ISSN: 2317-0840

Índice para identificação de regiões genômicas estáveis em GWAS utilizando múltiplos métodos estatísticos

Renata Dourado Roque[1], Brenda Vieira de Oliveira[1], Guilherme Flaviano Pereira[1], Camila Ferreira Azevedo[1†]

[1] *Universidade Federal de Viçosa; Centro de Ciências Exatas e Tecnológicas; Programa de Pós-Graduação em Estatística Aplicada e Biometria; Viçosa – Minas Gerais, Brasil.*

Resumo: *A Associação Genômica Ampla (Genome Wide Association Studies – GWAS) visa identificar associações entre loci de características quantitativas (Quantitative Trait Loci – QTL) e fenótipos. Esses estudos são de interesse para programas de melhoramento genético, pois possibilitam a identificação de marcadores associados a fenótipos de importância agronômica. Os métodos comumente utilizados na GWAS incluem Mixed Linear Model (MLM), Compressed MLM (CMLM), General Linear Model (GLM), Settlement of MLM Under Progressively Exclusive Relationship (SUPER), Multiple Locus Mixed Linear Model (MLMM), e FarmCPU. Esses métodos diferem em suas bases teóricas, resultando em variações nas regiões genômicas detectadas. Este estudo teve como objetivo desenvolver dois índices, um sem ponderação e outro ponderado pela importância de cada região, para identificar regiões genômicas significativas e estáveis entre os métodos estatísticos aplicados à GWAS. O intuito foi minimizar a detecção de regiões potencialmente falsas positivas. A análise foi realizada em um conjunto de dados composto por 413 indivíduos de arroz asiático (Oryza sativa), genotipados com 36.901 SNPs (Single Nucleotide Polymorphisms). Foram avaliadas 11 características fenotípicas, e os seis métodos mencionados foram comparados. Entre as características analisadas, quatro se destacaram por apresentarem mais regiões genômicas significativas detectadas por um maior número de métodos. Os métodos SUPER e GLM se mostraram os mais eficazes na detecção de associações genômicas. Os índices detectaram regiões genômicas descritas na literatura, associadas às características como resistência à brusone, fertilidade e altura. O índice ponderado demostrou maior sensibilidade na detecção de regiões genômicas significativas, devido à atribuição dos pesos, que priorizou as regiões mais importantes.*

Palavras-chave: *genotipagem; marcadores moleculares; melhoramento genético.*

Introdução

Os avanços na biotecnologia, especialmente no desenvolvimento e na aplicação de marcadores moleculares do tipo SNP (*Single Nucleotide Polymorphisms*), têm permitido a identificação de indivíduos com base em suas variações genéticas, impulsionando os estudos de associação genômica ampla (*Genome Wide Association Studes* - GWAS) (TIBBS CORTES; ZHANG; YU, 2021). A GWAS busca identificar associações entre loci de características quantitativas (*Quantitative Trait Loci* - QTL) e valores genéticos associados a características de interesse. Na prática, essas associações são realizadas entre SNPs e fenótipos devido à pressuposição de desequilíbrio de ligação (*Linkage Desequilibrium* - LD) entre marcadores e QTL. Essa abordagem fornece informações detalhadas sobre regiões genômicas que influenciam diretamente nas características alvos, sendo, portanto, de grande interesse para os programas de melhoramento genético.

No contexto do melhoramento genético, a GWAS se destaca como uma ferramenta essencial para compreender a arquitetura genética de características complexas, possibilitando a identificação de genes associados (*gene mining*) e o entendimento dos mecanismos moleculares que controlam essas características. Além disso, a GWAS viabiliza a aplicação da seleção assistida por marcadores (SAM), que utiliza marcadores moleculares para identificar e selecionar, de forma eficiente e com um menor custo em relação à predição genômica, indivíduos com maior potencial genético (BOOPATHI, 2020).

[†] **Autor correspondente: camila.azevedo@ufv.br**

Manuscrito recebido em: 23/12/2024; Revisado em: 14/03/2025; Aceito em: 25/03/2025.

Diversas abordagens estatísticas têm sido propostas no contexto da GWAS, e os métodos baseados em modelos lineares mistos (MLM) têm se destacado pela capacidade de controlar fatores que podem afetar o surgimento de falsos positivos nas análises, como estrutura de população e parentesco familiar, pois indivíduos dentro de uma mesma população frequentemente compartilham blocos genômicos extensos devido à ancestralidade comum (ZHANG *et al.*, 2010). O *General Linear Model* (GLM) é o método mais simples de GWAS que inclui correção para estrutura populacional sem a inclusão de efeitos aleatórios no modelo. Já os métodos amplamente utilizados e baseados em modelos mistos incluem: *Mixed Linear Model* (MLM), *Compressed MLM* (CMLM), *Settlement of MLM Under Progressively Exclusive Relationship* (SUPER), *Multiple Locus Mixed Linear Model* (MLMM) e *FarmCPU*.

Cada um desses métodos apresenta distinções teóricas e diferentes abordagens para corrigir a estrutura populacional e o parentesco, influenciando, consequentemente, o controle de falsos positivos e o tempo computacional (TIBBS CORTES; ZHANG; YU, 2021). Como resultado, há variações nas regiões genômicas detectadas. Esses métodos têm sido aplicados a diversas características fenotípicas e culturas, como rendimento de grãos e características da mostarda indiana (AKHATAR; BANGA, 2015), tolerância ao sal em linhas de germoplasma de soja (ZENG *et al.*, 2017), característica de rendimento em trigo (MALIK *et al.*, 2021), quantidade de clorofila em milho (XIONG *et al.*, 2023), mecanismos genéticos que regulam a arquitetura da panícula no arroz (ZHONG *et al.*, 2021) e produção de leite em raças de gado leiteiro francês (TEISSIER *et al.*, 2018).

As diferentes regiões genômicas detectadas por metodologias estatísticas distintas podem dificultar a seleção, por parte do pesquisador, das regiões genômicas a serem utilizadas nas rotinas dos programas de melhoramento. Assim, torna-se necessária a criação de procedimentos que facilitem a tomada de decisões pelos melhoristas. Neste contexto, inspirado nos índices de seleção desenvolvidos para a escolha de indivíduos geneticamente superiores (HAZEL, 1943; SMITH, 1936), foi proposto um índice para seleção de regiões genômicas. Esse índice visa identificar regiões genômicas significativas e estáveis, reduzindo esforços, tempo e recursos em análises que possam levar a falsos positivos, e priorizando aquelas que realmente merecem uma investigação mais aprofundada.

A presente pesquisa utilizou um conjunto de dados composto por 413 indivíduos de arroz asiático (*Oryza sativa*), genotipados para 36.901 marcadores do tipo SNP. Foram avaliadas 11 características fenotípicas relacionadas à produtividade, morfologia, qualidade do grão e resistência a doenças para a avaliação dos índices de seleção propostos.

Materiais e Métodos

Dados

O conjunto de dados utilizado nesta pesquisa refere-se a 413 indivíduos de arroz asiático *Oryza sativa*, genotipados para 36.901 marcadores do tipo SNPs. Os dados são públicos, fazem parte do Projeto OryzaSNP e do Projeto OMAP) (AMMIRAJU *et al.*, 2006; ZHAO *et al.*, 2011) e estão disponíveis em https://ricediversity.org/data/. O controle de qualidade dos marcadores foi realizado considerando um call rate inferior a 70% e uma baixa frequência do alelo mais raro inferior a 1%. Os experimentos foram conduzidos em Arkansas, Estados Unidos, durante o período de maio ao outono, nos anos de 2006 e 2007. Para o experimento, foram utilizadas duas repetições por ano em um delineamento em blocos completos casualizados. Cada parcela consistia em fileiras de 5 metros de comprimento, com espaçamento de 25 cm entre plantas e 50 cm entre fileiras (ZHAO *et al.*, 2011).

As características fenotípicas avaliadas são relacionadas à produtividade, morfologia, qualidade do grão e resistência a doenças, incluindo as dimensões da folha bandeira (comprimento e largura), número de flores e sementes por panícula, fertilidade e comprimento da panícula, número de panículas por planta, altura das plantas, número de ramos primários na panícula, teor de proteínas e resistência à brusone.

Foram avaliados seis métodos estatísticos, o *Mixed Linear Model* (MLM), *Compressed MLM* (CMLM), *General Linear Model* (GLM), *Settlement of MLM Under Progressively Exclusive Relationship* (SUPER), *Multiple Locus Mixed Linear Model* (MLMM) e *FarmCPU*, ajustados no pacote GAPIT do *software* R (WANG; ZHANG, 2021). A escolha desses métodos foi baseada em suas diferenças teóricas e modelagens específicas para controlar a taxa de falsos positivos e aumentar a precisão na detecção de regiões genômicas significativas. O GLM e o MLM são amplamente utilizados, pois consideram a estrutura populacional, reduzindo a taxa de falsos positivos (PRICE *et al.*, 2006). O CMLM otimiza o MLM ao agrupar indivíduos similares, reduzindo o esforço computacional (ZHANG *et al.*, 2010). O MLMM permite mapear características complexas que são controladas por múltiplos lócus genéticos (KALER *et al.*, 2020). O SUPER seleciona um subconjunto de marcadores para definir a matriz de parentesco, aumentando o poder estatístico (WANG *et al.*, 2014). Já o FarmCPU separa os efeitos fixos e aleatórios do modelo, melhorando o controle de fatores de confusão entre marcadores de teste e parentesco aumentando a eficiência computacional (LIU *et al.*, 2016).

Métodos utilizados para análise da GWAS

Modelo Linear Geral

O Modelo Linear Geral (*General linear Model* – GLM) é uma abordagem de regressão aplicada a marcadores individuais, que inclui correção para a estrutura populacional por meio da extração da matriz de parentesco genômica os primeiros componentes principais, para redução de associações espúrias (PRICE *et al.*, 2006). Definido por:

$$y = 1\mu + Qq + M_i m_i + e$$

em que y é o vetor de observações fenotípicas, 1 é o vetor cujos elementos são iguais a 1, μ é a média geral, q é o vetor com os primeiros q componentes principais da matriz de parentesco dos indivíduos (K) obtida conforme proposto por Vanraden (2008) e m_i é o efeito fixo do i-ésimo marcador. Q e $M_i$são as matrizes de incidência de seus respectivos efeitos. A distribuição de probabilidade dos erros aleatórios $e\, N(0, I\sigma_e^2)$, em que I é a matriz identidade e σ_e^2 é a variância residual.

Modelo Linear Misto

O Modelo Linear Misto (*Mixed Linear Model* - MLM) é uma abordagem de regressão aplicada a marcadores individuais, que inclui correção para a estrutura populacional por meio da extração dos primeiros componentes principais da matriz de parentesco genômica, reduzindo associações espúrias (PRICE *et al.*, 2006). Além disso, os efeitos poligênicos aleatórios são incluídos para controlar as diferenças de parentesco entre indivíduos, contribuindo ainda mais para a redução de falsos positivos (YU *et al.,* 2006). Esse modelo é definido por:

$$y = 1\mu + Qq + Zu + M_i m_i + e$$

em que y é o vetor de observações fenotípicas, 1 é o vetor cujos elementos são iguais a 1, μ é a média geral, q é o vetor com os primeiros q componentes principais da matriz de parentesco dos indivíduos (K), u é o vetor de efeitos poligênicos e m_i é o efeito fixo do i-ésimo marcador. Q, M_ie Z são as matrizes de incidência de seus respectivos efeitos. As distribuições de probabilidade dos efeitos aleatórios são $u\,N(0, K\sigma_u^2)$ e $e\,N(0, I\sigma_e^2)$, em que I é a matriz identidade**,** K é a matriz de parentesco genômico**,** σ_u^2 é a variância dos efeitos poligênicos e σ_e^2 é a variância residual.

Modelo Linear Misto Comprimido

O Modelo Linear Misto Comprimido (*Compressed MLM* – CMLM) é uma abordagem desenvolvida para evitar a dupla contagem das informações da matriz de parentesco no MLM, contabilizadas tanto nos componentes principais quanto nos efeitos poligênicos (LI *et al.*, 2014). Nesse método, é realizada uma análise de agrupamento utilizando a matriz de parentesco como medida de similaridade, empregando o método UPGMA (*Unweighted Pair Group Method With Arithmetic Mean*) para agrupar indivíduos semelhantes. Após a atribuição das linhas a grupos, estatísticas resumidas do parentesco entre e dentro dos grupos são usadas como elementos de uma matriz de parentesco reduzida. Esse procedimento é utilizado para criar uma matriz de parentesco reduzida para cada nível de compressão e criar uma matriz de parentesco reduzida ($K_{reduzida}$).

Modelo Linear Misto de Múltiplos Lócus

O Modelo Linear Misto de Múltiplos Lócus (*Multiple Locus Mixed Linear Model* – MLMM) combina a regressão do MLM com um processo interativo de seleção passo a passo (*forward-backward*) (SEGURA *et al.*, 2012). Essa abordagem permite identificar múltiplos lócus genéticos ao incorporar os efeitos de marcadores associados como covariáveis no modelo, contribuindo para o controle de falsos positivos.

SUPER

O método SUPER (*Settlement of MLM Under Progressively Exclusive Relationship*) foi desenvolvido com o objetivo de melhorar a detecção de regiões genômicas significativas, controlar a taxa de falsos positivos e a redução da inflamação dos valores de p-valor (KALER *et al.*, 2020). Nesse método, genoma é dividido em segmentos de tamanhos iguais, e cada segmento é representado pelo marcador mais significativo presente nele. Além disso, o tamanho e o número de segmentos selecionados são estimados utilizando o método de máxima verossimilhança em um modelo aleatório, com a matriz de parentesco derivada dos segmentos. O SUPER utiliza uma matriz de parentesco derivada dos marcadores associados, excluindo aqueles que apresentam forte desequilíbrio de ligação (*linkage disequilibrium – LD*) com os marcadores testados.

FarmCPU

O modelo FarmCPU (*Fixed and Random Model Circulating Probability Unification*) (LIU et al., 2016), desenvolvido com base no método MLMM, é uma abordagem que otimiza os marcadores associados, conhecidos como pseudo-nucleotídeos de características quantitativas (*pseudo Quantitative Trait Nucleotides* - pseudo-QTNs), e os utiliza como covariáveis para reduzir parcialmente a confusão entre os marcadores teste e os assumidos como covariáveis. O método

opera em duas etapas principais. Na primeira etapa, chamada Modelo de Efeito Fixo (MEF), cada marcador é testado individualmente, incorporando os marcadores associados previamente como covariáveis, o que contribui para o controle de falsos positivos. Na segunda etapa, denominada Modelo de Efeito Aleatório (MEA), os efeitos dos marcadores são estimados, e essas estimativas são utilizadas para definir a matriz de parentesco. Essas etapas são executadas de forma iterativa, permitindo o ajuste progressivo dos marcadores associados e melhorando a precisão das análises ao longo do processo.

Índice de seleção

O índice I é inspirado nos índices de seleção desenvolvidos para a escolha de indivíduos geneticamente superiores (HAZEL, 1943; SMITH, 1936) e é expresso por uma combinação linear dos métodos estatísticos aplicados à GWAS que foram avaliados, conforme descrito por:

$$I_j = b_{1j}m_{1j} + b_{2j}m_{2j} + b_{3j}m_{3j} + b_{4j}m_{4j} + b_{5j}m_{5j} + b_{6j}m_{6j}$$

em que I_j é o valor do índice para o j-ésimo marcador ($j = 1, \ldots, 36.901$) os b_{ij} são os coeficientes de ponderação do i-ésimo marcador e j-ésimo método ($i = 1, \ldots, 6$). E m_{ij} é uma variável indicadora definida como:

$$m_{ij} = \begin{cases} 1, \text{se o } SNP_j \text{ foi detectado pelo método } j \\ 0, \text{se } SNP_j \text{ não foi detectado pelo método } j \end{cases}$$

Foram criados dois índices de seleção das regiões genômicas. O primeiro utilizou todos os coeficientes iguais a 1 ($b_{ij} = 1$), considerando apenas a frequência simples de detecção das regiões. Já no segundo índice, foram atribuídos pesos que representam a importância de cada região, definido como:

$$b_{ij} = \frac{-\log\left(p - valor_{ij}\right)}{\sum_{i=1}^{6} \log\left(p - valor_{ij}\right)}$$

em que $p - valor_{ij}$ é o p-valor associado ao i-ésimo marcador e j-ésimo método.

As regiões genômicas foram definidas nos 12 cromossomos com tamanho 0.69 Mb de acordo com o avaliado Suela *et al.* (2022).

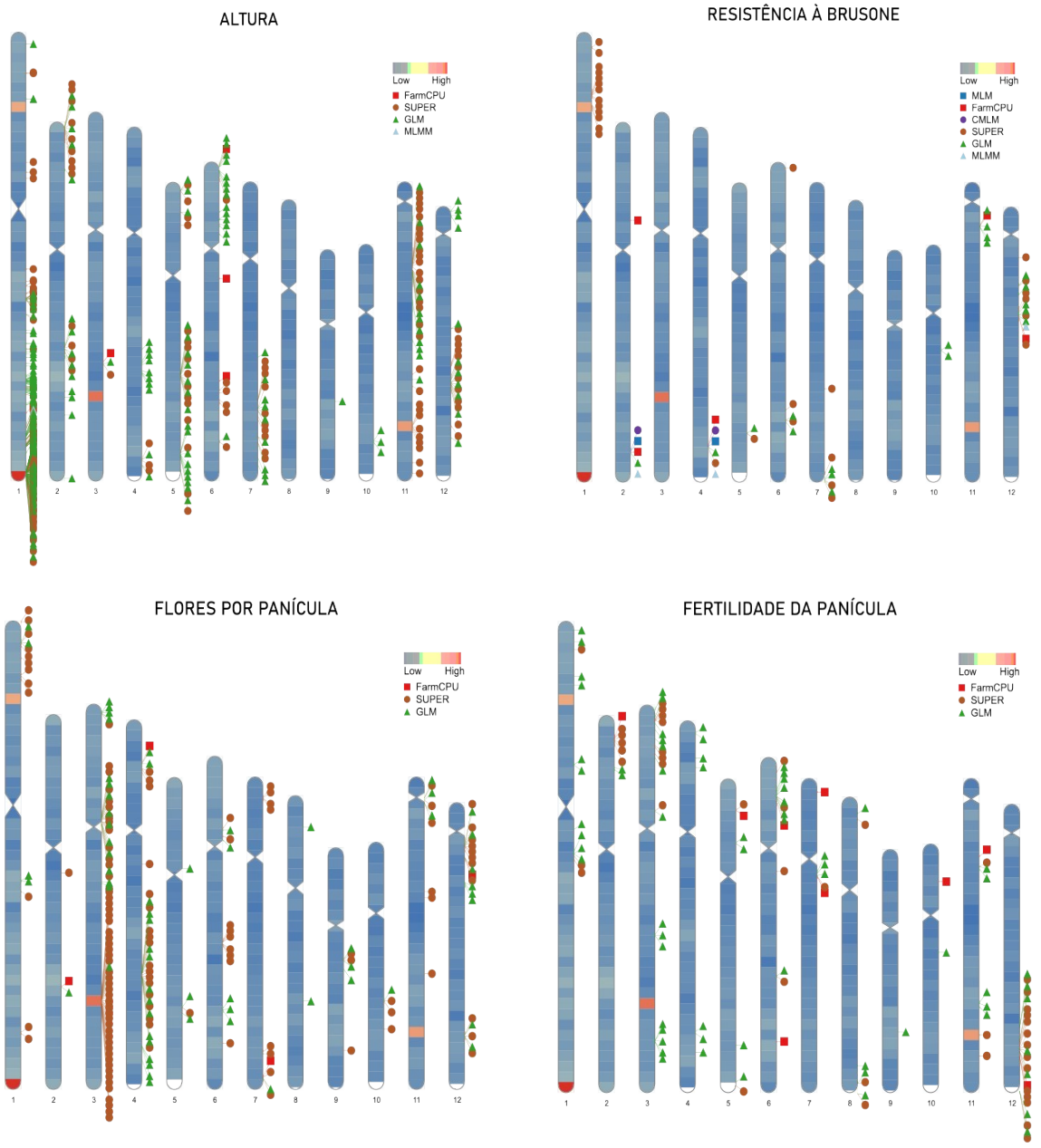
Resultados e Discussões

Ao comparar o desempenho dos métodos mencionados, GLM e SUPER foram os que identificaram o maior número de associações, enquanto MLM e CMLM detectaram um número inferior (Figura 1). Um LD elevado em uma determinada região genômica pode resultar em estimativas inflacionadas dos efeitos dos SNPs (TRYNKA *et al.*, 2015). Desta forma, SNPs podem ser erroneamente identificados como associados a um fenótipo, mesmo na ausência de uma relação causal real. Isso acorre porque os SNPs podem estar próximos de um QTL (*Quantitative Trait Locus* - Lócus de Caráter Quantitativo), mas a associação observada resulta, na verdade, da correlação entre os marcadores e não de uma causalidade direta. De acordo com Suela *et al.* (2022), o LD médio para os marcadores dentro de uma região de 0,69 Mb é de aproximadamente 0,20 neste conjunto de dados genômicos. Por exemplo, para as características altura da planta e número de flores por panícula, observou-se uma elevada detecção de regiões nos cromossomos 1 e 3, respectivamente, pelos métodos GLM e SUPER.

Outro fator biológico que contribui para uma elevada taxa de falsos positivos é a heterogeneidade populacional, decorrente da estrutura de população e do parentesco entre indivíduos (HALDAR; GHOSH, 2012). Diferenças genéticas entre subgrupos dentro de uma população podem levar a associações espúrias, onde as associações observadas são causadas por variações populacionais e não por uma relação genuína entre o SNP e a característica fenotípica. O GLM corrige apenas para a estrutura populacional, enquanto o SUPER elimina os marcadores com alto LD, o que pode ser insuficiente para minimizar falsos positivos.

Entre as características fenotípicas analisadas, quatro se destacaram por apresentarem um número maior de regiões genômicas significativas. Em particular, uma região localizada no cromossomo 4, associado à resistência à brusone, foi identificada por todos os métodos, conforme ilustrado na Figura 1.

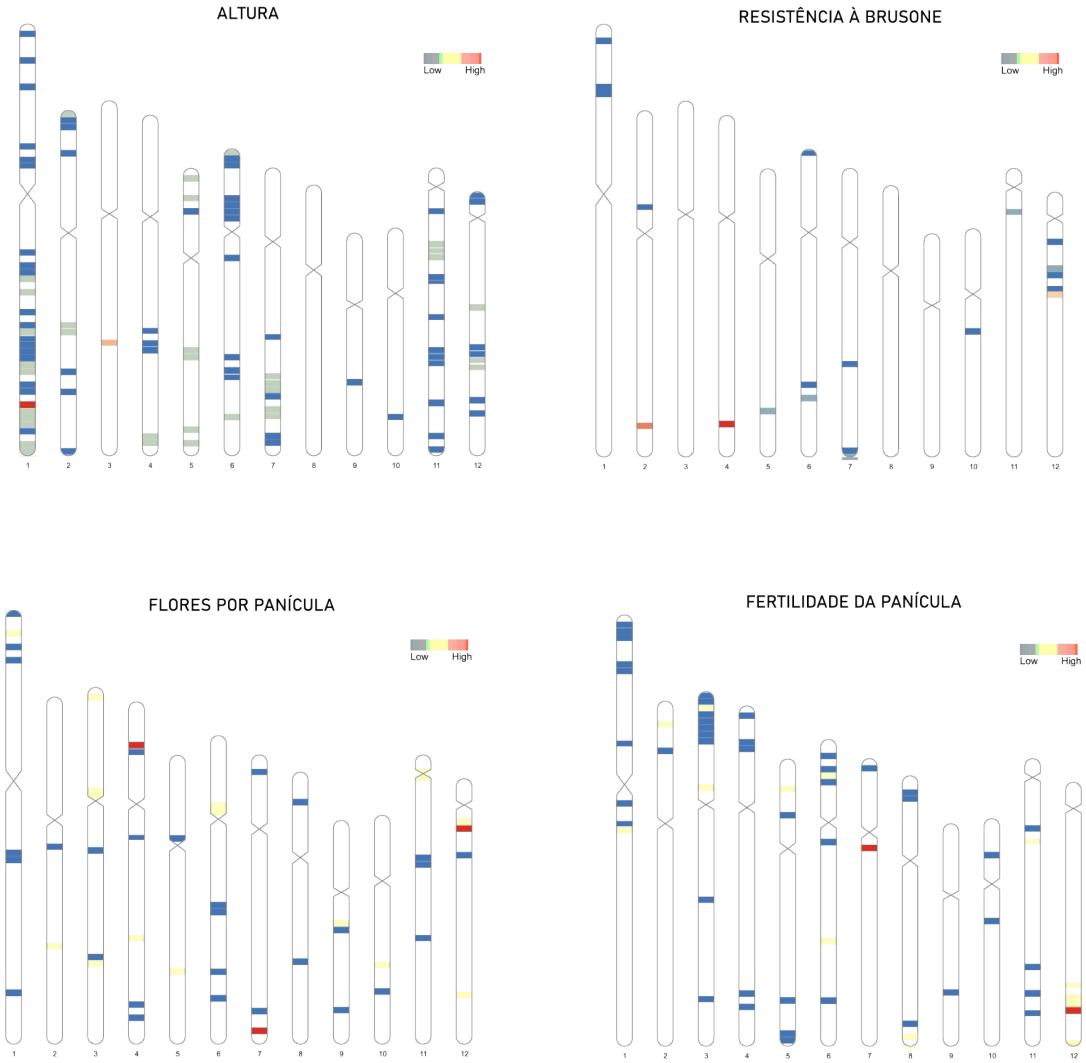
Figura 1: Densidade de SNPs exibida em uma escala de alta (vermelho) a baixa (azul), com formas geométricas representando as regiões significativas detectadas por cada método.



Fonte: dos autores (2025).

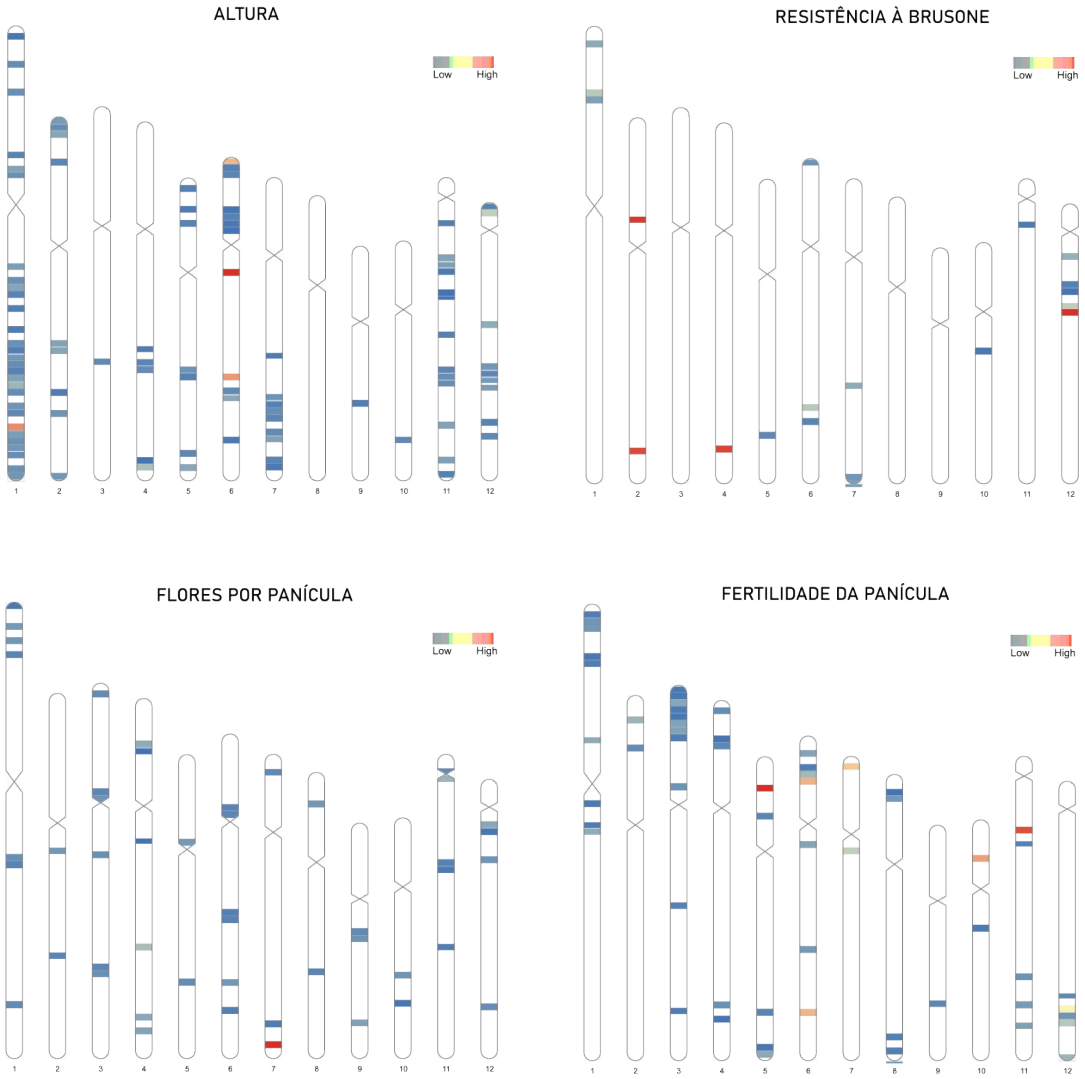
As Figuras 2 e 3 apresentam, por meio das cores, as intensidades dos índices de cada região do genoma. A cor de cada região reflete sua relevância de acordo com cada índice, sendo que a cor vermelha indica as regiões que os melhoristas, conforme cada índice, devem observar com mais atenção. Ao comparar os padrões de cores/intensidades dos índices, observou-se que as intensidades variaram entre as características. Para resistência à brusone e número de flores por panícula, apenas uma região apresentou a cor vermelha em ambos os índices, localizada nas porções iniciais dos cromossomos 4 e 7, respectivamente. Já para a característica fertilidade da panícula, as regiões destacadas em vermelho divergiram. No índice sem ponderação, essas regiões foram identificadas nos cromossomos 7 e 12, enquanto no índice ponderado, foram localizadas nos cromossomos 5 e 11.

Figura 2: Intensidade do índice sem ponderar cada região genômica ao longo dos 12 cromossomos, considerando um tamanho de 0,69 Mb, exibida em uma escala de alta (vermelho) a baixa (azul).



Fonte: dos autores (2025).

Figura 3: Intensidade do índice ponderada para cada região genômica ao longo dos 12 cromossomos, considerando um tamanho de 0,69 Mb, exibida em uma escala de alta (vermelho) a baixa (azul).



Fonte: dos autores (2025).

Contudo, o fato de uma região genômica ser detectada pelos métodos estatísticos não está necessariamente associada à sua importância para a variação genética da característica. O p-valor, conforme já utilizado na literatura (ESPOSITO *et al.*, 2023; SU *et al.*, 2014), é um indicador importante da relevância dos marcadores para a variação daquela característica fenotípica. Os marcadores identificados por meio de GWAS, que representam possíveis QTLs, podem ser usados pelos melhoristas na seleção assistida por marcadores somente se uma proporção razoável da variação genética da característica for explicada por esses marcadores (O'CONNOR *et al.*, 2020). Somente nestes casos, é possível discriminar indivíduos e categorizá-los em classes de selecionados e não selecionados com base em suas dosagens alélicas.

Os índices aplicados nesta pesquisa detectaram regiões genômicas de alta relevância (em vermelho), previamente descritas na literatura. As características que reportaram regiões já

catalogadas foram resistência à brusone, fertilidade e altura, utilizando o índice sem ponderação (Figura 2). Por sua vez, o índice ponderado identificou um maior número de regiões genômicas de alta relevância associadas às mesmas características (Figura 3).

Para a característica resistência à brusone, foram identificadas regiões nos cromossomos 4 e 12 (FUKUOKA; OKUNO, 2001; WANG *et al.*, 1994), sendo que esta última região apresentou maior relevância apenas no índice ponderado. Vale destacar que a resistência à brusone é uma das características mais importantes, devido ao impacto dessa doença na produtividade.

Em relação à altura da planta, as associações foram observadas nos cromossomos 1 e 6, com maior relevância identificada exclusivamente no índice ponderado. Essas regiões foram previamente descritas por Marri *et al.* (2005) e Mei *et al.* (2003), destacando a importância desse fenótipo no rendimento do arroz (ZHANG *et al.*, 2017).

Além disso, foram mapeadas regiões de alta intensidade relacionadas à fertilidade nos cromossomos 5, 8 e 11. No cromossomo 5, a relevância foi identificada apenas pelo índice ponderado, assim como no cromossomo 11. Essas regiões já haviam sido relatadas por He Yu-Qing (2000) e Mei *et al.* (2003).

Por outro lado, nenhuma das metodologias empregadas foi capaz de detectar regiões mais estáveis e de maior relevância associadas à característica número de flores por panícula.

Neste estudo, apresentaram-se e desenvolveram-se índices que permitiram identificar regiões genômicas significativas e estáveis com base nos diferentes métodos avaliados. Os índices foram ponderados por meio dos p-valores dos efeitos dos marcadores no fenótipo, seguindo a ideia de que estes estão associados à importância de cada marcador para a característica. No entanto, outras ponderações podem ser feitas em trabalhos futuros, como o coeficiente de determinação, a variância genética atribuída à região, o efeito da região no fenótipo, entre outras.

A estabilidade a que nos referimos está relacionada aos métodos estatísticos e às teorias por eles utilizadas. No entanto, também existe a estabilidade genética. Para expandir nosso estudo e conhecimento na área de associação genômica ampla, futuros trabalhos devem considerar não apenas uma única geração de dados, como, por exemplo, um ano de avaliação. Isso se deve ao fato de que populações distintas podem apresentar variações genéticas específicas, influenciando diretamente as regiões genômicas significativas detectadas, as quais podem variar, pois são afetadas pela estrutura genética e por fatores ambientais (WOJCIK *et al.*, 2019). Para uma análise mais robusta, é necessário aplicar os índices em diversas populações, a fim de verificar se as regiões genômicas importantes se mantêm estáveis, tanto em relação aos métodos quanto à genética.

Conclusões

Esta pesquisa destaca a relevância de utilizar múltiplos métodos estatísticos em estudos de associação para uma identificação mais eficiente das possíveis regiões significativas, reduzindo os esforços com aquelas que possam ser falsos positivos. Os índices com e sem ponderação apresentam padrões distintos em relação à relevância das regiões, porém, sugere-se o uso da ponderação das regiões no índice para indicar que essa região explique uma proporção razoável da variação genética das características. Em geral, a criação dos índices proporcionou uma melhor visualização das possíveis regiões genômicas associadas, sendo que o índice com ponderação apresentou maior sensibilidade na detecção de regiões significativas, devido aos diferentes pesos atribuídos, valorizando as regiões mais relevantes. Esses resultados oferecem *insights* que podem ser explorados em investigações futuras e aplicados nos programas de melhoramento genéticos, facilitando a identificação de indivíduos com fenótipos de interesse. Em estudos futuros, novas ponderações devem ser avaliadas e os índices aplicados em diferentes gerações de um programa de

melhoramento para verificar a estabilidade das regiões ao longo do tempo, contribuindo para a compreensão da arquitetura genética das características.

Agradecimentos

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pelo apoio financeiro.

Referências

AKHATAR, J.; BANGA, S. S. Genome-wide association mapping for grain yield components and root traits in Brassica juncea (L.) Czern & Coss. **Molecular Breeding**, v. 35, n. 1, p. 48, 2015.

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, jan. 2006.

BOOPATHI, N. M. Genetic Mapping and Marker Assisted Selection: Basics, Practice and Benefits. In: Singapore: Springer, 2020. p. 343–388.

ESPOSITO, S. *et al.* Simultaneous improvement of grain yield and grain protein concentration in durum wheat by using association tests and weighted GBLUP. **TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik**, v. 136, n. 12, 1 dez. 2023.

FUKUOKA, S.; OKUNO, K. QTL analysis and mapping of pi21, a recessive gene for field resistance to rice blast in Japanese upland rice. **Theoretical and Applied Genetics**, v. 103, n. 2, p. 185–190, 2001.

HALDAR, T.; GHOSH, S. Effect of Population Stratification On False Positive Rates Of Population-based Association Analyses Of Quantitative Traits. **Annals of Human Genetics**, v. 76, n. 3, p. 237, maio 2012.

HAZEL, L. N. THE GENETIC BASIS FOR CONSTRUCTING SELECTION INDEXES. **Genetics**, v. 28, n. 6, p. 476–490, 20 nov. 1943.

KALER, A. S. *et al.* Comparing Different Statistical Models and Multiple Testing Corrections for Association Mapping in Soybean and Maize. **Frontiers in Plant Science**, v. 10, 2020.

LI, M. *et al.* Enrichment of statistical power for genome-wide association studies. **BMC Biology**, v. 12, n. 1, p. 73, 2014.

LIU, X. *et al.* Iterative Usage of Fixed and Random Effect Models for Powerful and Efficient Genome-Wide Association Studies. **PLOS Genetics**, v. 12, n. 2, p. e1005767, 1 fev. 2016.

MALIK, P. *et al.* Single-trait, multi-locus and multi-trait GWAS using four different models for yield traits in bread wheat. **Molecular Breeding**, v. 41, n. 7, p. 46, 2021.

MARRI, P. R. *et al.* Identification and mapping of yield and yield related QTLs from an Indian accession of Oryza rufipogon. **BMC Genetics**, v. 6, n. 1, p. 33, 2005.

MEI, H. W. *et al.* Gene actions of QTLs affecting several agronomic traits resolved in a recombinant inbred rice population and two testcross populations. **Theoretical and Applied Genetics**, v. 107, n. 1, p. 89–101, 2003.

O'CONNOR, K. *et al.* Genome-wide association studies for yield component traits in a macadamia breeding population. **BMC Genomics**, v. 21, n. 1, p. 199, 2020.

PRICE, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics 2006 38:8**, v. 38, n. 8, p. 904–909, 23 jul. 2006.

SEGURA, V. *et al.* An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. **Nature Genetics**, v. 44, n. 7, p. 825–830, 2012.

SMITH, H. F. A DISCRIMINANT FUNCTION FOR PLANT SELECTION. **Annals of Eugenics**, v. 7, n. 3, p. 240–250, 1 nov. 1936.

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

SUELA, M. M. *et al.* Regional heritability mapping and genome-wide association identify loci for rice traits. **Crop Science**, v. 62, n. 2, p. 839–858, 1 mar. 2022.

TEISSIER, M. *et al.* Use of meta-analyses and joint analyses to select variants in whole genome sequences for genomic evaluation: An application in milk production of French dairy cattle breeds. **Journal of Dairy Science**, v. 101, n. 4, p. 3126–3139, 2018.

TIBBS CORTES, L.; ZHANG, Z.; YU, J. Status and prospects of genome-wide association studies in plants. **The Plant Genome**, v. 14, n. 1, p. e20077, 1 mar. 2021.

TRYNKA, G. *et al.* Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. **The American Journal of Human Genetics**, v. 97, n. 1, p. 139–152, 2 jul. 2015.

VANRADEN, P. M. Efficient Methods to Compute Genomic Predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

WANG, G. L. *et al.* RFLP mapping of genes conferring complete and partial resistance to blast in a durably resistant rice cultivar. **Genetics**, v. 136, n. 4, p. 1421–1434, 1 abr. 1994.

WANG, J.; ZHANG, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. **Genomics, Proteomics & Bioinformatics**, v. 19, n. 4, p. 629–640, 1 ago. 2021.

WANG, Q. *et al.* A SUPER Powerful Method for Genome Wide Association Study. **PLoS ONE**, v. 9, n. 9, p. e107684, 23 set. 2014.

WOJCIK, G. L. *et al.* Genetic analyses of diverse populations improves discovery for complex traits. **Nature**, v. 570, n. 7762, p. 514–518, 2019.

XIONG, X. *et al.* Genetic dissection of maize (Zea mays L.) chlorophyll content using multi-locus genome-wide association studies. **BMC Genomics**, v. 24, n. 1, p. 384, 2023.

YU, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. **Nature Genetics**, v. 38, n. 2, p. 203–208, 2006.

ZENG, A. *et al.* Genome-wide association study (GWAS) of salt tolerance in worldwide soybean germplasm lines. **Molecular Breeding**, v. 37, n. 3, p. 30, 2017.

ZHANG, Y. *et al.* Os MPH1 regulates plant height and improves grain yield in rice. **PLOS ONE**, v. 12, n. 7, p. e0180825, 14 jul. 2017.

ZHANG, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. **Nature Genetics**, v. 42, n. 4, p. 355–360, 2010.

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. 467, 2011.

ZHONG, H. *et al.* Uncovering the genetic mechanisms regulating panicle architecture in rice with GPWAS and GWAS. **BMC Genomics**, v. 22, n. 1, p. 86, 2021.