GABRIELA FRANÇA OLIVEIRA

TAMANHO POPULACIONAL NA DETECÇÃO DE QTL UTILIZANDO REGRESSÃO QUANTÍLICA EM ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA

Tese 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 *Doctor Scientiae*.

Orientador: Ana Carolina Campana Nascimento

Coorientadores: Camila Ferreira Azevedo
Moysés Nascimento

VIÇOSA - MINAS GERAIS
2023

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

T
O48t
2023

Oliveira, Gabriela França, 1991-
Tamanho populacional na detecção de QTL utilizando regressão quantílica em estudos de associação genômica ampla / Gabriela França Oliveira. – Viçosa, MG, 2023.
1 tese eletrônica (51 f.): il. (algumas color.).

Texto em português e inglês.
Orientador: Ana Carolina Campana Nascimento.
Tese (doutorado) - Universidade Federal de Viçosa, Departamento de Estatística, 2023.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2023.511
Modo de acesso: World Wide Web.

1. Análise de regressão. 2. Marcadores genéticos - Métodos estatísticos. 3. Melhoramento genético. I. Nascimento, Ana Carolina Campana, 1983-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.536

Bibliotecário(a) responsável: Euzébio Luiz Pinto CRB-6/3317

GABRIELA FRANÇA OLIVEIRA

TAMANHO POPULACIONAL NA DETECÇÃO DE QTL UTILIZANDO REGRESSÃO QUANTÍLICA EM ESTUDOS DE ASSOCIAÇÃO GENÔMICA AMPLA

Tese 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 *Doctor Scientiae*.

APROVADA: 15 de junho de 2023.

Assentimento:

Documento assinado digitalmente
gov.br
GABRIELA FRANCA OLIVEIRA
Data: 24/08/2023 15:06:49-0300
Verifique em https://validar.iti.gov.br

Gabriela França Oliveira
Autora

Documento assinado digitalmente
gov.br
ANA CAROLINA CAMPANA NASCIMENTO
Data: 24/08/2023 15:26:43-0300
Verifique em https://validar.iti.gov.br

Ana Carolina Campana Nascimento
Orientadora

Dedico essa tese à minha mãe, mulher guerreira que me ensinou a ser forte e batalhar para conquistar todos os meus objetivos. E à minha irmã, Fernanda, por estar sempre presente em todos os momentos e por ser meu porto seguro.

AGRADECIMENTOS

Agradeço imensamente a Deus por guiar meus passos e por sempre colocar pessoas especiais em meu caminho. Sou grata a Ele por todas as conquistas alcançadas.

À minha mãe, Rochele, por ser meu exemplo. Obrigada por estar ao meu lado em todas as minhas escolhas, por ser meu porto seguro, pelo carinho, pelo amor incondicional, pelas orações, amizade, confiança e por fazer meus sonhos os seus.

À minha melhor amiga e irmã, Fernanda, por estar sempre ao meu lado. Obrigada pelo incentivo, pelos valiosos conselhos e por cuidar tão bem de mim. À minha irmã, Nágela, por ter me dado um dos melhores presentes da minha vida, o Rafael e o João Henrique.

À minha família, em especial ao João, por ser uma presença constante em minha vida, agradeço de coração.

Ao Felipe, por todo companheirismo, amor, paciência, carinho, incentivo, compreensão e por tornar meus dias mais alegres.

À minha orientadora, Ana Carolina, gostaria de expressar minha profunda gratidão pela confiança que depositou em mim. Obrigada pelos conselhos, pelo incentivo, paciência, preocupação, dedicação constante e pelos conhecimentos valiosos que adquiri. Você foi muito mais do que uma orientadora, você se tornou uma grande amiga. Agradeço por me inspirar e ser referência de profissionalismo e excelência em tudo o que faz.

Aos meus coorientadores, Camila e Moysés, agradeço pela disponibilidade, ensinamentos, dedicação, amizade e conselhos. Sou grata por ter tido a oportunidade de aprender com vocês, pois vocês se tornaram uma inspiração para mim em termos de profissionalismo e competência.

Agradeço aos professores do Programa de Pós Graduação em Estatística Aplicada e Biometria pela minha formação acadêmica.

Aos funcionários do Departamento de Estatística, em especial ao Junior, pela amizade, paciência, conselhos e incentivo.

Aos membros da banca examinadora Camila Azevedo Ferreira, Laís Mayara Azevedo Barroso, Leísa Pires Lima e Moysés Nascimento pela disponibilidade e pelas sugestões que tanto contribuíram para o presente trabalho.

Ao Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE), por tornar o ambiente de estudos e trabalho mais prazeroso, produtivo e divertido ao longo

desses anos. Agradeço também pelo aprendizado adquirido e pelo meu desenvolvimento profissional e pessoal.

Aos meus colegas e amigos, em especial à Cynthia, Ithalo, Guilherme, Isabella, Jaquicele, Kátia, Leísa, Maurício, Pedro e Vivian, meu profundo agradecimento pelos momentos de descontração, companheirismo, conselhos e incentivo.

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

Enfim, a todos que colaboraram de alguma forma, alguns deles distantes, mas que conseguiram me confortar com um abraço ou com uma palavra de incentivo, para concretização deste trabalho.

“Feliz aquele que transfere o que sabe e aprende o que ensina”. (Cora Coralina)

RESUMO

OLIVEIRA, Gabriela França, D.Sc., Universidade Federal de Viçosa, junho de 2023. **Tamanho populacional na detecção de QTL utilizando regressão quantílica em estudos de associação genômica ampla**. Orientadora: Ana Carolina Campana Nascimento. Coorientadores: Camila Ferreira Azevedo e Moysés Nascimento.

Estudos de associação genômica (*Genome-Wide Association Studies* - GWAS) são aqueles que buscam identificar marcadores significativos que podem estar relacionadas às características de interesse nos programas de melhoramento. O Modelo Linear Geral (*General Linear Model* - GLM) é um dos principais procedimentos de avaliação de associações significativas entre marcadores e QTLs (*Quantitative Trait Locus*). A estimação dos efeitos dos marcadores por meio do GLM é baseada em médias condicionais. No entanto, esta estimação pode ser inadequada quando os erros não seguem distribuição normal e/ou não possuem variâncias homogêneas. Uma metodologia alternativa e que recentemente vem sendo explorada em estudos de associação genômica é a Regressão Quantílica (RQ), a qual possibilita a estimação do efeito dos marcadores ao longo de toda distribuição dos valores fenotípicos. A RQ já foi avaliada com sucesso em estudos de GWAS em um conjunto de dados reais que apresentava um número reduzido de indivíduos. Porém, a performance da técnica para diferentes tamanhos populacionais ainda não foi estudada. Diante do exposto, o objetivo deste estudo, foi avaliar a performance da RQ em estudos de GWAS quanto à capacidade de detectar QTLs associados as características fenotípicas de interesse, considerando diferentes tamanhos populacionais. Para isso, foram utilizados dados simulados, com características de diferentes níveis de herdabilidade ($h^2 = 0{,}30$ e $0{,}50$), controlados por 3 e 100 QTLs. Foi simulada uma população de 1000 indivíduos e posteriormente foram realizadas reduções aleatórias de 100 indivíduos até atingir uma população de tamanho 200. O poder de detecção de QTLs e a taxa de falsos positivos foram obtidos por meio do GLM e também por meio da RQ considerando três quantis diferentes ($\tau = 0{,}10;\ 0{,}50$ e $0{,}90$). Como resultado, observou-se que os modelos RQ, apresentaram maior poder de detecção de QTLs em todos os cenários avaliados e taxa de falsos positivos relativamente baixa em cenários com maior número de indivíduos. Os modelos de RQ nos quantis extremos ($\tau = 0{,}1$ e $0{,}90$) foram aqueles que obtiveram maior poder de detecção de QTLs verdadeiros. Em contrapartida, a análise baseada no GLM detectou poucos (cenários com maior tamanho populacional) ou nenhum QTL nos cenários avaliados. Nos cenários com baixa herdabilidade, o RQ obteve um alto poder de detecção. Dessa forma, verificou-se que a utilização da RQ em GWAS é eficaz, permitindo a detecção de QTLs

associados a características de interesse, mesmo em cenários com poucos indivíduos genotipados e fenotipados.

Palavras-chave: GWAS. Melhoramento genético. Modelo Linear Geral. Simulação. Quantis condicionais.

ABSTRACT

OLIVEIRA, Gabriela França, D.Sc., Universidade Federal de Viçosa, June, 2023. **Population size in QTL detection using quantile regression in genome-wide association studies.** Adviser: Ana Carolina Campana Nascimento. Co-advisers: Camila Ferreira Azevedo and Moysés Nascimento.

Genome-wide association studies (GWAS) are those that seek to identify significant markers that may be related to traits of interest in breeding programs. The General Linear Model (GLM) is one of the main procedures for evaluating significant associations between markers and QTLs (Quantitative Trait Locus). The estimation of the effects of the markers through the GLM is based on conditional means. However, this estimation may be inadequate when the errors do not follow a normal distribution and/or do not have homogeneous variances. An alternative methodology that has recently been explored in genomic association studies is Quantile Regression (QR), which makes it possible to estimate the effect of markers along the entire distribution of phenotypic values. QR has already been successfully evaluated in GWAS studies on a real dataset that had a reduced number of individuals. However, the performance of the technique for different population sizes has not yet been studied. Given the above, the objective of this study was to evaluate the performance of RQ in GWAS studies regarding the ability to detect QTLs associated with the phenotypic traits of interest, considering different population sizes. For this, simulated data was used, with traits of different levels of heritability ($h^2 = 0{,}30$ and $0{,}50$), and controlled by 3 and 100 QTLs. A population of 1000 individuals was simulated and then random reductions of 100 individuals were performed until reaching a population of size 200. The power of detection of QTLs and the false positive rate were obtained using the GLM and also using the QR considering three different quantiles ($\tau = 0.10, 0.50$ and 0.90). As a result, it was observed that the QR models showed greater QTL detection power in all evaluated scenarios and a relatively low false positive rate in scenarios with a greater number of individuals. The QR at the extreme quantiles ($\tau = 0.1$ and 0.90) were the models that obtained the greatest power to detect true QTLs. In contrast, the analysis based on the GLM detected few (scenarios with larger population size) or no QTL in the evaluated scenarios. In scenarios with low heritability, the QR obtained a high detection power. Thus, it was verified that using QR in GWAS is effective, allowing the detection of QTLs associated with characteristics of interest, even in scenarios with few genotyped and phenotyped individuals.

Keywords: GWAS. Genetic breeding. General linear model. Simulation. Conditional quantiles.

SUMÁRIO

INTRODUÇÃO GERAL

Os avanços da biologia molecular e das tecnologias de genotipagem de baixo custo baseadas em marcadores moleculares do tipo SNP (*Single Nucleotide Polymorphism*) e a utilização de métodos estatísticos adequados, têm permitido aos melhoristas identificar e localizar QTLs (*Quantitative Trait loci*) por meio dos estudos de associação genômica ampla (*Genome-Wide Association Studies* - GWAS) (RESENDE et al., 2018). A GWAS tem por objetivo identificar associações entre os marcadores moleculares e a característica fenotípica de interesse, devido ao desequilíbrio de ligação entre o marcador e o QTL.

Em estudos de GWAS, um método tradicional muito utilizado para estimação dos efeitos dos marcadores é o de regressão via marcas únicas. Neste método, estima-se o efeito individual de cada marcador sobre o fenótipo de interesse e, posteriormente, realizam-se múltiplos testes de hipóteses com intuito detectar quais os efeitos de marcadores são estatisticamente significativos (DE RESENDE et al., 2011).

A estimação do efeito de um marcador, pode ser realizada por meio de um modelo estatístico assumindo o marcador como efeito fixo. Entretanto, se os indivíduos em estudo possuem estrutura de população ou parentesco genético, a consideração apenas do efeito do marcador pode ser inadequada, pois a omissão desta informação pode ocasionar problemas de associações espúrias, isto é, associações falso positivas (AULCHENKO; DE KONING; HALEY, 2007; AZEVEDO et al., 2017; JABBARI et al., 2021; LI et al., 2014; MANCIN et al., 2021; PRITCHARD et al., 2000; RESENDE; SILVA; AZEVEDO, 2014; SUL; MARTIN; ESKIN, 2018). Uma alternativa para contornar este problema seria a utilização de métodos estatísticos que corrigem o modelo para estrutura de população. Um método que pode ser utilizado é o Modelo Linear Geral (*General Linear Model* - GLM). O GLM é um modelo de regressão via marcas únicas com a inserção da correção para estrutura de população via componentes principais (LIPKA et al., 2012; WANG; ZHANG, 2021).

Além da questão de considerar a estrutura de população, as associações falso positivas também podem estar relacionadas ao nível de significância adotado devido aos múltiplos testes aplicados, em virtude da alta densidade de marcadores (RESENDE; SILVA; AZEVEDO, 2014). Existem alguns métodos estatísticos que visam contornar este problema, dentre eles, pode-se destacar a correção de Bonferroni e a taxa de falsas descobertas (*False Discovery Rate* -FDR) (SILVA; VENCOVSKY, 2002; ZENG et al., 2015).

Nos métodos tradicionais de GWAS, a estimação dos parâmetros é baseada em médias condicionais. No entanto, a estimação obtida pode ser inadequada quando os erros não seguem

distribuição Normal e/ou não possuem variâncias homogêneas (GALARZA; LACHOS; BANDYOPADHYAY, 2017). Uma metodologia alternativa e ainda pouco explorada para estudos de GWAS é a Regressão Quantílica (RQ) (KOENKER; BASSETT, 1978). Essa metodologia, diferentemente dos métodos tradicionais baseados em médias, permite ajustar modelos de regressão para diferentes níveis (quantis) da distribuição do fenótipo de interesse, não requer pressuposições sobre a distribuição do erro e é robusta a pontos discrepantes (OLIVEIRA et al., 2021).

A regressão quantílica já foi aplicada com êxito em estudos de GWAS em dados reais por Nascimento et al (2018). Os autores utilizaram a RQ para identificar SNPs associados as características fenológicas (Dias para a primeira flor - DFF, Dias para floração – DTF e Dias até o final da floração - DEF) em feijoeiro comum. Neste estudo, foram avaliados 80 genótipos de feijão comum e 384 marcadores do tipo do tipo SNP. Como resultados, os autores observaram que o GLM não foi capaz de detectar nenhuma associação SNP-característica. Em contrapartida, ao utilizar a metodologia RQ considerando o quantil extremo ($\tau = 0,10$) foram encontrados, respectivamente, 1 e 7 SNPs significativos associados às características fenológicas DFF e DTF. Neste trabalho, a quantidade de genótipos disponíveis para avaliação foi relativamente pequena para estudos de GWAS, no entanto a metodologia de RQ conseguiu detectar associações significativas.

Diante do exposto, o objetivo deste trabalho foi avaliar a performance da metodologia de Regressão Quantílica em estudos de Associação Genômica por meio de dados simulados com diferentes tamanhos populacionais. Para tanto, este trabalho está dividido em dois Capítulos. O Capítulo 1 consiste em uma revisão de literatura sobre GWAS apresentando sua definição e importância para o melhoramento genético. Além disso, é apresentada uma abordagem tradicional em estudos de associação, o Modelo Linear Geral (GLM). Posteriormente, é discutida a metodologia alternativa, a Regressão Quantílica (RQ). E o Capítulo 2 é um artigo científico que tem por objetivo a avaliar performance da RQ em estudos de GWAS considerando um conjunto de dados simulados com diferentes tamanhos populacionais e arquiteturas genéticas, quanto à sua capacidade de detecção de QTLs associados a característica fenotípicas de interesse.

CAPÍTULO 1

REVISÃO DE LITERATURA

1. Associação genômica ampla

1.1 Definição e importância

Os estudos de Associação Genômica Ampla (*Genome-Wide Association Studies*-GWAS) é uma abordagem de crescente utilização em programas de melhoramento, pois utiliza informações diretamente do DNA através dos marcadores do tipo SNPs (*Single Nucleotide Polymorphism*) com intuito de identificar variações genéticas que possam estar associadas com características fenotípicas de interesse (HIRSCHHORN; DALY, 2005). A detecção dos marcadores significativos é realizada por meio de teste de hipóteses e, de posse desse marcadores, investiga-se se os mesmos podem estar relacionados ao controle de determinada característica e suas respectivas funções biológicas (MIQUELONI; RESENDE; DE ASSIS, 2019; PERS et al., 2015).

Inicialmente, a GWAS foi desenvolvida para aplicação em estudos epidemiológicos em humanos (MCCARTHY et al., 2008), entretanto, foi expandida para o melhoramento genético animal e vegetal. Esta metodologia tem sido aplicada com sucesso em diversas culturas, incluindo feijão (NASCIMENTO et al., 2018), milho (JAISWAL et al., 2019; KUKI et al., 2018; OLUKOLU et al., 2016), cevada (JABBARI et al., 2021; MWANDO et al., 2020), arroz (QUERO et al., 2018; SUELA et al., 2022), trigo (ARORA et al., 2019) e soja (MALLE et al., 2020; ZHANG et al., 2018, 2021), por exemplo.

Nos estudos de GWAS, a estrutura de população e o parentesco entre os indivíduos são fatores que devem ser corrigidos com intuito de evitar problemas de associações espúrias (AULCHENKO; DE KONING; HALEY, 2007; AZEVEDO et al., 2017; JABBARI et al., 2021; LI et al., 2014; MANCIN et al., 2021; PRITCHARD et al., 2000; RESENDE; SILVA; AZEVEDO, 2014; SUL; MARTIN; ESKIN, 2018; ZENG et al., 2015). A inserção da correção para estrutura de população pode ser empregada por meio de métodos estatísticos apropriados. Dentre os quais podem ser destacados a análise de componentes principais (DAETWYLER et al., 2012; NASCIMENTO et al., 2018), análise de covariância via autovetores (YANG et al., 2011) e modelos lineares mistos com inserção da correção para estrutura de população e efeitos poligênicos (MEYER; TIER, 2012; XIE et al., 2012; ZHANG et al., 2010).

Outro fator importante na análise de GWAS é adoção do nível de significância adequado para seleção correta dos marcadores SNPs, uma vez que serão realizados diversos testes de hipóteses devido à alta densidade de marcadores (RESENDE; SILVA; AZEVEDO, 2014). A adoção de um nível de significância inadequado pode também ocasionar problemas de associações falso positivas. Na literatura, existem métodos que permitem estabelecer limiares de seleção de marcadores com intuito de reduzir este problema. Dentre esses métodos pode-se destacar a correção de Bonferroni e a taxa de falsas descobertas (FDR) (SILVA; VENCOVSKY, 2002; ZENG et al., 2015).

1.2 Regressão via marcas únicas

As análises de regressão via marcas únicas, têm por objetivo estimar o efeito aditivo do i-ésimo marcador sobre o fenótipo de interesse. Estes modelos podem ser utilizados visando posteriormente avaliar existência de associação entre marcador (SNP) e o fenótipo, e podem ser definidos por (RESENDE; SILVA; AZEVEDO, 2014):

$$\boldsymbol{Y} = \boldsymbol{J}\mu + \boldsymbol{X_i}\beta_i + \boldsymbol{\varepsilon},$$

em que $\boldsymbol{Y}$ refere-se ao vetor de informações fenotípicas com dimensão $n \times 1$, sendo n é o número de indivíduos;

$\boldsymbol{J}$ é um vetor com valores iguais a 1;

μ refere-se a média geral;

$\boldsymbol{X_i}$ refere-se ao vetor de incidência do i-ésimo marcador;

β_i é o escalar referente ao efeito fixo do i-ésimo marcador;

$\boldsymbol{\varepsilon}$ refere-se ao vetor de erros aleatórios do modelo, com dimensão $(n \times 1)$, sendo que $\boldsymbol{\varepsilon} \sim N(0, I\sigma_\varepsilon^2)$ e σ_ε^2 a variância do erro.

A estimação do efeito do i-ésimo marcador é baseada em médias condicionais e pode ser realizada pelo método dos Mínimos Quadrados Ordinários (MQO). Este método consiste em estimar o vetor de parâmetros ($\boldsymbol{\theta} = [\mu, \beta_i]^t$) resolvendo o seguinte problema de otimização:

$$\hat{\theta} = argmin_{\hat{\theta}} \left[\sum_{i=1}^{n} \varepsilon_i^2 \right].$$

Após a estimação, a significância do efeito do marcador pode ser avaliada por meio de testes *t-student*.

1.3 Modelo linear geral

O Modelo Linear Geral (*General Linear Model* - GLM) é o modelo de regressão via marcas únicas com inserção da correção para estrutura de população (LIPKA et al., 2012; WANG; ZHANG, 2021). A correção para estrutura de população por meio de componentes principais (CP) é amplamente utilizada em GWAS (NASCIMENTO et al., 2018; PRICE et al., 2006), visando a redução da taxa de falsos positivos. Essa correção fundamenta-se na incorporação de covariáveis de efeitos fixos no modelo de regressão via marcas únicas. Essas covariáveis correspondem aos componentes principais derivados da matriz de parentesco genômico (G) (AZEVEDO et al., 2017). Este modelo pode ser utilizado para estimação dos efeitos dos marcadores, para posteriormente avaliar existência de associação entre marcador (SNP) e o fenótipo, e pode ser definido por:

$$Y = \boldsymbol{J}\mu + \boldsymbol{X_i}\beta_i + \sum_{k=1}^{K} \gamma_k \boldsymbol{CP_k} + \boldsymbol{\varepsilon}$$

em que $\boldsymbol{Y}$ refere-se ao vetor de informações fenotípicas com dimensão $n \times 1$, sendo n é o número de indivíduos;

$\boldsymbol{J}$ é um vetor com valores iguais a 1;

μ refere-se a média geral;

$\boldsymbol{X_i}$ refere-se ao vetor de incidência do i-ésimo marcador;

β_i é o escalar referente ao efeito fixo do i-ésimo marcador;

K: é o número de componentes principais;

γ_k é o efeito fixo do k-ésimo componente principal, ajustado como uma covariável;

$\boldsymbol{CP_k}$ é o vetor do k-ésimo componente principal da matriz de parentesco genômico (G);

$\boldsymbol{\varepsilon}$ refere-se ao vetor de erros aleatórios do modelo, com dimensão $(n \times 1)$, sendo que $\boldsymbol{\varepsilon} \sim N(0, I\sigma_\varepsilon^2)$ e σ_ε^2 a variância do erro.

Existem diversas expressões para a obtenção da matriz de parentesco genômico ($\boldsymbol{G}$), no entanto, a principal delas foi proposta por VanRaden (2008) e é dada por:

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\sum_{i=1}^{n} 2p_i q_i} \quad (1)$$

em que p_j e q_j são as frequências dos alelos A e a do i-ésimo marcador, respectivamente, e $\boldsymbol{W}$ é a matriz de incidência de marcadores codificada (centrada na média) como apresentado a seguir (VITEZICA et al., 2013):

$$\boldsymbol{W} = \begin{cases} Se\ AA, então\ X = 2 - 2p \rightarrow 2q \\ Se\ Aa, então\ X = 1 - 2p \rightarrow q - p \\ Se\ aa, então\ X = \ 0 - 2p \rightarrow -2p \end{cases}$$

sendo $\boldsymbol{W_i}$ a i-ésima coluna da matriz $\boldsymbol{W}$.

A estimação do efeito do i-ésimo marcador no GLM é, assim como nos modelos de marcas únicas, baseada em médias condicionais e pode ser realizada pelo método dos Mínimos Quadrados Ordinários (MQO). Este método consiste em estimar o vetor de parâmetros $\left(\boldsymbol{\theta} = \left[\mu,\ \beta_i, \gamma_{1,\cdots,}\gamma_k \right]^t \right)$ resolvendo o seguinte problema de otimização:

$$\hat{\theta} = argmin_{\hat{\theta}} \left[\sum_{i=1}^{n} \varepsilon_i^2 \right].$$

Após a estimação, a significância do efeito do marcador pode ser avaliada por meio de testes *t-student*.

1.4 Regressão quantílica

A Regressão Quantílica foi proposta na década de 70 por dois economistas Koenker e Bassett (KOENKER; BASSETT, 1978) como alternativa à regressão via Mínimos Quadrados Ordinários. Essa metodologia, permite ajustar modelos de regressão para diferentes níveis (quantis) da distribuição de probabilidade.

Nos modelos tradicionais baseados em médias, estima-se somente uma relação funcional entre as variáveis para explicar todo o conjunto de dados. Já nos modelos de RQ, pode-se estimar várias relações funcionais entre as variáveis ao longo de toda distribuição de probabilidade de acordo com o quantil de interesse, fornecendo assim uma visão mais completa

do estudo (CADE; NOON, 2003). Na Figura 1A, pode -se observar que o modelo tradicional fornece apenas a relação estimada em termos médios entre as variáveis estudadas. Já na Figura 1B é possível observar a relação funcional entre as variáveis estudadas em diferentes quantis da variável independente.

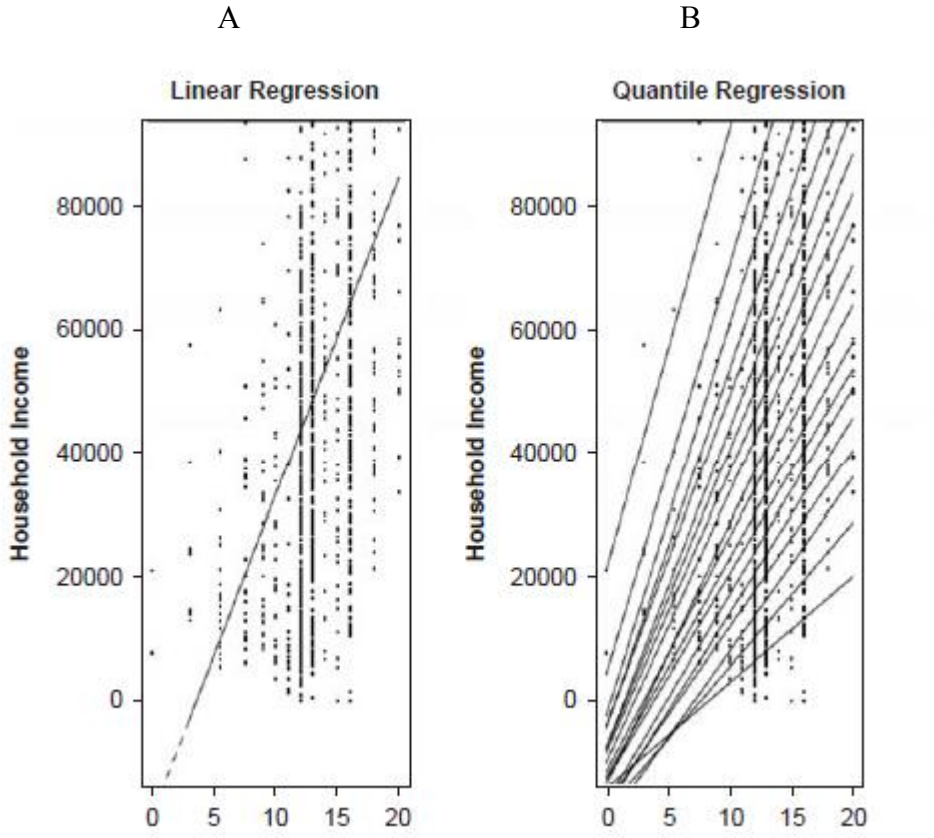

Figura 1: (A) Ajuste de um modelo tradicional (modelo linear). (B) Diferentes ajustes baseados em Regressão Quantílica.
Fonte: Hao Naiman (2007)

Geralmente, a definição dos quantis estudados depende da finalidade do estudo, isto significa que podemos estudar toda distribuição de probabilidade da variável de interesse ou somente alguns quantis específicos. Outras vantagens da RQ é que a metodologia não requer pressuposições sobre a distribuição do erro e é robusta a pontos discrepantes (OLIVEIRA et al., 2021).

A RQ tem-se mostrado uma metodologia bastante eficiente em diversas áreas da ciência. Especificamente no melhoramento genético animal e vegetal, pode-se citar estudos de Seleção e Associação Genômica, em que essa metodologia apresentou resultados melhores ou

iguais aos métodos tradicionais (BARROSO et al., 2017; DOS SANTOS et al., 2018; NASCIMENTO et al., 2017, 2018; OLIVEIRA et al., 2021).

O modelo de RQ que descreve a relação funcional entre o fenótipo e os marcadores do tipo SNP em diferentes níveis da distribuição de probabilidade, pode ser descrito conforme a equação:

$$\boldsymbol{Y} = \boldsymbol{J}\mu + \boldsymbol{X_i}\beta^{(\tau)} + \boldsymbol{\varepsilon}^{(\tau)},$$

em que $\boldsymbol{Y}$ refere-se ao vetor de informações fenotípicas com dimensão $n \times 1$, sendo n é o número de indivíduos;

$\boldsymbol{J}$ é um vetor com valores iguais a 1;

μ refere-se a média geral;

$\boldsymbol{X_i}$ refere-se ao vetor de incidência do i-ésimo marcador;

$\beta^{(\tau)}$é o escalar referente ao efeito fixo do i-ésimo marcador no quantil de ordem τ;

$\boldsymbol{\varepsilon}^{(\boldsymbol{\tau})}_{\boldsymbol{n\times 1}}$ refere-se ao vetor de efeitos aleatórios independentes e identicamente distribuídos com o quantil de ordem τ igual a zero.

Já o modelo de RQ que descreve a relação funcional entre o fenótipo e os marcadores do tipo SNP em diferentes níveis da distribuição de probabilidade do fenótipo, com a inserção da correção para estrutura de população via componentes principais, pode ser descrito conforme a equação (NASCIMENTO et al., 2018):

$$\boldsymbol{Y} = \boldsymbol{J}\mu + \boldsymbol{X_i}\beta_i^{(\tau)} + \sum_{k=1}^{K} \gamma_k^{(\tau)} \boldsymbol{CP_k} + \boldsymbol{\varepsilon}^{(\tau)},$$

em que $\boldsymbol{Y}$ refere-se ao vetor de informações fenotípicas com dimensão $n \times 1$, sendo n é o número de indivíduos;

$\boldsymbol{J}$ é um vetor com valores iguais a 1;

μ refere-se a média geral;

$\boldsymbol{X_i}$ refere-se ao vetor de incidência do i-ésimo marcador;

$\beta_i^{(\tau)}$ é o escalar referente ao efeito fixo do i-ésimo marcador no quantil de ordem τ;

K é o número de componentes principais;

$\gamma_k^{(\tau)}$ é o efeito fixo do k-ésimo componente principal no quantil de ordem τ, ajustado como uma covariável;

$\boldsymbol{CP_k}$ é o vetor do k-ésimo componente principal da matriz de parentesco genômico (**G**). A matriz **G** é calculada conforme a com a equação 1;

$\boldsymbol{\varepsilon_{n\times 1}^{(\tau)}}$ refere-se ao vetor de efeitos aleatórios independentes e identicamente distribuídos com o quantil de ordem τ igual a zero.

A estimação dos parâmetros na RQ é baseada no método da minimização dos erros absolutos ponderados. Essa metodologia consiste em estimar o vetor de parâmetros $\left(\hat{\theta}_\tau = \left[\mu,\ \beta_i, \gamma_{1,\cdots,}\gamma_k\right]^t\right)$ no quantil τ, resolvendo o seguinte problema de otimização:

$$\hat{\theta}_\tau = argmin_{\hat{\theta}_\tau}\left[\sum_{i=1}^{N} \rho_\tau|\varepsilon_i|\right],$$

em que, ρ_τ é chamada função check (KOENKER, 2005; NASCIMENTO et al., 2018; OLIVEIRA et al., 2021) e é definida como:

$$\rho_\tau(\varepsilon) = \begin{cases} \tau\varepsilon, & se\ \varepsilon_i \geq 0, \\ (1-\tau)\varepsilon, & se\ \varepsilon_i < 0 \end{cases}.$$

A minimização desta função exige a utilização de algoritmos de programação linear como o *Método Simplex*, que é apresentado em detalhes por Koenker (2005). Após a minimização da função em algum quantil de interesse, obtém-se as estimativas dos coeficientes da regressão naquele quantil. A significância do efeito do marcador em cada quantil de interesse pode ser avaliada por meio de testes *t-student*.

1.5 Correção para múltiplos testes

Após a estimação dos efeitos dos marcadores, para analisar a existência de associações significativas entre marcador e o fenótipo de interesse, são realizados múltiplos testes *t-student*. De forma que, a hipótese de nulidade (H_0) a ser testada é definida como "o i-ésimo marcador não apresenta efeito sobre o fenótipo" conta a hipótese alternativa (H_a) que é definida como sendo "o i-ésimo marcador tem efeito sobre o fenótipo"(SUL; MARTIN; ESKIN, 2018).

No entanto, devido à alta densidade de marcadores, são realizados múltiplos testes de hipóteses. A realização de múltiplos testes pode acarretar problemas de associações falso positivas, ou seja, declarar que o efeito do marcador é significativo, mas na verdade, o marcador não está em desequilíbrio de ligação com QTLs que controlam a característica de interesse (SUELA et al., 2022). Este problema pode estar relacionado ao nível de significância adotado (SILVA; VENCOVSKY, 2002). Assim, determinar um nível de significância adequado é essencial para discriminar os verdadeiros positivos dos falsos positivos. Na literatura existem diversos procedimentos estatísticos para adoção do limiar mais adequado, sendo a correção de Bonferroni e a taxa de falsas descobertas (FDR) as mais comumente utilizadas (KALER; PURCELL, 2019).

A correção de Bonferroni é um dos mecanismos mais simples para se aplicar em múltiplas comparações (NOBLE, 2009). Esta correção é obtida ajustando o nível de significância individual (definido pelo pesquisador) de acordo com o número de testes (m) realizados. Dessa forma, o nível de significância ajustado é obtido pela razão entre o nível de significância individual e o número de testes realizados $\left(\alpha' = \frac{\alpha}{m}\right)$ (ZENG et al., 2015). Um marcador é considerado significativo se seu *p-value* for menor que o nível de significância ajustado (α'). O problema em utilizar a correção de Bonferroni é a suposição de independência entre os testes, o que não ocorre nas análises de associação genômica, uma vez que os SNPs podem estar em desequilíbrio de ligação. Este método também é altamente conservador, isto é, existe uma grande dificuldade em rejeitar a hipótese de nulidade (KALER; PURCELL, 2019; SILVA; VENCOVSKY, 2002; ZENG et al., 2015).

A correção por FDR proposta Benjamini e Hochberg (1995) é baseada na proporção do número de falsos positivos em relação ao número total de resultados positivos. Uma forma de inserir FDR no teste de significância é por meio de uma correção do *p-value* associado ao teste, obtendo o que é denominado *q-value* (STOREY; TIBSHIRANI, 2003). O cálculo do *q-value* se dá da seguinte forma: primeiramente deve-se ordenar os *p-values* em ordem crescente (P_i) e, posteriormente realizar o cálculo do *q-value* por meio da seguinte expressão:

$$q_{(i)} = \frac{P_{(i)} \times n}{i},$$

em que $P_{(i)}$ é o *p-value* na posição i, n é o número de marcadores e i é a posição que o *p-value* se encontra. Após o cálculo de $q_{(i)}$ deve-se realizar a correção desses valores que será denominado q_i^c, seguindo a seguinte restrição:

i) $q_{(i)} < q_{(i+1)}$, então $q_{(i)}^c < q_{(i)}$

ii) $q_{(i)} > q_{(i+1)}$, então $q_{(i)}^c < q_{(i+1)}$

Neste caso, de acordo com Benjamini e Hochberg (1995) um marcador será considerado significativo se seu *q-value* corrigido ($q_{(i)}^c$) for menor que o nível de significância (α) adotado pelo pesquisador, ou seja $q_{(i)}^c < \alpha$.

REFERÊNCIAS BIBIOGRÁFICAS

ARORA, S. et al. Genome-wide association mapping of grain micronutrients concentration in aegilops tauschii. **Frontiers in Plant Science**, v. 10, n. February, p. 1–14, 2019.

AULCHENKO, Y. S.; DE KONING, D. J.; HALEY, C. Genomewide rapid association using mixed model and regression: A fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. **Genetics**, v. 177, n. 1, p. 577–585, 2007.

AZEVEDO, C. F. et al. Population structure correction for genomic selection through eigenvector covariates. **Crop Breeding and Applied Biotechnology**, v. 17, n. 4, p. 350–358, 2017.

BARROSO, L. M. A. et al. Regularized quantile regression for SNP marker estimation of pig growth curves. **Journal of Animal Science and Biotechnology**, v. 8, p. 1–9, 2017.

BENJAMINI, Y.; HOCHBERG, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. **Journal of the Royal statistical society: series B (Methodological)**, v. 57, n. 1, p. 289–300, 1995.

CADE, B. S.; NOON, B. R. A gentle introduction to quantile regression for ecologists.

Frontiers in Ecology and the Environment, v. 1, n. 8, p. 412–420, 2003.

DAETWYLER, H. D. et al. Components of the accuracy of genomic prediction in a multi-breed sheep population. **Journal of Animal Science**, v. 90, p. 3375–3384, 2012.

DE RESENDE, M. D. V. et al. Métodos estatísticos na seleção genômica ampla. 2011.

DOS SANTOS, P. M. et al. Use of regularized quantile regression to predict the genetic merit of pigs for asymmetric carcass traits. **Pesquisa Agropecuaria Brasileira**, v. 53, n. 9, p. 1011–1017, 2018.

GALARZA, C. E.; LACHOS, V. H.; BANDYOPADHYAY, D. Quantile regression in linear mixed models: a stochastic approximation EM approach. **Statistics and its Interface**, v. 10, n. 3, p. 471, 2017.

HIRSCHHORN, J. N.; DALY, M. J. Genome-wide association studies for common diseases and complex traits. **Nature Reviews Genetics**, v. 6, p. 95–108, 2005.

HAO, L.; NAIMAN, D. Q. **Quantile regression**. Sage, 2007.

JABBARI, M. et al. Association analysis of physiological traits in spring barley (Hordeum vulgare L.) under water-deficit conditions. **Food Science and Nutrition**, v. 9, p. 1761–1779, 2021.

JAISWAL, V. et al. Genome-wide association study (GWAS) delineates genomic loci for ten nutritional elements in foxtail millet (Setaria italica L.). **Journal of Cereal Science**, v. 85, n. November 2018, p. 48–55, 2019.

KALER, A. S.; PURCELL, L. C. Estimation of a significance threshold for genome-wide association studies. **BMC Genomics**, v. 20, n. 1, p. 1–8, 2019.

KOENKER, R. Quantile Regression. Cambridge ed. New York, 2005.

KOENKER, R.; BASSETT, G. Regression Quantiles. **Econometrica**, v. 46, n. 1, p. 33–50,

1978.

KUKI, M. C. et al. Genome wide association study for gray leaf spot resistance in tropical maize core. **PLoS ONE**, v. 13, n. 6, p. 1–13, 2018.

LI, M. et al. Enrichment of statistical power for genome-wide association studies. **BMC biology**, v. 12, p. 1–10, 2014.

LIPKA, A. E. et al. GAPIT: Genome association and prediction integrated tool. **Bioinformatics**, v. 28, n. 18, p. 2397–2399, 2012.

MALLE, S. et al. Genome-wide association identifies several QTLs controlling cysteine and methionine content in soybean seed including some promising candidate genes. **Scientific Reports**, v. 10, n. 1, p. 1–14, 2020.

MANCIN, E. et al. Accounting for population structure and phenotypes from relatives in association mapping for farm animals: A simulation study. **Frontiers in Genetics**, v. 12, p. 642065, 2021.

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

MEYER, K.; TIER, B. "SNP Snappy": A strategy for fast genome-wide association studies fitting a full mixed model. **Genetics**, v. 190, p. 275–277, 2012.

MIQUELONI, D. P.; RESENDE, R. M. S.; DE ASSIS, G. M. L. Seleção genômica ampla (GWS) e associação genômica ampla (GWAS) no melhoramento de forrageiras: abordagem conceitual, genética quantitativa e aplicações. 2019.

MWANDO, E. et al. Genome-Wide Association Study of Salinity Tolerance During Germination in Barley (Hordeum vulgare L.). **Frontiers in Plant Science**, v. 11, p. 1–15, 2020.

NASCIMENTO, M. et al. Regularized quantile regression applied to genome-enabled prediction of quantitative traits. **Genetics and Molecular Research**, v. 16, n. 1, p. 1–12, 2017.

NASCIMENTO, M. et al. Quantile regression for genome-wide association study of flowering time-related traits in common bean. **PLoS ONE**, v. 13, n. 1, p. 1–14, 2018.

NOBLE, W. S. How does multiple testing correction work? **Nature Biotechnology**, v. 27, n. 12, p. 1135–1137, 2009.

OLIVEIRA, G. F. et al. Quantile regression in genomic selection for oligogenic traits in autogamous plants: A simulation study. **PLoS ONE**, v. 16, n. 1, p. 1–12, 2021.

OLUKOLU, B. A. et al. A genome-wide association study for partial resistance to maize common rust. **Phytopathology**, v. 106, n. 7, p. 745–751, 2016.

PERS, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. **Nature Communications**, v. 6, p. 1–9, 2015.

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

PRITCHARD, J. K. et al. Association mapping in structured populations. **American Journal of Human Genetics**, v. 67, n. 1, p. 170–181, 2000.

QUERO, G. et al. Genome-Wide Association Study Using Historical Breeding Populations Discovers Genomic Regions Involved in High-Quality Rice. **The Plant Genome**, v. 11, n. 3, p. 1–12, 2018.

RESENDE, M. D. V. DE; SILVA, F. F.; 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**. Suprema ed. Viçosa, 2014.

RESENDE, R. T. et al. Genome-wide association and regional heritability mapping of plant architecture, lodging and productivity in Phaseolus vulgaris. **G3: Genes, Genomes, Genetics**, v. 8, n. 8, p. 2841–2854, 2018.

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.

SILVA, H. D.; VENCOVSKY, R. Poder de detecção de "quantitative trait loci", da análise de marcas simples e da regressão linear múltipla. **Scientia Agricola**, v. 59, n. 4, p. 755–762, 2002.

STOREY, J. D.; TIBSHIRANI, R. Statistical significance for genomewide studies. **PNAS**, v. 100, n. 16, p. 9440–9445, 2003.

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

SUL, J. H.; MARTIN, L. S.; ESKIN, E. Population structure in genetic studies: Confounding factors and mixed models**. PLoS Genetics**, 2018.

VANRADEN, P. M. Efficient methods to compute genomic predictions. **Journal of Dairy Science**, v. 91, n. 11, p. 4414–4423, 2008.

VITEZICA, Z. G.; VARONA, L.; LEGARRA, A. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics**, v. 195, n. 4, p. 1223–1230, 2013.

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

XIE, L. et al. Genome-wide association study identified a narrow chromosome 1 region associated with chicken growth traits. **PLoS ONE**, v. 7, n. 2, p. 1–9, 2012.

YANG, J. et al. GCTA: A tool for genome-wide complex trait analysis. **American Journal of Human Genetics**, v. 88, n. 1, p. 76–82, 2011.

ZENG, P. et al. Statistical analysis for genome-wide association study. **Journal of Biomedical Research**, v. 29, n. 4, p. 285–297, 2015.

ZHANG, J. et al. Genome-wide Scan for Seed Composition Provides Insights into Soybean Quality Improvement and the Impacts of Domestication and Breeding. **Molecular Plant**, v. 11, p. 460–472, 2018.

ZHANG, W. et al. Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean. **Theoretical and Applied Genetics**, 2021.

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

CAPÍTULO 2

Published as original paper in Scientific Reports

POPULATION SIZE IN QTL DETECTION USING QUANTILE REGRESSION IN GENOME-WIDE ASSOCIATION STUDIES

Gabriela França Oliveira[1,#a*], Ana Carolina Campana Nascimento[1,#a], Camila Ferreira Azevedo[1,#a] , Maurício de Oliveira Celeri[1,#a], Laís Mayara Azevedo Barroso[2], Isabela de Castro Sant'anna[3], José Marcelo Soriano Viana[4], Marcos Deon Vilela de Resende[5], Moysés Nascimento[1,#a]

[1]Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[2]Federal Institute of Education, Science and Technology of Mato Grosso, Sorriso, Mato Grosso, Brazil

[3]Rubber Tree and Agroforestry Systems Research Center, Campinas Agronomy Institute (IAC), Votuporanga, São Paulo, Brazil

[4]Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[5]Brazilian Agricultural Research Corporation, Embrapa Coffee, Brasília, DF- Brazil.

[#a] *Current address: Department of Statistics, Federal University of Viçosa, Av. Peter Henry Rolfs, s/n, Campus Universitário, 36570.900, Viçosa, Minas Gerais, Brazil*

*Corresponding author

E-mail: gabriela.franca@ufv.br (GFO)

ABSTRACT

The aim of this study was to evaluate the performance of Quantile Regression (QR) in Genome-Wide Association Studies (GWAS) regarding the ability to detect QTLs (Quantitative Trait Locus) associated with phenotypic traits of interest, considering different population sizes. For this, simulated data was used, with traits of different levels of heritability (0.30 and 0.50), and controlled by 3 and 100 QTLs. Populations of 1000 to 200 individuals were defined, with a random reduction of 100 individuals for each population. The power of detection of QTLs and the false positive rate were obtained by means of QR considering three different quantiles (0.10, 0.50 and 0.90) and also by means of the General Linear Model (GLM). In general, it was observed that the QR models showed greater power of detection of QTLs in all scenarios evaluated and a relatively low false positive rate in scenarios with a greater number of individuals. The models with the highest detection power of true QTLs at the extreme quantils (0.10 and 0.90) were the ones with the highest detection power of true QTLs. In contrast, the analysis based on the GLM detected few (scenarios with larger population size) or no QTLs in the evaluated scenarios. In the scenarios with low heritability, QR obtained a high detection power. Thus, it was verified that the use of QR in GWAS is effective, allowing the detection of QTLs associated with traits of interest even in scenarios with few genotyped and phenotyped individuals.

KEY WORDS: GWAS; Genetic breeding; General linear model; simulation; conditional quantiles.

INTRODUCTION

The world's population reached 7.7 billion inhabitants in 2019 and may reach 9.7 billion by 2050[1]. To the increase in population is added the growing concern about environmental impacts and the limitations of arable areas, which culminates in the demand for increased productivity of agronomic species [2]. In recent years, it is estimated that about 50% of the increase in productivity of several species was driven by genetic breeding, which has been seeking new strategies to obtain more adapted, resistant, and productive cultivars [3,4].

In this context, Genome-Wide Association Studies (GWAS) have been conducted in order to identify genetic variations that may be associated with phenotypic traits of interest [5–9]. The potentials of GWAS have already been successfully explored in traits of economic interest

and in different crops, such as barley [10,11], maize [12–14], soybean [15,16], rice [17–20], wheat [21–23] e arabica coffea [24–26].

In GWAS, a classic and widely used statistical method is single markers regression. This method estimates the individual effect of each marker on the phenotype of interest, and, subsequently, multiple hypothesis tests are performed in order to detect which marker effects are statistically significant [27]. When the correction for population structure is added to the single markers regression model , this model is called General Linear Model (GLM)[28].However, the estimation of parameters via single markers and GLM are based on conditional means, which may be inadequate when the errors do not follow a normal distribution [29] and in the presence of heteroscedasticity. An alternative and still little explored methodology for GWAS studies is Quantile Regression (QR)[30]. This methodology, unlike methods based on means, allows adjusting regression models for different levels (quantiles) of the distribution of the phenotype of interest, does not require assumptions about the error distribution, and is robust to discrepant points [31]. QR has already been successfully applied in GWAS studies on real data by[32] for traits related to the flowering time of common beans. These authors evaluated 80 common bean genotypes and 384 SNP markers (*single nucleotide polymorphism*) in order to identify genomic regions for three phenological traits. As a result, the authors found no significant associations using the general linear model. In contrast, when using QR at the extreme quantile $(\tau = 0{,}10)$, it was possible to detect 7 significant associations between SNPs and the phenological traits studied. In this study, the number of available genotypes was relatively small for GWAS studies, but it was still possible to detect significant associations using QR in this setting.

Although QR has already been applied to real data sets and has obtained interesting and promising results, the effect of population size on the ability to detect QTLs *(Quantitative Trait Locus)* has not yet been evaluated. To this end, it is possible to use data simulation since this strategy aims to reproduce the conditions of a biological system, facilitating the understanding of its real functioning and allowing prediction of the performance and recommendations before starting field studies [33,34]. In addition, simulation studies are especially convenient for testing and comparing methodologies because they demand fewer resources, time, human efforts, and the possibility of replication, thus generating greater efficiency in inferences [34,35].

In view of the above, this study evaluated the use of QR in GWAS regarding the power of QTL detection through SNP markers for simulated data with different levels of heritabilities, trait loci, and population sizes. The results of QR were compared with those obtained by GLM.

MATERIAL AND METHODS

Aiming to access the power of QTL detection and false positives rates in a genome-wide association study was performed a simulation study.

Genome and simulated populations

An advanced generation composite was obtained from two random mating populations in linkage equilibrium, which were crossed to generate a population of 5000 elements from 100 families using linkage disequilibrium (LD), subjected to five generations of random mating without mutation, selection, or migration.

From the advanced generation of the composite, 1000 individuals from the same generation and from 20 families of full siblings, each consisting of 50 individuals, were simulated. The simulated genome was composed of ten chromosomes with a size of 200 centimorgans (cM) each and comprised 2000 bi-allelic single nucleotide polymorphisms (SNPs) separated by 0.1 cM across the ten chromosomes. The LD value in a composite population is $\Delta_{ab} = \left(\frac{1\text{-}2\theta_{ab}}{4}\right)(p_a^1\text{-}p_a^2)(p_b^1\text{-}p_b^2)$, where a and b are two SNPs, two QTLs, or one SNP and one QTL, θ is the frequency of recombinant gametes, and p^1 and p^2 are the allele frequencies in the parental populations (1 and 2). The LD value depends on the allele frequencies in the parental populations. Thus, regardless of the distance between the SNPs and/or QTLs, if the allele frequencies are equal in the parental population, $\Delta = 0$. The LD is maximized ($|\Delta| = 0.25$) when $\theta = 0$ and $|p^1\text{-}p^2| = 1$. In this case, the LD value is positive with coupling and negative with repulsion[36] .

Simulation of traits and the phenotypic values

Two genetic architectures were simulated, representing different scenarios, with heritabilities of 0.30 and 0.50 and with 100 and 3 numbers of quantitative trait loci (QTLs), distributed randomly in the regions covered by the SNPs. The first scenario follows the infinitesimal model and the other (second scenario) with three major effects genes accounting for 50% of the genetic variability. For the former, to each of 100 QTLs one additive effect of small magnitude on the phenotype was assigned (under the Normal Distribution setting). For the latter, small additive effects were assigned to the remaining 97 loci. The effects were

normally distributed with zero mean and variance, allowing the desired heritability level. The phenotypic value was obtained by adding to the genotypic value a random deviate from a normal distribution $N(0, \sigma_e^2)$, where the variance σ_e^2 was defined according to two levels of broad-sense heritability, 0.30 and 0.50.

The data set was simulated using the Real Breeding program [37]. More information can be found detailed in[38].

Subsequently, in order to evaluate the effect of population size reduction, populations were defined with numbers of individuals ranging from 1000 to 200 individuals. According to[39], 200 individuals are considered as being sufficient for the construction of reasonably accurate genetic maps. A random reduction of 100 individuals was defined in each scenario, respecting the proportionality of individuals removed from each family. Thus, in all, thirty-six distinct scenarios were evaluated. These scenarios correspond to the combination of two levels of heritability, two genetic architectures, and nine variations in population size.

Linkage disequilibrium

A linkage disequilibrium (LD) analysis was performed to determine the markers associated with QTLs. Specifically, the LD decay pattern between marker pairs across the genome was obtained using a figure in which the square values of the correlation coefficient r^2 were plotted against the genetic distance between markers (in cM). Subsequently, a local polynomial regression (LOESS)[40–42] was fitted to the data and a horizontal straight line was plotted with a critical value of $r^2 = 0.20$ [43,44]. The window distance, defined as the intersection of the fitted LOESS curve and the horizontal straight line, will be used to determine which markers are associated with QTLs. Thus, all markers that distance the value of the window obtained (depending on the scenario evaluated) in relation to each QTL are considered as markers associated with the QTLs. The square of the correlation coefficient (r^2) was estimated using the *LD.decay* function of the *sommer* package [45] and the fit of the polynomial regression model using the *loess* function, both from the R software[46].

Genome-wide association study

To perform the genome-wide association analysis, first, the correction for population structure was performed through principal component analysis (PCA) of the genomic relatedness matrix (**G**) [20,47,48]. The number of principal components adopted was obtained using

STRUCTURE 2.3.4 software [49], selecting 300 markers in linkage equilibrium, aiming to ensure that these markers are not associated. A cluster number (K) ranging from 1 to 21 was tested, with ten independent replicates for each K value. In order to identify the optimal number of K, 10000 iterations were run, with 1000 burn-in. Then, the ΔK index [50] implemented in Structure Harvester software [51] was calculated to determine the choice of the most likely value of K. Subsequently, the K first principal components (CP) were used as fixed effect covariates in the GWAS model.

The GWAS model was defined by:

$$Y = \mu + \alpha_j SNP_j + \sum_{k=1}^{K} \beta_k CP_k + \varepsilon$$

where Y is the vector of phenotypic information; μ is the population mean; α_j is the effect of the j-th marker considered as fixed, $j = 1, \ldots, 2000$; SNP_j is the incidence vector of the j-th SNP marker; β_k is the fixed effect of the k-th principal component, adjusted as a covariate; CP_k is the vector of the k-th principal component; ε is the vector of random errors. The vector $\theta = [\mu, \alpha_j, \beta_{1,\ldots,}\beta_{k,}]'$ represents the unknown parameters, being estimated by means of QR and the GLM.

The methods estimate the individual effect of each marker on the phenotype of interest and then perform multiple hypothesis tests in order to detect which marker effects are statistically significant. The parameters were estimated via QR for different levels (quantiles) of the distribution of the phenotype of interest [30,32]. This methodology consists of estimating the parameters at the τ quantile by solving the following optimization problem:

$$\hat{\theta}_\tau = \text{argmin}_{\hat{\theta}_\tau} \left[\sum_{i=1}^{N} \rho_\tau |\varepsilon_i| \right],$$

where $\tau \in (0,1)$ indicating the quantile of interest, N indicates the population size evaluated, and $\rho_\tau(\cdot)$, denoted *check* function by[30], is defined by:

$$\rho_\tau(\epsilon_i) = \begin{cases} \tau\varepsilon_i & \text{if } \varepsilon_i \geq 0, \\ (1\text{-}\tau)\varepsilon_i, & \text{if } \varepsilon_i < 0 \end{cases}$$

In this study, three quantiles ($\tau = 0.10, 0.50$ and 0.90) were evaluated. For model fitting, the *rq* function from the *quantreg* package [52] of the R software was used. The individual coefficients (effects) of each marker are estimated by summing the weighted absolute errors. For estimation, it is necessary to use linear programming algorithms. One of the methods used is the Simplex Method [53].

The parameters were also estimated using GLM. This methodology consists of estimating the parameters in average terms and solving the following optimization problem:

$$\hat{\theta} = \text{argmin}_{\theta} \left[\sum_{i=1}^{N} \varepsilon_i^2 \right].$$

For model fitting, the individual coefficients (effects) of each marker were estimated by minimizing the sum of squared errors by the ordinary least squares method using the GAPIT R package[54] of the R software[46].

Hypothesis testing

After estimating the effects of individual markers through QR and GLM, multiple *t-student* tests were performed according to the methodology used, in order to analyze the existence of significant associations between the marker and the phenotype of interest. In the general linear model, the standard error estimate used was the usual, while in the quantile regression it was based on *rank* [53,55,56]. However, due to the high density of markers, performing multiple tests can lead to an increase in false positive associations [27]. An alternative to controlling this rate is the *False Discovery Rate* (FDR) [57,58]. One way to consider the FDR in hypothesis testing is through a correction in the p-value associated with the test, called the q-value [59]. In this study, a significance level of 0.01 ($\alpha = 1\%$) corrected by the FDR was used.

Comparison between methodologies

In order to evaluate the efficiency of the analyzed methodologies, the QTL detection power and the false positive rate were calculated and defined below:

i) The power of QTL detection corresponds to the proportion of pre-established windows (intervals) (by means of LD analysis) that contain at least one marker considered significant by means of the statistical methods evaluated.

ii) The false positive rate corresponds to the ratio between the number of markers that were significant by the evaluated statistical methods and are not associated with QTLs and the number of markers that are not associated with QTLs.

RESULTS AND DISCUSSION

Population structure

According to the method of[50], ΔK was plotted against the number of clusters (k). The maximum value of ΔK occurred at K = 19 and K = 18 for the scenarios of 3 QTLs and 100 QTLs, respectively (Fig. 1). Thus, 19 and 18 principal components were used as covariates in the GWAS analyses. According to the principal component analysis, 19 and 18 PCs accounted for explanation percentages of the variance present in the genotypic data between 85% and 96%, depending on the scenario evaluated. This result is in agreement with the simulated data of this study, where populations were simulated from 20 full sib families.

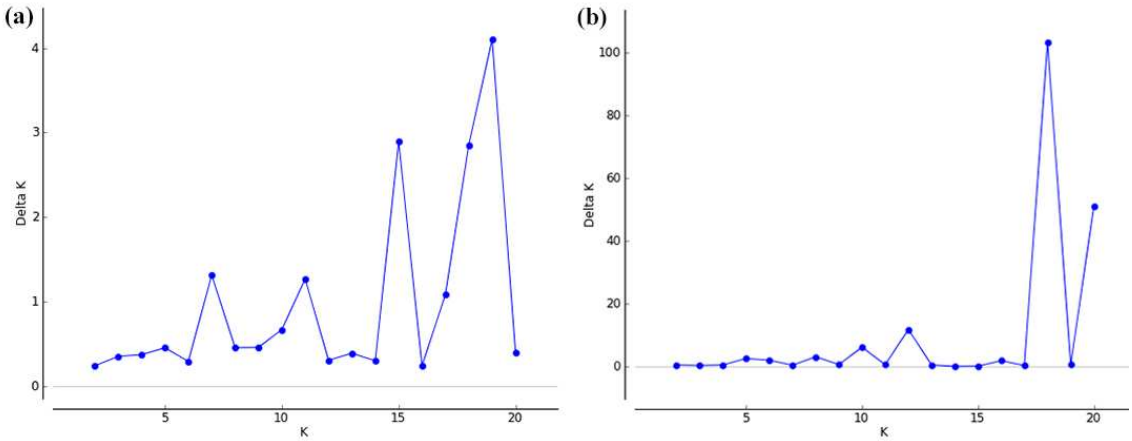

Fig.1 Graph ΔK versus number of clusters K. (A) Scenario with 3 QTLs. (B) Scenario with 100 QTLs

Linkage disequilibrium

The LD was calculated for all marker pairs in the same linkage group by means of r^2. Figures 2 and 3 graphically represent the decay of LD as a function of genetic distance

according to the number of QTLs evaluated. The critical value of $r^2 = 0.20$ was adopted, which according to [43], it is expected that values of $r^2 < 0.20$, the LD is corrupted, that is, there is a tendency of linkage equilibrium between the markers. The intersection of the LOESS curve with the horizontal straight line $(r^2 = 0.20)$ for the scenarios (different population sizes) of 3 QTLs, with a reduction in the number of individuals from 1000 to 200, was 0.924 cM, 0.994 cM, 1.085 cM, 1.161 cM, 1.302 cM, 1.444 cM, 1.617 cM, 1.830 cM and 2.158 cM, respectively (Fig. 2).

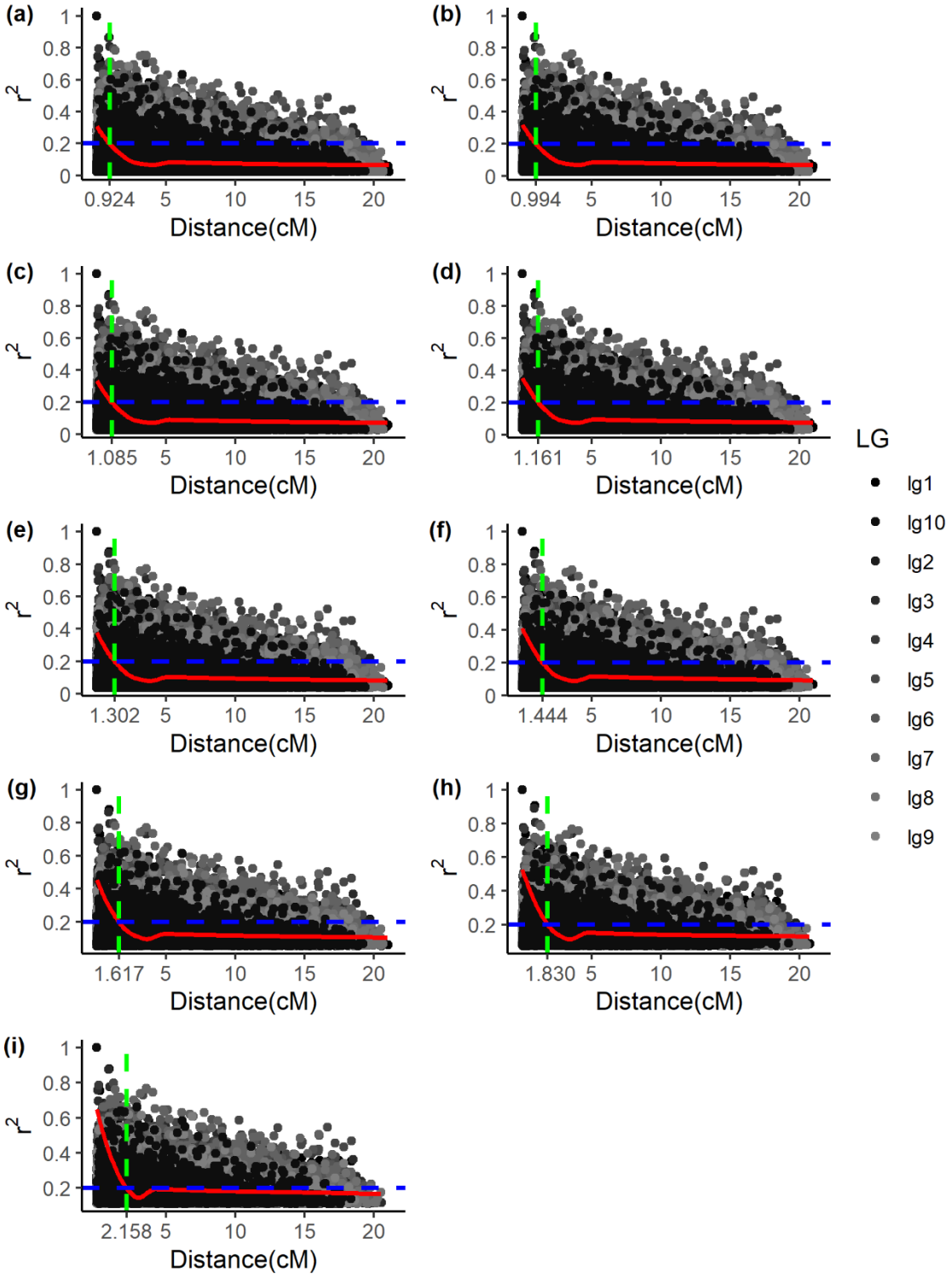

Fig. 2 Decay of linkage disequilibrium (r^2) as a function of genetic distance in the 10 linkage groups in the scenario with 3 QTLs. (a) Scenario: 1000 individuals (b) Scenario: 900 individuals (c) Scenario: 800 individuals (d) Scenario: 700 individuals (e) Scenario: 600 individuals (f) Scenario: 500 individuals (g) Scenario: 400 individuals (h) Scenario: 300 individuals (i) Scenario: 200 individuals

As for the scenario with 100 QTLs, the intersections obtained were: 0.943 cM, 1.019 cM, 1.101 cM, 1.196 cM, 1.312 cM, 1.452 cM, 1.620 cM, 1.820 cM, and 2.150 cM (Fig 3).

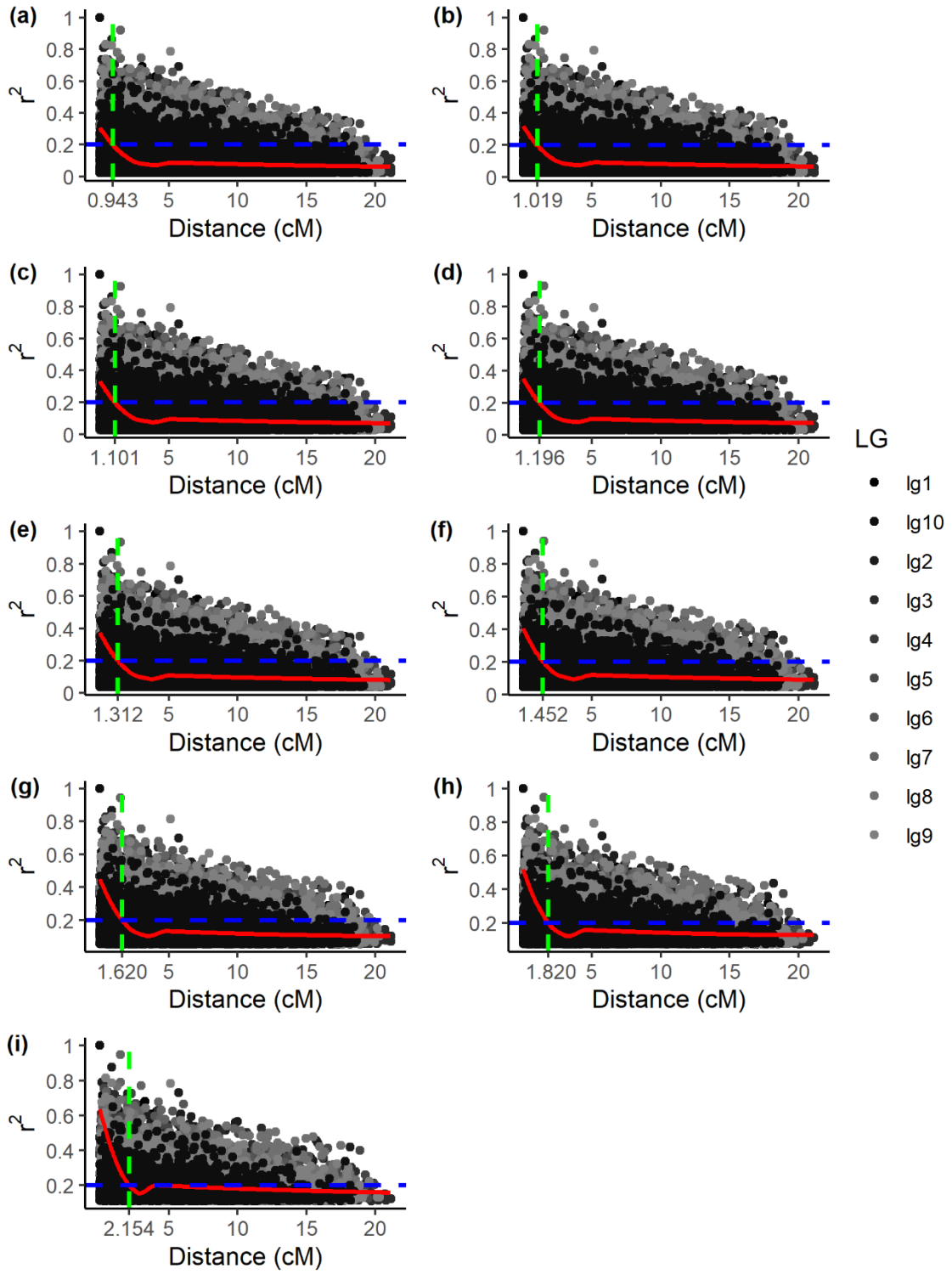

Fig. 3 Decay of linkage disequilibrium (r^2) as a function of genetic distance in the 10 linkage groups in the scenario with 100 QTLs. (a) Scenario: 1000 individuals (b) Scenario: 900 individuals (c) Scenario: 800 individuals (d) Scenario: 700 individuals (e) Scenario: 600

individuals (f) Scenario: 500 individuals (g) Scenario: 400 individuals (h) Scenario: 300 individuals (i) Scenario: 200 individuals

After obtaining these values, it was determined that all markers that are less than the distances mentioned above (depending on the scenario evaluated) from each QTL are considered as markers associated with the QTLs.

Genome-wide association

The general linear model obtained a low power of detection of QTLs in all scenarios evaluated (Table 1). In the scenarios with 3 QTLs, regardless of heritability and population size, this methodology showed power values equal to or less than 0.03 (Table 1). In the scenarios with 100 QTLs with 1000 individuals and a heritability of 0.30, the GLM obtained a power of detection on average of 0.21 ± 0.07 and with heritability 0.50, the power of detection was on average 0.56 ± 0.09. As the population size was reduced, the detection power was reduced until it reached zero in all scenarios evaluated (Table 1). This result was already expected and can be corroborated by several studies in the literature. For example, in the study by[60], in which the authors evaluated the effect of population size in GWAS, considering data from barley germplasm. In this study, the authors used a base population consisting of 766 individuals, and population size reduction was achieved by random resampling without replacement, forming populations with 96, 192, 288, 384, 480, 576, and 672 individuals, and observed that the detection power of QTLs decreased according to population size reduction. [61]Also evaluated the power of GWAS to identify true significant associations using simulated *Arabidopsis* data set with 200, 400, and 800 individuals. As a result, the authors observed that the power of identifying true associations decreased as the number of individuals decreased. In addition to these, [62]evaluated the influence of sample size in GWAS using simulated data from a Chinese soybean germplasm population consisting of 200, 400, 600, and 800 individuals randomly sampled from an ideal base population. As a result, the authors observed that the detection power of true significant associations decreased, and the false positive rate increased with decreasing sample size. Furthermore, according to[63] and[64], the efficiency of GWAS requires large population sizes.

However, the pattern reported by the authors mentioned above and those observed here for the GLM was not observed when using the QR models. In general, the QR, in all scenarios evaluated, obtained high detection power (Table 1). Additionally, unlike the results obtained

using GLM, the detection power of QTLs did not reduce with the decrease in population size (Table 1). This result may be related to the way in which the standard error is calculated by the two methodologies. In the GLM, the standard error estimate used was the usual one, while in the QR it was based on the rank statistic. The rank statistic is greatly influenced by the sample size[53,55]. Thus, the statistic of the test used generally presents higher values and, therefore, a greater number of QTLs being considered significant.

In scenarios with 3 QTLs, at quantiles of 0.10 and 0.90, regardless of heritability and population size variation, QR detected almost all simulated QTLs (Table 1). As for the scenarios with 100 QTLs, QR at the extreme quantiles (τ = 0.10 and 0.90) obtained higher or equal QTL detection power when compared to QR (τ = 0.50) (Table 1). In terms of population size, independent of heritability and quantile evaluated, QR detected all QTLs of interest considering population sizes equal to that of 200 and 300 individuals to QR (Table 1).

In general, the use of QR obtained a high QTL detection power independent of the population size, and especially in the extreme quantiles. This result is reasonable since QR uses the same idea of sampling for extremes [65]. Sampling extreme phenotypes samples individuals at the extremes in the hope that rare causal variants will be enriched among them [32]. However, unlike the extreme phenotype sampling approach, the use of QR does not require any assumptions about the distributions of traits, is robust to outliers, and uses all individuals in the estimation process, avoiding some problems related to extreme phenotype sampling, as an example, sampling bias and the assumption of normality [31,32].

The detection of significant SNPs with a small population size and at the extreme quantile has already been observed by[32]. The authors evaluated 80 genotypes and 384 SNP markers of common bean, aiming to identify genomic regions for three phenological traits (Days to first flowering-DPF; Days to flowering-DTF; and Days to end of flowering-DFF). As a result, the authors found no significant associations using GLM. On the other hand, when using QR at the 0.10 quantile, one and six significant SNPs were found for DPF and DTF, respectively. Although the work of [66] and[67] was not conducted in the context of genome-wide association, the authors also evaluated the performance of QR on simulated data set with small population sizes and concluded that QR is a robust technique in these situations. This result is very promising in breeding programs that have a reduced number of available genotypes.

Regarding the rate of false positives, we have found that the GLM, in all scenarios evaluated, presented low values for this rate. This result may be related to the low detection power of QTls by this methodology (Table 2). The false positive rate obtained by the QR methodology is relatively low in the scenarios with a higher number of individuals. QR (τ =

0.50) was the methodology that presented lower false positive rates. In scenarios where the QR detection power in the three quantiles evaluated was equal, the QR (τ = 0.50) showed better results than in the extreme quantiles QR (τ = 0.10 and 0.90) since the false positive rate was lower (Table 2). Regarding the reduction in the number of individuals, the false positive rate increased substantially according to the reduction in population size, a result that may be related to the observed increase in the number of QTLs detected in these scenarios.

Finally, it was observed that the decrease in the heritability of the trait implies a lower power of detection of QTLs when using the GLM in all scenarios evaluated (Table 1). This result is similar to that found by[62], in which the authors compared the detection power of true significant associations using five GWAS methods. This was done using simulated data from a Chinese soybean germplasm population with different levels of heritability ($h^2 = 0.20, 0.50$ and 0.90) and two genetic architectures with 10 and 100 QTls. As a result, the authors observed that the detection power was dramatically reduced for all methods and scenarios evaluated when the heritability of the trait was reduced. On the other hand, this behavior was not observed when using the QR methodology. The QR obtained greater or equal powers of detection of true significant associations in scenarios with lower heritability ($h^2 = 0.30$) regardless of the number of QTLs and sample size (Table 1). This result is interesting since it indicates that QR is an interesting methodology for GWAS studies in both low and moderate heritability scenarios.

Overall, these results indicate that using quantile regression to perform GWAS in the identification of QTLs is an interesting approach. QR proved to be efficient both in scenarios with many individuals and in scenarios with a reduced population size. Additionally, this methodology also proved to be interesting for GWAS studies in which the traits have low and moderate heritabilities.

Table 1. Means and standard errors (10 replicates) of QTL detection power against two methodologies.

			Population Size								
Nº. QTL	**h^2**	**Methods**	**1000**	**900**	**800**	**700**	**600**	**500**	**400**	**300**	**200**
3	0.30	QR (0.10)	1.00 ± 0.00	0.97 ± 0.03	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.50)	0.80 ± 0.07	0.90 ± 0.05	0.93 ± 0.04	0.96 ± 0.03	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.90)	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		GLM	0.03 ± 0.03	0.03 ± 0.03	0.03 ± 0.03	0.03 ± 0.03	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0 .00 ± 0.00	0.03 ± 0.03
	0.50	QR (0.10)	0.87 ± 0.07	0.90 ± 0.05	0.93 ± 0.04	0.93 ± 0.06	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.50)	0.70 ± 0.10	0.70 ± 0.10	0.50 ± 0.06	0.77 ± 0.05	0.90 ± 0.05	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.90)	0.80 ± 0.09	0.97 ± 0.03	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		GLM	0.03 ± 0.07	0.23 ± 0.07	0.23 ± 0.07	0.23 ± 0.07	0.20 ± 0.07	0.07 ± 0.04	0.03 ± 0.03	0.03 ± 0.03	0.00 ± 0.00
100	0.30	QR (0.10)	0.92 ± 0.02	0,97 ± 0.02	0.97 ± 0.02	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.50)	0.54 ± 0.09	0.72 ± 0.07	0.82 ± 0.05	0.92 ± 0.03	0.96 ± 0.03	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.90)	0.95 ± 0.02	0.98 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		GLM	0.21 ± 0.07	0.00 ± 0. 00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.10	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
	0.50	QR (0.10)	0.61 ± 0.06	0.77 ± 0.05	0.78 ± 0.07	0.93 ± 0.03	0.98 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		QR (0.50)	0.15 ± 0.06	0.23 ± 0.06	0.31 ± 0.08	0.57 ± 0.07	0.72 ± 0.06	0.85 ± 0.04	0.94 ± 0.02	1.00 ± 0.00	1.00 ± 0.00
		QR (0.90)	0.55 ± 0.06	0.64 ± 0.07	0.66 ± 0.07	0.85 ± 0.04	0.93 ± 0.02	0.98 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
		GLM	0.56 ± 0.09	0.07 ± 0.02	0.03 ± 0.01	0.04 ± 0.02	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.00 ± 0.00

N^O. QTL: number of loci controlling the trait; h^2: heritability; QR: quantile regression; GLM: general linear model.

Table 2. Averages and standard errors (10 repetitions) of the false positive rate against two methodologies.

Nº. QTL	h^2	Métodos	Population Size 1000	900	800	700	600	500	400	300	200
3	0.30	QR (0.10)	0.35 ± 0.04	0.36 ± 0.04	0.37 ± 0.03	0.42 ± 0.03	0.49 ± 0.03	0.54 ± 0.04	0.58 ± 0.03	0.65 ± 0.03	0.67 ± 0.01
		QR (0.50)	0.12 ± 0.02	0.13 ± 0.02	0.17 ± 0.02	0.20 ± 0.02	0.27 ± 0.02	0.34 ± 0.02	0.4 1± 0.02	0.49 ± 0.02	0.55 ± 0.02
		QR (0.90)	0.33 ± 0.02	0.40 ± 0.02	0.42 ± 0.02	0.50 ± 0.02	0.53 ± 0.02	0.53 ± 0.02	0.61 ± 0.02	0.69 ± 0.02	0.66 ± 0.03
		GLM	0.0006 ± 0.0003	0.0012 ± 0.0004	0.0003 ± 0.0003	0.0001 ± 0.0001	0.0001 ± 0.0001	0.0004 ± 0.0002	0.0002 ± 0.0002	0.0001 ± 0.0001	0.0001 ± 0.0001
	0.50	QR (0.10)	0.13 ± 0.04	0.18 ± 0.05	0.20 ± 0.04	0.24 ± 0.05	0.26 ± 0.04	0.33 ± 0.05	0.46 ± 0.04	0.55 ± 0.03	0.57 ± 0.02
		QR (0.50)	0.03 ± 0.01	0.04 ± 0.02	0.04 ± 0.02	0.06 ± 0.02	0.10 ± 0.02	0.14 ± 0.04	0.21 ± 0.04	0.30 ± 0.03	0.42 ± 0.03
		QR (0.90)	0.12 ± 0.03	0.18 ± 0.03	0.22 ± 0.03	0.26 ± 0.02	0.31 ± 0.03	0.36 ± 0.02	0.44 ± 0.03	0.57 ± 0.03	0.57 ± 0.03
		GLM	0.0082 ± 0.0021	0.0063 ± 0.0024	0.0022 ± 0.0005	0.0019 ± 0.0006	0.0016 ± 0.0007	0.0008 ± 0.0003	0.0004 ± 0.0002	0.0001 ± 0.0001	0.0001 ± 0.0001
100	0.30	QR (0.10)	0.18 ± 0.02	0.22 ± 0.02	0.24 ± 0.03	0.30 ± 0.02	0.33 ± 0.03	0.43 ± 0.03	0.47 ± 0.03	0.56 ± 0.03	0.67 ± 0.03
		QR (0.50)	0.07 ± 0.02	0.08 ± 0.02	0.10 ± 0.02	0.16 ± 0.03	0.21 ± 0.04	0.24 ± 0.02	0.36 ± 0.02	0.44 ± 0.03	0.56 ± 0.02
		QR (0.90)	0.23 ± 0.03	0.25 ± 0.02	0.26 ± 0.02	0.33 ± 0.03	0.41 ± 0.04	0.46 ± 0.03	0.54 ± 0.03	0.63 ± 0.03	0.73 ± 0.03
		GLM	0.0192 ± 0.0068	0.0003 ± 0.0003	0.0001 ± 0.0001	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000	0.0000 ± 0.0000
	0.50	QR (0.10)	0.06 ± 0.02	0.09 ± 0.02	0.11 ± 0.03	0.15 ± 0.02	0.22 ± 0.03	0.27 ± 0.03	0.36 ± 0.03	0.49 ± 0.04	0.64 ± 0.03
		QR (0.50)	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	0.04 ± 0.01	0.06 ± 0.01	0.08 ± 0.01	0.15 ± 0.02	0.24 ± 0.03	0.38 ± 0.03
		QR (0.90)	0.06 ± 0.01	0.06 ± 0.01	0.08 ± 0.02	0.12 ± 0.03	0.14 ± 0.02	0.21 ± 0.03	0.34 ± 0.04	0.46 ± 0.03	0.65 ± 0.03
		GLM	0.0680 ± 0.0147	0.0055 ± 0.0028	0.0016 ± 0.0007	0.0011 ± 0.0005	0.0004 ± 0.0002	0.0000 ± 0.0000	0.0007 ± 0.0007	0.0009 ± 0.0006	0.0001 ± 0.0001

N^O. QTL: number of loci controlling the trait; h^2: heritability; QR: quantile regression; GLM: general linear model.

CONCLUSION

The use of Quantile Regression models in genomic association studies on simulated data proved to be effective. Since its use, it allows a high power of detection of QTLs in all the scenarios analyzed in relation to the GLM. In scenarios with larger population sizes, the QR in the extreme quantiles (τ=0.1 and 0.9) were the most efficient models in the simulated conditions because they were the ones that obtained the highest QTL detection powers. In the scenario where the detection power of the QR in the three evaluated quantiles was equal, the QR (0.5) was more efficient, as the false positive rate was lower. In the low heritability scenarios, QR obtained a high detection power of QTLs. The false positive rate obtained by the QR methodology in the scenarios with many individuals is relatively low. QR proved to be efficient both in scenarios with many individuals and in scenarios with a small population size.

REFERENCES

1. Organização das Nações Unidas (ONU). População mundial deve chegar a 9,7 bilhões de pessoas em 2050, diz relatório da ONU. https://brasil.un.org/pt-br/83427-populacao-mundial-deve-chegar-97-bilhoes-de-pessoas-em-2050-diz-relatorio-da-onu.

2. Hunter, M. C., Smith, R. G., Schipanski, M. E., Atwood, L. W. & Mortensen, D. A. Agriculture in 2050: recalibrating targets for sustainable intensification. *Bioscience* **67**, 386–391 (2017).

3. Borém, A., Fritsche-Neto, R. & Miranda, G. V. *Melhoramento de plantas*. (2017).

4. Ramalho, M. A. P. *et al. Genética na Agropecuária*. (Editora UFLA, 2012).

5. Huang, X. & Han, B. Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* **65**, 531–551 (2014).

6. Nordborg, M. & Weigel, D. Next-generation genetics in plants. *Nature* **456**, 720–723 (2008).

7. Resende, R. T. *et al.* Genome-wide association and regional heritability mapping of plant

architecture, lodging and productivity in Phaseolus vulgaris. *G3 Genes, Genomes, Genet.* **8**, 2841–2854 (2018).

8. Wu, Z. & Zhao, H. Statistical power of model selection strategies for genome-wide association studies. *PLoS Genet.* **5**, e1000582 (2009).

9. Zhang, Z. *et al.* Improving the accuracy of whole genome prediction for complex traits using the results of genome wide association studies. *PLoS One* **9**, e93017 (2014).

10. Lorenz, A. J., Hamblin, M. T. & Jannink, J.-L. Performance of single nucleotide polymorphisms versus haplotypes for genome-wide association analysis in barley. *PLoS One* **5**, e14079 (2010).

11. Mwando, E. *et al.* Genome-Wide Association Study of Salinity Tolerance During Germination in Barley (Hordeum vulgare L.). *Front. Plant Sci.* **11**, 1–15 (2020).

12. Jaiswal, V. *et al.* Genome-wide association study (GWAS) delineates genomic loci for ten nutritional elements in foxtail millet (Setaria italica L.). *J. Cereal Sci.* **85**, 48–55 (2019).

13. Kuki, M. C. *et al.* Genome wide association study for gray leaf spot resistance in tropical maize core. *PLoS One* **13**, 1–13 (2018).

14. Olukolu, B. A., Tracy, W. F., Wisser, R., De Vries, B. & Balint-Kurti, P. J. A genome-wide association study for partial resistance to maize common rust. *Phytopathology* **106**, 745–751 (2016).

15. Malle, S., Eskandari, M., Morrison, M. & Belzile, F. Genome-wide association identifies several QTLs controlling cysteine and methionine content in soybean seed including some promising candidate genes. *Sci. Rep.* **10**, 1–14 (2020).

16. Zhang, W. *et al.* Comparative selective signature analysis and high-resolution GWAS reveal a new candidate gene controlling seed weight in soybean. *Theor. Appl. Genet.*

(2021) doi:10.1007/s00122-021-03774-6.

17. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).

18. Quero, G. *et al.* Genome-Wide Association Study Using Historical Breeding Populations Discovers Genomic Regions Involved in High-Quality Rice. *Plant Genome* **11**, 1–12 (2018).

19. Suela, M. M., Azevedo, C. F., Nascimento, M., Nascimento, A. C. C. & de Resende, M. D. V. Regional heritability mapping and genome-wide association identify loci for rice traits. *Crop Sci.* **62**, 839–858 (2022).

20. Zhao, K. *et al.* Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. *Nat. Commun.* **2**, 1–10 (2011).

21. Arora, S., Cheema, J., Poland, J., Uauy, C. & Chhuneja, P. Genome-wide association mapping of grain micronutrients concentration in Aegilops tauschii. *Front. Plant Sci.* **10**, 54 (2019).

22. Crossa, J. *et al.* Genomic Selection in Plant Breeding: Methods, Models, and Perspectives. *Trends Plant Sci.* **22**, 961–975 (2017).

23. Lin, Y. *et al.* Genome-wide association study of pre-harvest sprouting resistance in Chinese wheat founder parents. *Genet. Mol. Biol.* **40**, 620–629 (2017).

24. Gimase, J. M. *et al.* Genome-Wide Association Study identify the genetic loci conferring resistance to Coffee Berry Disease (Colletotrichum kahawae) in Coffea arabica var. Rume Sudan. *Euphytica* **216**, 1–17 (2020).

25. Sant'Ana, G. C. *et al.* Genome-wide association study reveals candidate genes influencing lipids and diterpenes contents in Coffea arabica L. *Sci. Rep.* **8**, 1–12 (2018).

26. Tran, H. T. M. *et al.* SNP in the Coffea arabica genome associated with coffee quality.

Tree Genet. Genomes **14**, (2018).

27. Resende, M. D. V. de, Silva, F. F. & 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*. (2014).

28. Wang, J. & Zhang, Z. GAPIT Version 3: Boosting Power and Accuracy for Genomic Association and Prediction. *Genomics, Proteomics Bioinforma.* **19**, 629–640 (2021).

29. Galarza, C. E., Lachos, V. H. & Bandyopadhyay, D. Quantile regression in linear mixed models: a stochastic approximation EM approach. *Stat. Interface* **10**, 471 (2017).

30. Koenker, R. & Bassett, G. Regression Quantiles. *Econometrica* **46**, 33–50 (1978).

31. Oliveira, G. F. *et al.* Quantile regression in genomic selection for oligogenic traits in autogamous plants: A simulation study. *PLoS One* **16**, 1–12 (2021).

32. Nascimento, M. *et al.* Quantile regression for genome-wide association study of flowering time-related traits in common bean. *PLoS One* **13**, 1–14 (2018).

33. Liu, H. *et al.* ADAM-Plant: A software for stochastic simulations of plant breeding from molecular to phenotypic level and from simple selection to complex speed breeding programs. *Front. Plant Sci.* **9**, 1–15 (2019).

34. Sun, X., Peng, T. & Mumm, R. H. The role and basics of computer simulation in support of critical decisions in plant breeding. *Mol. Breed.* **28**, 421–436 (2011).

35. Wang, J. Modelling and simulation of plant breeding strategies. in *Plant Breeding* 19–40 (IntechOpen, 2012).

36. Viana, J. M. S. Quantitative genetics theory for non-inbred populations in linkage disequilibrium. *Genet. Mol. Biol.* **27**, 594–601 (2004).

37. Viana, J. M. S. Programa para análises de dados moleculares e quantitativos Real Breeding. (2013).

38. Azevedo, C. F. *et al.* Population structure correction for genomic selection through eigenvector covariates. *Crop Breed. Appl. Biotechnol.* **17**, 350–358 (2017).

39. Ferreira, A., da Silva, M. F., da Costa e Silva, L. & Cruz, C. D. Estimating the effects of population size and type on the accuracy of genetic maps. *Genet. Mol. Biol.* **29**, 187–192 (2006).

40. Campoy, J. A. *et al.* Genetic diversity, linkage disequilibrium, population structure and construction of a core collection of Prunus avium L. landraces and bred cultivars. *BMC Plant Biol.* **16**, 1–15 (2016).

41. Jia, Z. *et al.* Genetic dissection of root system architectural traits in spring barley. *Front. Plant Sci.* **10**, 400 (2019).

42. Niu, S. *et al.* Genetic diversity , linkage disequilibrium , and population structure analysis of the tea plant (Camellia sinensis) from an origin center , Guizhou plateau , using genome- wide SNPs developed by genotyping-by- sequencing. *BMC Plant Biol.* **19**, 1–12 (2019).

43. Otyama, P. I. *et al.* Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. *BMC Genomics* **20**, 1–17 (2019).

44. Vos, P. G. *et al.* Evaluation of LD decay and various LD-decay estimators in simulated and SNP-array data of tetraploid potato. *Theor. Appl. Genet.* **130**, 123–135 (2017).

45. Covarrubias-Pazaran, G. Genome-assisted prediction of quantitative traits using the R package sommer. *PLoS One* **11**, e0156744 (2016).

46. Team, R. C. R: A language and environment for statistical computing. R Foundation for Statistical Computing. (2020).

47. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

48. Racedo, J. *et al.* Genome-wide association mapping of quantitative traits in a breeding population of sugarcane. *BMC Plant Biol.* **16**, 1–16 (2016).

49. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

50. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

51. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).

52. Koenker, R. quantreg: Quantile regression. (2015).

53. Koenker, R. *Quantile Regression*. (2005).

54. Lipka, A. E. *et al.* GAPIT: Genome association and prediction integrated tool. *Bioinformatics* **28**, 2397–2399 (2012).

55. Koenker, R. & Machado, J. A. F. Goodness of fit and related inference processes for quantile regression. *J. Am. Stat. Assoc.* **94**, 1296–1310 (1999).

56. Koenker, R. Confidence intervals for regression quantiles. in *Asymptotic statistics* 349–359 (Springer, 1994).

57. Fernando, R. L. *et al.* Controlling the proportion of false positives in multiple dependent tests. *Genetics* **166**, 611–619 (2004).

58. Silva, H. D. & Vencovsky, R. Poder de detecção de 'quantitative trait loci', da análise de marcas simples e da regressão linear múltipla. *Sci. Agric.* **59**, 755–762 (2002).

59. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *PNAS* **100**, 9440–9445 (2003).

60. Wang, H. *et al.* Effect of population size and unbalanced data sets on QTL detection using genome-wide association mapping in barley breeding germplasm. *Theor. Appl. Genet.* **124**, 111–124 (2012).

61. Korte, A. & Farlow, A. The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods* **9**, 1–9 (2013).

62. He, J. *et al.* An innovative procedure of genome-wide association analysis fits studies on germplasm population and plant breeding. *Theor. Appl. Genet.* **130**, 2327–2343 (2017).

63. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat. Genet.* **42**, 355–360 (2010).

64. Cantor, R. M., Lange, K. & Sinsheimer, J. S. Prioritizing GWAS results: a review of statistical methods and recommendations for their application. *Am. J. Hum. Genet.* **86**, 6–22 (2010).

65. Wang, K. *et al.* A genome-wide association study on obesity and obesity-related traits. *PLoS One* **6**, e18939 (2011).

66. Tarr, G. Small sample performance of quantile regression confidence intervals. *J. Stat. Comput. Simul.* **82**, 81–94 (2012).

67. Ismail, E. A.-R. Behavior of lasso quantile regression with small sample sizes. *J. Multidiscip. Eng. Sci. Technol.* **2**, 388–394 (2015).

STATEMENTS AND DECLARATIONS

Acknowledgments

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (process number 001) and CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico (process numbers 306772/2020-5 and 307798/2019-4), for the financial support and the grant conceded.

Author contributions

Conceptualization: [GFO, ACCN, CFA, MN]; Data curation: [JMSV, MDVR]; Methodology: [GFO, ACCN, CFA, MN]; Formal analysis and investigation: [GFO, ACCN, CFA, MOC, MN]; Writing - original draft preparation: [GFO, ACCN, CFA, MOC, MN]; Writing - review and editing: [GFO, ACCN, CFA, MOC, LMAB, ICS, MN]; Supervision: [ACCN, CFA, MN]; Software: [GFO, ACCN, CFA, MOC, JMSV, MDVR, MN];

Data Availability statement

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.