,26)
14000	-0,20	(1,17 ; 1,24)	-0,13	(1,10 ; 1,15)	-0,16	(1,13 ; 1,18)	-0,15	(1,12 ; 1,18)	-0,24	(1,21 ; 1,28)
16000	-0,22	(1,18 ; 1,25)	-0,14	(1,11 ; 1,16)	-0,17	(1,14 ; 1,20)	-0,16	(1,13 ; 1,19)	-0,26	(1,21 ; 1,30)
18000	-0,23	(1,19 ; 1,27)	-0,15	(1,12 ; 1,17)	-0,18	(1,15 ; 1,21)	-0,17	(1,13 ; 1,21)	-0,27	(1,23 ; 1,31)
20000	-0,24	(1,21 ; 1,28)	-0,16	(1,13 ; 1,18)	-0,20	(1,16 ; 1,23)	-0,18	(1,14 ; 1,22)	-0,29	(1,24 ; 1,33)
22000	-0,26	(1,22 ; 1,30)	-0,16	(1,13 ; 1,19)	-0,22	(1,18 ; 1,25)	-0,19	(1,15 ; 1,23)	-0,31	(1,26 ; 1,36)
24000	-0,29	(1,24 ; 1,33)	-0,18	(1,15 ; 1,21)	-0,22	(1,19 ; 1,26)	-0,20	(1,16 ; 1,24)	-0,33	(1,28 ; 1,39)
26000	-0,31	(1,26 ; 1,36)	-0,19	(1,16 ; 1,23)	-0,25	(1,21 ; 1,29)	-0,21	(1,16 ; 1,25)	-0,36	(1,30 ; 1,41)
28000	-0,33	(1,28 ; 1,38)	-0,20	(1,17 ; 1,23)	-0,26	(1,22 ; 1,30)	-0,22	(1,17 ; 1,26)	-0,39	(1,33 ; 1,45)
30000	-0,35	(1,29 ; 1,40)	-0,22	(1,18 ; 1,25)	-0,28	(1,24 ; 1,33)	-0,22	(1,18 ; 1,27)	-0,40	(1,34 ; 1,46)
32000	-0,37	(1,31 ; 1,43)	-0,23	(1,19 ; 1,27)	-0,30	(1,25 ; 1,35)	-0,23	(1,18 ; 1,28)	-0,44	(1,37 ; 1,51)
34000	-0,40	(1,33 ; 1,46)	-0,25	(1,21 ; 1,30)	-0,34	(1,29 ; 1,40)	-0,25	(1,20 ; 1,30)	-0,47	(1,40 ; 1,55)
36000	-0,42	(1,35 ; 1,48)	-0,27	(1,23 ; 1,32)	-0,36	(1,30 ; 1,42)	-0,26	(1,20 ; 1,31)	-0,49	(1,41 ; 1,57)

Comprimento da folha da bandeira (CFB); Largura da folha da bandeira (LFB); Número de panículas por planta (NPP); Altura da planta (AP); Comprimento da panícula (CP); Número do ramo primário da panícula (NRPP); Número de sementes por panícula (NSPP); Flores por panículas (FPP); Fertilidade da panícula (FP); Resistência à brusone (RB); Teor de proteína (TP).

Com exceção da característica teor de proteína, que apresentou capacidade preditiva de 0,79 e herdabilidade 0,81, as demais características demonstraram alta capacidade preditiva, variando de 0,84 até 0,93, e herdabilidade de alta magnitude variando de 0,83 a 0,91 para o grupo de 2.000 marcadores mais importantes.

Observa-se também que, à medida que o número de marcadores menos importantes/PIP aumenta, a capacidade preditiva e herdabilidade diminuem gradativamente. Portanto, a combinação dos métodos BayesA e BayesDπ resulta em um melhor desempenho no modelo de predição em comparação com a utilização dos métodos isoladamente.

As características comprimento da folha da bandeira, fertilidade da panícula e teor de proteína apresentaram o maior decréscimo da capacidade preditiva do modelo e da média da herdabilidade, enquanto a largura da folha da bandeira, número de panículas por planta e altura da planta apresentaram o menor decréscimo. Esse comportamento de decaimento pode ser explicado por Wray et al. (2013), que mostram que, à medida que marcadores irrelevantes para a predição da característica de interesse são incluídos, a capacidade preditiva teórica diminui.

No que diz respeito ao viés, pode-se observar que a seleção de marcadores resultou em um menor viés de predição em comparação com os modelos ajustados pelos métodos BayesA e BayesDπ separadamente. Além disso, os intervalos de confiança para o coeficiente da regressão têm menor amplitude, o que porporciona uma estimativa mais precisa para o coeficiente da regressão (CECON et al., 2012).

Os resultados obtidos neste estudo corroboram com os encontrados por Sousa et al. (2019), que utilizaram o banco de dados de arroz do programa de melhoramento de arroz irrigado do International Rice Research Institute (IRRI) para estimar os efeitos os marcadores pelo método RR-BLUP. Eles selecionaram os 10.000 marcadores de maiores valores absolutos. Após, ordenaram os marcadores, em ordem decrescente, em subconjuntos de 5.000, 2.500, 1.000, 500 e 100 SNP e reestimaram seus efeitos para calcular a capacidade preditiva, herdabilidade e viés de predição.

Na Tabela 2, pode-se observar que a seleção de marcadores resultou em um menor viés de predição em comparação com os modelos ajustados pelos métodos BayesA e BayesDπ separadamente (Tabela 1). Além disso, os intervalos de confiança para o coeficiente de regressão têm menor amplitude, o que

proporciona uma estimativa mais precisa do coeficiente de regressão (CECON et al., 2012).

4. Conclusão

A seleção dos marcadores mais importantes para a predição genômica, com base em sua probabilidade *a posteriori* de inclusão, demonstrou ter uma capacidade preditiva superior em relação aos métodos BayesA e BayesDπ sem a prévia seleção de marcadores. Além disso, essa abordagem mostrou-se menos tendenciosa ao apresentar um menor viés de predição.

A probabilidade *a posteriori* de inclusão no modelo de predição genômica também se mostrou eficaz no entendimento da arquitetura genética das características em estudo. Além disso, a seleção de marcadores contribui para a redução da alta dimensionalidade e do esforço computacional, que são desafios comuns enfrentados na seleção genômica.

Em resumo, essa estratégia de seleção de marcadores com base na probabilidade *a posteriori* de inclusão pode melhorar a precisão da predição genômica e facilitar a interpretação dos resultados, tornando-a uma ferramenta valiosa em estudos de genômica.

5. Referências

AMMIRAJU, J. S. S. et al. 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**, v. 16, n. 1, p. 140-147, 2006.

AZEVEDO, C. F. et al. Bayesian methods for genomic association of chromosomic regions considering the additive-dominance model. **Crop Breeding and Applied Biotechnology**, v. 22, 2022.

AZEVEDO, C. F. et al. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC genetics**, v. 16, n. 1, p. 1-13, 2015.

BORÉM, A.; MIRANDA, G. V.; FRITSCHE-NETO, R. **Melhoramento de plantas**. 8ª edição. São Paulo: Oficina de Textos, 2021.

BUKSZÁR, J.; MCCLAY, J. L.; VAN DEN OORD, E. J. C. G. Estimating the posterior probability that genome-wide association findings are true or false. **Bioinformatics**, v. 25, n. 14, p. 1807-1813, 2009.

CONAB - COMPANHIA NACIONAL DE ABASTECIMENTO. Acompanhamento da Safra Brasileira de Grãos, Brasília, DF, v. 10, safra 2022/23, n. 5 quinto levantamento, fevereiro 2023. Disponível em: <https://www.conab.gov.br/info-agro/safras/graos/boletim-da-safra-de-graos>. Acesso em: 07 de mar. 2023.

CECON, P. R. et al. **Métodos estatísticos (Série Didática)**. Viçosa, MG: Editora da UFV, 2012.

CHILDS, N.; LEBEAU, B. Rice Outlook: February 2023, RCS-23B, U.S. Department of Agriculture, Economic Research Service, February 10, 2023. Disponível em: <https://www.ers.usda.gov/webdocs/outlooks/105815/rcs-23b.pdf?v=3093.5 >. Acesso em: 07 de mar. 2023.

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

GUO, Z. et al. The impact of population structure on genomic prediction in stratified populations. **Theoretical and applied genetics**, v. 127, p. 749-762, 2014.

GUAN, Y.; STEPHENS, M. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. **The Annals of Applied Statistics**, Vol. 5, n. 3, 2011.

HABIER, D. et al. **Extension of the bayesian alphabet for genomic selection**. BMC Bioinformatics, 12:186. 2011.

HAYES, B. J. et al. Accuracy of genomic breeding values in multi-breed dairy cattle populations. **Genetics Selection Evolution**, v. 41, n. 1, p. 1-9, 2009.

HASAN, M. M et al. Marker-assisted backcrossing: a useful method for rice improvement. **Biotechnology & Biotechnological Equipment**, v. 29, n. 2, p. 237-254, 2015.

HAILE, T. A. et al. Genomic selection for lentil breeding: Empirical evidence. **The Plant Genome**, v. 13, n. 1, p. e20002, 2020.

HESLOT, N.; JANNINK, J. L.; SORRELLS, M. E. Perspectives for genomic selection applications and research in plants. **Crop Science**, v. 55, n. 1, p. 1-12, 2015.

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

JILO, T. et al. Genetic variability, heritability and genetic advance of maize (Zea mays L.) inbred lines for yield and yield related traits in southwestern Ethiopia. **Journal of plant breeding and crop science**, v. 10, n. 10, p. 281-289, 2018.

LOPEZ, M. A.; XAVIER, A.; RAINEY, K. M. Phenotypic variation and genetic architecture for photosynthesis and water use efficiency in soybean (Glycine max L. Merr). **Frontiers in plant science**, v. 10, p. 680, 2019.

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

NEVES, H. H. R.; CARVALHEIRO, R.; QUEIROZ, S. A. A comparison of statistical methods for genomic selection in a mice population. **BMC genetics**, v. 13, n. 1, p. 1-17, 2012.

Nações Unidas, Departamento de Assuntos Econômicos e Sociais, Divisão de População (2022). Perspectivas da População Mundial 2022: Nota de lançamento.

PAULINO, C. D.; TURKMAN, M. A. A.; MURTEIRA, B.; SILVA, G. L. **Estatística Bayesiana**. Lisboa, Portugal: Fundação Calouste Gulbenkian, 2018.

PEIXOTO, L. A.; BHERING, L. L.; CRUZ, C. D. **Seleção Genômica Aplicada ao Melhoramento Genético**. Viçosa, MG: Editora UFV, 2022.

PEREZ, P.; DE LOS CAMPOS, G. Genome-Wide Regression and Prediction with the BGLR Statistical Package. **Genetics**, v. 198, n. 2, p. 483-495, 2014.

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. de; SILVA, F. F. e; AZEVEDO, C. F. **Estatística matemática, biométrica e computacional: Modelos mistos, multivariados, categóricos e generalizados (REML/BLUP), inferência bayesiana, regressão aleatória, seleção genômica, QTL-GWAS, estatística espacial e temporal, competição, sobrevivência**. Viçosa: UFV, 2014.

SPINDEL, J. et al. Genomic selection and association mapping in rice (Oryza sativa): effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS genetics**, v. 11, n. 2, p. e1004982, 2015.

SUELA, M. M. et al. Combined index of genomic prediction methods applied to productivity. **Ciência Rural**, v. 49, 2019.

SOUSA, M. B. et al. Increasing accuracy and reducing costs of genomic prediction by marker selection. **Euphytica**, v. 215, p. 1-14, 2019.

WALTER, M.; MARCHEZAN, E.; ÁVILA, L. A. de. Arroz: composição e características nutricionais. **Ciência Rural**, v. 38, p. 1184-1192, 2008.

WANG, X. et al. Predicting rice hybrid performance using univariate and multivariate GBLUP models based on North Carolina mating design II. **Heredity**, v. 118, n. 3, p. 302-310, 2017.

WANG, G. et al. A simple new approach to variable selection in regression, with application to genetic fine mapping. **Journal of the Royal Statistical Society Series B: Statistical Methodology**, v. 82, n. 5, p. 1273-1300, 2020.

WISSER, R. J. et al. Identification and characterization of regions of the rice genome associated with broad-spectrum, quantitative disease resistance. **Genetics**, v. 169, n. 4, p. 2277-2293, 2005.

WRAY, N. R. et al. Pitfalls of predicting complex traits from SNPs. **Nature Reviews Genetics**, v. 14, n. 7, p. 507-515, 2013.

YANG, D. et al. Inference of posterior inclusion probability of QTLs in Bayesian shrinkage analysis. **Genetics Research**, v. 97, p. e6, 2015.

ZHAO, K. et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. **Nature communications**, v. 2, n. 1, p. 1-10, 2011.