LEÍSA PIRES LIMA

MÉTODOS APLICADOS AOS ESTUDOS DE ASSOCIAÇÃO GENÔMICA VIA REGIÕES CROMOSSÔMICAS CONSIDERANDO EFEITOS ADITIVOS E DE DOMINÂNCIA

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*.

Orientadora: Camila Ferreira Azevedo

Coorientadores: Marcos Deon Vilela de Resende
Moysés Nascimento

VIÇOSA - MINAS GERAIS
2021

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

T
L732m
2021

Lima, Leísa Pires, 1993-
Métodos aplicados aos estudos de associação genômica via regiões cromossômicas considerando efeitos aditivos e de dominância / Leísa Pires Lima. – Viçosa, MG, 2021.
115 f. : il. ; 29 cm.

Orientador: Camila Ferreira Azevedo.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Sequenciamento de nucleotídeo. 2. Teoria bayesiana de decisão estatística. 3. Varição genética. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22. ed. 519.542

LEÍSA PIRES LIMA

MÉTODOS APLICADOS AOS ESTUDOS DE ASSOCIAÇÃO GENÔMICA VIA REGIÕES CROMOSSÔMICAS CONSIDERANDO EFEITOS ADITIVOS E DE DOMINÂNCIA

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: 01 de março de 2021.

Assentimento:

Leísa Pires Lima
Autor(a)

Camila Ferreira Azevedo
Orientador(a)

A Deus por ser meu amparo e
aos meus pais, Rossir e Yêda,
por toda dedicação e apoio.

AGRADECIMENTOS

A Deus, meu guia, por me dar força e amparo para passar por todos os obstáculos, desanimo, cansaço e desespero. Sem Ele eu não chegaria até aqui.

A Universidade Federal de Viçosa e ao Programa de Pós-Graduação em Estatística Aplicada e Biometria, pela oportunidade concedida para realização do curso.

Aos meus pais, Yêda e Rossir, pelo amor incondicional, pelos conselhos, ensinamentos, orações, pela dedicação e confiança.

Aos meus irmãos, Osvair, Marisa e Junior, pelo incentivo, amizade e apoio e por sempre estarem ao meu lado.

Ao meu namorado Sillas, pelo carinho, paciência, incentivo, amor e companheirismo.

À minha família, pela torcida e apoio.

Aos meus amigos do PPESTBIO em especial Jaquicele, Gabriely, Gabriela, Carol e Roberta pelos momentos de descontração, pelas trocas de experiência, pelas palavras de conforto e constantes incentivos.

À Doutora e orientadora Camila Ferreira Azevedo, pela paciência, confiança, pelos ensinamentos, conselhos, críticas e sugestões que foram primordiais para o meu crescimento tanto profissional quanto pessoal. Fica aqui registrada minha eterna gratidão.

Aos Doutores e coorientadores Marcos Deon Vilela de Resende e Moysés Nascimento, pela disponibilidade, confiança, incentivo e pelos saberes transmitidos.

Aos membros da banca examinadora, Cosme Damião Cruz, Leonardo Siqueira Glória, Isabela de Castro Sant'anna e Moysés Nascimento, pela disponibilidade e pelas sugestões para o enriquecimento deste trabalho.

Aos professores do Programa de Pós-Graduação em Estatística Aplicada e Biometria, por contribuírem para minha formação acadêmica.

Aos secretários do Departamento de Estatística, Júnior e Anita, por sempre estarem dispostos a ajudar e também pela amizade, incentivo e carinho.

Ao Laboratório de Inteligência Computacional e Aprendizado Estatístico (LICAE) e ao Grupo de Estudos em Estatística Aplicada e Biometria (GESTBIO) por terem contribuído para meu crescimento, tanto pessoal quanto profissional.

A CAPES, pela concessão da bolsa de estudos.

Enfim, muito obrigada a todos aqueles que de certa forma contribuíram para a concretização deste trabalho.

RESUMO

LIMA, Leísa Pires, D.Sc., Universidade Federal de Viçosa, março de 2021. **Métodos aplicados aos estudos de associação genômica via regiões cromossômicas considerando efeitos aditivos e de dominância.** Orientadora: Camila Ferreira Azevedo. Coorientadores: Marcos Deon Vilela de Resende e Moysés Nascimento.

Os avanços na biologia molecular e as inovações nas tecnologias de sequenciamento e de genotipagem permitiram o desenvolvimento de novos marcadores moleculares favorecendo os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS). A análise via marcas únicas se destaca como o principal procedimento para estudar a associação entre marcas e QTL (*Quantitative Trait Loci*), porém metodologias que consideraram grupos de marcadores para flanquear regiões genômicas vem elucidando importantes resultados para estudos de associação. Várias abordagens estatísticas veem sendo propostas no âmbito da GWAS, no entanto, estudos comparativos revelam que os métodos bayesianos são superiores em termos do poder em detectar marcadores com associações significativas. Entre os critérios existentes de seleção de regiões se destacam, a seleção pela porcentagem da variância explicada por regiões genômicas (%var), o critério de seleção de *tag*SNPs (*tag*SNPs) e a seleção com base na probabilidade *a posteriori* da associação de regiões genômicas (WPPA - *Window Posterior Probability of Association*). Para também detectar regiões potencialmente associadas, foi proposto o critério baseado na probabilidade *a posteriori* do intervalo (*Posterior Probability of Interval* - PP_{int}) que visa selecionar regiões com base nos marcadores de maiores efeitos estimados via método bayesiano, neste estudo o BayesDπ. Além disso, uma metodologia alternativa, denominada mapeamento de herdabilidades regionais (*Regional Heritability Mapping* - RHM) vem apresentando importantes resultados. Dessa forma, o primeiro capítulo deste trabalho consiste em uma revisão de literatura sobre a GWAS apresentando sua definição e importância no melhoramento genético e abordando detalhes teóricos acerca dos critérios citados acima. Já o capítulo 2 visa propor a medida PP_{int} e compará-la às demais abordagens, *tag*SNP, %var, WPPA conjuntamente ao BayesDπ e metodologia de marcas únicas, quanto a eficiência em selecionar e identificar marcadores ou regiões associados a QTL. Para isso, utilizou-se dados simulados considerando seis cenários diferentes, sendo os SNPs alocados em regiões genômicas não sobrepostas. Os resultados do segundo capítulo indicaram que para características com herança oligogênica, o critério WPPA seguido dos critérios %var e PP_{int} se mostraram superiores, apresentando maiores valores de poder de detecção, capturando maiores

porcentagens de variância genética e maiores áreas. Para características com herança poligênica, os critérios PP_{int} e WPPA foram considerados superiores aos demais. Ademais, o capítulo 3 avalia os critérios, PP_{int} e WPPA, que se mostraram superiores no capítulo 2 junto aos métodos de análise via marcas únicas e o RHM. No entanto, a eficiência em termos de poder de detecção e de falsos positivos destes métodos foi avaliada considerando ou não a inclusão dos efeitos de dominância nos modelos estatísticos. Para isso, foram utilizados dados simulados em dezoito cenários com diferentes níveis de herdabilidade, arquitetura genética e grau médio de dominância. Os resultados indicaram que para os efeitos aditivos considerando características com arquitetura genética oligogênica, os critérios WPPA, RHM e PP_{int} se mostraram superiores para todos os graus de dominância analisados. Já para características com herança poligênica, os critérios PP_{int} e WPPA podem ser considerados superiores aos demais. Considerando apenas os efeitos devido à dominância, os critérios WPPA, RHM, análise via marcas únicas e PP_{int} apresentaram resultados relevantes com relação as medidas de eficiência para as características controladas por 3 QTL.

Palavras-chave: Regiões Genômicas. Marcadores Moleculares. Métodos Bayesianos. Variância Genética.

ABSTRACT

LIMA, Leísa Pires, D.Sc., Universidade Federal de Viçosa, March, 2021. **Methods applied to genomic association studies via chromosomal regions considering additive and dominance effects.** Adviser: Camila Ferreira Azevedo. Co-advisers: Marcos Deon Vilela de Resende and Moysés Nascimento.

Advances in molecular biology and innovations in sequencing and genotyping technologies have allowed the development of new molecular markers favoring genome-wide association studies (GWAS). The single-marker analyses stand out as the central procedure to study the association between markers and quantitative trait loci (QTL). However, methodologies that considered groups of markers to flank genomic regions have elucidated important results for association studies. Several statistical approaches are being proposed within the scope of GWAS. However, comparative analyses reveal that Bayesian methods are superior in terms of the power to detect markers with significant associations. Among the existing criteria for the region selection, the selection by the percentage of variance explained (%var), the selection criteria for tag single nucleotide polymorphisms (*tag*SNPs), and the selection based on the window posterior probability of association (WPPA). To also detect potentially associated regions, a criterion based on the a posteriori probability of interval (PP_{int}) was proposed, aiming to select regions based on the markers of greatest effects estimated via the Bayesian method. In this study, the BayesDπ. An alternative methodology, called regional heritability mapping (RHM), has been shown substantial results. Thus, the first chapter of this work consists of a literature review on GWAS presenting its definition and importance in genetic improvement and addressing theoretical details about the criteria mentioned above. Chapter 2 aims to propose the PP_{int} measure and compare it to the other approaches, *tag*SNP, %var, WPPA together with BayesDπ and single-marker analyses, regarding the efficiency in selecting and identifying markers or regions associated with QTL. For this, simulated data was used considering six different scenarios, with SNPs being allocated in non-overlapping genomic regions. The second chapter results indicated that for traits with oligogenic inheritance, the WPPA criterion followed by the %var and PP_{int} criteria were superior, presenting higher values of detection power, capturing higher percentages of genetic variance and larger areas. The criteria PP_{int} and WPPA were considered superior to the others. Also, chapter 3 evaluates the criteria, PP_{int} and WPPA, which proved to be superior in chapter 2 together with the single-marker analyses and RHM. However, the detection power and the false positives was assessed

considering whether (or not) the inclusion of the dominance effects in the statistical models. For that, simulated data were used in eighteen scenarios with different heritability levels, genetic architecture, and degree of dominance. The results indicated that for the additive effects considering traits with oligogenic genetic architecture, the WPPA, RHM and PP_{int} criteria were superior for all analyzed degrees of dominance. For traits with polygenic inheritance, the PP_{int} and WPPA criteria can be considered superior. Considering only the effects due to dominance, the criteria WPPA, RHM, single-marker analyses and PP_{int} presented relevant results regarding the efficiency measures for the traits controlled by 3 QTL.

Keywords: Genomic Regions. Molecular Markers. Bayesian Methods. Genetic Variance.

SUMÁRIO

INTRODUÇÃO GERAL

Os avanços na biologia molecular e as inovações nas tecnologias de sequenciamento e de genotipagem resultaram em reduções nos custos e rapidez na obtenção de informações sobre marcadores moleculares. Essas inovações biotecnológicas permitiram o desenvolvimento de marcadores moleculares SNPs (*Single Nucleotide Polymorphisms*) proporcionando a obtenção de informações sobre variabilidade genética, identificação e localização de genes específicos favorecendo os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS) (MEUWISSEN et al., 2001). A GWAS visa identificar as associações entre os locos de características quantitativas (*Quantitative Trait Loc*i - QTL) e os valores genéticos dos indivíduos e, posteriormente, identificar em que posição do genoma está o QTL, buscando, assim, a compreensão da influência genética sobre a expressão fenotípica. No entanto, como não se tem acesso direto ao QTL e aos valores genéticos e devido ao desequilíbrio de ligação (*Linkage Disequilibrium* – LD) entre os marcadores e QTL considera-se que na prática, os estudos destas associações podem ser realizados entre os marcadores moleculares e os fenótipos.

Várias abordagens estatísticas veem sendo propostas no âmbito da GWAS, como as análises via marcas únicas, análises via modelos lineares mistos, modelos de haplótipos, modelos mistos baseados em genealogia e modelos de seleção de marcadores via abordagens bayesianas (DASHAB et al., 2012; FERNANDO e GARRICK, 2013; ZHOU et al., 2014; WU et al., 2014; GUO et al., 2016; BENNEWITZ et al., 2017; FERNANDO et al., 2017; BRAZ et al., 2019). No entanto, estudos comparativos revelam que os métodos baseados em modelos mistos e utilizando a abordagem bayesiana são superiores em termos do poder em detectar marcadores com associações significativas nas características de interesse (SAHANA et al., 2010; DASHAB et al., 2012; FERNANDO et al., 2017).

A análise via marcas únicas se destaca como o principal procedimento para estudar a associação entre marcas e QTL. Essa metodologia tradicionalmente utilizada, estima o efeito individual de cada marcador no fenótipo e, posteriormente, são realizados testes de hipóteses para detectar os efeitos de marcadores com significância estatística. Porém, este método além de superestimar os efeitos dos marcadores sofre também com a alta taxa de falsos positivos (Fernando et al., 2004), que é declarar o efeito do SNP como significativo quando na verdade não é, assumindo que o marcador está em LD com o QTL, quando ele não está.

Modelos utilizando abordagem bayesiana têm sido propostos e vêm se mostrando eficientes para a estimação simultânea dos efeitos dos marcadores e para a seleção de grupos de SNPs relevantes no melhoramento animal e vegetal. As metodologias bayesianas tais como BayesA, BayesB, BayesCπ e BayesDπ têm propiciado novas perspectivas para as questões relacionadas à estimação de componentes de variância e de parâmetros genéticos, pois diferentes graus de incerteza relativos a determinado parâmetro podem ser inferidos por meio de um modelo estatístico *a priori*, o qual considera tanto as observações, quanto os parâmetros, como quantidades aleatórias. Além disso, essas metodologias permitem investigar a arquitetura genética envolvida na expressão de diversas características de interesse de um programa de melhoramento. O principal procedimento bayesiano para a identificação de associação entre marcador e QTL, de modo similar a análise via marcas únicas, é o Fator de Bayes (*Bayes Factor* – BF) proposto por Kass e Raftery (1995) (HEATH, 1997; VARONA et al. 2001; HABIER et al. 2011; LEGARRA et al. 2015). No entanto, na prática como são considerados milhares de marcadores nas análises e devido a utilização de cadeias de *Markov*, em que são realizadas muitas iterações até atingir a convergência em cada análise, o BF se torna inviável pois exige uma tarefa computacional complexa, elevado tempo de execução bem como problemas de armazenamento de informações.

Dispondo-se de milhares de SNPs a fim de inferir associações, é esperado que marcadores próximos uns aos outros estejam altamente correlacionados. Dessa forma, ao considerar um grupo de marcadores para encontrar uma determinada região genômica, espera-se que a maior parte da variabilidade de determinado loco de característica seja explicada conjuntamente (MOORE et al. 2010). Ademais, estudos como os de Onteru et al. (2010), Fernando e Garrick (2013) e Fernando et al. (2017) têm indicado que marcadores únicos explicam uma pequena fração da variância genética quando se considera características quantitativas e assim, podem não mostrar uma associação forte entre marcador e fenótipo. As regiões que contribuírem com maiores variâncias genéticas são consideradas aquelas mais associadas à característica (PETERS et al., 2012).

Estudos como o de Sollero et al. (2016) têm mostrado grandes vantagens nos estudos de GWAS ao utilizar metodologias bayesianas para selecionar essas regiões genômicas. A partir disso, metodologias estatísticas alternativas para o melhoramento animal e vegetal, utilizando abordagem bayesiana, que não exigem grandes esforços computacionais e elevados tempos computacionais veem se mostrando eficientes para a seleção de grupos ótimos de SNPs, quando comparados com a análise via marcas únicas, indicando genes com relevantes efeitos no

desempenho dos indivíduos (STRAM, 2004; NAGAMINE et al., 2012; SOLLERO et al, 2016; RESENDE et al., 2017). Dentre elas se destacam a seleção pela porcentagem da variância explicada por regiões genômicas (%var), o critério de seleção de *tag*SNPs proposto por Sollero et al. (2017) (*tag*SNPs) e a seleção com base na probabilidade *a posteriori* da associação de regiões genômicas (WPPA - *Window Posterior Probability of Association*) implementada por Fernando e Garrick (2013) e Fernando et al. (2017). As regiões significativas encontradas por esses procedimentos, possivelmente, estão em maiores níveis de LD com o QTL e assim capturam uma proporção maior da variância genética explicada pelos SNPs.

A seleção de grupos de SNPs por meio da proporção da variância genética explicada por regiões genômicas, representada aqui por %var, foi proposta inicialmente por Wang et al. (2014) utilizando dados de peso de frango de corte para a comparação de alguns métodos implementados nos estudos de GWAS e utilizada por de Oliveira et al. (2017) e Soares et al. (2017) considerando dados de bovinos da raça *Nelore* e *Brahman*, respectivamente. Essa metodologia, sob a abordagem bayesiana, visa selecionar regiões que possuem variância genética superior a um *threshold* pré-estabelecido, baseado na porcentagem da média *a posteriori* da variância genética total explicada pelos SNPs. Esse critério se mostrou eficiente em selecionar regiões indicando genes associados a características de interesse (OLIVEIRA et al., 2017).

O critério para seleção de *tag*SNPs tem intuito, inicialmente, de também identificar regiões que expliquem maior porcentagem da média *a posteriori* da variância genética total utilizando a mesma abordagem proposta pelo critério %var. No entanto, posteriormente, esse critério visa selecionar, nessas regiões, SNPs que potencialmente possuem alto LD com o QTL associados a características de interesse. Essa seleção é feita utilizando as frequências de inclusão (FI) dos SNPs no modelo que é calculada por meio da razão entre o número de iterações salvas das cadeias MCMC (*Markov chain Monte Carlo*) que incluem o SNP em questão no modelo e o número total de iterações salvas e a estatística *t-like* (TL) que é a razão do efeito médio *a posteriori* somente das cadeias que incluíram determinado marcador no modelo, dividido pelo desvio-padrão destes efeitos. Sollero et al. (2017) em estudos de seleção de SNPs para a característica de resistência à carrapatos em raças brasileiras de bovinos *Braford* e *Hereford* verificou que essa metodologia se mostrou eficiente para a detecção de marcadores associados a genes com importantes efeitos no desempenho desses indivíduos.

Já a probabilidade *a posteriori* da associação da região genômica baseia-se também na proporção da variância genética explicada pelos marcadores em cada região obtida por meio

dos efeitos estimados nas cadeias provenientes dos algoritmos MCMC. A medida WPPA é obtida pela razão entre o número de amostras em que a proporção de explicação da variância genética regional é maior do que a proporção esperada com base na média *a posteriori* da variância genética total (PETERS et al., 2012; BENNEWITZ et al., 2017) e o número total de iterações salvas da cadeia.

Para também detectar regiões potencialmente associadas utilizando a abordagem bayesiana, no segundo capítulo desta tese foi proposta a medida PP_{int} (*Posterior Probability of Interval*). Essa medida é calculada pela razão entre o número de iterações das cadeias MCMC em que a região possui pelo menos um SNP com magnitude de efeito superior ao valor do terceiro quartil, considerando toda a distribuição dos efeitos absolutos naquela iteração, e o número total de iterações salvas. As regiões com valores de PP_{int} superiores a um *threshold* especificado são selecionadas e consideradas como regiões associadas.

Além dos métodos bayesianos, uma metodologia estatística alternativa para a utilização de regiões do genoma, denominada mapeamento de herdabilidades regionais (*Regional heritability mapping* - RHM), foi proposta por Nagamine et al. (2012) e visa também determinar as regiões do genoma que estão associadas ao valor genético. O método RHM visa determinar as regiões do genoma que estão associadas ao fenótipo. Essa metodologia foi utilizada por Shirali et al. (2014) em seres humanos, por Riggio et al. (2013), Usai et al. (2014) e Matika et al. (2016) em estudos de animais domésticos e recentemente foi aplicada por Resende et al. (2017) em estudos de GWAS para espécies florestais e por Resende et al. (2018) em feijão. O RHM utiliza uma matriz de relacionamentos genômicos entre os indivíduos baseada em variantes de SNPs comuns e raros encontrados em segmentos curtos do genoma para estimar a variância explicada das características por essas regiões (RESENDE et al., 2017).

Geralmente, as metodologias para associação genômica anteriormente citadas são aplicadas considerando apenas os efeitos aditivos dos marcadores. No entanto, os efeitos de dominância são fontes de variação de características complexas e que não deveriam ser negligenciados (VAN TASSEL et al., 2000; SU et al., 2012; ERTL et al., 2014; BENNEWITZ et al., 2017). As vantagens em considerar efeitos de dominância em estudos de associação foram elucidadas por Bennewitz et al. (2017), onde incluíram a dominância no modelo utilizando o enfoque bayesiano, que permitiu um aumento no poder de detecção para genes causadores que explicavam mais de 2,5% da variância genética, considerando dados simulados e dados reais referentes a gado de leite. No entanto, o método bayesiano utilizado pressupunha variância

comum a todos os marcadores, o que não é coerente com a prática para características oligogênicas e, em geral, não é desejável para estudos de associação genômica.

O capítulo 1 desta tese consiste em uma revisão de literatura sobre a GWAS apresentando sua definição e importância no melhoramento genético. Além disso, neste capítulo detalhes teóricos são apresentados acerca da análise via marcas únicas, o qual consiste no método mais utilizado em estudos de associação. Além disso, são apresentadas as vantagens teóricas de se considerar regiões genômicas utilizando metodologias bayesianas no estudo de GWAS com ênfase no método BayesDπ, pois este apresenta propriedades estatísticas importantes para a associação, e uma descrição detalhada dos critérios citados acima que visam selecionar regiões genômicas.

Já o capítulo 2 visa propor a medida PP_{int} e compará-la às demais abordagens, *tag*SNP, %var e WPPA, quanto a eficiência em selecionar e identificar marcadores ou regiões que estão localizadas dentro ou próximas a genes associados às características de interesse. Este estudo também teve como propósito comparar os resultados, de poder de detecção e de falsos positivos, obtidos por estas metodologias com a análise via marcas únicas visto que essa abordagem é a mais usual na prática nos estudos de GWAS. Para isso, utilizou-se dados simulados considerando seis cenários diferentes com ausência de dominância (dois níveis de herdabilidade distintas em cada cenário × três arquiteturas genéticas), sendo os SNPs alocados em regiões genômicas não sobrepostas de tamanhos definidos com base na extensão do LD entre marcador e QTL em cada cenário.

Ademais, o capítulo 3 avalia os métodos superiores encontrados no capítulo 2 junto ao método RHM, em detectar regiões do genoma, que estão localizadas dentro ou próximas a genes associados as características simuladas. No entanto, a eficiência em termos de poder de detecção e de falsos positivos destes métodos são avaliadas considerando ou não a inclusão dos efeitos de dominância nos modelos estatísticos para os estudos de GWAS. Para a avaliação e a apresentação das abordagens descritas anteriormente foram simulados dezoito cenários. Os cenários diferem de acordo com os níveis de herdabilidade (dois níveis distintos em cada cenário), níveis de grau médio de dominância ($gmd = 0{,}00; 0{,}50; 1{,}00$) e arquiteturas genéticas, características oligogênicas (controladas por poucos locos) e poligênicas (controladas por muitos locos).

CAPÍTULO 1

REVISÃO DE LITERATURA

1 Associação Genômica Ampla (*Genome Wide Association Studies* – GWAS)

1.1 Definição e Importância

Os estudos da associação genômica ampla (*Genome Wide Association Studies* – GWAS) ganharam destaque nos programas de melhoramento animal e vegetal devido ao desenvolvimento de tecnologias de sequenciamento e de genotipagem em larga escala, possibilitando o estudo de características complexas (GODDARD, HAYES, 2009; AGUILAR et al., 2010; SCHMID e BENNEWITZ, 2017; OTYAMA et al., 2019). Devido a essa evolução foi possível o desenvolvimento de marcadores moleculares proporcionando a obtenção de informações sobre variabilidade genética, identificação e localização de genes específicos e suas associações com características fenotípicas.

Por sua vez, a GWAS visa identificar estas associações entre os loci de características quantitativas (*Quantitative Trait Loci* - QTL) e os valores genéticos dos indivíduos e, posteriormente, identificar em que região do genoma está o QTL, buscando, assim, a compreensão da influência genética sobre a expressão fenotípica (RESENDE et al., 2014). No entanto, como não se tem acesso direto ao QTL e aos valores genéticos e devido a pressuposição de desequilíbrio de ligação (*Linkage Disequilibrium* – LD) entre os marcadores e QTL, considera-se que na prática, os estudos destas associações podem ser realizados entre os marcadores moleculares e os fenótipos, explorando todo o genoma por meio dos marcadores e associando os QTL à característica fenotípica de interesse. Os marcadores moleculares que mais se destacam, neste contexto, são os Polimorfismos de nucleotídeo único (*Single Nucleotide Polymorphisms* - SNPs), pois são amplamente distribuídos no genoma, são codominantes e devido sua genotipagem em larga escala.

As técnicas da GWAS foram inicialmente validadas por McCarthy et al. (2008) em estudos epidemiológicos em humanos se mostrando como uma poderosa abordagem de identificação de genes envolvidos com caracteres de interesse. A partir disso, tem sido cada vez mais utilizada no melhoramento genético animal e vegetal (WEI et al., 2016; SANTANA et al., 2014; WOLFE et al, 2016). Estes estudos tiveram um caráter inovador, permitindo a

identificação de associações entre milhares de locos genômicos e características complexas, aumentando o entendimento sobre a base genética das mesmas.

Uma das características dos estudos de associação genômica é a possibilidade de se testar um grande número de marcadores e várias características quantitativas, tendo como resultado disso um grande número de posições genômicas associadas.

Várias abordagens estatísticas que buscam identificar as associações no âmbito da GWAS veem sendo propostas, como as análises via marcas únicas, análises via modelos lineares mistos, modelos de haplótipos, modelos mistos baseados em genealogia e modelos de seleção de marcadores via abordagens bayesianas (DASHAB et al., 2012; FERNANDO e GARRICK, 2013; ZHOU et al., 2014; WU et al., 2014; GUO et al., 2016; BENNEWITZ et al., 2017; FERNANDO et al., 2017; BRAZ et al., 2019). No entanto, estudos comparativos revelam que os métodos estatísticos baseados em modelos mistos e utilizando a abordagem bayesiana são superiores em termos do poder em detectar marcadores com associações relevantes (SAHANA et al., 2010; DASHAB et al., 2012; FERNANDO et al., 2017).

Na prática, o método estatístico mais utilizado na GWAS é a análise via marcas únicas, em que se estima o efeito individual de cada marcador no fenótipo e, posteriormente, realizam-se testes de hipóteses para avaliar os efeitos de marcadores com significância estatística. Desta forma, são apresentadas a seguir descrições metodológicas básicas sobre a análise via marcas únicas.

1.2 Análise via marcas únicas

O modelo linear misto em marcas simples visa estimar o efeito aditivo do j-ésimo marcador no fenótipo e é definido por:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{W_j} m_j + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$ é o vetor de fenótipos com dimensão $N \times 1$ sendo N o número de indivíduos;

μ é a média geral da característica;

$\mathbf{1}$ é um vetor de mesma dimensão de $\boldsymbol{y}$ com todos os elementos iguais a 1;

$\boldsymbol{a}$ $(N \times 1)$ é o vetor de efeitos genéticos aditivos poligênicos com matriz de incidência $\boldsymbol{Z_1}$ $(N \times N)$, sendo $\boldsymbol{a} \sim N(0, \boldsymbol{G}\sigma_a^2)$ em que $\boldsymbol{G}$ é a matriz de parentesco genômico aditiva e σ_a^2 é a variância aditiva poligênica, m_j é o escalar referente ao efeito fixo do j-ésimo marcador;

$\boldsymbol{W_j}$ é o vetor de incidência do j-ésimo marcador;

$\boldsymbol{e}$ $(N \times 1)$ é o vetor de erros do modelo com $\boldsymbol{e} \sim N(0; \sigma_e^2)$ sendo σ_e^2 a variância do erro.

A matriz de parentesco genômica aditiva é dada por (VAN RADEN, 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\sum_{j=1}^{n} 2p_j q_j}.$$

em que p_j e q_j são as frequências dos alelos A e a do j-ésimo marcador, respectivamente, e $\boldsymbol{W}$ ($N \times n$ sendo n o número de marcadores) é a matriz de incidência de marcadores codificada como apresentado a seguir (VITEZICA et al., 2013; RESENDE et al., 2014):

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

sendo $\boldsymbol{W_j}$ a j-ésima coluna da matriz $\boldsymbol{W}$.

Para a estimação dos efeitos genéticos aditivos poligênicos e o efeito do j-ésimo marcador pode-se utilizar as equações de modelos mistos (HENDERSON, 1973) dadas por:

$$\begin{bmatrix} \boldsymbol{1'1} & \boldsymbol{1'Z_1} & \boldsymbol{1'W_j} \\ \boldsymbol{Z_1'1} & \boldsymbol{Z_1'Z_1} + \boldsymbol{G^{-1}}\frac{\sigma_e^2}{\sigma_a^2} & \boldsymbol{Z_1'W_j} \\ \boldsymbol{W_j'1} & \boldsymbol{W_j'Z_1} & \boldsymbol{W_j'W_j} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\boldsymbol{a}} \\ \widehat{m}_j \end{bmatrix} = \begin{bmatrix} \boldsymbol{1'y} \\ \boldsymbol{Z_1'y} \\ \boldsymbol{W_j'y} \end{bmatrix},$$

em que os componentes de variância, σ_a^2 e σ_e^2, são estimados via o método REML (*Restricted maximum likelihood*) proposto por Patterson e Thompson (1971).

Após a estimação do efeito do j-ésimo marcador m_j é realizado o teste de Wald para o mesmo, visando testar a existência de associação com significância estatística entre o j-ésimo marcador e o fenótipo. Dessa forma, a hipótese de nulidade (H_0) é definida como "o j-ésimo marcador não apresenta qualquer efeito sobre o fenótipo", e a hipótese alternativa (H_a) definida como "o j-ésimo marcador afeta o fenótipo", ou seja, o j-ésimo marcador e o QTL encontram-se em LD. No entanto, essa análise estatística sofre com a ocorrência de uma alta taxa de falsos positivos, o qual consiste em declarar o efeito de um marcador como significativo, quando na verdade este marcador não está em LD com o QTL, devido a ocorrência de testes múltiplos. Uma alternativa para controlar esse fato é monitorar o número de falsos positivos em relação ao número total de resultados positivos por meio da taxa de descobertas falsas (*False Discovery Rate* - FDR) conforme apresentado por Fernando et al. (2004). Além disso, uma maneira de se considerar a FDR no teste de significância é por meio de uma correção no *p-value* associado ao teste, denominado de *q-value* (STOREY e TIBSHIRANI, 2003).

Além da elevada taxa de falsos positivos, este método pode sofrer com a superestimação dos efeitos. Ademais, estudos como os de Onteru et al. (2010), Fernando e Garrick (2013) e Fernando et al. (2017) também têm indicado que marcadores únicos explicam uma pequena fração da variância genética quando se considera características quantitativas e, assim, podem não mostrar uma associação forte entre o marcador e o fenótipo, indicando baixo poder em detecção. Dessa forma, considerar grupos ou regiões de marcadores conjuntamente tem sido apontado como uma grande vantagem pois esses conjuntos tendem a capturar uma proporção maior da variância genética explicada pelos marcadores e, assim, identificar relações mais complexas entre eles (MOORE et al., 2010).

Dentre os critérios existentes de seleção de regiões genômicas associadas às características de interesse podemos citar a seleção pela porcentagem da variância explicada por regiões genômicas (%var), o critério de seleção de *tag*SNPs proposto por Sollero et al. (2017) (*tag*SNPs) e a seleção com base na probabilidade *a posteriori* da associação de regiões genômicas (WPPA - *Window Posterior Probability of Association*) implementada por Fernando e Garrick (2013) e Fernando et al. (2017). Todas estas abordagens selecionam regiões considerando a variância genética obtida por meio dos efeitos de marcadores estimados via métodos bayesianos e se diferenciam na maneira pela qual as regiões são selecionadas e nos *thresholds* estabelecidos para esta seleção.

Para também detectar regiões potencialmente associadas, foi proposta no segundo capítulo desta tese a medida PP_{int} (*Posterior Probability of Interval*). Essa medida é calculada pela razão entre o número de iterações das cadeias MCMC (*Markov chain Monte Carlo*) em que a região possui pelo menos um SNP com magnitude de efeito superior ao valor do terceiro quartil, considerando toda a distribuição dos efeitos absolutos naquela iteração, e o número total de iterações salvas. As regiões com valores de PP_{int} superiores a um *threshold* especificado são selecionadas e consideradas como regiões associadas.

São apresentadas a seguir descrições metodológicas básicas sobre a importância da bayesiana nos estudos de associação, o método bayesiano BayesDπ e os critérios de seleção de regiões genômicas citados acima.

1.3 Métodos Bayesianos

1.3.1 Importância nos estudos de GWAS

Na atualidade, a maioria dos estudos em GWAS apresentam o número de marcadores disponíveis para análise muito maior que o número de observações fenotípicas. Assim, a utilização de modelos de regressão múltipla conjuntamente com o método de estimação de mínimos quadrados ordinários não podem ser usados diretamente para estimar simultaneamente os efeitos de todos os marcadores, o que consiste no cenário ideal para o processo de estimação em GWAS uma vez que se considera o LD entre todos os marcadores. Para resolver esse problema, vários métodos de regularização já propostos foram estendidos ao contexto da GWAS (SAMPSON et al., 2013; JIANG et al., 2016; BAO e WANG, 2017; HUANG et al., 2017) tais como os métodos LASSO (*Least Absolute Shrinkage and Selection Operator*) introduzido por Tibshirani (1996), *Elastic net* proposto por Zou e Hastie (2005) e MCP (*Minimax Concave Penalty*) apresentado por Zhang et al. (2010).

Embora os métodos utilizando abordagem bayesiana tenham sido propostos originalmente para os estudos de seleção genômica ampla (*Genome Wide Selection* - GWS), eles também podem ser usados vantajosamente na GWAS (FAN et al., 2011; SUN et al., 2011). Isso ocorre pois os métodos bayesianos podem estimar conjuntamente os efeitos dos marcadores sobre os fenótipos de características observadas e por meio deles também é possível incorporar informações *a priori* sobre estes efeitos (MEUWISSEN et al., 2001).

Comparados aos métodos de regularização, para selecionar SNPs ou grupos de SNPs associados à característica de interesse, os métodos utilizando abordagem bayesiana são preferíveis, uma vez que para os métodos clássicos, podem zerar os efeitos de muitos marcadores a depender do número de indivíduos, as inferências em regiões genômicas são computacionalmente muito desafiadoras pois, geralmente, requerem análises repetidas dos dados com *bootstrap* ou amostras permutadas para obter níveis de significância para os testes (HAYES et al., 2010; FAN et al., 2011; FERNANDO et al., 2017). Além disso, os métodos bayesianos têm a vantagem de quantificar a incerteza e combinar informações *a priori* (ZHAO et al., 2019). Sua aplicação recente no contexto de GWAS também mostrou um maior poder de detecção, ajustando simultaneamente os efeitos dos marcadores, corrigindo implicitamente estruturas biológicas e capturando maiores porcentagens de variância genética, consequentemente, sendo superiores para fazer inferências (SAHANA et al., 2010; DASHAB et al. 2012; FERNANDO e GARRICK, 2013).

Os métodos bayesianos mais indicados para a associação genômica são aqueles que apresentam variância específica para cada marcador, seleção de variáveis e aprendizado bayesiano (RESENDE et al., 2014). Metodologias, tais como regressão *ridge* bayesiana

(*Bayesian Ridge Regression* - BRR), BayesA, BayesB e BayesDπ, vêm se mostrando eficientes para a seleção de grupos de SNPs relevantes no melhoramento animal e vegetal. Porém, quando comparado a outros métodos, o método BayesDπ apresenta vantagens teóricas para GWAS, as quais serão apresentadas a seguir.

1.3.2 Método BayesD$\boldsymbol{\pi}$

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

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{m} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$, μ e $\boldsymbol{e}$ já foram definidos anteriormente;

$\boldsymbol{m}$ é o vetor de efeitos genéticos aditivos dos marcadores com dimensão $n \times 1$ sendo n o número de marcadores;

$\boldsymbol{W}$ $(N \times n)$ é a matriz de incidência que relaciona os efeitos aditivos dos marcadores aos fenótipos contidos em $\boldsymbol{y}$ e é codificada como apresentado anteriormente.

A distribuição dos dados é denotada por meio de $\boldsymbol{y}|\mu, \boldsymbol{m}, \sigma_e^2 \sim \text{MVN}(\mathbf{1}\mu + \boldsymbol{W}\boldsymbol{m}, \boldsymbol{I}\sigma_e^2)$ sendo que MVN representa a distribuição normal multivariada. No entanto, a inferência Bayesiana trata o vetor de parâmetros desconhecidos do modelo como quantidades aleatórias e qualquer informação inicial sobre elas pode ser representada por meio de modelos probabilísticos. Assim, é assumido distribuições de probabilidade, que neste contexto são denominadas distribuições *a priori*, a todas as quantidades desconhecidas, tais como:

$p(\mu) \propto k$ (distribuição *a priori flat* ou não informativa);

$$\boldsymbol{e}|\sigma_e^2 \sim MVN(0, \boldsymbol{I}\sigma^2);$$

$$\sigma_e^2 \sim v_e S_e^2 \chi^{-2};$$

em que k é uma constante e $v_e S_e^2 \chi^{-2}$ representa a distribuição qui-quadrado invertida escalada com os parâmetros v_e e S_e^2.

O vetor de parâmetros $\boldsymbol{m}$ do modelo que representa o vetor de efeitos de marcadores também assumem distribuições de probabilidade *a priori*. As diferentes distribuições *a priori* assumidas para os efeitos de marcadores dão origem aos diversos métodos bayesianos propostos. Desta forma, ao considerar *a priori* que os efeitos apresentam normalidade e que: (i) todos os marcadores dispõem de uma mesma variância genética σ_m^2, o método é denominado de regressão *ridge* bayesiana (BRR); (ii) cada marcador dispõe de uma variância genética

específica $\sigma^2_{m_j}$ (variância do j-ésimo marcador), o método é denominado de BayesA; (iii) uma porcentagem π de marcadores apresentam variância genética específica sendo π escolhido subjetivamente pelo pesquisador e uma porcentagem $1-\pi$ de marcadores apresentam variância genética igual a zero ($\sigma^2_{m_j}=0$), o método é denominado de BayesB. A variância comum σ^2_m no método BRR apresenta distribuição qui-quadrado invertida escalada com os hiperparâmetros -2 e S^2_m, sendo estes escolhidos também subjetivamente. Já a variância específica $\sigma^2_{m_j}$ nos métodos BayesA e BayesB apresentam distribuição qui-quadrado invertida escalada com os hiperparâmetros υ e S^2_m, sendo estes escolhidos também subjetivamente.

Já o método BayesDπ, que nesta tese será o método bayesiano utilizado, apresenta as seguintes distribuições de probabilidade *a priori*:

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

$$\sigma^2_{m_j}|\pi \sim Escalada-inv-\chi^2(\upsilon,S^2_m),$$

em que $\pi|m_j,\sigma^2_{m_j},\sigma^2_e,S^2_m \sim U[0,1]$, $S^2_m \sim Gama(\alpha,\beta)$ sendo α e β os hiperparâmetros da distribuição *a priori* e a média *a posteriori* da variância genética aditiva total dos marcadores é dada por $\hat{\sigma}^2_a = \sum_j 2p_jq_j\hat{\sigma}^2_{m_j}$ sendo $\hat{\sigma}^2_{m_j}$ a média *a posteriori* da variância genética associada a cada marcador.

Note que neste método considera-se que uma fração $(1-\pi)$ dos marcadores não possui efeito sobre a característica e a fração restante π possui efeitos com distribuição *a priori* dada por uma normal com variância específica $\sigma^2_{m_j}$ para cada marcador sendo a variância de cada marcador e no BayesDπ a quantidade π apresenta distribuição de probabilidade não sendo mais escolhida subjetivamente.

Já o hiperparâmetro S^2_m pode conduzir a estimativas dos efeitos viesadas dependendo da sua magnitude. Sendo assim, valores elevados de S^2_m conduzem a uma superestimação das estimativas dos efeitos e um pequeno valor de S^2_m leva a uma subestimação das mesmas, devido ao demasiado encolhimento dos efeitos. Desta forma no BayesDπ, o hiperparâmetro S^2_m é considerado como parâmetro desconhecido.

O método BayesDπ possibilita que uma porcentagem $(1-\pi)$ de efeitos de marcadores sejam iguais à zero, conduzindo a um menor número de efeitos de marcadores a serem estimados, tornando mais acurado o processo de estimação uma vez que muitos dos marcadores não possuem efeitos genéticos, ou sejam não estão em LD com o QTL (HABIER et al., 2011).

Consequentemente, o BayesDπ pode ser vantajosamente utilizado na GWAS em relação às outras abordagens bayesianas pois, como assume variância específica para cada locos, permite a obtenção de informações a respeito da arquitetura genética da característica, estima o valor da probabilidade π que é considerada como desconhecida e considera que a maioria dos marcadores possuem efeitos zero, exceto por alguns mais próximos à mutação causal que teriam efeitos maiores, dando mais sentido biológico para as análises (HABIER et al., 2011).

Após a estimação simultânea dos efeitos dos marcadores é necessário a identificação da associação entre marcador e QTL e o principal procedimento bayesiano para esta identificação é o Fator de Bayes (*Bayes Factor* – BF) proposto por Kass e Raftery (1995) (HEATH, 1997; VARONA et al. 2001; HABIER et al. 2011; LEGARRA et al. 2015). No entanto, na prática como são considerados milhares de marcadores nas análises e devido a utilização de cadeias de *Markov,* em que são realizadas muitas iterações até atingir a convergência em cada análise, o BF se torna inviável pois exige uma tarefa computacional complexa, elevado tempo de execução bem como problemas no armazenamento de informações. A partir disso, critérios utilizando abordagem bayesiana para seleção de SNPs e regiões associados a QTL e que não exigem grandes esforços computacionais veem sendo propostos e são apresentados a seguir.

1.4 Critérios de Seleção de Regiões ou SNPs associados

De acordo com Fernando et al. (2017) em um painel de SNPs de alta densidade, espera-se que muitos SNPs dentro de uma pequena região genômica sejam altamente correlacionados entre si e com qualquer QTL que esteja próximo deles e, portanto, qualquer SNP sozinho pode contribuir pouco para explicar a variabilidade de um QTL. Mesmo que cada SNP muito próximo a outro esteja fracamente associado a um QTL, espera-se que SNPs conjuntamente possam explicar muito mais a variabilidade de um QTL do que qualquer SNP por si só (ONTERU et al., 2010; FAN et al., 2011). Além disso, grupos de marcadores podem estar em maior desequilíbrio de ligação com QTL do que os marcadores únicos e devido a isso o LD entre o QTL e as regiões genômicas é aumentado, aumentando assim o poder de detecção dessas variantes causais (HAYES, 2013). De acordo com Zhao et al. (2019), vários SNPs causais podem estar localizados em uma única região, cada um com um pequeno efeito e para aumentar o poder de mapeá-los, reforçando a ideia de que é desejável considerá-los simultaneamente e realizar análises com este conjunto de SNPs.

Adicionalmente, segundo Beissenger et al. (2015), além de ocasionar aumento no poder estatístico, considerar regiões genômicas para os estudos de GWAS pode simplificar as análises computacionais, reduzir o ruído de amostragem e reduzir o número total de testes a serem realizados. Portanto, considerar e selecionar grupos de marcadores para detectar QTL nos estudos de GWAS tem sido apontado como uma grande vantagem para os programas de melhoramento (HAYES et al., 2010; MOORE et al., 2010; SAHANA et al. 2010; VISSCHER et al. 2010; FAN et al. 2011; FERNANDO et al., 2017).

Em estudos de simulação, os tamanhos das regiões genômicas podem ser determinados com base no desequilíbrio de ligação médio entre os marcadores e o próprio QTL (ZHAO et al., 2007; VIANA et al. 2016). Os valores de LD podem variar entre zero e um, referindo-se, respectivamente, a ausência e ao LD completo. A menor distância que fornece LD de acordo com uma extensão estabelecida pelo pesquisador pode ser utilizada para definir o tamanho da região dentro de cada cromossomo. A partir disso, visto a importância de seleção de regiões genômicas nos estudos de GWAS e devido a superioridade em utilizar os métodos bayesianos nessas análises são apresentados abaixo critérios de seleção de regiões/SNPs utilizando os efeitos de SNPs estimados via métodos bayesianos.

1.4.1 Proporção da variância genética explicada por regiões genômicas - %var

A seleção de grupos de SNPs por meio da proporção da variância genética explicada por regiões genômicas, representada por %var, foi proposta inicialmente por Wang et al. (2014) como medida de eficiência para comparar métodos de seleção de regiões na GWAS sob enfoque bayesiano. Para os efeitos de SNPs estimados via abordagem bayesiana tem-se que a variância genética associada à k-ésima região pode ser calculada por meio de:

$$\hat{\sigma}^2_{g_k} = \sum_{j \epsilon k} 2p_j q_j \hat{m}_j^2$$

em que que $\hat{m}_j$ é a média *a posteriori* do efeito de substituição alélica do j-ésimo SNP pertencente a k-ésima região estimado via método bayesiano. A porcentagem de explicação da variância genética de cada região é obtida da seguinte forma:

$$\%var = \frac{\hat{\sigma}^2_{g_k}}{\hat{\sigma}^2_g} \times 100$$

sendo $\hat{\sigma}^2_g$ a média *a posteriori* da variância genética total, ou seja, considerando todos os marcadores.

As regiões que apresentarem valores de proporção da variância genética superiores à razão entre a média *a posteriori* da variância genética total e o número de regiões considerado são selecionadas como regiões associadas e podem, posteriormente, serem utilizadas para explorar e determinar possíveis locos de características quantitativas.

1.4.2 Seleção de *tag*SNPs

Essa abordagem consiste em utilizar métodos bayesianos para identificar regiões associadas no genoma e, posteriormente, selecionar SNPs nessas regiões que supostamente possuem alto LD com o QTL da característica de interesse. Neste método, os SNPs são alocados em regiões genômicas, não sobrepostas e assim, as regiões que potencialmente contêm QTL associados com a característica de interesse são denominadas de *top-windows*.

As *top-windows* são identificadas com base em um limiar definido em termos da contribuição da variância genética dos marcadores que pode ser obtida por meio dos efeitos estimados de todos os SNPs via método bayesiano. Sollero et al (2017) em estudos de seleção de SNPs para a característica poligênica de resistência à carrapatos em raças brasileiras de bovinos *Braford* e *Hereford* consideraram como *top-windows* todas as regiões que explicaram proporções da variância genética maiores que cinco vezes o limiar descrito por Schurink et al. (2012), que é obtido por meio do quociente entre a média *a posteriori* da variância genética total pelo número de regiões consideradas. De acordo com Onteru et al. (2013), em seus estudos considerando características de ganho de gordura e área muscular do lombo em suínos, QTL candidatos podem ser considerados se a porcentagem esperada de variância genética contida em uma região genômica for pelo menos cinco vezes maior do que a razão de 100% pelo número de janelas propostas considerando todo o genoma. No entanto, o limiar considerado para a proporção da variância genética em um estudo pode depender da espécie e da característica analisada.

Os efeitos de SNPs ditos associados dentro de cada *top-window* são denominados de *tag*SNPs, e podem ser selecionados utilizando as frequências de inclusão (FI) dos SNPs no modelo, ou seja, razão entre o número de iterações salvas das cadeias MCMC que incluem o SNP em questão no modelo e o número total de iterações salvas. Assim, para cada *top-window*, calcula-se os valores de FI para todos os SNPs e os *tag*SNPs são selecionados como os SNPs que possuem o maior valor de FI.

Para também selecionar os *tag*SNPs dentro de cada *top-window* a estatística *t-like* (TL) pode ser utilizada considerando uma abordagem semelhante ao FI para avaliar a consistência dos efeitos dos SNPs. Essa medida é dada pelo valor absoluto dos efeitos médios *a posteriori* dos marcadores (apenas para as cadeias que incluíram o SNP no modelo) dividido pelos respectivos desvios-padrão desses efeitos, conduzindo a uma aproximação da estatística *t* de *student*. Com base nessa estatística, os SNPs considerados significativos (*p-valor* < 0,05) dentro de cada *top-window* são também estabelecidos como *tag*SNPs.

1.4.3 Seleção pela probabilidade *a posteriori* da associação da região genômica – WPPA

A medida WPPA pode ser implementada utilizando os efeitos de SNPs estimados via método bayesiano e é obtida com base na proporção da variância genética explicada pelos marcadores de cada região genômica. A variância genômica associada a *k*-ésima região pode ser estimada, neste contexto, por meio de:

$$\hat{\sigma}^2_{g_k} = \sum_{j \epsilon k} 2p_j q_j \hat{m}^2_j$$

em que $\hat{m}_j$ é a média *a posteriori* do efeito de substituição alélica do j-ésimo SNP pertencente a k-ésima região estimado pelo método bayesiano. A partir disso, a proporção da variância genética explicada pelos marcadores na k-ésima região, denotada por q_k, é definida como:

$$q_k = \frac{\hat{\sigma}^2_{g_k}}{E(\sigma^2_{g_k})}$$

em que $E\left(\sigma^2_{g_k}\right) = \sum_{j \epsilon k} 2p_j\, q_j\, E(m^2_j)$ (na ausência de dominância), sendo $E\left(m^2_j\right) = \frac{(\sigma^2_g)}{n\bar{H}}$ e σ^2_g a variância genética dos marcadores, n o número de SNPs e $\bar{H}$ a média de $2p_j q_j$. Caso $q_k > 1$, existe a presença de uma mutação causativa dentro da *k*-ésima região, uma vez que apresenta efeito maior do que o esperado sob a hipótese de uma distribuição igual da variância genética ao longo do genoma (PETERS et al., 2012; BENNEWITZ et al., 2017). Dessa forma, a medida WPPA é obtida pela razão entre o número de amostras em que q_k é maior que 1 e o número total de iterações salvas.

As regiões que possuem WPPA acima de um *threshold* pré-estabelecido são selecionadas como regiões associadas. Segundo Fernando e Garrick (2013) e Fernando et al. (2017), caso se utilize valores de WPPA superiores a 0,95 para declarar regiões associadas, isto resultará em uma proporção de falsos positivos inferior a 0,05. Já Bennewitz et al. (2017)

considerou os níveis 0,85, 0,95 e 0,99 e verificou que o poder em detectar uma região associada diminuiu com o aumento desses níveis.

Além das abordagens bayesianas, uma metodologia estatística alternativa, denominada mapeamento de herdabilidades regionais (*Regional heritability mapping* - RHM), foi proposta por Nagamine et al. (2012) e visa também determinar as regiões do genoma que estão associadas ao valor genético. O RHM vem mostrando maior poder para a detecção de QTL verdadeiros e reduzidas taxas de falsos positivos, quando comparada com as metodologias tradicionais de GWAS (USAI et al., 2014).

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

O método de mapeamento de herdabilidades regionais (*Regional Heritability Mapping* - RHM) foi proposto por Nagamine et al. (2012) e por Riggio et al. (2013) e utiliza o seguinte modelo para a estimação do efeito da k-ésima região no fenótipo:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_{2k} r_k} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$, μ, $\boldsymbol{a}$, $\boldsymbol{Z_1}$ e $\boldsymbol{e}$ já foram definidos anteriormente;

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

A matriz $\boldsymbol{G_{reg_k}}$ utiliza o subconjunto da matriz de incidência de marcadores $\boldsymbol{W}$ referente a k-ésima região e pode ser construída analogamente à matriz $\boldsymbol{G}$ dada anteriormente.

As equações de modelos mistos propostas por Henderson (1973) são utilizadas para estimar os efeitos genéticos poligênicos e o efeito da k-ésima região do genoma no fenótipo são dadas por:

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'Z_1} & \mathbf{1'Z_{2k}} \\ \boldsymbol{Z_1'1} & \boldsymbol{Z_1'Z_1} + \boldsymbol{G^{-1}}\frac{\sigma^2_e}{\sigma^2_a} & \boldsymbol{Z_1'Z_{2k}} \\ \boldsymbol{Z_{2k}'1} & \boldsymbol{Z_{2k}'Z_1} & \boldsymbol{Z_{2k}'Z_{2k}} + \boldsymbol{G^{-1}_{reg_k}}\frac{\sigma^2_e}{\sigma^2_{r_k}} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{a}} \\ \widehat{\boldsymbol{r_k}} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \boldsymbol{Z_1'y} \\ \boldsymbol{Z_{2k}'y} \end{bmatrix},$$

em que os componentes de variância σ^2_a, $\sigma^2_{r_k}$, e σ^2_e, são estimados via método REML (PATTERSON e THOMPSON, 1971).

O modelo definido acima considera a presença da k-ésima região ($\boldsymbol{r_k}$) e por isso é definido como modelo completo. Já o modelo sem a presença da região, ou seja, o modelo $\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{e}$, é considerado como modelo restrito. Dessa forma, para testar o efeito da k-ésima região no fenótipo, os modelos podem ser comparados utilizando o teste da razão de verossimilhança (TRV) considerando o logaritmo da função de verossimilhança ($ln\, L$), como apresentado a seguir:

$$TRV = 2ln\left(\frac{L_1}{L_0}\right)$$

sendo L_1 a função de verossimilhança para o modelo completo e L_0 a função de verossimilhança para o modelo restrito. A hipótese de nulidade (H_0) para este teste é de que os modelos não diferirem entre si, indicando que a k-ésima região não apresenta efeito sobre o fenótipo. Enquanto, que a hipótese alternativa (H_a) é definida como os modelos diferem entre si, determinando que a k-ésima região e o QTL encontram-se em LD. Dado que foram determinadas K regiões no genoma tem-se que os procedimentos de estimação via equações de modelos mistos e TRV devem ser realizados K vezes.

2 GWAS considerando efeitos de dominância

2.1 Definição e Importância

Os efeitos de dominância muitas vezes não são considerados nos estudos de GWAS para o melhoramento genético animal e vegetal. Um dos motivos pode ser a falta de informações genômicas disponíveis e também devido a indisponibilidade de grandes conjuntos de dados com proporções suficientes de indivíduos com níveis de efeito de dominância não nulos (ERTL et al., 2014). No entanto, os efeitos de dominância são fontes de variação de características complexas e que não deveriam ser negligenciados (VAN TASSEL et al., 2000; SU et al., 2012; ERTL et al., 2014; BENNEWITZ et al., 2017). Para Mao et al. (2020) entender a arquitetura genética na presença de efeitos de dominância é útil para planejar estratégias de melhoramento e aumentar os ganhos genéticos. Já segundo Lopes et al. (2014) e Zhou et al. (2012) os principais benefícios dos efeitos da dominância no melhoramento genético animal e vegetal são esperados no cruzamento, uma vez que a dominância tem sido sugerida como um dos mecanismos genéticos que explicam a heterose.

Os efeitos de dominância ocorrem quando os efeitos dos alelos de um determinado loco não são somente aditivos, mas interagem entre si de modo que o valor do genótipo heterozigoto desvia-se da média dos valores dos genótipos homozigotos (FALCONER e MACKAY, 1996). A variância genética atribuída aos desvios de dominância, muitas vezes, também é ignorada, normalmente por não possuir aplicabilidade para prever a resposta à seleção. De acordo com Varona et al. (1998) em seus estudos para característica de estatura em gado bovino, embora as estimativas sejam escassas, essa variância, geralmente, representa porcentagens consideráveis da variação fenotípica total. Além disso, Huang e Mackay (2016) relatam que as ações gênicas dominantes além de contribuírem para os componentes da variância genética devido à dominância podem também contribuir nas estimativas dos componentes de variância genética aditiva. No entanto, segundo esses autores a quantidade específica de variação genética que cada tipo de ação gênica (aditiva, dominante e epistática) contribui depende da arquitetura genética ou pode até ser imensurável porque as ações gênicas podem não ser independentes uma das outras.

A disponibilidade de painéis de alta densidade de SNPs ofereceu também novas oportunidades para detecção e utilização dos efeitos de dominância (LOPES et al., 2014). Do ponto de vista computacional, a investigação do efeito de dominância usando uma abordagem baseada em genômica é muito mais simples do que usando abordagens baseadas em *pedigree* devido ao fato dos heterozigotos e homozigotos serem diretamente distinguidos por informações genômicas (MAO et al., 2020).

No melhoramento vegetal há uma carência em estudos genômicos com inclusão de dominância, no entanto, a contribuição da dominância para a variação genética das características é essencial em espécies de propagação vegetativa e em populações cruzadas como os híbridos (DENIS e BOUVET, 2013). O estudo de Yang et al. (2014) em milho enfatiza essa importância mostrando um aumento na proporção da herdabilidade explicada ao considerar a dominância, permitindo assim uma melhor visão geral da heterose. No melhoramento animal essa situação também não é diferente, Bolormaa et al. (2015) conduziram um estudo de associação incluindo a dominância utilizando regressão via marcas únicas e encontraram SNPs associados aos efeitos de dominância em características de crescimento, de carcaça e de fertilidade de bovinos de corte. Enquanto que Bennewitz et al. (2017) realizaram um estudo de associação incluindo a dominância utilizando um método bayesiano, o qual denominaram de BayesD, e verificaram um aumento no poder de detecção para genes causadores que explicam mais de 2,5% da variância genética, considerando dados simulados e dados reais referentes a

gado de leite. Adicionalmente, Mao et al. (2020) em estudos de GWAS em novilhas verificou que os efeitos de dominância são importantes para a arquitetura genética da característica fertilidade. Neste estudo, os autores concluíram que ambos, efeitos aditivos e devido à dominância, foram significativos e que provavelmente eles afetam a característica estudada conjuntamente.

Considerando a inclusão da dominância nos estudos de GWAS são descritos abaixo, no contexto aditivo-dominante, as metodologias já apresentadas acima.

2.2 Métodos considerando o modelo aditivo-dominante

2.2.1 Análise via Marcas Únicas

O seguinte modelo misto em marcas simples pode ser empregado, visando estimar os efeitos aditivos e devido à dominância do j-ésimo marcador no fenótipo:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{w_j} m_{aj} + \boldsymbol{s_j} m_{dj} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$, μ, $\boldsymbol{a}$, $\boldsymbol{Z_1}$ e $\boldsymbol{e}$ já foram definidos anteriormente;

$\boldsymbol{d}$ é o vetor de efeitos genéticos poligênicos devido à dominância com matriz de incidência $\boldsymbol{Z_2}$ ($N \times N$), sendo $\boldsymbol{d} \sim N(0, \boldsymbol{D}\sigma_d^2)$ em que $\boldsymbol{D}$ é a matriz de parentesco genômico devido à dominância e σ_d^2 é a variância poligênica devido à dominância;

m_{aj} é o escalar referente ao efeito fixo aditivo do j-ésimo marcador com vetor de incidência do j-ésimo marcador dado por $\boldsymbol{w_j}$ ($N \times 1$);

m_{dj} é o escalar referente ao efeito fixo devido à dominância do j-ésimo marcador com vetor de incidência do j-ésimo marcador dado por $\boldsymbol{s_j}$ ($N \times 1$).

O vetor de incidência aditivo do j-ésimo marcador $\boldsymbol{w_j}$ representa a j-ésima coluna da matriz $\boldsymbol{W}$, enquanto o vetor de incidência devido à dominância do j-ésimo marcador $\boldsymbol{s_j}$ representa a j-ésima coluna da matriz $\boldsymbol{S}$ definida a seguir:

$$\boldsymbol{S} = \begin{cases} Se\ AA, então\ 0 \rightarrow 2q^2 \\ Se\ Aa, então\ 1 \rightarrow 2pq \\ Se\ aa, então\ 0 \rightarrow -2p^2 \end{cases}.$$

A partir disso, a matriz de parentesco genômica devido à dominância é dada conforme Vitezica et al. (2013):

$$\boldsymbol{D} = \frac{\boldsymbol{SS'}}{\sum_{j=1}^{n}\left(2p_jq_j\right)^2},$$

O vetor de efeitos genéticos poligênicos, os efeitos aditivos e devido a dominância do j-ésimo marcador podem ser estimados via equações de modelos mistos (HENDERSON, 1973) dadas por:

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'Z_1} & \mathbf{1'Z_2} & \mathbf{1'w_j} & \mathbf{1's_j} \\ \mathbf{Z_1'1} & \mathbf{Z_1'Z_1} + \boldsymbol{G^{-1}}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{Z_1'Z_2} & \mathbf{Z_1'w_j} & \mathbf{Z_1's_j} \\ \mathbf{Z_2'1} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \boldsymbol{D^{-1}}\frac{\sigma_e^2}{\sigma_d^2} & \mathbf{Z_2'w_j} & \mathbf{Z_2's_j} \\ \mathbf{w_j'1} & \mathbf{w_j'Z_1} & \mathbf{w_j'Z_2} & \mathbf{w_j'w_j} & \mathbf{w_j's_j} \\ \mathbf{s_j'1} & \mathbf{s_j'Z_1} & \mathbf{s_j'Z_2} & \mathbf{s_j'w_j} & \mathbf{s_j's_j} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \widehat{\boldsymbol{a}} \\ \widehat{\boldsymbol{d}} \\ \widehat{m}_{aj} \\ \widehat{m}_{dj} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \\ \mathbf{w_j'y} \\ \mathbf{s_j'y} \end{bmatrix},$$

Após a estimação dos efeitos aditivos e dominantes do j-ésimo marcador é realizado o teste de Wald para os mesmos visando testar a existência de associação com significância estatística entre o j-ésimo marcador e o QTL. A hipótese nula (H_0) é definida como o j-ésimo marcador não apresenta efeito (aditivo ou devido à dominância) sobre o fenótipo e a hipótese alternativa (H_a) definida como o j-ésimo marcador afeta o fenótipo, ou seja, o j-ésimo marcador e o QTL encontram-se em LD. Por fim, pode ser também considerado os casos em que tanto os efeitos aditivos quanto os devido a dominância foram simultaneamente ditos como significativos.

2.2.2 Método Bayes Dπ com inclusão de efeitos devido à dominância

Meuwissen et al. (2001) apresentaram o seguinte modelo linear básico em nível de marcas:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Wm_a} + \boldsymbol{Sm_d} + \boldsymbol{e},$$

em que:

$\boldsymbol{y}$, $\mathbf{1}$, μ, $\boldsymbol{W}$, $\boldsymbol{S}$ e $\boldsymbol{e}$ já foram definidos anteriormente;

$\boldsymbol{m_a}$ e $\boldsymbol{m_d}$ são, respectivamente, os vetores de efeitos genéticos aditivos e devido à dominância dos marcadores.

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

$$\boldsymbol{y}|m_{a1}, m_{a2}, \ldots, m_{an}, m_{d1}, m_{d2}, \ldots, m_{dn} \sim N(\mathbf{1}\mu + \boldsymbol{Wm_a} + \boldsymbol{Sm_d}, \sigma_e^2);$$

$\sigma_e^2 \sim v_e s_e^2 \chi^{-2}$ (distribuição qui-quadrado invertida escalada)

sendo v_e e s_e^2 os hiperparâmetros.

A especificação das distribuições *a priori* com a inclusão dos efeitos devido á dominância é dada por:

$$m_{a_j}|\pi_a, \sigma^2_{m_{a_j}} \sim \pi_a N(0, \sigma^2_{m_{a_j}}) + (1-\pi_a)N(0, \sigma^2_{m_{a_j}} = 0);$$

$$m_{d_j}|\pi_d, \sigma^2_{m_{d_j}} \sim \pi_d N(0, \sigma^2_{m_{d_j}}) + (1-\pi_d)N(0, \sigma^2_{m_{d_j}} = 0);$$

$$\sigma^2_{m_{a_j}}|\pi_a \sim Scale - inv - X^2(v_a, S^2_{m_a}),$$

$$\sigma^2_{m_{d_j}}|\pi_d \sim Scale - inv - X^2(v_d, S^2_{m_d}),$$

em que $\pi_a|m_{a_j}, \sigma^2_{m_{a_j}}, \sigma^2_e, S^2_{m_a} \sim U[0,1]$, $\pi_d|m_{d_j}, \sigma^2_{m_{d_j}}, \sigma^2_e, S^2_{m_d} \sim U[0,1]$, $S^2_{m_a} \sim Gama(a,b)$ e $S^2_{m_d} \sim Gama(c,d)$ sendo $v_a, v_d, a, b, c\ e\ d$ os hiperparâmetros das distribuições *a priori*.

2.2.3 Seleção pela Probabilidade *a Posteriori* da Associação da Janela – WPPA

Para os efeitos de SNPs estimados via método bayesiana tem-se que a variância genética total associada à k-ésima região pode ser calculada por meio de:

$$\hat{\sigma}^2_{g_k} = \sum_{j=1}^{n} (2p_j q_j \hat{m}^2_{a_j} + (2p_j q_j)^2 \hat{m}^2_{d_j})$$

em que n é o número de marcadores contidos nesta região, $\hat{m}_{a_j}$e $\hat{m}_{d_j}$são, respectivamente, as médias *a posteriori* dos efeitos aditivos e devido à dominância do j-ésimo SNP pertencente a k-ésima região. A partir disso, tendo as cadeias de *Markov* de cada efeito de marcador ($m_{a_j}\ e\ m_{d_j}$) é possível estimar a probabilidade *a posteriori* de um grupo de marcadores (ou região) estar associada a característica analisada. Esta probabilidade foi proposta por Fernando e Garrick (2013) e é dada por:

$$q_k = \frac{\hat{\sigma}^2_{gk}}{\sum_{j=1}^{m} \left(H_j \frac{\hat{\sigma}^2_a}{n\overline{H}} + H_j^2 \frac{\hat{\sigma}^2_d}{n\overline{H^2}} \right)}$$

em que $\hat{\sigma}^2_a$ e $\hat{\sigma}^2_d$ são, respectivamente, as variâncias genéticas aditiva e devido à dominância estimadas considerando todos os marcadores, $H_j = 2p_j q_j$, $\overline{H} = \frac{1}{n}\sum_{j=1}^{n} H_j$ e $\overline{H^2} = \frac{1}{n}\sum_{j=1}^{n} H_j^2$. Como já mencionado anteriormente, a medida WPPA é obtida pela razão entre o número de iterações em que q_k é maior que 1 (um) e o número total de iterações salvas. As regiões que

possuem WPPA acima de um *threshold* pré-estabelecido são selecionadas como regiões significativas.

2.2.4 Probabilidade a Posteriori do Intervalo – $\boldsymbol{PP_{int}}$

Como já mencionado acima para apenas efeitos aditivos, a medida PP_{int} foi proposta para detectar regiões associadas às características de interesse baseada nos efeitos de SNPs obtidos nas amostras MCMC. Além dos efeitos aditivos, para também selecionar SNPs com grandes efeitos de dominância a PP_{int} é calculada pela razão entre o número de iterações em que determinada região possui pelo menos um SNP com magnitude de efeito de dominância superior ao valor do terceiro quartil, considerando toda a distribuição dos efeitos absolutos de dominância naquela iteração, e o número total de iterações salvas.

As regiões com valores de PP_{int}, considerando efeitos aditivos e/ou dominantes superiores a *thresholds* pré-especificados são escolhidas como regiões associadas. Este limiar pode ser escolhido pelo investigador e reflete diretamente na probabilidade *a posteriori* de um QTL estar na região.

2.2.5 Mapeamento de herdabilidade regionais – RHM

O método de mapeamento de herdabilidades regionais utiliza o seguinte modelo linear para as estimações dos efeitos da k-ésima região no fenótipo:

$$\boldsymbol{y} = \boldsymbol{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{Z_{3k} r_{dk}} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$, μ, $\boldsymbol{a}$, $\boldsymbol{d}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ e $\boldsymbol{e}$ já foram definidos anteriormente;

$\boldsymbol{r_{ak}}$ é o vetor de efeitos genômicos aditivos referente a k-ésima região do genoma $(N \times 1)$ com matriz de incidência $\boldsymbol{Z_{2k}}$ $(N \times N)$, sendo $\boldsymbol{r_{ak}} \sim N\left(0, \boldsymbol{G_{reg_k}} \sigma^2_{ar_k}\right)$ em que $\boldsymbol{G_{reg_k}}$ é a matriz de parentesco genômico aditiva e $\sigma^2_{ar_k}$ é a variância aditiva referentes a k-ésima região ($k = 1,2, \dots, K$, em que K é o número total de regiões determinadas);

$\boldsymbol{r_{dk}}$ é o vetor de efeitos genômicos devido à dominância referente a k-ésima região do genoma $(N \times 1)$ com matriz de incidência $\boldsymbol{Z_{3k}}$ $(N \times N)$, sendo $\boldsymbol{r_{dk}} \sim N(0, \boldsymbol{D_{reg_k}} \sigma^2_{dr_k})$ em que $\boldsymbol{D_{reg_k}}$ é a matriz de parentesco genômico devido à dominância e $\sigma^2_{dr_k}$ é a variância devido à dominância referentes a k-ésima região ($k = 1,2, \dots, K$).

As matrizes $\boldsymbol{G_{reg_k}}$ e $\boldsymbol{D_{reg_k}}$ utilizam, respectivamente, um subconjunto da matriz de incidência de marcadores W e S referente a k-ésima região e pode ser construída analogamente as matrizes G e D dadas anteriormente.

O vetor de efeitos genéticos poligênicos e os vetores de efeitos genômicos aditivos e devido à dominância, referente a k-ésima região do genoma, podem ser estimados via equações de modelos mistos (HENDERSON, 1973) dadas por:

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'Z_1} & \mathbf{1'Z_2} & \mathbf{1'Z_{2k}} & \mathbf{1'Z_{3k}} \\ \mathbf{Z_1'1} & \mathbf{Z_1'Z_1} + \mathbf{G^{-1}}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{Z_1'Z_2} & \mathbf{Z_1'Z_{2k}} & \mathbf{Z_1'Z_{3k}} \\ \mathbf{Z_2'1} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{D^{-1}}\frac{\sigma_e^2}{\sigma_d^2} & \mathbf{Z_2'Z_{2k}} & \mathbf{Z_2'Z_{3k}} \\ \mathbf{Z_{2k}'1} & \mathbf{Z_{2k}'Z_1} & \mathbf{Z_{2k_j}'Z_2} & \mathbf{Z_{2k}'Z_{2k}} + \mathbf{G_{reg_k}^{-1}}\frac{\sigma_e^2}{\sigma_{ar_k}^2} & \mathbf{Z_{2k}'Z_{3k}} \\ \mathbf{Z_{3k}'1} & \mathbf{Z_{3k}'Z_1} & \mathbf{Z_{3k_j}'Z_2} & \mathbf{Z_{3k}'Z_{2k}} & \mathbf{Z_{3k}'Z_{3k}} + \mathbf{D_{reg_k}^{-1}}\frac{\sigma_e^2}{\sigma_{dr_k}^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{d}} \\ \hat{\boldsymbol{r}}_{\boldsymbol{ak}} \\ \hat{\boldsymbol{r}}_{\boldsymbol{dk}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{1'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \\ \mathbf{Z_{2k}'y} \\ \mathbf{Z_{3k}'y} \end{bmatrix},$$

sendo os componentes de variância, σ_a^2, σ_d^2, $\sigma_{ar_k}^2$, $\sigma_{dr_k}^2$ e σ_e^2, estimados via o método REML.

A descrição dos modelos (completo e reduzido) utilizados no RHM são apresentadas na Tabela 2.

Tabela 2 – Descrição dos modelos (completo e reduzido) para respectivos efeitos aditivo (A) e aditivo-dominante (AD) utilizados no mapeamento de herdabilidades regionais (RHM).

Modelo	Efeito	
AD	A	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}$
	D	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{3k} r_{dk}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}$
	AD	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{Z_{3k} r_{dk}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}$
A	A	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{e}$

$\boldsymbol{y}$**:** vetor de fenótipos; μ: média geral da característica; **1:** vetor de mesma dimensão de $\boldsymbol{y}$ com todos os elementos iguais a 1; $\boldsymbol{a}$**:** vetor de efeitos genéticos aditivos poligênicos com matriz de incidência $\boldsymbol{Z_1}$; $\boldsymbol{d}$: vetor de efeitos genéticos poligênicos devido à dominância com matriz de incidência $\boldsymbol{Z_2}$; $\boldsymbol{r_{ak}}$: vetor de efeitos genômicos aditivos referente a k-ésima região do genoma com matriz de incidência $\boldsymbol{Z_{2k}}$; $\boldsymbol{r_{dk}}$: vetor de efeitos genômicos devido à dominância referente a k-ésima região do genoma com matriz de incidência $\boldsymbol{Z_{3k}}$**;** $\boldsymbol{e}$**:** vetor de erros.

Os modelos são comparados, e consequentemente, é testada a significância do efeito da k-ésima região, pela mudança no logaritmo da função de verossimilhança ($ln\, L$), por meio do teste da razão de verossimilhança (TRV). A estatística TRV realizada entre os modelos (completo e reduzidos) é obtida pela seguinte expressão:

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

sendo L_1 a função de verossimilhança para o modelo completo e L_0 a função de verossimilhança para o modelo restrito. A hipótese de nulidade (H_0) para este teste foi de que os modelos não diferiram entre si, indicando que a k-ésima região não apresenta efeito (aditivo e/ou devido à dominância) sobre o fenótipo. Enquanto, que a hipótese alternativa (H_a) é definida como os modelos diferem entre si, determinando que a k-ésima região e o QTL encontram-se em LD. Dado que foram determinadas K regiões no genoma tem-se que os procedimentos de estimação via equações de modelos mistos e TRV devem ser realizados K vezes. Os *p-valores* associados aos efeitos das *K* regiões e encontrados por meio do TRV também podem ser corrigidos, conforme feito no procedimento anterior.

Referências

AGUILAR, I. et al. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. **Journal of dairy science**, v. 93, n. 2, p. 743-752, 2010.

BAO, M.; WANG, K. Genome-wide association studies using a penalized moving-window regression. **Bioinformatics**, v. 33, n. 24, p. 3887-3894, 2017.

BEISSINGER, T. M. et al. Defining window-boundaries for genomic analyses using smoothing spline techniques. **Genetics Selection Evolution**, v. 47, n. 1, p. 30, 2015.

BENNEWITZ, J. et al. Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. **Genetics Selection Evolution**, v. 49, n. 1, p. 7, 2017.

BOLORMAA, S. et al. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. **Genetics Selection Evolution**, v. 47, n. 1, p. 26, 2015.

BRAZ, C. U. et al. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. **BMC genetics**, v. 20, n. 1, p. 8, 2019.

DASHAB, G. R. et al. Comparison of linear mixed model analysis and genealogy-based haplotype clustering with a Bayesian approach for association mapping in a pedigreed population. In: BMC proceedings. **BioMed Central**, 2012. p. S4.

DE OLIVEIRA SILVA, R. M. et al. Genome-wide association study for carcass traits in an experimental Nelore cattle population. **PloS one**, v. 12, n. 1, p. e0169860, 2017.

DENIS, M.; BOUVET, J.M. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genetics and Genomes**, 9:37–51, 2013.

ERTL, Johann et al. Genomic analysis of dominance effects on milk production and conformation traits in Fleckvieh cattle. **Genetics Selection Evolution**, v. 46, n. 1, p. 40, 2014.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to quantitative genetics** (4th edition). Longman,1996.

FAN, B. et al. Genome-wide association study identifies loci for body composition and structural soundness traits in pigs. **PloS one**, v. 6, n. 2, p. e14726, 2011.

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

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

FERNANDO, R. L.; GARRICK, D. Bayesian methods applied to GWAS. In: Genome-wide association studies and genomic prediction. **Humana Press, Totowa**, NJ, 2013. p. 237-274.

GODDARD, M. E.; HAYES, B. J. Mapping genes for complex traits in domestic animals and their use in breeding programmes. **Nature Reviews Genetics**, v. 10, n. 6, p. 381-391, 2009.

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

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

HAYES, B. Overview of statistical methods for genome-wide association studies (GWAS). In: **Genome-wide association studies and genomic prediction**. Humana Press, Totowa, NJ, 2013. p. 149-169.

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

HEATH, S. C. Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. **American journal of human genetics**, v. 61, n. 3, p. 748, 1997.

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

HUANG, C. et al. FGWAS: Functional genome wide association analysis. **NeuroImage**, v. 159, p. 107-121, 2017.

HUANG, W.; MACKAY, T. F. C. The genetic architecture of quantitative traits cannot be inferred from variance component analysis. **PLoS genetics**, v. 12, n. 11, p. e1006421, 2016.

JIANG, Y.; HE, Y.; ZHANG, H. Variable selection with prior information for generalized linear models via the prior LASSO method. **Journal of the American Statistical Association**, v. 111, n. 513, p. 355-376, 2016.

KASS, R. E.; RAFTERY, A. E. Bayes factors. **Journal of the american statistical association**, v. 90, n. 430, p. 773-795, 1995.

LEGARRA, A. et al. A comparison of methods for whole-genome QTL mapping using dense markers in four livestock species. **Genetics Selection Evolution**, v. 47, n. 1, p. 6, 2015.

LOPES, M. S. et al. A genome-wide association study reveals dominance effects on number of teats in pigs. **PloS one**, v. 9, n. 8, 2014.

MAO, X. et al. Genome-wide association mapping for dominance effects in female fertility using real and simulated data from Danish Holstein cattle. **Scientific Reports**, v. 10, n. 1, p. 1-9, 2020.

MATIKA, O. et al. Genome-wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. **Genetics Selection Evolution**, v. 48, n. 1, p. 11, 2016.

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

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

MEUWISSEN, T.; HAYES, B.; GODDARD, M. Genomic selection: A paradigm shift in animal breeding. **Animal frontiers**, v. 6, n. 1, p. 6-14, 2016.

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bioinformatics challenges for genome-wide association studies. **Bioinformatics**, v. 26, n. 4, p. 445-455, 2010.

NAGAMINE, Y. et al. Localising loci underlying complex trait variation using regional genomic relationship mapping. **PloS one**, v. 7, n. 10, 2012.

ONTERU, S. K. et al. Whole-genome association analyses for lifetime reproductive traits in the pig. **Journal of Animal Science**, v. 89, n. 4, p. 988-995, 2011.

ONTERU, S. K. et al. Whole genome association studies of residual feed intake and related traits in the pig. **PloS one**, v. 8, n. 6, 2013.

OTYAMA, P. I. et al. Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. **BMC genomics**, v. 20, n. 1, p. 481, 2019.

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

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

RESENDE, M. D. V.; 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, Visconde do Rio Branco, 2014.

RESENDE, R. T. et al. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. **New Phytologist**, v. 213, n. 3, p. 1287-1300, 2017.

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.

RIGGIO, V. et al. Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. **Heredity**, v. 110, n. 5, p. 420-429, 2013.

SAHANA, G. et al. Comparison of association mapping methods in a complex pedigreed population. **Genetic epidemiology**, v. 34, n. 5, p. 455-462, 2010.

SAMPSON, J. N. et al. Controlling the local false discovery rate in the adaptive Lasso. **Biostatistics**, v. 14, n. 4, p. 653-666, 2013.

SANTANA, M. H. A. et al. Genome-wide association study for feedlot average daily gain in Nellore cattle (Bos indicus). **Journal of animal breeding and genetics**, v. 131, n. 3, p. 210-216, 2014.

SCHMID, M.; BENNEWITZ, J. Invited review: Genome-wide association analysis for quantitative traits in livestock–a selective review of statistical models and experimental designs. **Archives Animal Breeding**, v. 60, n. 3, p. 335-346, 2017.

SCHURINK, A. et al. Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. **Genetics Selection Evolution**, v. 44, n. 1, p. 31, 2012.

SHIRALI, M. et al. Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. **Heredity**, v. 116, n. 3, p. 333-338, 2016.
SOARES, A. C. C. et al. Multiple-trait genomewide mapping and gene network analysis for scrotal circumference growth curves in Brahman cattle. **Journal of animal science**, v. 95, n. 8, p. 3331-3345, 2017.

SOLLERO, B. P. et al. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. **Genetics Selection Evolution**, v. 49, n. 1, p. 49, 2017.

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

STRAM, D. O. Tag SNP selection for association studies. **Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society**, v. 27, n. 4, p. 365-374, 2004.

SU, G. et al. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. **PloS one**, v. 7, n. 9, 2012.

SUN, X. et al. Genomic breeding value prediction and QTL mapping of QTLMAS2010 data using Bayesian methods. In: **BMC proceedings**. BioMed Central, 2011. p. S13.

TIBSHIRANI, R. Regression shrinkage and selection via the lasso. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 58, n. 1, p. 267-288, 1996.

USAI, M. G. et al. XVI th QTLMAS: simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation. In: **BMC proceedings**. BioMed Central, 2014. p. S1.

VAN RADEN, P. M. Efficient Methods to compute genomic predictions. **Journal of Dairy Science**, 91: 4414-4423, 2008.

VAN TASSELL, C. P.; MISZTAL, I.; VARONA, L. Method R estimates of additive genetic, dominance genetic, and permanent environmental fraction of variance for yield and health traits of Holsteins. **Journal of dairy science**, v. 83, n. 8, p. 1873-1877, 2000.

VARONA, L. et al. Effect of full sibs on additive breeding values under the dominance model for stature in United States Holsteins. **Journal of dairy science**, v. 81, n. 4, p. 1126-1135, 1998.

VARONA, L.; GARCÍA-CORTÉS, L. A.; PÉREZ-ENCISO, M. Bayes factors for detection of quantitative trait loci. **Genetics Selection Evolution**, v. 33, n. 2, p. 133, 2001.

VISSCHER, P. M.; YANG, J.; GODDARD, M. E. A commentary on 'common SNPs explain a large proportion of the heritability for human height'by Yang et al.(2010). **Twin Research and Human Genetics**, v. 13, n. 6, p. 517-524, 2010.

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, H. et al. Genome-wide association mapping including phenotypes from relatives without genotypes in a single-step (ssGWAS) for 6-week body weight in broiler chickens. **Frontiers in Genetics**, v. 5, p. 134, 2014.

WANG, W. Y. S. et al. Genome-wide association studies: theoretical and practical concerns. **Nature Reviews Genetics**, v. 6, n. 2, p. 109-118, 2005.

WEI, L. et al. Genome-wide association analysis and differential expression analysis of resistance to Sclerotinia stem rot in Brassica napus. **Plant Biotechnology Journal**, v. 14, n. 6, p. 1368-1380, 2016.

WOLFE, M. D. et al. Genome-wide association and prediction reveals genetic architecture of cassava mosaic disease resistance and prospects for rapid genetic improvement. **The plant genome**, v. 9, n. 2, 2016.

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

YANG, J. et al. Dominant gene action accounts for much of the missing heritability in a gwas and provides insight into heterosis. **Genome-wide association studies to dissect the genetic architecture of yield-related traits in maize and the genetic basis of heterosis**, v. 1001, p. 44, 2014.

ZHANG, Cun-Hui et al. Nearly unbiased variable selection under minimax concave penalty. **The Annals of statistics**, v. 38, n. 2, p. 894-942, 2010.

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

ZHOU, G. et al. Genetic composition of yield heterosis in an elite rice hybrid. **Proceedings of the National Academy of Sciences**, v. 109, n. 39, p. 15847-15852, 2012.

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

ZOU, H.; HASTIE, T. Regularization and variable selection via the elastic net. **Journal of the royal statistical society: series B (statistical methodology)**, v. 67, n. 2, p. 301-320, 2005.

CAPÍTULO 2

Accepted as original paper to Scientia Agricola

EVALUATION OF BAYESIAN METHODS OF GENOMIC ASSOCIATION VIA CHROMOSOMIC REGIONS USING SIMULATED DATA

Leísa Pires Lima[1*], Camila Ferreira Azevedo[1], Marcos Deon Vilela de Resende[2], Moysés Nascimento[1], Fabyano Fonseca e Silva[3]

[1]Federal University of Viçosa – Dept. of Statistics, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG - Brazil.

[2] Embrapa Café, Federal University of Viçosa, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG - Brazil.

[3] Federal University of Viçosa – Dept. of Animal Science, Av. Peter Henry Rolfs, s/n - 36570-000 – Viçosa, MG - Brazil.

Abstract

The development of efficient methods for genome-wide association studies (GWAS) between quantitative trait loci (QTL) and genetic values are extremely important for animal and plant breeding programs. Bayesian approaches that aim to select regions of single nucleotide polymorphisms (SNPs) prove to be efficient, indicating genes with important effects. Among the selection criteria for SNPs or regions, selection criterion by the percentage of variance can be explained by genomic regions (%var), selection of *tag*SNPs, and selection based on the window posterior probability of association (WPPA). To also detect potentially associated regions, we propose to measure posterior probability of the interval (PP_{int}), which aims to select regions based on the markers of greatest effects. Therefore, the objective of this work was to evaluate these approaches, regarding efficiency in selecting and identifying markers or regions that are located within or close to genes associated with traits. This study also aimed to compare these methodologies with single-marker analyses. For that, simulated data were used in six scenarios, with SNPs allocated in non-overlapping genomic regions. Considering traits with oligogenic inheritance, WPPA criterion followed by %var and PP_{int} criteria were shown to be superior, presenting higher values of detection power, capturing higher percentages of genetic variance and larger areas. For traits with polygenic inheritance, PP_{int} and WPPA criteria were considered superior. Single-marker analyses identified SNPs associated only in oligogenic inheritance scenarios and was lower than the other criteria.

Keywords: Genomic regions, Bayesian methods, genetic variance.

1 Introduction

The development of new sequencing and genotyping technologies has promoted the growth of molecular genetics, enabling breeding programs to carry out genome-wide association studies (GWAS) between quantitative trait loci (QTL) and genetic values of individuals. The selection of marker groups has been identified as a major advantage for GWAS because these groups tend to capture a greater proportion of the genetic variance, identifying more complex relationships between markers (Moore et al., 2010). According to Fan et al. (2011) and Fernando et al. (2017), to select the associated regions, methods using Bayesian approach are preferable showing great advantages, such as the possibility of incorporating a prior knowledge and simultaneous estimation of marker effects.

Based on this, criteria using the Bayesian approach to select markers and associated regions that don't require major computational efforts are being proposed. Among the existing criteria, selection by the percentage of variance explained by genomic regions (%var), selection criteria of tag single nucleotide polymorphisms (*tag*SNPs), and selection based on window posterior probability of association (WPPA) are the most notable. These approaches consider the genetic variance and differ in the criteria used for selecting the regions and in the thresholds that are established for selection.

To also detect potentially associated regions and to select SNPs with greater effects, we propose to measure the posterior probability of the interval (PP_{int}). The selection of these SNPs becomes viable, since biologically it is expected that SNPs that are close to a QTL will have a greater effect because they are close to the causal mutation (Habier et al., 2011; Meuwissen et al., 2016). Thus, PP_{int} is based on the number of iterations of the Markov Chain Monte Carlo (MCMC), in which the region has at least one SNP with an effect magnitude greater than the value of the third quartile, considering the entire distribution of the absolute effects in that iteration.

Given the above, this paper aims to propose the measure of PP_{int} and compare it to the *tag*SNP, %var, and WPPA approaches to determine the efficiency in selecting and identifying markers or regions that are located within or near genes associated with traits of interest. This study also aimed to compare the results obtained by these methodologies with single-marker analyses. For that, we used simulated data considering six different scenarios, with SNPs allocated in non-overlapping genomic regions.

2 Materials and methods

2.1 Simulated Data

The data set was simulated using Genes software (Cruz, 2013). The genome was composed of 10 linking groups, with 20 centimorgans (cM) and 200 markers in each group. The SNPs were considered to be distributed approximately equidistant in the genome. In the analysis of gene linkage, an F1 generation was simulated, in which one parent was a dominant homozygote and the other, a recessive homozygote. From the genotypes of the F1 population, a population of F2 mapped with 1000 individuals was generated, considering 5000 gametes, causing entire linkage disequilibrium (LD) to be caused by the linkage group. Quantitative traits were simulated considering a zero degree of dominance, mean equal to 100, and heritability levels were chosen to represent traits with high ($h_a^2 = 0.50$ and $h_a^2 = 0.60$), moderate ($h_a^2 = 0.30$ and $h_a^2 = 0.40$), and low ($h_a^2 = 0.10$ and $h_a^2 = 0.20$) heritabilities. Three genetic architectures were generated using three, ten, and one hundred loci controlling the trait, which explained equal parts of the genetic variance, these QTL being distributed in the regions covered by the markers. In the first architecture, a case was considered in which three QTL were randomly distributed among the ten chromosomes. In the second, ten controlling loci of the trait were assumed, in which one QTL was assigned to each of the ten chromosomes. In the third, traits controlled by many genes with small effects were considered, in which ten QTL were distributed in each of the ten chromosomes, totaling 100 QTL. Additionally, according to Goddard et al. (2011), the proportion of genetic variation associated with the QTL explained by the markers (r_{mq}^2) was obtained by:

$$r_{mq}^2 = \frac{n}{n + n_{QTL}}, \quad (1)$$

where n was the number of SNPs and n_{QTL} was the number of QTL.

In this way, six different scenarios were used in the analyses: three genetic architecture × two different levels of heritability in each architecture. The description of the scenarios is presented in Table 1. Each type of scenario was simulated ten times to assess the efficiency of the methods, according Lima et al. (2019). Thus, the measures used were calculated in each repetition of the simulation and then the mean and standard error of these values were obtained.

2.2 BayesDπ

The BayesDπ method allowed a π percentage of marker effects to be equal to zero, leading to a lower number of marker effects to be estimated, making the estimation process more accurate since many of the markers do not have genetic effects or are not in LD with QTL (Habier et al., 2011). Consider the following linear model proposed by Meuwissen et al. (2001):

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{m} + \boldsymbol{e} \quad (2)$$

where $\boldsymbol{y}$ was the vector of phenotypes ($N \times 1$, N was the number of individuals), μ was the general mean of the trait, $\mathbf{1}$ was a vector of the same dimension of y with all elements equal to 1, $\boldsymbol{m}$ was the vector of additive genetic effects of the markers ($n \times 1$, n was the number of markers), $\boldsymbol{W}$ ($N \times n$) was the additive incidence matrix and $\boldsymbol{e}$ ($N \times 1$) was the vector of errors of the model with $\boldsymbol{e} \sim N\,(0, \sigma_e^2)$ with σ_e^2 being the error variance. The W matrix is coded according to Vitezica et al., 2013.

In this method it was considered that a fraction $(1 - \pi)$ of the markers had no effect on the trait and the remaining fraction π had effects with a prior distribution given by a normal with specific variance $\sigma_{m_j}^2$ for each marker being the variance of each marker from a scaled inverse chi-square distribution with υ degrees of freedom and scale parameter S_m^2. Thus, the equation used was:

$$m_j|\pi \sim \pi N\left(0, \sigma_{m_j}^2\right) + (1-\pi)N\left(0, \sigma_{m_j}^2 = 0\right) \quad (3)$$

$$\sigma_{m_j}^2|\pi \sim Scale - inv - \chi^2(\upsilon, S_m^2) \quad (4)$$

where $\pi|m_j, \sigma_m^2, \sigma_e^2, S_m^2 \sim U[0,1]$ and $S_m^2 \sim Gamma(\alpha, \beta)$ being α and β, respectively, as the hyperparameters of a prior distribution. Note that this method allowed the specification of a prior distribution for the π probability and S_m^2 hyperparameter, considering them unknown parameters of the model to limit the influence of subjectivity in the selection of markers and the shrinkage factor. The a posterior mean of the total genetic variance of the markers was given by $\hat{\sigma}_g^2 = \sum_j 2p_j q_j \hat{\sigma}_{m_j}^2$, where p_j and q_j were the allele frequencies associated with "A" and "a" alleles, respectively, of the j-th marker.

BayesDπ can be advantageously used in GWAS in relation to other Bayesian approaches because it assumes specific variance for each locus to obtain information about the genetic architecture of the trait and estimate the value of probability π, which is considered unknown and considers that most of the markers have small effects (or zero effect), except for those closer

to the causal mutation that would have greater effects, giving more biological meaning to the analyses (Habier et al., 2011).

For inference about the posterior distribution of the estimated effects of SNPs, 320,000 iterations were used for the MCMC algorithms, of which 20,000 were discarded (burn-in) to guarantee the heating of the chain and selection of one in ten iterations (thin). Convergence analysis was performed using the criterion proposed by Geweke (1992).

2.3 Formation of Regions

On each chromosome, SNPs were allocated to non-overlapping genomic regions of defined sizes, according to the mean LD between the markers and the QTL itself. Linkage disequilibrium values vary between zero and one, referring to the absence or complete LD, respectively. According to Zhao et al. (2007) and Viana et al. (2016), the effectiveness of locating QTL using LD between markers and QTL depends on the extent of LD and how it decreases with the distance between markers and QTL in a population. In this study, two extensions of LD with values between 0.64 and 0.81 were considered thresholds to determine the size of the regions. These values were chosen because they represent a high correlation of 0.80 and 0.90, respectively, between the QTL and markers. The shortest distance that LD provided between the QTL and marker following these two established LD extensions was used to define the size of the region in the scenarios. Thus, for each scenario, two sizes of regions were used according to the extent of LD considered. For the LD extension of 0.64, sizes of 4 cM, 5 cM, and 2.7 cM were established for scenarios 1 and 2, 3 and 4, and 5 and 6, respectively. For LD of 0.81, sizes found for the respective scenarios were 2 cM, 1.5 cM, and 1 cM. The six scenarios were analyzed considering these region sizes and the five approaches to selecting SNPs/regions, single-marker analyses, %var, *tag*SNP, WPPA, and PP_{int}.

2.4 Comparison of methodologies

To verify the efficiency of the analyzed criteria, the following measures described below were calculated:

i) False positive (FP) consisted of declaring a marker/region as associate, when in fact this marker/region was not in LD with the QTL and was defined by the ratio between

the number of SNPs/regions considered associated and that don't affect the trait and the number of SNPs/regions that don't affect the trait.

ii) Detection power (PD) consisted of declaring a marker/region effect associated, when this marker/region was actually in LD with the QTL and was defined as the ratio between the number of SNPs/regions considered associated and that affect the trait and the number of SNPs/regions that affect the trait.

iii) Percentage of genetic variance recovered was based on the percentage of genetic variance captured by the SNPs/regions and was obtained by the ratio between the genetic variance of the SNPs/regions considered associated and a posterior mean of total genetic variance. According to Peters et al. (2012), genomic regions that contribute with greater genetic variances were considered those most associated with the trait of interest.

iv) Area under the curve obtained between false positive rates and detection power was calculated using the receiver operating characteristic curve (ROC) proposed by Metz (1978) to also compare the criteria. Previous studies (Gage et al., 2018; Liu et al., 2016) used ROC curves or similar visual aids to assess the effectiveness of different methods in GWAS. In a ROC curve, the detection power values are plotted against the false positive rate and thus, the criterion that provides the highest area value below the curve is considered superior. The use of the area to compare the results in a single statistic allows direct comparison of the results of GWAS of traits with different simulation parameters (Gage et al., 2018).

Thus, the methodology that presents lower rates of false positives, greater detection power, that captures a greater proportion of the genetic variance, and that has a larger area under the ROC curve was considered the most suitable for GWAS.

2.5 Selection Criteria for Regions or SNPs

2.5.1 Selection by proportion of genetic variance explained by genomic regions -%var

The selection of SNP groups using the proportion of genetic variance explained by genomic regions (%var), was initially proposed by Wang et al. (2014) as an efficiency measure to compare methods of selecting regions in GWAS. For the effects of estimated SNPs, the genetic variance associated with the k-th region was calculated using:

$$\hat{\sigma}^2_{g_k} = \sum_{j \epsilon k} 2p_j q_j \hat{m}^2_j \quad (5)$$

where $\hat{m}_j$ was the allelic substitution effect of the j-th SNP belonging to the k-th region as estimated by BayesDπ. The explanation percentage of the genetic variance for each region was obtained as follows:

$$\%var = \frac{\hat{\sigma}^2_{g_k}}{\hat{\sigma}^2_g} \times 100 \quad (6)$$

with $\hat{\sigma}^2_g$ being the posterior mean of the total genetic variance considering all markers.

The regions that presented proportion values of the genetic variance higher than the ratio between the posterior mean of the total genetic variance and the number of regions considered were selected as associated regions and subsequently used to explore and determine possible QTL. A grid of values ranging from 100% to 200% of the ratio between the a posterior mean of the total genetic variation and the number of regions was considered and the one that returned the least difference between the detection power and confidence level (an associated region was not declared when this region was not actually in LD with QTL) was considered in the results.

2.5.2 Selection of *tag*SNPs using Bayesian methods

This approach consisted of identifying associated regions in the genome and subsequently, selecting SNPs in those regions that supposedly had high LD with the QTL. In this method, the SNPs were allocated in genomic regions but not overlapping and thus, the regions that potentially contained QTL associated with the trait of interest were called top windows. Top windows were identified based on a threshold defined in terms of the contribution of the genetic variance of the markers that could be obtained through the estimated effects of all SNPs via BayesDπ. Sollero et al (2017) considered all regions that explained proportions of genetic variance greater than five times the threshold described by Schurink et al. (2012) as top windows, which is obtained through the quotient between the a posterior mean of the total genetic variance by the number of regions considered. In this criterion, the same grid of values of the percentage of the posterior mean of the total genetic variance used in the %var criterion was considered and thus, the regions that presented genetic variances above this threshold were considered top windows. Again, the percentage used and considered in the analyses were those that provided the least difference between the power to detect an associated region and the level of confidence.

The effects of SNPs considered associated within each top window were called *tag*SNPs and were selected using the inclusion frequencies (IF) of the SNPs in the model, that is, the ratio between the number of saved iterations of the MCMC that include the SNP in question in the model and the total number of iterations saved. Thus, for each top window, the IF values were calculated for all SNPs and the *tag*SNPs were selected as the SNPs that had the highest IF value. To also select the *tag*SNPs within each top window, *t*-like statistics (TL) were used, considering an approach similar to IF to assess the consistency of the SNPs effects. This measure is given by the absolute value of the posterior mean effects of the markers (only for the chains that included the SNP in the model) divided by the respective standard deviations of these effects, an approximation of the student's t-statistic. Based on this statistic, SNPs considered significant ($p < 0.05$) within each top window were also established as *tag*SNPs.

2.5.3 Selection by the window posterior probability of association – WPPA

The WPPA measure was implemented using the effects of SNPs estimated via BayesDπ and obtained based on the proportion of the genetic variance explained by the markers of each genomic region. The genomic variance associated with the k-th region was estimated, in this context, by means of:

$$\hat{\sigma}^2_{g_k} = \sum_{j \epsilon k} 2 p_j q_j \hat{m}^2_j \tag{7}$$

where $\hat{m}_j$ was the allelic substitution effect of the j-th SNP belonging to k-th region estimated by BayesDπ and p_j and q_j were allele frequencies. From this, the proportion of the genetic variance explained by the markers in the k-th region, denoted by q_k, was defined as:

$$q_k = \frac{\hat{\sigma}^2_{g_k}}{E\left(\sigma^2_{g_k}\right)} \tag{8}$$

where $E\left(\sigma^2_{g_k}\right) = \sum_{j \epsilon k} 2 p_j \, q_j \, E(m^2_j)$ (in the absence of dominance), with $E\left(m^2_j\right) = \frac{(\sigma^2_g)}{n\bar{H}}$ and σ^2_g the genetic variance of the markers, n the number of SNPs, and $\bar{H}$ the mean of $2p_j q_j$. If $q_k > 1$, there was a causative mutation within the k-th region, since it had a greater than expected effect under the hypothesis of an equal distribution of genetic variance across the genome (Bennewitz et al., 2017; Peters et al., 2012). Thus, the WPPA measure was obtained by the ratio between the number of samples, where q_k was greater than one and the number of samples saved.

Regions that had a WPPA above a pre-established threshold were selected as associated regions. According to Fernando and Garrick (2013) and Fernando et al. (2017), if WPPA values greater than 0.95 are used to declare associated regions, this will result in a proportion of false positives below 0.05. Bennewitz et al. (2017) considered the levels 0.85, 0.95, and 0.99, and found that the power to detect an associated region decreased with the increase of these levels. In this study, several threshold levels ranging from 0.50 to 1.00, with an increment of 0.01, were tested in the analyses and the value that provided the least difference between the power to detect a region and the confidence level was considered.

2.5.4 Selection by a posterior probability of interval - PP_{int}

Biologically, it is expected that SNPs close to a QTL will have a greater effect being close to the causal mutation (Habier et al., 2011; Meuwissen et al., 2016) and for this reason it becomes viable to select these SNPs in the search for associations between markers and QTL through the PP_{int} measure that represents the probability SNPs with great effects are included in the region. In addition, according to Resende et al (2008), QTL can be located by adding the absolute effects of SNPs within each region and the regions with the largest sums of these absolute effects are likely to contain a QTL or be adjacent to a region containing a QTL and thus, the position of the QTL can be found and the discovery of QTL with great effect is facilitated. According to these authors, if there is no QTL in a given region, all estimates of the effects of SNPs within it will be small in magnitude. Thus, to also detect regions associated with the traits of interest, a new selection approach based on the effects of SNPs obtained in the MCMC samples was proposed and called a posterior probability of interval (PP_{int}). PP_{int} represented the probability of SNPs with large effects being included in the region and was calculated by the ratio between the number of iterations in which a given region had at least one SNP with an effect magnitude greater than the value of the third quartile, considering the entire distribution of the absolute effects in that iteration and the number of samples saved.

Regions with PP_{int} values greater than a pre-specified threshold were chosen as associated. In this study, threshold values also varying from 0.50 to 1.00 with an increment of 0.01 were tested and thus, the value that provided the least difference between the detection power and the confidence level was chosen. This threshold was chosen by the researcher and directly reflects the posterior probability of a QTL being in the region.

2.5.5 Single-marker analyses

The mixed linear model of single-markers was used to estimate the effect of the j-th marker on the phenotype and was defined by:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 g} + \boldsymbol{W_j} m_j + \boldsymbol{e} \tag{9}$$

where $\boldsymbol{y}$, μ, and $\boldsymbol{e}$ have been defined previously; $\boldsymbol{g}$ ($N \times 1$) was the vector of polygenic genetic effects with an incidence matrix $\boldsymbol{Z_1}$ ($N \times N$), being $\boldsymbol{g} \sim N\left(0, \boldsymbol{G}\sigma_g^2\right)$ where $\boldsymbol{G}$ was the additive genomic relationship matrix, σ_g^2 was the polygenic variance, m_j was the scalar, referring to the fixed effect of the j-th marker, and $\boldsymbol{W_j}$ was the incidence vector of the j-th marker. The matrix of additive genomic relationship was given by (VanRaden, 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\sum_{j=1}^{n} 2p_j q_j}. \tag{10}$$

To estimate the polygenic genetic effects and the effect of the j-th marker, the mixed model equations (Henderson, 1973) were given by:

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'Z_1} & \mathbf{1'W_j} \\ \boldsymbol{Z_1}'\mathbf{1} & \boldsymbol{Z_1}'\boldsymbol{Z_1} + \boldsymbol{G}^{-1}\dfrac{\sigma_e^2}{\sigma_g^2} & \boldsymbol{Z_1}'\boldsymbol{W_j} \\ \boldsymbol{W_j}'\mathbf{1} & \boldsymbol{W_j}'\boldsymbol{Z_1} & \boldsymbol{W_j}'\boldsymbol{W_j} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{g}} \\ \hat{m}_j \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \boldsymbol{Z_1}'\boldsymbol{y} \\ \boldsymbol{W_j}'\boldsymbol{y} \end{bmatrix}, \tag{11}$$

where the components of variance, σ_g^2 and σ_e^2, were estimated via the restrict maximum likelihood (REML).

After estimating the effect of the j-th marker, the Wald test was performed to test the existence of an association between the marker and QTL. Thus, the null hypothesis (H_0) was defined as when "the j-th marker had no effect on the phenotype", and the alternative hypothesis (H_a) defined as "the j-th marker affected the phenotype", that is, the j-th marker and the QTL were found in LD. However, this statistical analysis suffers from the occurrence of a high rate of false positives due to the occurrence of multiple tests. An alternative to control this fact is to monitor the number of false positives in relation to the total number of positive results through the false discovery rate (FDR) as presented by Fernando et al. (2004). One way of considering FDR in the significance test is through correction in the p-value, called q-value (Storey and Tibshirani, 2003).

The sizes of regions obtained based on the LD between the marker and QTL were used to define a region that determined the number of SNPs that really affected the trait. Thus, the SNPs that were distant to the QTL on the chromosome, lower than these thresholds, were

considered associated SNPs and used to calculate false positive rates, detection power, percentage of the variance, and area under the ROC curve.

2.6 Computational Resources

The entire implementation of the methods used was performed based on R software (R Development Core Team, 2019) through the GenomicLand visual interface (Azevedo et al., 2019). Convergence analysis of the effects of SNPs was estimated via the BayesDπ and was performed using the *coda* package (Plummer et al., 2006). In single-marker analysis, the *sommer* package was used (Covarrubias-Pazaran, 2016). The codes and data are available at https://www.licae.ufv.br/codes-for-association-analysis/.

3 Results and discussion

The results, considering the LD extension of 0.64 for determining the sizes of the regions, are shown in Table 2. In this simulation study, to identify superior procedures for the selection of associated regions, first, we chose the criteria that presented the highest value points for detection power. These criteria were considered preferable, since they informed the true proportion of regions that had been detected and that were actually associated (Bennewitz et al., 2017). The results in Table 2 revealed that for scenarios with oligogenic genetic inheritance (traits controlled by a few genes with greater effects - scenarios 1, 2, 3, and 4), the WPPA criterion followed by the %var and PP_{int} criteria were higher than the *tag*SNP criterion, presenting higher and similar point values for the power to detect associated regions and, consequently, capturing higher percentages of explanation of the genetic variance. For scenarios 1 and 2, the power to detect associated regions and the percentage of explanation of the variance were highest for the WPPA and %var criteria, elucidating the superiority of these methods in this genetic architecture.

Using a more general choice procedure, just like the one made by Gage et al. (2018), WPPA criterion stood out in relation to the area under the ROC curve in scenarios with oligogenic inheritance, providing higher values compared to other methods. For false positive rates in scenarios 1 and 2, all criteria were similar, providing values of zero or close to zero. However, in scenarios 3 and 4, *tag*SNP was the method that showed superiority. Initially, to select *tag*SNPs, the frequency of inclusion of SNPs in the model (IF) and the *t*-like statistic

(TL) were used. However, TL measures did not identify significant SNPs within the top windows and for this reason, only the results referring to the *tag*SNPs selected by the IF measure are shown in Table 2.

For the scenarios considering polygenic inheritance (traits controlled by many genes with small effects - scenarios 5 and 6), the PP_{int} and WPPA criterion were more efficient, presenting maximum values of power in detecting regions and percentage of variance explained. These criteria were also superior with respect to the area under the ROC curve, providing larger areas compared to the %var and *tag*SNP criteria for these scenarios. Note that the %var criterion was lower in these scenarios, proving efficient only for traits controlled by a few genes. With regard to the rate of false positives, in these scenarios, all methods analyzed obtained the lowest possible values (zero), indicating that no SNP/region was declared associated when it was not.

The WPPA and PP_{int} criteria showed lower power values in scenarios 3 and 4 than in other scenarios, which may have been influenced by the size of the region, which was the largest among all scenarios. Notably, in the smallest distance size considered (scenarios 5 and 6), these criteria stood out. Braz et al. (2019) and Bennewitz et al. (2017) reported on the influence of the size of the windows on detection power. In their studies considering sliding windows to select haplotypes associated with the bovine genome, Braz et al. (2019) used a mixed linear model, showing that smaller window sizes detect more associated regions and that larger window sizes may be more likely to introduce analytical problems, resulting in an excessive number of haplotypes, creating noise and computer memory problems. Furthermore, the results obtained by Braz et al. (2019) corroborate those obtained above, in which power decreased with increasing window size, however, they contradict the results obtained by Bennewitz et al. (2017), where they verified an increase in power with an increase in the size of the windows.

For the PP_{int} criterion, detection power and the percentage of explanation of the variance increased with the increase of heritability for all scenarios with oligogenic inheritance (scenarios 1 for 2 and scenarios 3 for 4). However, the area under the ROC curve decreased with the increase of this heritability in these scenarios for this method. The same trend occurred with respect to power and percentage verified for WPPA criterion in scenarios 3 and 4 and for the *tag*SNP in scenarios 1 and 2. In the other scenarios, the values of power and percentage of variance were similar in relation to the increase in heritability for all criteria, including in scenarios with polygenic inheritance (scenario 5 to scenario 6). These results are in agreement with Shin and Lee (2015), in which they compared the statistical power according to the

heritability for oligogenic and polygenic traits and found that the power difference between a heritability of 0.30 and 0.50 increased in the scenario containing twenty causal variants but decreased when there were 100. According to Shin and Lee (2015), the power estimated empirically from the simulation study would be applicable to GWAS for quantitative traits with known genetic parameters, predicting the degree of false negative associations.

The power values for %var criterion decreased according to the increase in the number of loci controlling the trait, indicating again that this method can be used advantageously for scenarios with oligogenic inheritance, however, it becomes inferior when considering traits with polygenic inheritance. The PP_{int} criterion stood out in scenarios controlled by many loci (polygenic inheritance), presenting lower rates of false positives, higher values of detection power, higher percentages of explanation of variance, and larger areas next to the WPPA criterion, which was also superior. Thus, the PP_{int} and WPPA criteria can be widely used in GWAS for inheritance, especially considering that the inheritance of most agronomically important traits are controlled by many genes, which individually have small or rare alleles (Yang et al., 2010).

Bennewitz et al. (2017) reported that the WPPA criterion seemed to be an inadequate approach to control the rate of false positives, since it was not built for this purpose and for this reason, it sometimes presents values at very high levels for this measure. Conversely, as observed in this study, this method captured greater proportions of the genetic variance and, in most cases, had greater power to detect associated regions. Fernando et al. (2017) also stated that the high threshold values for WPPA can compromise false positive rates, which corroborates the results found in some scenarios in this study (scenarios 3 and 4). The same can be observed for the thresholds considered in the PP_{int} criterion in the same scenarios. Note that for the calculation of PP_{int}, we used regions that had SNPs with effects greater than the third quartile, however, in future works, it should be verified to determine the possibility of an improvement in the results when considering other quartiles.

The WPPA measure, which was different from PP_{int} criterion, considers, for the calculation of the genetic variance of the regions, the allele frequency used to obtain the effects of the SNPs and the mean heterosis under the assumption of an equal distribution of the additive genetic variance and the dominance variance in all SNPs. The consideration of the allelic frequency in this criterion becomes feasible, since the detection power of SNPs is also determined by this measure (Shan and Purcell, 2014). According to these authors, a low allelic frequency influences a low detection power, unless there are relatively greater effects of SNPs.

In addition, considering heterosis in GWAS can also satisfactorily affect the detection power of SNPs (Vidoti et al., 2019).

For all considered scenarios, the *tag*SNP criterion obtained false positive rates equal to zero, however, this criterion was the one that provided the least power to detect associated SNPs, compared to the other criteria. These results corroborate the information reported by Schmid and Bennewitz (2017), where they stated that the decrease in the number of false positives in GWAS can compromise power. However, for scenarios 1, 2, 5, and 6, the criteria WPPA and PP_{int} also presented false positive rates equal to zero and in addition, they were efficient in terms of detection power. According to Li et al. (2014), the occurrence of false positives in GWAS can be controlled but this is only possible at the expense of reducing the power to detect true positives or statistical power. In other words, establishing a strict threshold for the association criterion is an effective way to control the rate of false positives. However, this also reduces the number of true positives detected. A desirable solution would be to reduce false positives, without compromising the detection power of the analysis, as performed by the WPPA and PP_{int} criteria in these studied scenarios.

The results also revealed that the π probability obtained by the BayesDπ method varied from 0.16 to 0.47 between the scenarios, indicating that the number of markers that are supposed to be in LD with the QTL varied from 320 to 940. According to Fernando and Garrick (2017), higher values of π may be more discriminatory for the identification of QTL with the greatest effect, which is an important factor for the selection of SNPs. Additionally, Sollero et al. (2017), in studies to select *tag*SNPs related to tick resistance in cattle breeds, found that the decrease in the π probability value may cause an increase in the proportion of genetic variance explained by the SNPs, in accordance with the results obtained here for *tag*SNP criterion.

The results, considering 0.81 as a threshold for determining the regions in LD, are shown in Table 3. Regarding the detection power and percentage of explanation of the variance, the results were similar to those found using the LD of 0.64 (Table 2), in which the criteria WPPA, %var, and PP_{int} were superior to the criteria of selection by *tag*SNPs. In scenarios 5 and 6, the PP_{int} criterion was the most efficient when compared to the others, once again showing its superiority in scenarios with polygenic inheritance. Regarding the areas on the ROC curve, the rates of false positives, and detection power according to the increase in heritability, the same results as those obtained previously were also observed. Results regarding the rate of false positives are in line with what was discussed by Moore et al. (2010), which highlighted the

advantage of considering groups of markers together, since these sets tend to capture a greater proportion of the genetic variance.

Comparing Tables 2 and 3, in relation to the LD extensions considered, the results revealed that for the criteria %var, WPPA, and PP_{int}, the detection power and the percentage of variance explained increased with the decrease in LD extension from 0.81 to 0.64, for the two scenarios analyzed with polygenic inheritance. The sizes of regions obtained, considering the LD extension of 0.64, were larger than those found with an extension of 0.81 and thus, the results reported when we increased the sizes of the regions there seems to be an increase in power for these three criteria. These results corroborate those obtained by Bennewitz et al. (2017), in which there was also an increase in power with the increase in the size of the regions. However, the optimum size of the genomic regions may differ between studies or between different QTL in the same study, depending on the extent of LD between the markers and QTL, the effective size of the population, and the detection power of each approach (Braz et al., 2019; Guo et al., 2016).

For scenarios with oligogenic inheritance, only the *tag*SNP criterion presented different power values between the two extensions of the LD, verifying an increase in these values with an increase of the thresholds. The results also revealed that the threshold obtained in the PP_{int} and WPPA criteria decreased according to the increase in the sizes of the regions for scenarios with polygenic inheritance. However, Guo et al. (2016), using a procedure to also select regions with GWAS in pigs, found that for regions with sizes above 5 Megabases, there was no increase in the values of this threshold.

The results for the single-marker analysis are shown in Table 4 and reveal that this procedure identified SNPs associated only in scenarios with oligogenic inheritance in which three QTL were randomly distributed among the ten chromosomes (scenarios 1 and 2). The detection power, considering an LD of 0.64, was always lower than that obtained using an LD of 0.81. Regarding the false positive rate, this method presented values equal to zero, indicating that the method considering oligogenic effects was efficient in identifying SNPs only when they are really associated with the traits of interest. However, the area obtained under the ROC curve was zero for both scenarios.

Compared to the criteria considered in this study, the single-marker analysis was lower than the %var, WPPA, and PP_{int} criteria, with less detection power, lower variance explained percentages, and smaller areas in scenarios 1 and 2 for both LD levels analyzed. However, this method showed higher values of power than the *tag*SNP criterion and similar percentages of

explanation for the genetic variance. According to Resende et al. (2017), the single-marker method can capture a larger percentage of the genetic variance due to the fact that it generally overestimates the effects of the tags, since the estimation process is not done simultaneously.

4 Conclusions

Considering traits with oligogenic genetic inheritance, the WPPA criteria, followed by the %var and PP_{int} criteria, were shown to be superior to the tag SNP criterion presenting higher values of detection power, capturing higher percentages of genetic variance, and larger areas under the ROC curve. For traits with polygenic inheritance, the PP_{int} and WPPA criteria were considered superior to the others for the LD extension of 0.64 and for the LD of 0.81, only PP_{int} stood out as being more efficient. The single-marker analysis method identified SNPs associated only in oligogenic inheritance scenarios and was lower than the %var, WPPA, and PP_{int} criteria. In general, the PP_{int} and WPPA criteria can be widely used in GWAS, especially considering that the inheritance of most agronomically important traits are controlled by many genes, which individually have small or rare alleles.

Acknowledgments

We thank the Brazilian funding organizations: Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

Authors' contributions

Conceptualization: Lima, L.P., Azevedo, C.F., Resende, M.D.V., Nascimento, M., Silva, F.F. Data acquisition: Lima, L.P., Azevedo, C.F., Resende, M.D.V. Data Analysis: Lima, L.P., Azevedo, C.F. Design of methodology: Lima, L.P., Azevedo, C.F., Resende, M.D.V., Nascimento, M., Silva, F.F. Software development: Lima, L.P., Azevedo, C.F. Writing and Editing: Lima, L.P., Azevedo, C.F., Resende, M.D.V., Nascimento, M.

References

Azevedo, C.F.; Nascimento, M.; Fontes, V.C.; Resende, M.D.V.D.; Cruz, C.D. 2019. GenomicLand: Software for genome-wide association studies and genomic prediction. **Acta Scientiarum. Agronomy** 41.

Bennewitz, J.; Edel, C.; Fries, R.; Meuwissen, T.H.; Wellmann, R. 2017. Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. **Genetics Selection Evolution** 49(1): 7.

Braz, C.U.; Taylor, J.F.; Bresolin, T.; Espigolan, R.; Feitosa, F.L.; Carvalheiro, R.; De Oliveira, H. N. 2019. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. **BMC genetics** 20(1): 1-12.

Covarrubias-Pazaran, G. 2016. Genome-assisted prediction of quantitative traits using the R package sommer. **PloS one** 11(6).

Cruz, C. D. 2013. Genes: a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum. Agronomy** 35(3): 271-276.

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

Fernando, R. L.; Garrick, D. 2013. Bayesian methods applied to GWAS. p. 237-274. In Genome-wide association studies and genomic prediction. **Humana Press, Totowa**, NJ.

Fernando, R.; Toosi, A.; Wolc, A.; Garrick, D.; Dekkers, J. 2017. Application of whole-genome prediction methods for genome-wide association studies: a Bayesian approach. **Journal of Agricultural, Biological and Environmental Statistics** 22(2): 172-193.

Fernando, R.L.; Nettleton, D.; Southey, B.R.; Dekkers, J.C.M.; Rothschild, M.F.; Soller, M. 2004. Controlling the proportion of false positives in multiple dependent tests. **Genetics** 166(1): 611-619.

Gage, J.L.; De Leon, N.; Clayton, M.K. 2018. Comparing genome-wide association study results from different measurements of an underlying phenotype. **G3: Genes, Genomes, Genetics** 8(11): 3715-3722.

Geweke, J. 1992. Evaluating the accuracy of sampling-based approaches to the calculations of posterior moments. **Bayesian statistics** 4: 641-649.

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

Guo, X.; Su, G.; Christensen, O.F.; Janss, L.; Lund, M.S. 2016. Genome-wide association analyses using a Bayesian approach for litter size and piglet mortality in Danish Landrace and Yorkshire pigs. **BMC genomics** 17(1): 468.

Habier, D.; Fernando, R.L.; Kizilkaya, K.; Garrick, D.J. 2011. Extension of the Bayesian alphabet for genomic selection. **BMC bioinformatics** 12(1): 186.

Henderson, C.R. 1973. Sire evaluation and genetic trends. **Journal of Animal Science** 1973(Symposium): 10-41.

Li, M.; Liu, X.; Bradbury, P.; Yu, J.; Zhang, Y.M.; Todhunter, R.J.; Zhang, Z. 2014. Enrichment of statistical power for genome-wide association studies. **BMC biology** 12(1): 73.

Lima, L. P.; Azevedo, C. F.; Resende, M. D. V. D.; Viana, J. M. S.; Oliveira, E. J. D. 2019. Triple categorical regression for genomic selection: application to cassava breeding. **Scientia Agricola** 76(5): 368-375.

Liu, X.; Huang, M.; Fan, B.; Buckler, E.S.; Zhang, Z. 2016. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. **PLoS genetics** 12(2): e1005767.

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

Meuwissen, T., Hayes, B., and Goddard, M. 2016. Genomic selection: A paradigm shift in animal breeding. **Animal Frontiers** 6:6–14. doi:10.2527/af.2016-0002

Meuwissen, T.H.E.; Hayes, B.J.; Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 157: 1819–1829.

Moore, J.H.; Asselbergs, F.W.; Williams, S.M. 2010. Bioinformatics challenges for genome-wide association studies. **Bioinformatics** 26(4): 445-455.

Peters, S.O.; Kizilkaya, K.; Garrick, D.J.; Fernando, R.L.; Reecy, J.M.; Weaber, R.L.; Silver, G.A.; Thomas, M. G. 2012. Bayesian genome-wide association analysis of growth and yearling ultrasound measures of carcass traits in Brangus heifers. **Journal of animal science** 90(10): 3398-3409.

Plummer, M.; Best, N.; Cowles, K.; Vines, K. 2006. CODA: convergence diagnosis and output analysis for MCMC. **R news** 6(1): 7-11.

R Core Team. 2019. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Resende, M. D. V. de; Lopes, P. S.; Silva, R. L. da; Pires, I. E. 2008. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa florestal brasileira** n. 56, p. 63.

Resende, R.T.; Resende, M.D.V.; Silva, F.F.; Azevedo, C.F.; Takahashi, E.K.; Silva-Junior, O.B.; Grattapaglia, D. 2017. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. **New Phytologist** 213(3): 1287-1300.

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

Schurink, A.; Wolc, A.; Ducro, B.J.; Frankena, K.; Garrick, D.J.; Dekkers, J.C.; van Arendonk, J.A. 2012. Genome-wide association study of insect bite hypersensitivity in two horse populations in the Netherlands. **Genetics Selection Evolution** 44(1): 31.

Shin, J.; Lee, C. 2015. Statistical power for identifying nucleotide markers associated with quantitative traits in genome-wide association analysis using a mixed model. **Genomics** 105(1): 1-4.

Sollero, B.P.; Junqueira, V.S.; Gomes, C.C.; Caetano, A.R.; Cardoso, F.F. 2017. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. **Genetics Selection Evolution** 49(1): 49.

Storey, J.D.; Tibshirani, R. 2003. Statistical significance for genomewide studies. **Proceedings of the National Academy of Sciences** 100(16): 9440-9445.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. **Journal of dairy science** 91(11): 4414-4423.

Viana, J.M.S.; Piepho, H.P. 2016. Quantitative genetics theory for genomic selection and efficiency of breeding value prediction in open-pollinated populations. **Scientia Agricola** 73(3): 243-251.

Vitezica, Z.G.; Varona, L.; Legarra, A. 2013. On the additive and dominant variance and covariance of individuals within the genomic selection scope. **Genetics** 195(4): 1223-1230.

Wang, Q.; Tian, F.; Pan, Y.; Buckler, E.S.; Zhang, Z. 2014a. A SUPER powerful method for genome wide association study. **PloS one** 9(9): e107684.

Yang, J.; Benyamin, B.; McEvoy, B.P.; Gordon, S.; Henders, A.K.; Nyholt, D. R; Madden, P.A.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; Goddard, M.E. 2010. Common SNPs explain a large proportion of the heritability for human height. **Nature genetics** 42(7): 565.

Zhao, H.; Nettleton, D.; Dekkers, J.C. 2007. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between single nucleotide polymorphisms. **Genetics Research** 89(1): 1-6.

Table 1 - Description of scenarios with the proportion of QTL variation explained by the SNPs (r^2_{mq}), genetic architecture, number of QTL and narrow-sense heritability (h^2).

Scenarios	r^2_{mq}	Genetic architecture	Number of QTL	h^2
Scenario 1	0.99	3 QTL on 10 chromosomes	3	0.50
Scenario 2		3 QTL on 10 chromosomes	3	0.60
Scenario 3	0.99	1 QTL in each of the 10 chromosomes	10	0.30
Scenario 4		1 QTL in each of the 10 chromosomes	10	0.40
Scenario 5	0.95	10 QTL on each of the 10 chromosomes	100	0.10
Scenario 6		10 QTL on each of the 10 chromosomes	100	0.20

Table 2 - Size of the regions (distance) in cM found through the LD between the markers and QTL in each scenario using the LD extension of 0.64 and the means and standard errors of the estimated π probability via BayesDπ, false positive rates (FP), detection power (PD), percentage of genetic variance recovered (PE), area under the ROC curve and threshold for selecting regions obtained by criteria %var, *tag*SNP, WPPA and PP_{int}.

Scenarios	π	Distance	Criterion	FP	Power	PE	Area	Threshold
1	0.32±0.03	4	%var	0.02±0.01	1.00±0.00	1.00±0.00	0.06±0.01	2.90±0.20
			*tag*SNP	0.00±0.00	0.02±0.00	0.15±0.03	0.01±0.00	0.66±0.09
			WPPA	0.01±0.01	1.00±0.00	1.00±0.00	0.99±0.00	0.90±0.03
			PP_{int}	0.06±0.06	0.77±0.10	0.85±0.07	0.85±0.04	0.93±0.03
2	0.16±0.04		%var	0.01±0.01	1.00±0.00	1.00±0.00	0.02±0.01	3.77±0.33
			*tag*SNP	0.00±0.00	0.02±0.00	0.33±0.04	0.00±0.00	1.82±0.29
			WPPA	0.00±0.00	1.00±0.00	1.00±0.00	0.60±0.16	0.75±0.07
			PP_{int}	0.00±0.00	1.00±0.00	1.00±0.00	0.60±0.16	0.77±0.07
3	0.45±0.00	5	%var	0.16±0.02	0.80±0.03	0.90±0.02	0.10±0.01	1.12±0.04
			*tag*SNP	0.00±0.00	0.01±0.00	0.02±0.00	0.00±0.00	0.58±0.03
			WPPA	0.18±0.02	0.79±0.02	0.89±0.01	0.89±0.02	0.96±0.00
			PP_{int}	0.40±0.07	0.62±0.06	0.68±0.06	0.47±0.04	1.00±0.00
4	0.46±0.00		%var	0.14±0.02	0.82±0.03	0.91±0.01	0.09±0.02	1.87±0.12
			*tag*SNP	0.00±0.00	0.01±0.00	0.02±0.00	0.00±0.00	0.94±0.04
			WPPA	0.17±0.03	0.87±0.02	0.94±0.01	0.91±0.02	0.96±0.00
			PP_{int}	0.49±0.09	0.78±0.07	0.84±0.05	0.44±0.07	1.00±0.00
5	0.46±0.00	2.7	%var	0.00±0.00	0.48±0.01	0.63±0.01	0.00±0.00	0.73±0.02
			*tag*SNP	0.00±0.00	0.02±0.00	0.02±0.00	0.00±0.00	0.76±0.02
			WPPA	0.00±0.00	1.00±0.00	1.00±0.00	0.89±0.01	0.85±0.01
			PP_{int}	0.00±0.00	1.00±0.00	1.00±0.00	0.89±0.04	0.84±0.01
6	0.47±0.00		%var	0.00±0.00	0.49±0.01	0.65±0.01	0.00±0.00	1.36±0.02
			*tag*SNP	0.00±0.00	0.02±0.00	0.03±0.00	0.00±0.00	1.39±0.02
			WPPA	0.00±0.00	1.00±0.00	1.00±0.00	0.94±0.01	0.84±0.00
			PP_{int}	0.00±0.00	1.00±0.00	1.00±0.00	0.94±0.01	0.84±0.00

Scenarios with oligogenic inheritance – 3 QTL: Scenario 1 ($h^2 = 0.50$) and Scenario 2 ($h^2 = 0.60$), 10 QTL: Scenario 3 ($h^2 = 0.30$) and Scenario 4 ($h^2 = 0.40$). Scenarios with polygenic inheritance – 100 QTL: Scenario 5 ($h^2 = 0.10$) and Scenario 6 ($h^2 = 0.20$). %var: proportion of genetic variance

explained by genomic regions, WPPA: window posterior probability of association and PP_{int}: posterior probability of interval.

Table 3 - Size of the regions (distance) in cM found through the LD between the markers and QTL in each scenario using the LD extension of 0.81 and the means and standard errors of the estimated π probability via BayesDπ, false positive rates (FP), detection power (PD), percentage of genetic variance recovered (PE), area under the ROC curve and threshold for selecting regions obtained by criteria %var, *tag*SNP, WPPA and PP_{int}.

Scenarios	π	Distance	Criterion	FP	Power	PE	Area	Threshold
1	0.32±0.03	2	%var	0.03±0.00	1.00±0.00	1.00±0.00	0.08±0.01	1.68±0.09
			*tag*SNP	0.00±0.00	0.05±0.00	0.21±0.03	0.00±0.00	0.27±0.05
			WPPA	0.01±0.00	1.00±0.00	1.00±0.00	0.98±0.01	0.89±0.03
			PP_{int}	0.22±0.11	1.00±0.00	1.00±0.00	0.89±0.06	0.91±0.03
2	0.16±0.04		%var	0.02±0.00	1.00±0.00	1.00±0.00	0.04±0.01	2.11±0.16
			*tag*SNP	0.00±0.00	0.04±0.00	0.38±0.03	0.00±0.00	0.86±0.15
			WPPA	0.00±0.00	1.00±0.00	1.00±0.00	0.60±0.16	0.72±0.06
			PP_{int}	0.00±0.00	1.00±0.00	1.00±0.00	0.59±0.16	0.73±0.06
3	0.45±0.00	1.5	%var	0.18±0.02	0.83±0.02	0.88±0.02	0.17±0.01	0.39±0.02
			*tag*SNP	0.01±0.00	0.06±0.00	0.07±0.00	0.00±0.00	0.06±0.00
			WPPA	0.19±0.04	0.76±0.06	0.82±0.05	0.85±0.02	0.96±0.00
			PP_{int}	0.07±0.01	0.53±0.06	0.63±0.06	0.75±0.03	0.99±0.00
4	0.46±0.00		%var	0.15±0.02	0.85±0.02	0.91±0.01	0.18±0.01	0.67±0.04
			*tag*SNP	0.01±0.00	0.06±0.00	0.07±0.00	0.00±0.00	0.11±0.01
			WPPA	0.12±0.02	0.83±0.04	0.89±0.03	0.91±0.01	0.96±0.00
			PP_{int}	0.10±0.01	0.74±0.06	0.82±0.04	0.83±0.02	0.99±0.00
5	0.46±0.00	1	%var	0.33±0.01	0.42±0.02	0.52±0.02	0.07±0.00	0.32±0.01
			*tag*SNP	0.02±0.00	0.04±0.00	0.05±0.00	0.00±0.00	0.30±0.01
			WPPA	0.46±0.04	0.56±0.05	0.62±0.04	0.59±0.01	0.93±0.00
			PP_{int}	0.54±0.04	0.75±0.05	0.79±0.05	0.63±0.01	0.95±0.00
6	0.47±0.00		%var	0.31±0.01	0.43±0.01	0.54±0.01	0.07±0.00	0.57±0.01
			*tag*SNP	0.01±0.00	0.04±0.00	0.05±0.00	0.00±0.00	0.54±0.01
			WPPA	0.42±0.03	0.56±0.04	0.65±0.03	0.60±0.01	0.92±0.00
			PP_{int}	0.53±0.05	0.76±0.06	0.80±0.05	0.64±0.00	0.95±0.00

Scenarios with oligogenic inheritance – 3 QTL: Scenario 1 ($h^2 = 0.50$) and Scenario 2 ($h^2 = 0.60$), 10 QTL: Scenario 3 ($h^2 = 0.30$) and Scenario 4 ($h^2 = 0.40$). Scenarios with polygenic inheritance – 100 QTL: Scenario 5 ($h^2 = 0.10$) and Scenario 6 ($h^2 = 0.20$). %var: proportion of genetic variance

explained by genomic regions, WPPA: window posterior probability of association and PP_{int}: posterior probability of interval.

Table 4 - Extensions of LD used to determine the distances between SNPs, means and respective standard errors of false positive rates (FP), detection power (PD), percentage of genetic variance recovered (PE), the area under the ROC curve estimated by the single-marker analysis and also the threshold used to select significant SNPs.

Scenarios	LD	Distance	FP	Power	PE	Area	Threshold
1	0.64	4	0.00±0.00	0.07±0.01	0.13±0.02	0.01±0.00	0.05±0.00
	0.81	2	0.00±0.00	0.15±0.02	0.13±0.02	0.00±0.00	0.05±0.00
2	0.64	4	0.00±0.00	0.11±0.01	0.20±0.02	0.00±0.00	0.05±0.00
	0.81	2	0.00±0.00	0.21±0.02	0.20±0.02	0.00±0.00	0.05±0.00

Scenarios with oligogenic inheritance – 3 QTL: Scenario 1 ($h^2 = 0.50$) and Scenario 2 ($h^2 = 0.60$).

CAPÍTULO 3

AVALIAÇÃO DE MÉTODOS BAYESIANOS DE ASSOCIAÇÃO GENÔMICA VIA REGIÕES CROMOSSÔMICAS CONSIDERANDO EFEITOS ADITIVOS E DE DOMINÂNCIA

Resumo

Considerar e selecionar regiões genômicas vem elucidando importantes resultados para os estudos de associação ampla do genoma (GWAS) entre loci de características quantitativas (QTL) e valores genéticos. Isso se deve ao fato de que maior parte da variabilidade de determinado QTL é explicada conjuntamente. Entre os critérios de seleção de regiões existentes se destacam, as abordagens baseadas na probabilidade *a posteriori* da associação de regiões genômicas (*Window Posterior Probability of Association* - WPPA) e na probabilidade *a posteriori* do intervalo (*Posterior Probability of Interval* - PP_{int}) que utilizam efeitos de marcadores estimados via abordagem bayesiana. Além disso, uma metodologia alternativa, denominada mapeamento de herdabilidades regionais (*Regional Heritability Mapping* - RHM) vem apresentando importantes resultados na busca de regiões associadas. Portanto, o objetivo deste trabalho foi avaliar essas abordagens, quanto à eficiência na identificação de regiões que estão localizados dentro ou próximos a genes associados a características de interesse e também compará-las com a análise via marcas únicas. A eficiência em termos de poder de detecção e de falsos positivos destes métodos foi avaliada considerando ou não a inclusão dos efeitos de dominância nos modelos estatísticos. Para isso utilizou-se dados simulados em dezoito cenários com diferentes níveis de herdabilidade, arquitetura genética e grau médio de dominância. Para os efeitos aditivos considerando características com arquitetura genética oligogênica, os critérios WPPA, RHM e PP_{int} se mostraram superiores à análise via marcas únicas para todos os graus de dominância analisados. Já para características com herança poligênica, os critérios PP_{int} e WPPA podem ser considerados superiores aos demais. Considerando apenas os efeitos devido à dominância, os quatro critérios analisados apresentaram resultados relevantes com relação as medidas de eficiência para as características controladas por 3 QTL e para os demais cenários novamente os critérios WPPA e PP_{int} incluindo também a análise via marcas únicas se mostraram superiores.

Palavras-chave: Janelas genômicas, marcadores moleculares, variância genética, dominância, GWAS.

Abstract

Considering and selecting genomic regions has elucidated important results for genome-wide association studies (GWAS) between quantitative trait loci (QTL) and genetic values. This is due to the fact that most of the variability of a given QTL is explained together Among the selection criteria for SNPs or regions, the approaches based on the selection based on the window posterior probability of association (WPPA) and the posterior probability of the interval (PP_{int}) that use effects stand out of estimated markers via the Bayesian approach. In addition, an alternative methodology, called regional heritability mapping (RHM) has been showing important results in the search for associated regions. Therefore, the objective of this work was to evaluate these approaches, regarding efficiency in the identification of regions that are located within or close to genes associated with traits of interest and also to compare them with the single-marker analyses. The efficiency in terms of detection power and false positives of these methods was assessed considering whether or not the inclusion of dominance effects in statistical models. For that, simulated data were used in eighteen scenarios with different levels of heritability, genetic architecture and degree of dominance. For the additive effects considering traits with oligogenic genetic inheritance, the WPPA, RHM and PP_{int} criteria were superior to the single-marker analyses for all analyzed degrees of dominance. For traits with polygenic inheritance, the PP_{int} and WPPA criteria can be considered superior to the others. Considering only the effects due to dominance, the four analyzed criteria presented relevant results regarding the efficiency measures for the traits controlled by 3 QTL and for the other scenarios again the criteria WPPA and PP_{int} also including the single-marker analyses were shown to be superior.

Keywords: Genomic window, molecular markers, genetic variance, dominance, GWAS.

1 Introdução

Os estudos de associação genômica ampla (*Genome Wide Association Studies* - GWAS) entre os loci de características quantitativas (*Quantitative Trait Loc*i - QTL) e os valores genéticos dos indivíduos são de extrema importância para os programas de melhoramento genético e podem ser realizados indiretamente entre os marcadores moleculares codominantes SNPs (*Single Nucleotide Polymorphisms*) e os fenótipos das características de interesse uma vez que se considera a existência de desequilíbrio de ligação (*Linkage Disequilibrium* - LD) entre marcador e QTL (MEUWISSEN et al., 2001).

Dispondo-se de milhares de SNPs a fim de inferir associações, é esperado que marcadores próximos uns aos outros estejam altamente correlacionados. Dessa forma, ao considerar um grupo de marcadores para encontrar uma determinada região genômica, espera-se que a maior parte da variabilidade de determinado loco de característica seja explicada conjuntamente (MOORE et al., 2010; LEE et al., 2014). Além disso, grupos de marcadores podem estar em maior LD com o QTL do que os marcadores únicos e devido a isso, o LD entre o QTL e as regiões genômicas é aumentado, aumentando assim o poder de detecção dessas variantes causais (HAYES, 2013). A partir disso, modelos utilizando abordagem bayesiana têm sido propostos e vêm se mostrando eficientes para a seleção de grupos de SNPs relevantes no melhoramento animal e vegetal (SAHANA et al., 2010; NAGAMINE et al., 2012; GODDARD et al.,2016; SOLLERO et al., 2016). As abordagens bayesianas apresentam a vantagem de estimar os efeitos de marcadores simultaneamente e de evidenciar as diferenças entre as proporções da variação explicada por cada marcador uma vez que incorporam informações *a priori* sobre esses efeitos. Além disso, propiciam informações sobre a arquitetura genética dos QTL e suas posições por modelagem da frequência dos SNPs não nulos se mostrando em geral superiores aos métodos regularização já propostos no contexto de GWAS (RESENDE et al., 2012; RESENDE et al., 2010).

Dentre os critérios de seleção de regiões associadas considerando efeitos de SNPs estimados por meio de metodologias bayesianas se destacam a seleção com base na probabilidade *a posteriori* da associação de regiões genômicas (WPPA - *Window Posterior Probability of Association*) implementada por Fernando e Garrick (2013) e Fernando et al. (2017) e a medida baseada na Probabilidade *a Posteriori* do Intervalo (PP_{int} - *Posterior Probability of Interval*) proposta por Lima et al. (2021) em modelos aditivos. Além dos métodos utilizando abordagem bayesiana, uma metodologia alternativa, denominada mapeamento de

herdabilidades regionais (*Regional heritability mapping* - RHM), que visa também determinar regiões do genoma que estão associadas ao fenótipo foi proposta por Nagamine et al. (2012) e vem elucidando importantes resultados com relação ao poder em detectar QTL verdadeiros e reduzidas taxas de falsos positivos (USAI et al., 2014; MATIKA et al., 2016).

A disponibilidade de painéis de alta densidade de SNPs ofereceu também novas oportunidades para detecção, utilização e investigação dos efeitos de dominância (LOPES et al., 2014). Estes efeitos são fontes de variação de características complexas e apresentam grande importância no controle dessas características, podendo ocasionar nos estudos de associação aumento no poder detecção de QTL (BENNEWITZ et al., 2017). Tem sido demonstrado que a dominância é um fator importante que contribui para a heterose sendo ela de grande importância na maioria dos estudos de melhoramento genético para identificação de regiões genômicas que controlam características de interesse (BONNAFOUS et al., 2018; MAO et al., 2020). Outra vantagem de adicionar efeitos de dominância a uma análise GWAS é que a possibilidade de se identificar variantes adicionais pode aumentar, ajudando a capturar uma fração maior da variância genética (DU et al., 2016; LU et al., 2017). Além disso, de acordo com Lundregan et al. (2020), a contribuição da variância genética aditiva pode ser superestimada quando qualquer variância de dominância existente não é contabilizada. No entanto, a maioria das metodologias consideradas na GWAS ainda consideram apenas os efeitos aditivos, negligenciando a possível eficiência dos métodos ao considerar também os efeitos de dominância.

Diante disso, o objetivo deste trabalho foi avaliar a eficiência das metodologias, WPPA, PP_{int} e RHM em detectar regiões associadas no genoma considerando a inclusão ou não dos efeitos de dominância nos modelos estatísticos na GWAS. Estes critérios também foram comparados com o método de análise via marcas únicas que é a metodologia mais usual nos estudos de associação. Para a avaliação e a apresentação das abordagens descritas anteriormente foram considerados dezoito cenários simulados que diferiam em níveis de grau médio de dominância, arquiteturas genéticas e níveis de herdabilidade em sentido amplo.

2 Materiais e Métodos

2.1 Dados Simulados

Para a avaliação e a apresentação das abordagens descritas anteriormente foram simulados dezoito cenários utilizando o *software* Genes (CRUZ, 2013). O genoma simulado foi

composto por 10 grupos de ligação, com 20 cM e 200 marcadores SNPs em cada grupo. Foram considerados marcadores distribuídos de forma aproximadamente equidistante no genoma. Na análise de ligação gênica foi simulado uma geração F1 no qual, um genitor era homozigoto dominante e o outro homozigoto recessivo. A partir dos genótipos da população F1 foi gerada uma população de mapeamento F2 com 1.000 indivíduos, considerando 5.000 gametas. Ao considerar a população F2, obtida do cruzamento de genitores homozigotos contrastantes tem-se que todos os alelos de cada loco marcador nessa população possuem frequências iguais a meio e valor máximo de desequilíbrio igual a 25% e assim, o LD máximo entre dois pares de loci corresponde à completa ligação fatorial (SCHUSTER I. e CRUZ, C.D., 2013; SANT'ANNA et al, 2019). Além disso, de acordo com Sánches et al. (2013), nas populações F2, os mapas de LD e de ligação fatorial são equivalentes. Esses autores também concluíram que o tipo de população utilizada tem impacto relevante sobre os padrões de LD e consequentemente sobre o número de marcadores necessários para identificar genes que controlam a característica de interesse ao melhoramento.

Características quantitativas foram simuladas considerando três graus médios de dominância ($d/a = 0$, 0,50 e 1, em que a e d são os valores genotípicos do homozigoto e do heterozigoto, respectivamente), média igual a 100 e herdabilidades de sentido amplo distintas entre os cenários. Os níveis de herdabilidade foram escolhidos para representar características com altas ($h^2 = 0{,}50$ e $h^2 = 0{,}60$), moderadas ($h^2 = 0{,}30$ e $h^2 = 0{,}40$) e baixas ($h^2 = 0{,}10$ e $h^2 = 0{,}20$) herdabilidades de acordo com a arquitetura genética das mesmas. Foram geradas três arquiteturas genéticas utilizando 3, 10 e 100 locos controladores da característica, que explicam partes iguais da variância genética, sendo esses QTL distribuídos nas regiões abrangidas pelos marcadores. Na primeira arquitetura foi considerado o caso no qual 3 QTL foram distribuídos aleatoriamente entre os 10 cromossomos, sendo que em cada cromossomo não houvesse mais que 1 QTL. Na segunda, assumiu-se 10 locos controladores da característica em que 1 QTL foi designado a cada um dos 10 cromossomos. Já na terceira, considerou-se características controladas por muitos genes de pequenos efeitos no qual 10 QTL foram distribuídos em cada um dos 10 cromossomos, totalizando 100 QTL.

Adicionalmente, de acordo com Goddard et al. (2011), foi obtida para cada cenário a proporção da variação genética associada ao QTL explicada pelos marcadores (r^2_{mq}) sendo dada por:

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

em que n é o número de SNPs e n_{QTL} é o número de QTL.

Dessa forma, dezoito cenários distintos foram utilizados nas análises: três níveis de grau médio de dominância × dois níveis de herdabilidade distintas em cada arquitetura × três arquiteturas genéticas. A descrição dos cenários é apresentada na Tabela 1.

Cada tipo de cenário foi simulado 10 vezes para a avaliar a eficiência dos métodos utilizados com base nas taxas de falsos positivos, no poder de cada metodologia em detectar SNPs/regiões realmente associadas, na porcentagem da variância genética capturada por cada critério e na área abaixo da curva obtida entre as taxas de falsos positivos e poder. Assim, essas medidas foram calculadas em cada repetição da simulação e depois foi obtida a média e o erro-padrão desses valores.

Tabela 1 – Descrição dos cenários com as respectivas proporções de variação dos QTL explicada pelos SNPs (r^2_{mq}), arquiteturas genéticas, número de QTL em cada cenário (Nº de QTL) e as herdabilidades em sentido amplo (h^2).

GMD	Cenários	r^2_{mq}	Arquitetura Genética	Nº de QTL	h^2
0,00	Cenário 1	0,99	3 QTL em 10 cromossomos	3	0,50
	Cenário 2		3 QTL em 10 cromossomos	3	0,60
	Cenário 3	0,99	1 QTL em cada um dos 10 cromossomos	10	0,30
	Cenário 4		1 QTL em cada um dos 10 cromossomos	10	0,40
	Cenário 5	0,95	10 QTL em cada um dos 10 cromossomos	100	0,10
	Cenário 6		10 QTL em cada um dos 10 cromossomos	100	0,20
0,50	Cenário 7	0,99	3 QTL em 10 cromossomos	3	0,50
	Cenário 8		3 QTL em 10 cromossomos	3	0,60
	Cenário 9	0,99	1 QTL em cada um dos 10 cromossomos	10	0,30
	Cenário 10		1 QTL em cada um dos 10 cromossomos	10	0,40
	Cenário 11	0,95	10 QTL em cada um dos 10 cromossomos	100	0,10
	Cenário 12		10 QTL em cada um dos 10 cromossomos	100	0,20
1,00	Cenário 13	0,99	3 QTL em 10 cromossomos	3	0,50
	Cenário 14		3 QTL em 10 cromossomos	3	0,60
	Cenário 15	0,99	1 QTL em cada um dos 10 cromossomos	10	0,30
	Cenário 16		1 QTL em cada um dos 10 cromossomos	10	0,40
	Cenário 17	0,95	10 QTL em cada um dos 10 cromossomos	100	0,10
	Cenário 18		10 QTL em cada um dos 10 cromossomos	100	0,20

2.2 Modelo linear sob enfoque bayesiano

Meuwissen et al. (2001) apresentaram o seguinte modelo linear básico em nível de marcas considerando efeitos aditivos e devido a dominância:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{W}\boldsymbol{m_a} + \boldsymbol{S}\boldsymbol{m_d} + \boldsymbol{e},$$

em que:

$\boldsymbol{y}$ é o vetor de fenótipos com dimensão $N \times 1$ sendo N o número de indivíduos;

μ é a média geral da característica;

$\mathbf{1}$ é um vetor de mesma dimensão de $\boldsymbol{y}$ com todos os elementos iguais a 1;

$\boldsymbol{m_a}$ e $\boldsymbol{m_d}$ são, respectivamente, os vetores de efeitos genéticos aditivos e devido à dominância dos marcadores;

$\boldsymbol{W}$ $(N \times n)$ e $\boldsymbol{S}$ $(N \times n)$ são as matrizes de incidência que relacionam respectivamente, os efeitos aditivos e devido à dominância dos marcadores aos fenótipos contidos em $\boldsymbol{y}$ e são codificadas como apresentado a seguir (VITEZICA et al., 2013; RESENDE et al., 2014):

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

em que p e q são respectivamente as frequências dos alelos A e a do marcador;

$\boldsymbol{e}$ $(N \times 1)$ é o vetor de erros do modelo com $\boldsymbol{e} \sim N(0; \sigma_e^2)$ sendo σ_e^2 a variância do erro.

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

$$\boldsymbol{y}|m_{a1}, m_{a2}, \dots, m_{an}, m_{d1}, m_{d2}, \dots, m_{dn} \sim N(\boldsymbol{1}\mu + \boldsymbol{W m_a} + \boldsymbol{S m_d}, \sigma_e^2);$$

$$\sigma_e^2 \sim \upsilon_e s_e^2 \chi^{-2} \text{ (distribuição qui-quadrado invertida escalada)}$$

sendo υ_e e s_e^2 os hiperparâmetros.

A especificação das distribuições *a priori* com a inclusão dos efeitos devido á dominância é dada por:

$$m_{a_j}|\pi_a, \sigma^2_{m_{a_j}} \sim \pi_a N(0, \sigma^2_{m_{a_j}}) + (1 - \pi_a) N(0, \sigma^2_{m_{a_j}} = 0);$$

$$m_{d_j}|\pi_d, \sigma^2_{m_{d_j}} \sim \pi_d N(0, \sigma^2_{m_{d_j}}) + (1 - \pi_d) N(0, \sigma^2_{m_{d_j}} = 0);$$

$$\sigma^2_{m_{a_j}}|\pi_a \sim Scale - inv - X^2(\upsilon_a, S^2_{m_a}),$$

$$\sigma^2_{m_{d_j}}|\pi_d \sim Scale - inv - X^2(\upsilon_d, S^2_{m_d}),$$

em que $\pi_a|m_{a_j}, \sigma^2_{m_{a_j}}, \sigma^2_e, S^2_{m_a} \sim U[0,1]$, $\pi_d|m_{d_j}, \sigma^2_{m_{d_j}}, \sigma^2_e, S^2_{m_d} \sim U[0,1]$, $S^2_{m_a} \sim Gama(a, b)$ e $S^2_{m_d} \sim Gama(c, d)$ sendo $\upsilon_a, \upsilon_d, a, b, c\ e\ d$ os hiperparâmetros das distribuições *a priori*. Na literatura a abordagem no qual admite as distribuições dadas acima é conhecida como BayesDπ.

Para a inferência sobre a distribuição *a posteriori* dos efeitos aditivos e/ou efeitos devido à dominância estimados dos SNPs foram utilizadas 320.000 iterações para os algoritmos MCMC (*Markov Chain Monte Carlo*), das quais 20.000 foram descartadas (*burn-in*) para garantir o aquecimento da cadeia e com a seleção de uma em cada 10 iterações (*thin*). A análise de convergência foi realizada via o critério proposto por Geweke (1992).

2.3 Formação das Regiões

Em cada cromossomo os SNPs foram alocados em regiões genômicas não sobrepostas de tamanhos definidos de acordo com o LD médio entre os marcadores e o próprio QTL. Os valores de LD podem variar entre zero e um, referindo-se, respectivamente, a ausência ou ao completo LD. Segundo Zhao et al. (2007), a efetividade da localização de QTL usando o LD entre marcadores e QTL depende da extensão do LD e de como ele diminui com a distância entre os marcadores e os QTL em uma população. Neste estudo, a extensão de LD considerada como limiar para determinar o tamanho das regiões foi a mesma utilizada por Lima et al. (2021) no valor de 0,64. Segundo estes autores este valor apresentou resultados relevantes e foi escolhido por representar alta correlação (acima de 0,80) entre o QTL e a marca. Para essa extensão de LD, estabeleceu o tamanho de 4 cM para os cenários considerando 3 QTL (cenários 1, 2, 7, 8, 13 e 14), 5 cM para os cenários com 10 QTL (cenários 3, 4, 9,10, 15 e 16) e 2,7 cM para os cenários com 100 QTL (cenários 5, 6, 11, 12, 17 e 18).

Os cenários simulados foram analisados considerando esses tamanhos de regiões e as abordagens de seleção de SNPs/regiões, WPPA, PP_{int}, RHM e análise via marcas únicas.

2.4 Seleção pela Probabilidade *a Posteriori* da Associação da Janela - WPPA

A medida WPPA pode ser implementada utilizando os efeitos de SNPs estimados via método bayesiano e é obtida com base na proporção da variância genética explicada pelos marcadores de cada região genômica. Para os efeitos de SNPs estimados tem-se que a variância genética total associada a k-ésima região pode ser calculada por meio de:

$$\hat{\sigma}^2_{g_k} = \sum_{j=1}^{n} (2p_j q_j \hat{m}^2_{a_j} + \left(2p_j q_j\right)^2 \hat{m}^2_{d_j})$$

em que n é o número de marcadores contidos nesta região, $\hat{m}_{a_j}$e $\hat{m}_{d_j}$ são, respectivamente, as médias *a posteriori* dos efeitos aditivos e devido à dominância do j-ésimo SNP pertencente a k-ésima região. A partir disso, tendo as cadeias de *Markov* de cada efeito de marcador (m_{a_j} *e* m_{d_j}) é possível estimar a probabilidade *a posteriori* de um grupo de marcadores (ou região) estar associada a característica analisada. Esta probabilidade foi proposta por Fernando e Garrick (2013) e é dada por:

$$q_k = \frac{\hat{\sigma}^2_{gk}}{\sum_{j=1}^{n}\left(H_j \frac{\hat{\sigma}^2_a}{n\overline{H}} + H_j^2 \frac{\hat{\sigma}^2_d}{n\overline{H^2}}\right)}$$

em que $\hat{\sigma}^2_a$ e $\hat{\sigma}^2_d$ são, respectivamente, as variâncias genéticas aditiva e devido à dominância estimadas considerando todos os marcadores, $H_j = 2p_jq_j$, $\overline{H} = \frac{1}{n}\sum_{j=1}^{n} H_j$ e $\overline{H^2} = \frac{1}{n}\sum_{j=1}^{n} H_j^2$. Caso $q_k > 1$, existe a presença de uma mutação causativa dentro da *k*-ésima região, uma vez que apresenta efeito maior do que o esperado sob a hipótese de uma distribuição igual da variância genética ao longo do genoma (PETERS et al., 2012; BENNEWITZ et al., 2017). Dessa forma, a medida WPPA é obtida pela razão entre o número de iterações em que q_k é maior que 1 e o número total de iterações salvas.

As regiões que possuem WPPA acima de um *threshold* pré-estabelecido são selecionadas como regiões associadas. Segundo Fernando e Garrick (2013) e Fernando et al. (2017), caso se utilize valores de WPPA superiores a 0,95 para declarar regiões associadas, isto resultará em uma proporção de falsos positivos inferior a 0,05. Já Bennewitz et al. (2017) considerou os níveis 0,85, 0,95 e 0,99 e verificou que o poder em detectar uma região associada diminuiu com o aumento desses níveis. Neste estudo, vários níveis de *threshold* variando de 0,50 a 1,00 com incremento de 0,01 foram testados nas análises e o valor que forneceu menor diferença entre o poder de detectar uma região e o nível de confiança (não declarar um marcador/região associado quando realmente este marcador/região não está em LD com o QTL) foi considerado.

2.5 Probabilidade a Posteriori do Intervalo – PP_{int}

A medida PP_{int} representa a probabilidade de SNPs com grandes efeitos serem incluídos na região e é calculada pela razão entre o número de iterações em que determinada região possui pelo menos um SNP com magnitude de efeito superior ao valor do terceiro quartil, considerando toda a distribuição dos efeitos absolutos naquela iteração, e o número total de iterações salvas. A seleção de SNPs de maiores efeitos na busca de associações entre marcas e QTL torna-se viável uma vez que biologicamente, pode-se esperar que SNPs, próximos a um QTL, tenham um maior efeito estando próximos à mutação causal (HABIER et al., 2011; MEUWISSEN et al., 2016). A medida PP_{int} considerando os efeitos devido a dominância foi calculada pela razão entre o número de iterações em que determinada região possui pelo menos um SNP com magnitude de efeito de dominância superior ao valor do terceiro quartil, considerando toda a

distribuição dos efeitos absolutos de dominância naquela iteração, e o número total de iterações salvas. Ademais, para o cálculo da medida PP_{int}, foram também considerados os efeitos aditivos de SNPs estimados via método BayesDπ negligenciando a presença dos efeitos devido à dominância.

As regiões com valores de PP_{int} considerando efeitos aditivos ou dominantes superiores a um *threshold* pré-especificado são escolhidas como regiões associadas. As regiões que simultaneamente foram ditas associadas para os efeitos aditivos e para os efeitos devido à dominância foram denominadas de regiões aditivo-dominantes. Neste estudo, valores de *threshold* variando também de 0,50 até 1,00 com incremento de 0,01 foram testados e assim o valor que forneceu menor diferença entre o poder de detectar uma região e o nível de confiança foi escolhido. Este limiar pode ser escolhido pelo investigador e reflete diretamente na probabilidade *a posteriori* de um QTL estar na região.

2.6 Mapeamento de herdabilidade regionais - RHM

O método de mapeamento de herdabilidades regionais utiliza o seguinte modelo linear misto para as estimações dos efeitos da k-ésima região no fenótipo:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{Z_{3k} r_{dk}} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$, μ, $\boldsymbol{a}$, $\boldsymbol{Z_1}$, $\boldsymbol{Z_2}$ e $\boldsymbol{e}$ já foram definidos anteriormente;

$\boldsymbol{r_{ak}}$ é o vetor de efeitos genômicos aditivos referente a k-ésima região do genoma $(N \times 1)$ com matriz de incidência $\boldsymbol{Z_{2k}}$ $(N \times N)$, sendo $\boldsymbol{r_{ak}} \sim N\left(0, \boldsymbol{G_{reg_k}} \sigma^2_{ar_k}\right)$ em que $\boldsymbol{G_{reg_k}}$ é a matriz de parentesco genômico aditiva e $\sigma^2_{ar_k}$ é a variância aditiva referentes a k-ésima região ($k = 1,2, \dots, K$, em que K é o número total de regiões determinadas);

$\boldsymbol{r_{dk}}$ é o vetor de efeitos genômicos devido à dominância referente a k-ésima região do genoma $(N \times 1)$ com matriz de incidência $\boldsymbol{Z_{3k}}$ $(N \times N)$, sendo $\boldsymbol{r_{dk}} \sim N(0, \boldsymbol{D_{reg_k}} \sigma^2_{dr_k})$ em que $\boldsymbol{D_{reg_k}}$ é a matriz de parentesco genômico devido à dominância e $\sigma^2_{dr_k}$ é a variância devido à dominância referentes a k-ésima região ($k = 1,2, \dots, K$).

As matrizes $\boldsymbol{G_{reg_k}}$ e $\boldsymbol{D_{reg_k}}$ utilizam, respectivamente, um subconjunto da matriz de incidência de marcadores W e S referente a k-ésima região e pode ser construída analogamente as matrizes G e D dadas anteriormente.

O vetor de efeitos genéticos poligênicos e os vetores de efeitos genômicos aditivos e devido à dominância, referente a k-ésima região do genoma, podem ser estimados via equações de modelos mistos (HENDERSON, 1973) dadas por:

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'Z_1} & \mathbf{1'Z_2} & \mathbf{1'Z_{2k}} & \mathbf{1'Z_{3k}} \\ \mathbf{Z_1'1} & \mathbf{Z_1'Z_1} + \mathbf{G^{-1}}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{Z_1'Z_2} & \mathbf{Z_1'Z_{2k}} & \mathbf{Z_1'Z_{3k}} \\ \mathbf{Z_2'1} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{D^{-1}}\frac{\sigma_e^2}{\sigma_d^2} & \mathbf{Z_2'Z_{2k}} & \mathbf{Z_2'Z_{3k}} \\ \mathbf{Z_{2k}'1} & \mathbf{Z_{2k}'Z_1} & \mathbf{Z_{2k}'Z_2} & \mathbf{Z_{2k}'Z_{2k}} + \mathbf{G_{reg_k}^{-1}}\frac{\sigma_e^2}{\sigma_{ar_k}^2} & \mathbf{Z_{2k}'Z_{3k}} \\ \mathbf{Z_{3k}'1} & \mathbf{Z_{3k}'Z_1} & \mathbf{Z_{3k}'Z_2} & \mathbf{Z_{3k}'Z_{2k}} & \mathbf{Z_{3k}'Z_{3k}} + \mathbf{D_{reg_k}^{-1}}\frac{\sigma_e^2}{\sigma_{dr_k}^2} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{a}} \\ \hat{\boldsymbol{d}} \\ \hat{\boldsymbol{r}}_{\boldsymbol{ak}} \\ \hat{\boldsymbol{r}}_{\boldsymbol{dk}} \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{1'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \\ \mathbf{Z_{2k}'y} \\ \mathbf{Z_{3k}'y} \end{bmatrix},$$

sendo os componentes de variância, σ_a^2, $\sigma_{ar_k}^2$, $\sigma_{dr_k}^2$ e σ_e^2, estimados via o método REML.

A descrição dos modelos (completo e reduzido) utilizados no RHM são apresentadas na Tabela 2.

Tabela 2 – Descrição dos modelos (completo e reduzido) para respectivos efeitos aditivo (A) e aditivo-dominante (AD) utilizados no mapeamento de herdabilidades regionais (RHM).

Modelo	Efeito	
AD	A	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}$
	D	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{3k} r_{dk}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}$
	AD	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{Z_{3k} r_{dk}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{e}$
A	A	Modelo Completo: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_{2k} r_{ak}} + \boldsymbol{e}$ Modelo Reduzido: $\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{e}$

$\boldsymbol{y}$**:** vetor de fenótipos; μ: média geral da característica; **1:** vetor de mesma dimensão de $\boldsymbol{y}$ com todos os elementos iguais a 1; $\boldsymbol{a}$**:** vetor de efeitos genéticos aditivos poligênicos com matriz de incidência $\boldsymbol{Z_1}$; $\boldsymbol{d}$: vetor de efeitos genéticos poligênicos devido à dominância com matriz de incidência $\boldsymbol{Z_2}$; $\boldsymbol{r_{ak}}$: vetor de efeitos genômicos aditivos referente a k-ésima região do genoma com matriz de incidência $\boldsymbol{Z_{2k}}$; $\boldsymbol{r_{dk}}$: vetor de efeitos genômicos devido à dominância referente a k-ésima região do genoma com matriz de incidência $\boldsymbol{Z_{3k}}$**;** $\boldsymbol{e}$**:** vetor de erros.

Os modelos são comparados, e consequentemente, é testada a significância do efeito da k-ésima região, pela mudança no logaritmo da função de verossimilhança ($ln\ L$), por meio do teste da razão de verossimilhança (TRV). A estatística TRV realizada entre os modelos (completo e reduzidos) é obtida pela seguinte expressão:

$$TRV = 2ln\left(\frac{L_1}{L_0}\right)$$

sendo L_1 a função de verossimilhança para o modelo completo e L_0 a função de verossimilhança para o modelo restrito. A hipótese de nulidade (H_0) para este teste foi de que os modelos não diferiram entre si, indicando que a k-ésima região não apresenta efeito (aditivo e/ou devido à dominância) sobre o fenótipo. Enquanto, que a hipótese alternativa (H_a) é definida como os modelos diferem entre si, determinando que a k-ésima região e o QTL encontram-se em LD. Dado que foram determinadas K regiões no genoma tem-se que os procedimentos de estimação via equações de modelos mistos e TRV devem ser realizados K vezes.

2.7 Análise via Marcas Únicas

O seguinte modelo misto em marcas simples pode ser empregado, visando estimar os efeitos aditivos e devido à dominância do j-ésimo marcador no fenótipo:

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Z_1 a} + \boldsymbol{Z_2 d} + \boldsymbol{w_j} m_{aj} + \boldsymbol{s_j} m_{dj} + \boldsymbol{e}$$

em que:

$\boldsymbol{y}$, μ e $\boldsymbol{e}$ já foram definidos anteriormente;

$\boldsymbol{a}$ é o vetor de efeitos genéticos aditivos poligênicos com matriz de incidência $\boldsymbol{Z_1}$ $(N \times N)$, sendo $\boldsymbol{a} \sim N(0, \boldsymbol{G}\sigma_a^2)$ em que $\boldsymbol{G}$ é a matriz de parentesco genômico aditiva e σ_a^2 é a variância aditiva poligênica, m_j é o escalar referente ao efeito fixo do j-ésimo marcador;

$\boldsymbol{d}$ é o vetor de efeitos genéticos poligênicos devido à dominância com matriz de incidência $\boldsymbol{Z_2}$ $(N \times N)$, sendo $\boldsymbol{d} \sim N(0, \boldsymbol{D}\sigma_d^2)$ em que $\boldsymbol{D}$ é a matriz de parentesco genômico devido à dominância e σ_d^2 é a variância poligênica devido à dominância;

m_{aj} é o escalar referente ao efeito fixo aditivo do j-ésimo marcador com vetor de incidência do j-ésimo marcador dado por $\boldsymbol{w_j}$ ($N\times$ 1);

m_{dj} é o escalar referente ao efeito fixo devido à dominância do j-ésimo marcador com vetor de incidência do j-ésimo marcador dado por $\boldsymbol{s_j}$ ($N\times$ 1).

O vetor de incidência aditivo do j-ésimo marcador $\boldsymbol{w_j}$ representa a j-ésima coluna da matriz $\boldsymbol{W}$, enquanto o vetor de incidência devido à dominância do j-ésimo marcador $\boldsymbol{s_j}$ representa a j-ésima coluna da matriz $\boldsymbol{S}$. A matriz de parentesco genômica aditiva é dada por (VAN RADEN, 2008):

$$\boldsymbol{G} = \frac{\boldsymbol{WW'}}{\sum_{j=1}^{n} 2p_j q_j}.$$

em que p_j e q_j são as frequências dos alelos A e a do j-ésimo marcador. Já a matriz de parentesco genômica devido à dominância é dada conforme Vitezica et al. (2013):

$$\boldsymbol{D} = \frac{\boldsymbol{SS'}}{\sum_{j=1}^{n} (2p_j q_j)^2},$$

O vetor de efeitos genéticos poligênicos, os efeitos aditivos e devido a dominância do j-ésimo marcador podem ser estimados via equações de modelos mistos (HENDERSON, 1973) dadas por:

$$\begin{bmatrix} \mathbf{1'1} & \mathbf{1'Z_1} & \mathbf{1'Z_2} & \mathbf{1'w_j} & \mathbf{1's_j} \\ \mathbf{Z_1'1} & \mathbf{Z_1'Z_1} + \mathbf{G^{-1}}\frac{\sigma_e^2}{\sigma_a^2} & \mathbf{Z_1'Z_2} & \mathbf{Z_1'w_j} & \mathbf{Z_1's_j} \\ \mathbf{Z_2'1} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{D^{-1}}\frac{\sigma_e^2}{\sigma_d^2} & \mathbf{Z_2'w_j} & \mathbf{Z_2's_j} \\ \mathbf{w_j'1} & \mathbf{w_j'Z_1} & \mathbf{w_j'Z_2} & \mathbf{w_j'w_j} & \mathbf{w_j's_j} \\ \mathbf{s_j'1} & \mathbf{s_j'Z_1} & \mathbf{s_j'Z_2} & \mathbf{s_j'w_j} & \mathbf{s_j's_j} \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{d}} \\ \hat{m}_{aj} \\ \hat{m}_{dj} \end{bmatrix} = \begin{bmatrix} \mathbf{1'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \\ \mathbf{w_j'y} \\ \mathbf{s_j'y} \end{bmatrix},$$

Após a estimação dos efeitos aditivos e dominantes do j-ésimo marcador é realizado o teste de Wald para os mesmos visando testar a existência de associação com significância estatística entre o j-ésimo marcador e o QTL. A hipótese nula (H_0) é definida como o j-ésimo marcador não apresenta efeito (aditivo ou devido à dominância) sobre o fenótipo e a hipótese alternativa (H_a) definida como o j-ésimo marcador afeta o fenótipo, ou seja, o j-ésimo marcador e o QTL encontram-se em LD. Por fim, para o cálculo das medidas de eficiência do método, foi também considerado e denominado de efeito aditivo-dominante os casos em que tanto os efeitos aditivos quanto os devido a dominância foram simultaneamente ditos como significativos.

As análises estatísticas descritas acima sofrem com a ocorrência de uma alta taxa de falsos positivos devido a ocorrência de testes múltiplos. Uma alternativa para controlar esse fato é monitorar o número de falsos positivos em relação ao número total de resultados positivos por meio da taxa de descobertas falsas (*False Discovery Rate* - FDR) conforme apresentado por Fernando et al. (2004). Uma maneira de se considerar a FDR no teste de significância é por meio de uma correção no *p-value* associado ao teste, denominado de *q-value* (STOREY e TIBSHIRANI, 2003).

Os tamanhos de regiões obtidas com base no LD entre marcador e QTL foram utilizadas para determinar os SNPs que realmente afetam a característica. Assim, os SNPs que possuem distâncias ao QTL, no cromossomo, menores que esses limiares foram considerados como SNPs associados e utilizados para o cálculo das taxas de falsos positivos, poder, porcentagem de explicação da variância e área sob a curva a curva ROC.

2.8 Comparação das metodologias

Com intuito de verificar a eficiência dos critérios analisados foram calculadas as taxas de falsos positivos, o poder de cada método em detectar marcadores ou regiões que realmente estão associadas ao QTL, a porcentagem da variância genética capturada por cada critério e a

área abaixo da curva obtida entre as taxas de falsos positivos e o poder de detecção. Essas medidas são descritas abaixo:

i) Falso positivo (FP) consiste em declarar um marcador (ou região) como associado, quando na verdade este marcador (ou região) não está em LD com o QTL e é definido pela razão entre o número de SNPs (ou regiões) ditos associados e que não afetam a característica e o número de SNPs (ou regiões) que não afetam a característica.

ii) O poder de detecção (PD) consiste em declarar um efeito de marcador (ou região) associado quando realmente este marcador (ou região) está em LD com o QTL e é definido pela razão do número de SNPs (ou regiões) ditos associados e que afetam a característica pelo número de SNPs (ou regiões) que afetam a característica.

iii) Porcentagem explicada da variância genética: as metodologias também foram comparadas com base na porcentagem da variância genética capturada pelos SNPs/regiões encontrados por cada critério e foi obtida pela razão entre a variância genética dos SNPs/regiões consideradas associadas e a média *a posteriori* da variância genética total. Segundo Peters et al. (2012), regiões ou janelas genômicas que contribuírem com maiores variâncias genéticas são consideradas aquelas mais associadas à característica de interesse.

iv) Área sob a curva obtida entre as taxas de falsos positivos e o poder de detecção: a curva ROC (*Receiver Operating Characteristic*) proposta por Metz (1978) também foi utilizada para comparar os critérios. Estudos anteriores (WANG et al., 2014; LIU et al., 2016; GAGE et al., 2018) usaram curvas ROC ou recursos visuais semelhantes para avaliar a eficácia de diferentes métodos na GWAS. Em uma curva ROC, os valores de poder de detecção são plotados contra a taxa de falsos positivos e assim, o critério que fornece maior valor de área abaixo da curva é considerado superior. O uso da área para comparar os resultados em uma única estatística permite a comparação direta dos resultados da GWAS de características com diferentes parâmetros de simulação (GAGE et al., 2018).

Desta forma, a metodologia que apresentar menores taxas de falsos positivos, maior poder de detecção, que capturar uma proporção maior da variância genética e que possuir maior área sob a curva ROC será considerada a mais adequada para a associação genômica.

2.9 Recursos Computacionais

Toda a implementação dos métodos utilizados foi realizada baseada no software R (R Development Core Team, 2018) por meio da interface visual *GenomicLand* (AZEVEDO et al., 2019). A análise de convergência dos efeitos dos SNPs estimados via método BayesDπ foi realizada pelo software R pelo pacote *coda* (PLUMMER et al., 2006). Na análise via marcas únicas e o RHM foi utilizado o pacote *sommer* (COVARRUBIAS-PAZARAN, 2016).

3 Resultados e Discussão

Os resultados considerando grau médio de dominância de zero são apresentados na Tabela 3. Para os cenários com características controladas por apenas 3 QTL de efeitos maiores (cenários 1 e 2) o critério WPPA seguido do RHM e do critério PP_{int} foram superiores a análise via marcas únicas apresentando maiores valores pontuais para o poder em detectar regiões associadas e, consequentemente, capturando maiores porcentagens de explicação da variância genética. Nestes cenários o poder em detectar regiões associadas e a porcentagem de explicação da variância foram máximos para o critério WPPA elucidando a superioridade deste método nesta arquitetura genética. O mesmo pode ser observado para o PP_{int} no cenário 2.

Tabela 3 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Análise via Marcas Únicas (MU) em cada cenário para grau médio de dominância de 0.

	Distância	Critério	FP	Poder	PE	*Threshold*	Área
Cenário 1	4	RHM	0,01±0,00	0,90±0,07	0,94±0,05	0,05±0,00	0,10±0,01
		WPPA	0,01±0,01	1,00±0,00	1,00±0,00	0,90±0,03	0,99±0,00
		PP_{int}	0,06±0,06	0,77±0,10	0,85±0,07	0,93±0,03	0,85±0,04
		MU	0,00±0,00	0,07±0,01	0,13±0,02	0,05±0,00	0,01±0,00
Cenário 2		RHM	0,01±0,00	0,97±0,03	0,99±0,01	0,05±0,00	0,09±0,01
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,75±0,07	0,60±0,16
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,77±0,07	0,60±0,16
		MU	0,00±0,00	0,11±0,01	0,20±0,02	0,05±0,00	0,00±0,00
Cenário 3	5	RHM	0,00±0,00	0,06±0,02	0,26±0,09	0,05±0,00	0,01±0,00
		WPPA	0,18±0,02	0,79±0,02	0,89±0,01	0,96±0,00	0,89±0,02
		PP_{int}	0,40±0,07	0,62±0,06	0,68±0,06	1,00±0,00	0,47±0,04
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
Cenário 4		RHM	0,01±0,00	0,01±0,01	0,04±0,04	0,05±0,00	0,00±0,00
		WPPA	0,17±0,03	0,87±0,02	0,94±0,01	0,96±0,00	0,91±0,02
		PP_{int}	0,49±0,09	0,78±0,07	0,84±0,05	1,00±0,00	0,44±0,07
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
Cenário 5	2,7	RHM	0,01±0,01	0,01±0,00	0,21±0,08	0,05±0,00	0,00±0,00
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,85±0,01	0,89±0,01
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,84±0,01	0,89±0,04
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
Cenário 6		RHM	0,00±0,00	0,00±0,00	0,03±0,03	0,05±0,00	0,00±0,00
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,84±0,00	0,94±0,01
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,84±0,00	0,94±0,01
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 1 ($h^2 = 0{,}50$) e Cenário 2 ($h^2 = 0{,}60$), 10 QTL: Cenário 3 ($h^2 = 0{,}30$) e Cenário 4 ($h^2 = 0{,}40$). Cenários com

arquitetura genética poligênica – 100 QTL: Cenário 5 ($h^2 = 0{,}10$) e Cenário 6 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

Ainda para os cenários com arquitetura genética oligogênica, porém com 10 QTL controlando a característica (cenários 3 e 4), os critérios que ganharam destaque com relação as medidas de eficiência consideradas foram o WPPA e o PP_{int}. O critério WPPA também foi eficiente com relação à área sob a curva ROC fornecendo maiores valores se comparados aos demais métodos.

Para os cenários considerando arquitetura poligênica (características controladas por muitos genes de pequenos efeitos – cenários 5 e 6), os critérios PP_{int} e o WPPA foram novamente mais eficientes, apresentando valores máximos de poder em detectar regiões e de porcentagem de explicação da variância. Estes critérios também foram superiores com relação à área sob a curva ROC fornecendo maiores áreas se comparados aos demais critérios para estes cenários.

As taxas de falsos positivos, para os cenários 1, 2, 5 e 6, de todos os critérios foram similares fornecendo valores zero ou próximos de zero. Já para os cenários 3 e 4, apenas o RHM e a análise via marcas únicas foram superiores com relação a essa medida de eficiência. Dessa forma, o RHM junto à análise via marcas únicas foram os critérios que apresentaram, para todos os cenários, baixos valores das taxas de falsos positivos. Esses resultados encontrados para o RHM corroboram aos obtidos por Usai et al. (2014) em que menciona que essa metodologia apresenta superioridade com relação a essas taxas. De acordo com Li et al. (2014), a ocorrência de falsos positivos em estudos de GWAS pode ser controlada, mas isso só é possível à custa da redução do poder de detecção de verdadeiros positivos ou do poder estatístico.

Os critérios WPPA e PP_{int} apresentaram menores valores de poder nos cenários 3 e 4 do que nos demais cenários o que pode ter sido influenciado pelo tamanho da região que foi a maior dentre todos os cenários. Note também que no menor tamanho de distância considerado (cenário 5 e 6) estes critérios se destacaram. Braz et al. (2019) e Bennewitz et al. (2017) relatam em seus estudos sobre a influência do tamanho das janelas no poder de detecção. Estes estudos consideram janelas deslizantes para selecionar haplótipos associados ao genoma bovino, Braz et al. (2019) utilizando o modelo linear misto mostraram que tamanhos de janelas menores detectam mais regiões associadas e que janelas de tamanhos maiores podem ter maiores probabilidades de introduzir problemas analíticos resultando em números excessivos de haplótipos, criando problemas de ruído e de memória de computador.

Para o critério PP_{int}, o poder de detecção de regiões e a porcentagem de explicação da variância aumentaram com o aumento da herdabilidade para todos os cenários com herança oligogênica (cenário 1 para o 2 e cenário 3 para o 4). No entanto, a área sob a curva ROC

diminuiu com o aumento dessa herdabilidade nestes cenários para este método. O mesmo com relação ao poder e a porcentagem foi verificado para o RHM nos cenários 1 e 2 e, e para o WPPA nos cenários 3 e 4. Nos cenários com arquitetura poligênica os valores de poder e porcentagem da variância foram similares em relação ao aumento da herdabilidade para todos os critérios. Esses resultados estão de acordo com Shin e Lee (2015) em que comparam o poder estatístico de acordo com a herdabilidade para características oligogênicas e poligênicas e verificaram que a diferença de poder entre as herdabilidades de 0,30 e 0,50 aumentou no cenário contendo 20 variantes causais, mas diminuiu quando havia 100 variantes causais.

O critério PP_{int} se destacou em cenários controlados por muitos locos (herança poligênica) apresentando menores taxas de falsos positivos, maiores valores de poder de detecção, maiores porcentagens de explicação da variância e maiores áreas junto ao critério WPPA que também foi superior. Já o RHM se destacou com relação ao poder de detecção e porcentagem de explicação da variância apenas nos cenários controlados por 3 QTL e também foi superior a análise via marcas únicas nos cenários controlados por 10 QTL. Nagamine et al. (2012) mostraram que o RHM teve um desempenho melhor do que a análise via marcas únicas, especialmente quando os SNPs associados não têm efeito grande o suficiente para serem declarados significativos corroborando com os resultados obtidos por Resende et al. (2017) ao avaliar características de eucalipto que supostamente eram controlados por muitos genes de pequenos efeitos. No entanto, neste trabalho, para características com arquitetura genética poligênica, esses dois métodos apresentaram valores similares, o que pode ter sido influenciado pelo tamanho das regiões que foi maior nestes cenários. Além disso, para essa arquitetura esses critérios foram inferiores comparados aos demais critérios com relação ao poder detecção, porcentagem de explicação da variância e área sob a curva ROC.

Os resultados associados aos efeitos aditivos ao considerar ou não a dominância para grau médio de dominância de 0,5 são apresentados na Tabela 4. Para a diferenciação dos resultados chamou-se de modelo aditivo o caso em que a dominância foi negligenciada e de modelo aditivo-dominante quando se considerou a dominância. Os resultados encontrados são similares aos obtidos considerando grau médio de dominância de 0 (Tabela 3), na qual para os cenários controlados com 3 QTL o critério WPPA seguido dos critérios PP_{int} e RHM foram superiores ao marcas únicas apresentando maiores e similares valores pontuais para o poder em detectar regiões associadas e, consequentemente, capturando maiores porcentagens de explicação da variância genética. Os critérios WPPA e PP_{int} também se destacaram com relação as medidas de eficiência utilizadas nos cenários controlados por 10 QTL.

Tabela 4 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Marcas Únicas (MU) em cada cenário associados aos efeitos aditivos considerando o modelo aditivo (A) e o modelo aditivo-dominante (AD) para grau médio de dominância de 0,5.

	Distância	Critério	Modelo	FP	Poder	PE	*Threshold*	Área
Cenário 7	4	RHM	A	0,02±0,00	0,90±0,05	0,96±0,02	0,05±0,00	0,06±0,01
			AD	0,03±0,00	0,87±0,05	0,93±0,03	0,05±0,00	0,07±0,01
		WPPA	A	0,05±0,01	1,00±0,00	1,00±0,00	0,94±0,02	0,97±0,01
			AD	0,03±0,01	0,93±0,04	0,98±0,02	0,73±0,07	0,49±0,16
		PP_{int}	A	0,23±0,13	0,73±0,12	0,75±0,12	0,87±0,06	0,75±0,07
			AD	0,22±0,13	0,93±0,04	0,98±0,02	0,70±0,07	0,42±0,13
		MU	A	0,00±0,00	0,08±0,02	0,14±0,03	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,28±0,01	0,32±0,02	0,05±0,00	0,02±0,00
Cenário 8	4	RHM	A	0,03±0,00	0,90±0,05	0,97±0,02	0,05±0,00	0,04±0,01
			AD	0,03±0,00	0,90±0,05	0,97±0,02	0,05±0,00	0,06±0,01
		WPPA	A	0,02±0,01	0,97±0,03	0,99±0,01	0,70±0,07	0,43±0,15
			AD	0,02±0,01	0,90±0,05	0,97±0,02	0,58±0,05	0,21±0,13
		PP_{int}	A	0,02±0,01	0,87±0,07	0,90±0,06	0,73±0,07	0,46±0,15
			AD	0,12±0,10	0,90±0,05	0,97±0,02	0,55±0,03	0,16±0,10
		MU	A	0,00±0,00	0,13±0,01	0,23±0,02	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,31±0,02	0,39±0,02	0,05±0,00	0,04±0,01
Cenário 9	5	RHM	A	0,00±0,00	0,03±0,02	0,05±0,03	0,05±0,00	0,01±0,00
			AD	0,00±0,00	0,04±0,02	0,14±0,07	0,05±0,00	0,01±0,00
		WPPA	A	0,21±0,05	0,83±0,03	0,93±0,01	0,97±0,00	0,83±0,06
			AD	0,17±0,03	0,80±0,02	0,92±0,01	0,95±0,00	0,89±0,02
		PP_{int}	A	0,39±0,11	0,73±0,08	0,81±0,06	1,00±0,00	0,50±0,08
			AD	0,57±0,05	0,76±0,07	0,85±0,05	1,00±0,00	0,36±0,03
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,23±0,01	0,29±0,01	0,05±0,00	0,03±0,01
Cenário 10	5	RHM	A	0,00±0,00	0,04±0,02	0,12±0,05	0,05±0,00	0,02±0,00

			AD	0,00±0,00	0,06±0,02	0,15±0,05	0,05±0,00	0,02±0,00
		WPPA	A	0,18±0,04	0,85±0,03	0,95±0,01	0,97±0,00	0,86±0,05
			AD	0,21±0,04	0,81±0,06	0,89±0,05	0,95±0,00	0,84±0,05
		PP_{int}	A	0,34±0,08	0,79±0,05	0,89±0,03	1,00±0,00	0,58±0,06
			AD	0,35±0,08	0,61±0,10	0,69±0,09	1,00±0,00	0,50±0,05
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,00±0,00	0,02±0,01	0,05±0,00	0,02±0,00
Cenário 11	2,7	RHM	A	0,01±0,01	0,00±0,00	0,08±0,05	0,05±0,00	0,00±0,00
			AD	0,01±0,01	0,01±0,00	0,11±0,05	0,05±0,00	0,00±0,00
		WPPA	A	0,07±0,01	0,96±0,01	0,98±0,01	0,91±0,01	0,93±0,01
			AD	0,09±0,00	0,93±0,01	0,96±0,01	0,88±0,00	0,91±0,02
		PP_{int}	A	0,08±0,01	1,00±0,00	1,00±0,00	0,84±0,01	0,95±0,00
			AD	0,09±0,00	1,00±0,00	1,00±0,00	0,84±0,00	0,95±0,00
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,02±0,01
Cenário 12	2,7	RHM	A	0,01±0,01	0,01±0,00	0,11±0,06	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,01±0,00	0,10±0,04	0,05±0,00	0,00±0,00
		WPPA	A	0,04±0,02	0,95±0,01	0,98±0,01	0,88±0,01	0,95±0,00
			AD	0,09±0,01	0,92±0,01	0,96±0,01	0,85±0,00	0,94±0,00
		PP_{int}	A	0,09±0,00	1,00±0,00	1,00±0,00	0,84±0,00	0,95±0,00
			AD	0,09±0,00	1,00±0,00	1,00±0,00	0,85±0,00	0,95±0,00
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,01±0,00	0,04±0,01	0,05±0,00	0,04±0,01

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 7 ($h^2 = 0{,}50$) e Cenário 8 ($h^2 = 0{,}60$), 10 QTL: Cenário 9 ($h^2 = 0{,}30$) e Cenário 10 ($h^2 = 0{,}40$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 11 ($h^2 = 0{,}10$) e Cenário 12 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

A identificação de marcadores por meio da análise via marcas únicas foi inferior aos demais critérios com relação ao poder de detecção e porcentagem de explicação para todos os cenários analisados, no entanto, a taxa de falsos positivos nessa metodologia foi sempre igual a zero, ou seja, destaca-se com relação a não classificar erroneamente um SNP como significativo quando realmente esse SNP não era. Porém, para os cenários 7, 8, 10 e 11, os critérios WPPA e PP_{int} apresentaram também baixas taxas de falsos positivos, além disso foram eficientes em termos do poder de detecção.

O critério PP_{int} novamente foi superior nos cenários com arquiteturas genéticas poligênicas no qual as características eram controladas por 100 QTL de pequenos efeitos. Nestes cenários, esse critério apresentou valores máximos de poder e porcentagem de explicação da variância e também obteve altos valores de área sob a curva ROC e baixas taxas de falsos positivos. Para os cenários com arquiteturas genéticas oligogênicas (cenários 7, 8, 9 e 10), o critério PP_{int} apresentou maiores taxas de falsos positivos em comparação com os demais critérios no qual apresentaram valores baixos e próximos de zero. Note que o RHM foi inferior com relação ao poder e porcentagem de explicação da variância nestes cenários mostrando-se eficiente novamente apenas para características controladas por poucos genes. Ainda assim, o RHM apresenta destaque com relação as taxas de falsos positivos obtendo, para todos os cenários, valores zero ou próximos de zero validando novamente o que foi abordado por Usai et al. (2014).

Em todos os cenários considerando grau médio de dominância de 0,5, o critério WPPA apresentou melhores resultados quando a dominância foi negligenciada (modelo aditivo). Para o RHM os valores foram bastantes similares com relação ao considerar ou não a dominância no modelo. Já o PP_{int} e a análise via marcas únicas obtiveram melhores resultados pontuais ao considerar a dominância para os cenários com herança genética oligogênica e, assim como o RHM, apresentaram valores similares para os cenários com herança poligênica. O procedimento de considerar ou não a dominância para estimar os efeitos aditivos na análise via marcas únicas foi utilizado também por Bolormaa et al. (2015) e elucidou importantes resultados nos estudos de GWAS encontrando muitos SNPs associados a características com um efeito de dominância. De acordo com Bennewitz et al (2017) mesmo que se estime apenas os efeitos aditivos, a dominância não deve ser negligenciada uma vez que esses dois efeitos não são completamente independentes e podem conjuntamente por muitas vezes aumentar o poder de detecção de regiões genômicas. Ademais, Wellmann e Bennewitz (2011) também mostraram que os efeitos de dominância e aditivos são dependentes um do outro, porém de maneira

complexa e que grandes efeitos de dominância são geralmente observados para genes com grandes efeitos aditivos.

Como já observado nos cenários sem dominância ($gmd = 0$), os critérios WPPA e PP_{int} apresentaram menores valores de poder nos cenários em que as características eram controladas por 10 QTL (cenários 9 e 10) do que nos demais cenários o que pode ter sido influenciado pelo tamanho da região que foi a maior dentre todos os cenários.

Para o critério PP_{int}, o poder de detecção de regiões e a porcentagem de explicação da variância aumentaram com o aumento da herdabilidade para todos os cenários com herança oligogênica (cenário 7 para o 8 e cenário 9 para o 10) considerando o modelo aditivo, porém, ao considerar a dominância, esses valores diminuíram com o aumento da herdabilidade nestes cenários. Com relação a área sob a curva ROC também foi observado uma diminuição em seus valores com o aumento da herdabilidade nesses cenários para este critério. Já para os cenários com herança poligênica, os valores de poder e porcentagem foram similares com relação ao aumento da herdabilidade de 0,10 para 0,20 para todos os critérios considerados.

Os resultados associados aos efeitos genéticos devido à dominância para grau médio de dominância de 0,5 são apresentados na Tabela 5. Para todos os cenários considerados, novamente o critério WPPA se destacou apresentado melhores resultados com relação ao poder em detectar regiões associadas, porcentagem de explicação da variância e área sob a curva ROC. Nos cenários com arquitetura genética poligênica, o critério PP_{int}, junto ao WPPA, também se mostrou vantajoso com relação a todas as medidas de eficiência utilizadas elucidando novamente sua superioridade nesses cenários.

Tabela 5 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Análise via Marcas Únicas (MU) em cada cenário associados aos efeitos devido à dominância considerando grau médio de dominância de 0,5.

	Distância	Critério	FP	Poder	PE	*Threshold*	Área
Cenário 7	4	RHM	0,03±0,00	0,53±0,05	0,79±0,06	0,05±0,00	0,08±0,01
		WPPA	0,06±0,03	0,87±0,05	0,96±0,02	0,74±0,07	0,50±0,15
		PP_{int}	0,22±0,13	0,87±0,05	0,96±0,02	0,69±0,07	0,46±0,14
		MU	0,71±0,01	0,65±0,01	0,69±0,01	0,05±0,00	0,31±0,01
Cenário 8		RHM	0,03±0,00	0,63±0,09	0,77±0,10	0,05±0,00	0,06±0,01
		WPPA	0,03±0,01	0,87±0,05	0,95±0,03	0,59±0,05	0,22±0,13
		PP_{int}	0,12±0,09	0,83±0,08	0,93±0,03	0,60±0,05	0,18±0,10
		MU	0,74±0,01	0,68±0,01	0,72±0,01	0,05±0,00	0,33±0,01
Cenário 9	5	RHM	0,01±0,01	0,02±0,01	0,11±0,07	0,05±0,00	0,01±0,00
		WPPA	0,87±0,07	0,97±0,02	0,98±0,01	0,70±0,08	0,55±0,03
		PP_{int}	0,55±0,08	0,57±0,09	0,61±0,10	1,00±0,00	0,32±0,04
		MU	0,73±0,02	0,75±0,01	0,74±0,01	0,05±0,00	0,40±0,01
Cenário 10		RHM	0,00±0,00	0,02±0,02	0,09±0,09	0,05±0,00	0,02±0,01
		WPPA	0,63±0,10	0,77±0,09	0,81±0,08	0,94±0,05	0,59±0,04
		PP_{int}	0,47±0,09	0,56±0,12	0,61±0,12	0,95±0,05	0,43±0,03
		MU	0,72±0,01	0,74±0,01	0,74±0,01	0,05±0,00	0,40±0,00
Cenário 11	2,7	RHM	0,00±0,00	0,00±0,00	0,06±0,05	0,05±0,00	0,00±0,00
		WPPA	0,09±0,00	1,00±0,00	1,00±0,00	0,92±0,01	0,95±0,00
		PP_{int}	0,09±0,00	1,00±0,00	1,00±0,00	0,84±0,00	0,95±0,00
		MU	0,70±0,02	0,72±0,01	0,72±0,01	0,05±0,00	0,40±0,02
Cenário 12		RHM	0,01±0,01	0,00±0,00	0,03±0,02	0,05±0,00	0,00±0,00
		WPPA	0,08±0,01	0,98±0,01	0,98±0,01	0,93±0,02	0,95±0,00
		PP_{int}	0,09±0,00	1,00±0,00	1,00±0,00	0,83±0,00	0,95±0,00
		MU	0,74±0,02	0,76±0,01	0,75±0,01	0,05±0,00	0,38±0,02

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 7 ($h^2 = 0{,}50$) e Cenário 8 ($h^2 = 0{,}60$), 10 QTL: Cenário 9 ($h^2 = 0{,}30$) e Cenário 10 ($h^2 = 0{,}40$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 11 ($h^2 = 0{,}10$) e Cenário 12 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

Outro método que também ganhou destaque com relação ao poder de detecção de marcadores associados aos efeitos devido a dominância foi a análise via marcas únicas. No entanto, esse método apresentou as maiores taxas de falsos positivos para todos os cenários corroborando com o que foi mencionado por Hayes (2013) e Resende et al. (2012) em que o nível de significância adotado nos testes poderia causar ocorrências de elevadas taxas falsos positivos. Para os cenários em que as características eram controladas por 10 QTL, as taxas de falsos positivos para o RHM foram sempre inferiores que as encontradas pelos critérios WPPA, análise via marcas únicas e PP_{int}, mostrando que este método é eficiente para controlar essas taxas na GWAS ao considerar os efeitos devido à dominância. O mesmo para este método pode ser observado nos demais cenários para estes efeitos.

Para o critério WPPA ocorreu diminuição do poder com o aumento da herdabilidade para todos os cenários em que 10 ou 100 QTL controlavam a característica (cenários 9, 10, 11 e 12). Já nos cenários 7 e 8 não ocorreu nenhuma mudança no poder com o aumento da herdabilidade para esse critério. Ao contrário do WPPA, para os critérios RHM e PP_{int} houve aumento nos valores de poder com o aumento dos níveis de herdabilidade nos cenários 7 e 8 sendo que, para os demais cenários, esses valores foram similares com relação aos níveis de herdabilidade. Isso foi observado para a análise via marcas únicas nos cenários com arquitetura genética oligogênica.

As taxas de falsos positivos dos efeitos de dominância para a análise via marcas únicas foram maiores do que as valores de falsos positivos dos efeitos aditivos para todos os cenários (Tabela 4 e 5). Isso foi observado para o WPPA nos cenários controlados por 10 QTL. De acordo com Bolormaa et al. (2015) isso sugere que os efeitos de dominância são menores do que os efeitos aditivos e/ou mais difíceis de estimar. Para os demais critérios não se verificou diferenças consideráveis nestas taxas com relação aos efeitos aditivos e devido a dominância. Para os critérios RHM, WPPA e PP_{int}, o poder em detectar efeitos de dominância foi menor do que o poder em detectar efeitos aditivos. Segundo Mao et al. (2020), em seu estudo de associação para dados reais e simulados para característica de fertilidade em vacas, isso pode ser explicado porque para detectar efeitos de dominância necessita-se de um LD entre marca e QTL muito mais forte do que para detectar efeitos aditivos.

Para o RHM, com relação a porcentagem de explicação da variância, os resultados sugerem que os efeitos aditivos contribuem mais para a variância genética do que os efeitos de dominância corroborando com os resultados obtidos por Lopes et al. (2014) em seu estudo de associação considerando efeitos de dominância na característica número de tetas em porcos.

Ademais, Lundregan et al. (2020), estudando resistência de aves à parasitas, mostraram que a contribuição da variância genética aditiva pode ser superestimada quando qualquer variância de dominância existente não é contabilizada. De acordo com esses autores embora a necessidade de considerar efeitos genéticos não aditivos possa não ser aplicável a todas as características, pode ser especialmente importante para características relacionadas à aptidão, como resistência a doenças, que devem exibir a variância de dominância.

Para a análise via marcas únicas os valores de porcentagem de explicação da variância para os efeitos de dominância foram maiores ou similares à porcentagem encontrada para os efeitos aditivos, contrariando os resultados já descritos acima obtidos por Lopes et al. (2014) e corroborando com os resultados reportados por Mao et al. (2020). Já para os demais critérios os resultados com relação a essa medida para os efeitos aditivos e devido a dominância foram similares para todos os cenários analisados.

De acordo com Wang et al. (2004), embora o progresso genético seja mais lento na presença de dominância em comparação com a situação em que apenas efeitos aditivos desempenham um papel, se os efeitos de dominância existirem e não forem devidamente considerados, o progresso genético pode ser ainda mais lento. Além disso, Bonnafous et al. (2018) ao comparar modelos de associação para identificar efeitos não aditivos para tempo de florescimento em girassóis relatam a importância de levar em consideração os desvios de dominância e, consequentemente, a heterose, para identificar regiões genômicas que controlam características de interesse.

Os resultados considerando-se, simultaneamente, os efeitos genéticos aditivos e devido à dominância para grau médio de dominância de 0,5 são apresentados na Tabela 6. Para todas as medidas de eficiência os mesmos resultados encontrados para os efeitos aditivos foram observados, sendo que o WPPA se destacou em todos os cenários, o RHM se destacou nos cenários controlados por 3 QTL e o PP_{int} foi superior nos cenários com herança poligênica.

Tabela 6 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Análise via Marcas Únicas (MU) em cada cenário associados aos efeitos aditivo e devido à dominância considerando grau médio de dominância de 0,5.

	Distância	Critério	FP	Poder	PE	*Threshold*	Área
Cenário 7	4	RHM	0,03±0,00	0,93±0,04	0,97±0,02	0,05±0,00	0,13±0,02
		WPPA	0,04±0,01	1,00±0,00	1,00±0,00	0,77±0,07	0,56±0,15
		PP_{int}	0,21±0,13	0,83±0,08	0,89±0,05	0,68±0,07	0,42±0,14
		MU	0,00±0,00	0,20±0,01	0,23±0,02	0,05±0,00	0,01±0,00
Cenário 8		RHM	0,04±0,01	0,97±0,03	0,99±0,01	0,05±0,00	0,09±0,02
		WPPA	0,04±0,01	0,97±0,03	0,99±0,01	0,67±0,06	0,25±0,12
		PP_{int}	0,10±0,09	0,77±0,10	0,82±0,07	0,58±0,05	0,16±0,11
		MU	0,00±0,00	0,23±0,01	0,29±0,01	0,05±0,00	0,03±0,01
Cenário 9	5	RHM	0,01±0,01	0,09±0,03	0,18±0,07	0,05±0,00	0,05±0,01
		WPPA	0,75±0,09	0,91±0,09	0,92±0,08	0,84±0,08	0,58±0,02
		PP_{int}	0,37±0,07	0,49±0,10	0,54±0,10	1,00±0,00	0,44±0,03
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
Cenário 10		RHM	0,00±0,00	0,13±0,03	0,32±0,06	0,05±0,00	0,06±0,01
		WPPA	0,54±0,12	0,86±0,09	0,89±0,08	0,89±0,07	0,68±0,04
		PP_{int}	0,38±0,11	0,56±0,12	0,59±0,13	0,90±0,07	0,54±0,02
		MU	0,00±0,00	0,00±0,00	0,01±0,01	0,05±0,00	0,01±0,00
Cenário 11	2,7	RHM	0,01±0,01	0,01±0,00	0,11±0,04	0,05±0,00	0,02±0,01
		WPPA	0,09±0,00	0,97±0,02	0,98±0,01	0,97±0,00	0,92±0,01
		PP_{int}	0,09±0,00	1,00±0,00	1,00±0,00	0,82±0,00	0,95±0,00
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,02±0,01
Cenário 12		RHM	0,02±0,01	0,02±0,00	0,12±0,04	0,05±0,00	0,01±0,00
		WPPA	0,08±0,01	0,94±0,01	0,97±0,01	0,97±0,00	0,94±0,01
		PP_{int}	0,09±0,00	1,00±0,00	1,00±0,00	0,82±0,00	0,95±0,00
		MU	0,00±0,00	0,01±0,00	0,03±0,01	0,05±0,00	0,03±0,01

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 7 ($h^2 = 0{,}50$) e Cenário 8 ($h^2 = 0{,}60$), 10 QTL: Cenário 9 ($h^2 = 0{,}30$) e Cenário 10 ($h^2 = 0{,}40$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 11 ($h^2 = 0{,}10$) e Cenário 12 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

Os resultados associados aos efeitos aditivos ao considerar ou não a dominância para grau médio de dominância de 1 são apresentados na Tabela 7. Em geral, os resultados encontrados com relação as medidas de eficiência consideradas para comparação dos critérios são similares aos obtidos considerando grau médio de dominância de 0 e 0,5 (Tabela 3 e 4) na qual o critério WPPA seguido dos critérios PP_{int} e RHM foram superiores a análise via marcas únicas para arquiteturas genéticas oligogênicas e o WPPA e PP_{int} foram superiores para características com herança poligênica. Zhang et al. (2010), utilizando métodos adaptados de modelos lineares mistos para estudos de associação, também verificaram que a superioridade dos métodos não foi alterada ao aumentar os níveis do grau médio de dominância e também ao utilizar diferentes níveis de herdabilidade.

Tabela 7 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Marcas Únicas (MU) em cada cenário associados aos efeitos aditivos considerando o modelo aditivo (A) e o modelo aditivo-dominante (AD) para grau médio de dominância de 1.

	Distância	Critério	Modelo	FP	Poder	PE	*Threshold*	Área
Cenário 13	4	RHM	A	0,02±0,01	0,63±0,08	0,72±0,06	0,05±0,00	0,09±0,01
			AD	0,02±0,00	0,83±0,06	0,88±0,04	0,05±0,00	0,09±0,01
		WPPA	A	0,06±0,01	1,00±0,00	1,00±0,00	0,96±0,01	0,97±0,00
			AD	0,01±0,01	1,00±0,00	1,00±0,00	0,79±0,06	0,41±0,15
		PP_{int}	A	0,15±0,10	0,63±0,11	0,72±0,09	0,93±0,05	0,73±0,06
			AD	0,06±0,06	0,87±0,07	0,9±0,07	0,80±0,06	0,33±0,12
		MU	A	0,00±0,00	0,04±0,03	0,05±0,04	0,05±0,00	0,00±0,00
			AD	0,01±0,00	0,14±0,03	0,16±0,04	0,05±0,00	0,08±0,02
Cenário 14	4	RHM	A	0,03±0,00	0,77±0,10	0,85±0,07	0,05±0,00	0,06±0,01
			AD	0,02±0,00	0,90±0,05	0,94±0,03	0,05±0,00	0,07±0,01
		WPPA	A	0,05±0,02	1,00±0,00	1,00±0,00	0,96±0,01	0,97±0,00
			AD	0,00±0,00	1,00±0,00	1,00±0,00	0,52±0,01	0,00±0,00
		PP_{int}	A	0,12±0,10	0,77±0,09	0,81±0,08	0,94±0,05	0,82±0,05
			AD	0,00±0,00	1,00±0,00	1,00±0,00	0,53±0,01	0,01±0,00
		MU	A	0,00±0,00	0,09±0,02	0,14±0,03	0,05±0,00	0,00±0,00
			AD	0,02±0,00	0,21±0,02	0,27±0,02	0,05±0,00	0,09±0,01
Cenário 15	5	RHM	A	0,01±0,00	0,01±0,01	0,06±0,06	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,01±0,01	0,06±0,06	0,05±0,00	0,00±0,00
		WPPA	A	0,30±0,03	0,75±0,03	0,86±0,02	0,96±0,00	0,77±0,03
			AD	0,23±0,02	0,75±0,04	0,85±0,04	0,96±0,00	0,82±0,02
		PP_{int}	A	0,34±0,07	0,57±0,07	0,62±0,07	1,00±0,00	0,50±0,04
			AD	0,71±0,06	0,86±0,04	0,90±0,03	1,00±0,00	0,27±0,05
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,03±0,01

Cenário 16	5	RHM	A	0,01±0,00	0,01±0,01	0,04±0,04	0,05±0,00	0,00±0,00
			AD	0,01±0,01	0,02±0,01	0,09±0,06	0,05±0,00	0,01±0,01
		WPPA	A	0,21±0,03	0,79±0,03	0,89±0,02	0,96±0,00	0,86±0,02
			AD	0,18±0,02	0,82±0,03	0,91±0,02	0,95±0,00	0,89±0,01
		PP_{int}	A	0,49±0,08	0,74±0,05	0,79±0,04	1,00±0,00	0,43±0,05
			AD	0,43±0,09	0,72±0,10	0,76±0,10	1,00±0,00	0,45±0,06
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,02±0,01	0,04±0,01	0,05±0,00	0,04±0,01
Cenário 17	2,7	RHM	A	0,01±0,01	0,00±0,00	0,07±0,04	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,00±0,00	0,05±0,03	0,05±0,00	0,00±0,00
		WPPA	A	0,00±0,00	1,00±0,00	1,00±0,00	0,85±0,01	0,86±0,04
			AD	0,00±0,00	1,00±0,00	1,00±0,00	0,86±0,01	0,93±0,02
		PP_{int}	A	0,00±0,00	1,00±0,00	1,00±0,00	0,84±0,01	0,90±0,02
			AD	0,00±0,00	1,00±0,00	1,00±0,00	0,85±0,01	0,94±0,02
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,02±0,01	0,01±0,00	0,02±0,01	0,05±0,00	0,05±0,01
Cenário 18	2,7	RHM	A	0,02±0,01	0,00±0,00	0,01±0,01	0,05±0,00	0,00±0,00
			AD	0,00±0,00	0,01±0,00	0,05±0,03	0,05±0,00	0,00±0,00
		WPPA	A	0,00±0,00	1,00±0,00	1,00±0,00	0,85±0,01	0,95±0,01
			AD	0,00±0,00	0,99±0,01	1,00±0,00	0,85±0,01	0,93±0,01
		PP_{int}	A	0,00±0,00	1,00±0,00	1,00±0,00	0,84±0,00	0,93±0,01
			AD	0,00±0,00	1,00±0,00	1,00±0,00	0,85±0,01	0,94±0,01
		MU	A	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,00±0,00
			AD	0,11±0,03	0,06±0,01	0,14±0,02	0,05±0,00	0,10±0,02

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 13 ($h^2 = 0{,}50$) e Cenário 14 ($h^2 = 0{,}60$), 10 QTL: Cenário 15 ($h^2 = 0{,}30$) e Cenário 16 ($h^2 = 0{,}40$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 17 ($h^2 = 0{,}10$) e Cenário 18 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

Ao contrário dos resultados obtidos para grau médio de dominância de 0,5 (Tabela 4), para o grau médio de dominância de 1 não se verificou para o critério WPPA diferença no poder de detecção ao considerar ou não a dominância. Já para os demais critérios os melhores resultados para os cenários com arquitetura oligogênicas foram obtidos ao considerar a dominância corroborando com o que for reportado por Bennewitz et al (2017).

Para os cenários com herança oligogênica, o poder de detecção para efeitos aditivos considerando ou não a dominância dos critérios RHM e marcas únicas diminuiu com o aumento do grau médio de dominância (Tabela 3, 4 e 7). Isso foi observado para o PP_{int} ao negligenciar a dominância e para o WPPA nos cenários em que a característica era controlada por 10 QTL.

Os resultados associados aos efeitos genéticos devido à dominância para grau médio de dominância de 1 são apresentados na Tabela 8 e se mostram equivalentes aos resultados obtidos para grau médio de dominância de 0,5, no entanto, a superioridade dos critérios WPPA e PP_{int} em comparação a análise via marcas únicas com relação ao poder de detecção e porcentagem de explicação da variância foi menos evidente apresentando valores similares, principalmente, em cenários com características controladas por 10 QTL.

Tabela 8 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Análise via Marcas Únicas (MU) em cada cenário associados aos efeitos devido à dominância considerando grau médio de dominância de 1.

	Distância	Critério	FP	Poder	PE	*Threshold*	Área
Cenário 13	4	RHM	0,01±0,00	0,87±0,05	0,94±0,03	0,05±0,00	0,08±0,02
		WPPA	0,01±0,00	0,97±0,03	1,00±0,00	0,74±0,07	0,42±0,15
		PP_{int}	0,17±0,11	0,87±0,07	0,92±0,06	0,69±0,07	0,31±0,10
		MU	0,76±0,01	0,71±0,01	0,73±0,01	0,05±0,00	0,34±0,00
Cenário 14		RHM	0,02±0,00	0,93±0,04	0,96±0,03	0,05±0,00	0,07±0,01
		WPPA	0,00±0,00	0,93±0,04	0,99±0,01	0,57±0,03	0,01±0,00
		PP_{int}	0,00±0,00	0,93±0,04	0,99±0,01	0,58±0,03	0,01±0,00
		MU	0,78±0,00	0,73±0,01	0,76±0,01	0,05±0,00	0,35±0,00
Cenário 15	5	RHM	0,00±0,00	0,01±0,01	0,03±0,03	0,05±0,00	0,00±0,00
		WPPA	0,32±0,03	0,72±0,04	0,85±0,03	0,98±0,00	0,77±0,02
		PP_{int}	0,73±0,05	0,80±0,05	0,84±0,05	1,00±0,00	0,23±0,04
		MU	0,71±0,01	0,76±0,01	0,76±0,01	0,05±0,00	0,44±0,00
Cenário 16		RHM	0,00±0,00	0,03±0,02	0,13±0,07	0,05±0,00	0,01±0,00
		WPPA	0,21±0,03	0,73±0,05	0,86±0,03	0,98±0,00	0,83±0,02
		PP_{int}	0,54±0,10	0,69±0,08	0,76±0,08	0,95±0,05	0,41±0,05
		MU	0,74±0,01	0,77±0,01	0,77±0,01	0,05±0,00	0,43±0,00
Cenário 17	2,7	RHM	0,00±0,00	0,01±0,00	0,19±0,09	0,05±0,00	0,00±0,00
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,86±0,00	0,89±0,03
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,83±0,00	0,90±0,02
		MU	0,77±0,02	0,77±0,01	0,76±0,01	0,05±0,00	0,41±0,02
Cenário 18		RHM	0,00±0,00	0,00±0,00	0,01±0,01	0,05±0,00	0,00±0,00
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,85±0,00	0,91±0,02
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,83±0,00	0,92±0,02

MU	0,80±0,00	0,82±0,01	0,80±0,01	0,05±0,00	0,37±0,03

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 13 ($h^2 = 0{,}50$) e Cenário 14 ($h^2 = 0{,}60$), 10 QTL: Cenário 15 ($h^2 = 0{,}30$) e Cenário 16 ($h^2 = 0{,}40$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 17 ($h^2 = 0{,}10$) e Cenário 18 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

Para os cenários considerando características controladas por 3 QTL, o poder de detecção e a porcentagem de explicação da variância de todos os critérios aumentou com o aumento do grau médio de dominância de 0,5 para 1 (Tabela 6 e 8). Isso foi observado para o critério PP_{int} nos cenários com características controladas por 10 QTL e para a análise via marcas únicas nos cenários com arquitetura poligênica. Resultados semelhantes já foram descritos por Mao et al. (2020) e segundo os autores isso é claramente demostrado uma vez que se houver mais peso na dominância na ação do gene, maior é o poder de detecção alcançado.

Como abordado acima, o poder de detecção associado aos efeitos devido a dominância para características, controladas por poucos genes de efeitos maiores, aumentou com relação ao aumento do grau médio de dominância. De acordo com Bennewitz et al. (2017), isso se deve ao fato de que à medida que o efeito aditivo de marcador aumenta, o grau médio de dominância provavelmente se torna mais positivo e, em média, aumenta também seu nível e assim os efeitos devido à dominância se tornam mais passíveis de serem detectados. Esse padrão corrobora também aos resultados de Caballero e Keightley (1994), que descobriram que genes com grandes efeitos aditivos, provavelmente, têm grau médio de dominância maiores.

Para tamanhos de regiões menores (cenários controlados por 3 e 100 QTL) a taxa de falsos positivos para os critérios que selecionam regiões (RHM, WPPA e PP_{int}) diminuiu ao aumentarmos o grau médio de dominância de 0,5 para 1, tanto para os efeitos aditivos considerando o modelo aditivo-dominante (Tabela 4 e 7) quanto para os efeitos devido à dominância (Tabela 5 e 8), evidenciando a necessidade de considerar a dominância nos estudos de associação em caraterísticas complexas.

Os resultados considerados simultaneamente os efeitos genéticos aditivos e devido à dominância para grau médio de dominância de 1 são apresentados na Tabela 9. Em geral, para todas as medidas de eficiência, os mesmos resultados encontrados para os efeitos aditivos foram observados.

Tabela 9 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão das taxas de falsos positivos (FP), do poder de detecção (PD), da porcentagem da variância genética capturada (PE), da área sob a curva ROC e do *threshold* de seleção de regiões obtidos pelos critérios RHM, WPPA, PP_{int} e Análise via Marcas Únicas (MU) em cada cenário associados aos efeitos aditivo e devido à dominância considerando grau médio de dominância de 1.

	Distância	Critério	FP	Poder	PE	*Threshold*	Área
Cenário 13	4	RHM	0,04±0,00	1,00±0,00	1,00±0,00	0,05±0,00	0,11±0,01
		WPPA	0,09±0,05	0,97±0,03	0,98±0,03	0,91±0,03	0,48±0,13
		PP_{int}	0,14±0,10	0,83±0,09	0,86±0,08	0,67±0,07	0,26±0,11
		MU	0,00±0,00	0,10±0,02	0,11±0,03	0,05±0,00	0,06±0,01
Cenário 14		RHM	0,04±0,00	1,00±0,00	1,00±0,00	0,05±0,00	0,08±0,01
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,68±0,04	0,04±0,02
		PP_{int}	0,00±0,00	0,93±0,04	0,96±0,03	0,50±0,00	0,00±0,00
		MU	0,01±0,00	0,15±0,01	0,20±0,02	0,05±0,00	0,07±0,01
Cenário 15	5	RHM	0,01±0,01	0,06±0,02	0,16±0,08	0,05±0,00	0,04±0,01
		WPPA	0,64±0,15	0,74±0,13	0,76±0,12	0,75±0,08	0,55±0,02
		PP_{int}	0,53±0,08	0,70±0,07	0,74±0,07	1,00±0,00	0,38±0,05
		MU	0,00±0,00	0,00±0,00	0,00±0,00	0,05±0,00	0,03±0,00
Cenário 16		RHM	0,02±0,01	0,07±0,03	0,19±0,08	0,05±0,00	0,05±0,01
		WPPA	0,58±0,11	0,84±0,11	0,86±0,09	0,99±0,00	0,63±0,02
		PP_{int}	0,46±0,11	0,69±0,09	0,70±0,09	0,90±0,07	0,52±0,04
		MU	0,00±0,00	0,02±0,01	0,04±0,01	0,05±0,00	0,03±0,01
Cenário 17	2,7	RHM	0,01±0,01	0,01±0,00	0,14±0,04	0,05±0,00	0,01±0,01
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,97±0,00	0,83±0,05
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,83±0,00	0,91±0,02
		MU	0,00±0,00	0,01±0,00	0,02±0,01	0,05±0,00	0,05±0,01
Cenário 18		RHM	0,00±0,00	0,01±0,00	0,05±0,03	0,05±0,00	0,00±0,00
		WPPA	0,00±0,00	1,00±0,00	1,00±0,00	0,96±0,00	0,87±0,03
		PP_{int}	0,00±0,00	1,00±0,00	1,00±0,00	0,82±0,00	0,86±0,03

MU	0,04±0,02	0,05±0,01	0,12±0,02	0,05±0,00	0,09±0,02

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 13 ($h^2 = 0{,}50$) e Cenário 14 ($h^2 = 0{,}60$), 10 QTL: Cenário 15 ($h^2 = 0{,}30$) e Cenário 16 ($h^2 = 0{,}40$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 17 ($h^2 = 0{,}10$) e Cenário 18 ($h^2 = 0{,}20$). Critérios – RHM: mapeamento de herdabilidades regionais, WPPA: probabilidade *a posteriori* da associação da região genômica e PP_{int}: probabilidade *a posteriori* do intervalo.

Os valores da probabilidade *a priori* π de um marcador ter um efeito não nulo estimada via método bayesiano Bayes Dπ são apresentados na Tabela 10. A probabilidade π obtida pelo método BayesDπ variou de 0,15 a 0,47 entre os cenários indicando que o número de marcadores que supostamente estão em LD com o QTL variaram de 300 a 940. De acordo com Resende et al. (2014), essa modelagem de π para previamente selecionar um grupo de marcadores torna-se necessária nos estudos de associação uma vez que, a maioria das marcas não estão em LD com os QTL.

Tabela 10 – Tamanho das regiões (distância) em cM encontrada por meio do desequilíbrio de ligação (LD) entre os marcadores e QTL em cada cenário e as respectivas médias e erros-padrão da probabilidade π estimada via BayesDπ considerando o modelo aditivo e o modelo aditivo-dominante.

Cenários	Distância	Modelo Aditivo	Modelo Aditivo-Dominante	
		Aditivo	Aditivo	Dominante
1	4	0,32±0,03	-	-
2		0,16±0,04	-	-
3		0,45±0,00	-	-
4		0,46±0,00	-	-
5		0,46±0,00	-	-
6		0,47±0,00	-	-
7	5	0,34±0,02	0,17±0,04	0,17±0,04
8		0,15±0,05	0,06±0,03	0,06±0,03
9		0,45±0,00	0,45±0,01	0,45±0,01
10		0,45±0,00	0,44±0,01	0,44±0,00
11		0,45±0,01	0,47±0,00	0,47±0,00
12		0,46±0,00	0,46±0,00	0,46±0,00
13	2,7	0,41±0,02	0,18±0,05	0,18±0,05
14		0,39±0,02	0,01±0,00	0,01±0,00
15		0,45±0,00	0,46±0,01	0,46±0,01
16		0,45±0,00	0,45±0,00	0,45±0,00
17		0,46±0,00	0,47±0,00	0,47±0,00
18		0,46±0,01	0,47±0,01	0,47±0,01

Cenários com arquitetura genética oligogênica – 3 QTL: Cenário 1 ($h^2 = 0{,}50$ e $gmd = 0$), Cenário 2 ($h^2 = 0{,}60$ e $gmd = 0$), Cenário 7 ($h^2 = 0{,}50$ e $gmd = 0{,}5$), Cenário 8 ($h^2 = 0{,}60$ e $gmd = 0{,}5$), Cenário 13 ($h^2 = 0{,}50$ e $gmd = 1$), Cenário 14 ($h^2 = 0{,}60$ e $gmd = 1$), 10 QTL: Cenário 3 ($h^2 = 0{,}30$ e $gmd = 0$), Cenário 4 ($h^2 = 0{,}40$ e $gmd = 0$), Cenário 9 ($h^2 = 0{,}30$ e $gmd = 0{,}5$), Cenário 10 ($h^2 = 0{,}40$ e $gmd = 0{,}5$), Cenário 15 ($h^2 = 0{,}30$ e $gmd = 1$) e Cenário 16 ($h^2 = 0{,}40$ e $gmd = 1$). Cenários com arquitetura genética poligênica – 100 QTL: Cenário 5 ($h^2 = 0{,}10$ e $gmd = 0$), Cenário 6 ($h^2 = 0{,}20$ e $gmd = 0$), Cenário 11 ($h^2 = 0{,}10$ e $gmd = 0{,}5$), Cenário 12 ($h^2 = 0{,}20$ e $gmd = 0{,}5$), Cenário 17 ($h^2 = 0{,}10$ e $gmd = 1$), Cenário 18 ($h^2 = 0{,}20$ e $gmd = 1$).

Menores valores de π foram encontrados para os cenários em que apenas 3 QTL controlavam a característica sendo sempre estes valores menores quando considerado o modelo aditivo-dominante tanto para os efeitos aditivos quanto para os efeitos devido à dominância. Para os demais cenários os valores de π foram similares independente do modelo analisado. De acordo com Fernando e Garrick (2017), valores mais altos de π podem ser mais discriminatórios para a identificação do QTL de maior efeito, o que é um fator importante para a seleção de SNPs. Não foi observada também nenhuma diferença nos valores de probabilidade para os diferentes graus médios de dominância considerados.

4 Conclusões

Considerando características com herança oligogênica, os critérios WPPA, PP_{int} e RHM se mostraram superiores para todos os graus médios de dominância analisados apresentando maiores valores de poder de detecção, capturando maiores porcentagens da variância genética e maiores áreas sob a curva ROC tanto para efeitos aditivos quando para efeitos devido à dominância. Já para características com herança poligênica, apenas os critérios PP_{int} e WPPA podem ser considerados superiores aos demais para encontrar regiões podendo ser utilizados vantajosamente nos estudos de associação. A análise via marcas únicas se destacou apenas quando foram considerados apenas os efeitos devido à dominância. Em geral, para tamanhos de regiões menores a taxa de falsos positivos para os critérios que selecionam regiões genômicas diminuiu ao aumentarmos o grau médio de dominância de 0,5 para 1 evidenciando a necessidade de considerar a dominância nos estudos de associação em caraterísticas complexas.

Referências

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

BENNEWITZ, J. et al. Application of a Bayesian dominance model improves power in quantitative trait genome-wide association analysis. **Genetics Selection Evolution**, v. 49, n. 1, p. 7, 2017.

BOLORMAA, S. et al. Non-additive genetic variation in growth, carcass and fertility traits of beef cattle. **Genetics Selection Evolution**, 124(7):1315-24, 2015.

BONNAFOUS, F. et al. Comparison of GWAS models to identify non-additive genetic control of flowering time in sunflower hybrids. **Theoretical and applied genetics**, v. 131, n. 2, p. 319-332, 2018.

BRAZ, C. U. et al. Sliding window haplotype approaches overcome single SNP analysis limitations in identifying genes for meat tenderness in Nelore cattle. **BMC genetics**, v. 20, n. 1, p. 8, 2019.

CABALLERO, A.; KEIGHTLEY, P. D. A pleiotropic nonadditive model of variation in quantitative traits. **Genetics**, v. 138, n. 3, p. 883-900, 1994.

COVARRUBIAS-PAZARAN, G. Genome-assisted prediction of quantitative traits using the R package sommer. **PloS one**, v. 11, n. 6, p. e0156744, 2016.

CRUZ, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**. v.35, n.3, p.271-276, 2013.

DU, Q. et al. Genetic architecture of growth traits in Populus revealed by integrated quantitative trait locus (QTL) analysis and association studies. **New Phytologist**, v. 209, n. 3, p. 1067-1082, 2016.

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

FERNANDO, R. L., et al. Controlling the proportion of false positives in multiple dependent tests. **Genetics**, 166.1: 611-619, 2004.

FERNANDO, R. L.; GARRICK, D. Bayesian methods applied to GWAS. In: Genome-wide association studies and genomic prediction. **Humana Press, Totowa**, NJ, 2013. p. 237-274.

GAGE, J. L. et al. Comparing genome-wide association study results from different measurements of an underlying phenotype. **G3: Genes, Genomes, Genetics**, v. 8, n. 11, p. 3715-3722, 2018.

GEWEKE, J. et al. **Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments**. Minneapolis, MN: Federal Reserve Bank of Minneapolis, Research Department, 1992.

GODDARD, M. E. et al. Genetics of complex traits: prediction of phenotype, identification of causal polymorphisms and genetic architecture. **Proceedings of the Royal Society B: Biological Sciences**, v. 283, n. 1835, p. 20160569, 2016.

GODDARD, M. E.; HAYES, B. J.; MEUWISSEN, T. H. E. Using the genomic relationship matrix to predict the accuracy of genomic selection. **Journal of animal breeding and genetics**, v. 128, n. 6, p. 409-421, 2011.

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

HAYES, B. Overview of statistical methods for genome-wide association studies (GWAS). In: Genome-wide association studies and genomic prediction. **Humana Press**, Totowa, NJ, 2013. p. 149-169.

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

LEE, S. et al. Rare-variant association analysis: study designs and statistical tests. **The American Journal of Human Genetics**, v. 95, n. 1, p. 5-23, 2014.

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**, 12.2: e1005767, 2016.

LOPES, M. S. et al. A genome-wide association study reveals dominance effects on number of teats in pigs. **PloS one**, v. 9, n. 8, 2014.

LU, M. et al. Association genetics of growth and adaptive traits in loblolly pine (Pinus taeda L.) using whole-exome-discovered polymorphisms. **Tree Genetics & Genomes**, v. 13, n. 3, p. 57, 2017.

LUNDREGAN, S. L. et al. Resistance to gapeworm parasite has both additive and dominant genetic components in house sparrows, with evolutionary consequences for ability to respond to parasite challenge. **Molecular Ecology**, 2020.

MAO, X. et al. Genome-wide association mapping for dominance effects in female fertility using real and simulated data from Danish Holstein cattle. **Scientific Reports**, v. 10, n. 1, p. 1-9, 2020.

MATIKA, O. et al. Genome-wide association reveals QTL for growth, bone and in vivo carcass traits as assessed by computed tomography in Scottish Blackface lambs. **Genetics Selection Evolution**, v. 48, n. 1, p. 1-15, 2016.

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

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

MOORE, J. H.; ASSELBERGS, F. W.; WILLIAMS, S. M. Bioinformatics challenges for genome-wide association studies. **Bioinformatics**, v. 26, n. 4, p. 445-455, 2010.

NAGAMINE, Y. et al. Localising loci underlying complex trait variation using regional genomic relationship mapping. **PLoS ONE**, 7: e46501, 2012.

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

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

R Core Team. R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/, 2018.

RESENDE, M. D. V. et al. **Computação da seleção genômica ampla (GWS)**. Embrapa Florestas-Documentos (INFOTECA-E), 2010.

RESENDE, M. D. V.; 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, Visconde do Rio Branco, 2014.

RESENDE, R. T. et al. Regional heritability mapping and genome-wide association identify loci for complex growth, wood and disease resistance traits in Eucalyptus. **New Phytologist**, v. 213, n. 3, p. 1287-1300, 2017.

SAHANA, G. et al. Comparison of association mapping methods in a complex pedigreed population. **Genetic Epidemiology**, 34(5):455-62, 2010.

SÁNCHEZ, C. F. B. et al. Seleção genômica ampla em populações derivadas de acasalamento ao acaso ou de autofecundação. Tese (Doutorado em Genética e Melhoramento), Universidade Federal de Viçosa, Viçosa, MG, 2013.

SANT'ANNA, I. C. et al. Multigenerational prediction of genetic values using genome-enabled prediction. **PloS one**, v. 14, n. 1, p. e0210531, 2019.

SCHUSTER, I; CRUZ, C. D. Estatística genômica aplicada a populações derivadas de cruzamentos controlados. **In: Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. 2004. p. 568-568.

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

SOLLERO, B. P. et al. Tag SNP selection for prediction of tick resistance in Brazilian Braford and Hereford cattle breeds using Bayesian methods. **Genetics Selection Evolution**, v. 49, n. 1, p. 49, 2017.

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

USAI, M.G. et al. XVIth QTLMAS: simulated dataset and comparative analysis of submitted results for QTL mapping and genomic evaluation. **BMC Proceedings**, 8: 1–9, 2014.

VAN RADEN, P.M. Efficient Methods to compute genomic predictions. **Journal of Dairy Science**, 91: 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. et al. Simulating the effects of dominance and epistasis on selection response in the CIMMYT Wheat Breeding Program using QuCim. **Crop Science**, v. 44, n. 6, p. 2006-2018, 2004.

WANG, Q. et al. A SUPER powerful method for genome wide association study. **PloS one**, 9.9: e107684, 2014.

WELLMANN, R.; BENNEWITZ, J. The contribution of dominance to the understanding of quantitative genetic variation. **Genetics Research**, v. 93, n. 2, p. 139-154, 2011.

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