LEONARDO DE AZEVEDO PEIXOTO

ABORDAGEM SOBRE MODELOS, COVARIÁVEIS E ACURÁCIA NA SELEÇÃO GENÔMICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

VIÇOSA
MINAS GERAIS - BRASIL
2016

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

T

P379a
2016

Peixoto, Leonardo de Azevedo, 1990-
Abordagem sobre modelos, covariáveis e acurácia na seleção genômica / Leonardo de Azevedo Peixoto. - Viçosa, MG, 2016.
xv, 167f. : il. ; 29 cm.

Inclui anexos.
Orientador: Leonardo Lopes Bhering.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Plantas - Melhoramento genético. 2. Plantas - Seleção. 3. Predição (Genética quantitativa). 4. Marcadores moleculares. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 631.52

LEONARDO DE AZEVEDO PEIXOTO

ABORDAGEM SOBRE MODELOS, COVARIÁVEIS E ACURÁCIA NA SELEÇÃO GENÔMICA

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Doctor Scientiae*.

APROVADA: 30 de novembro de 2016

Camila Ferreira Azevedo

Cosme Damião Cruz
(Coorientador)

Aparecida Célia Paula dos Santos

Pedro Crescêncio Souza Carneiro

Leonardo Lopes Bhering
(Orientador)

A Deus,
Por me indicar o caminho
entre as pedras

OFEREÇO

Aos meus pais Marise e Vanildo,
por toda dedicação, companheirismo
e carinho em todas as horas da minha vida

DEDICO

"Nunca deixem que lhe digam que não vale apena acreditar em um sonho que tem, ou que seus planos nunca vão dar certo, ou que você nunca vai ser alguém. Quem acredita sempre alcança". (Renato Russo)

AGRADECIMENTOS

A Deus, por sempre iluminar meu caminho e me ajudar a escolher o que é certo na minha vida.

Aos meus pais, Marise Aparecida Azevedo Peixoto e Vanildo Peixoto Lacerda por toda dedicação durante estes anos e por tudo que fizeram por mim até hoje, sempre torcendo e rezando para que eu conquiste aquilo que almejo.

As minhas afilhadas Lara e Laisa, que sempre me proporcionam momentos de eterna alegria.

A minha namorada Tássia Boeno de Oliveira por sempre estar ao meu lado nos momentos bons e ruins me apoiando, ajudando e torcendo.

A Universidade Federal de Viçosa (UFV), pela oportunidade oferecida para realização dos trabalhos.

Ao Conselho Nacional do Desenvolvimento Cientifico e Tecnológico (Cnpq), pela concessão da bolsa de estudos.

Ao professor Leonardo Lopes Bhering, pela sua orientação, amizade, disponibilidade, dedicação e ensinamento.

Ao professor Cosme Damião Cruz pela sua co-orientação, disponibilidade e dedicação para execução deste trabalho.

Ao professor Fabyano Fonseca e Silva pela sua co-orientação e amizade ao longo destes anos.

Ao professor Pedro Crescêncio Souza Carneiro pela sua amizade, ensinamento e disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

À professora Camila Ferreira Azevedo, pela sua amizade, ensinamento e disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

Á professora Aparecida Célia Paula dos Santos, pela disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

A todos os professores e colegas do Programa de Pós-Graduação em genética e Melhoramento, em especial aos secretários Marco Túlio e Odilon Junior que sempre me ajudaram a resolver inúmeras buracrácias, e a todos aqueles que contribuíram de alguma forma para a execução deste trabalho e para minha formação acadêmica.

Aos amigos dos laboratórios de Bioinformática e Biometria pelos momentos compartilhados: Nadson, Vinicius, Lidiane, Sara, Jâneo, Paulo, Bruno, Haroldo, Isabela, Luciano, Juan, Rafael.

Aos amigos Edson, Marcos, Vinicius, Edineia, Jetcelainy, Michael, Eduarda, Kelly, Roberta, Paulo Sérgio, Wanderley e Luiz Fernando que me proporcionam momentos muito felizes e sempre me apoiaram e torceram por mim em mais esta etapa da minha vida.

A todos que de alguma forma contribuíram com amizade, dedicação e apoio para que eu chegasse ao final de mais esta etapa da minha vida, deixo aqui meus sinceros e eternos agradecimentos.

BIOGRAFIA

LEONARDO DE AZEVEDO PEIXOTO, filho de Marise Aparecida Azevedo Peixoto e Vanildo Peixoto Lacerda, nasceu em Alegre, Espírito Santo, no dia 18 de abril de 1990.

No Município de Alegre, cursou o ensino primário na Escola Unidocente Benjamim Barros, de 1996 a 1999. Na Escola Estadual de Ensino Fundamental e Médio Aristeu Aguiar cursou o ensino fundamental de 2000 a 2003.

Em 2003, iniciou o ensino médio e o curso de Técnico em Agropecuária pela Escola Agrotécnica Federal de Alegre (EAFA), onde estudou até 2006.

Em 2007, iniciou a graduação em Agronomia pelo Centro de Ciências Agrárias da Universidade Federal do Espírito Santo (CCAUFES), colando grau em fevereiro de 2012.

Em março de 2012, iniciou o mestrado em Genética e Melhoramento pela Universidade Federal de Viçosa. Em fevereiro de 2013 foi sua defesa.

Em março de 2013, iniciou o doutorado em Genética e Melhoramento pela Universidade Federal de Viçosa, sendo que de fevereiro de 2015 a janeiro de 2016 ele fez doutorado sanduíche na Iowa State University, nos Estados Unidos. Em 30 de novembro de 2016 foi sua defesa.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

RESUMO

PEIXOTO, Leonardo de Azevedo, D.Sc., Universidade Federal de Viçosa, novembro de 2016. **Abordagem sobre modelos, covariáveis e acurácia na seleção genômica.** Orientador: Leonardo Lopes Bhering. Coorientadores: Cosme Damião Cruz e Fabyano Fonseca e Silva.

A seleção genômica (SG) tem se tormado uma ferramenta de grande potencial no melhoramento de plantas. Além dela, o estudo de associação genômica (EAGA) e a seleção assistida por marcadores moleculares (SAM) também são metodologias com aplicabilidade no melhoramento. A diferença básica entre essas metodologias é que enquanto a SAM utiliza mapas de ligação e o EAGA utiliza mapas de associação para identificar marcadores significativos, a SG utiliza todos os marcadores disponíveis sem a necessidade de nenhum tipo de mapa. Portanto os objetivos desta pesquisa foram: 1) avaliar modelos utilizando os SNPs significativos encontrados pelos SAM e EAGA como efeito fixo nos modelos comumente utilizados na SG, em que no modelo tradicional, todos os SNPs são estabelecidos como de efeito aleatório. Estes modelos foram comparados com o modelo padrão utilizado na SG (RRBLUP bayesiano); 2) comparar os métodos tradicionais de seleção genômica (todos os SNPs como efeito aleatório); 3) verificar como a herdabilidade e o número de QTLs que controlam a característica podem influenciar na predição do valor genético; 4) estabelecer uma equação de predição da correlação genética em função da correlação fenotípica; 5) estabelecer o número ideal de indivíduos para compor a população de treinamento e; 6) estabelecer a quantidade necessária de marcadores para obter máxima acurácia pelos métodos de seleção genômica. Foram simuladas populações F_2 com 1.000 indivíduos em diferentes cenários. As populações foram simuladas com 4.500 (objetivo 1) e 3.000 marcadores (demais objetivos). Foram simuladas características com diferentes herdabilidades (5, 20, 40, 60, 80 e 99%) e o número de QTLs (60, 120, 180 e 240) (objetivos 2, 3 e 4). Foram estimados para todos os cenários a capacidade preditiva fenotípica e genotípica, a acurácia fenotipica e genotípica, a herdabilidade genômica, a variância genética, o ganho com a seleção, o índice de coincidência e o tempo de processamento. Foi utilizado a cross validação 5-fold com 50 repetições. As principais conclusões desta pesquisa foram: 1) A utilização de um modelo de SG com as marcas significativas encontradas pelo EAGA como efeito fixo e as demais marcas como efeito aleatório é uma boa estratégia para selecionar indivíduos superiores com alta acurácia; 2) A introdução no modelo de SG de QTLs que já foram descritos

previamente para a característica em estudo, como efeito fixo, permite a seleção de indivíduos superiores de forma mais acurada; 3) os modelos de seleção genômica para predição em populações F_2 devem ser compostos por 200 a 900 marcadores de maior efeito sobre a característica e mais de 600 indivíduos na população de treinamento.

ABSTRACT

PEIXOTO, Leonardo de Azevedo, D.Sc., Universidade Federal de Viçosa, November, 2016. **Approach on models, covariables and accuracy in the genomic selection.** Adviser: Leonardo Lopes Bhering. Co-advisers: Cosme Damião Cruz and Fabyano Fonseca e Silva.

Genomic selection (GS) has become a high potential tool in plant breeding. Moreover, genomic wide association study (GWAS) and marker-assisted selection (MAS) are also methodologies with great potential in plant breeding. The basic difference among them is while MAS requires linkage mapping and GWAS requires association mapping to identify significant markers, GWS performs all available markers without any mapping. Therefore, the objectives in this research were: 1) to evaluate models using significant SNPs found by GWAS and MAS as fixed effect in the widely GS models, which, in the traditional model all SNPs are treated as random effect. These models were compared with the standart GS model (Bayesian RRBLUP); 2) To compare the most GS traditional models (all SNPs as random effect); 3) to verify how the heritability and number of QTLs which control a specific trait can influence for predicting genetic value; 4) to establish a prediction equation to estimate the genetic correlation based on phenotypic correlation; 5) to establish the optimal number of individuals to compose the training population and; 6) to establish the number of markers needed to obtain the maximum accuracy by the genomic selection methods. F_2 population was simulated with 1,000 individuals in several scenarios. Populations were simulated with 4,500 (objective 1) and 3,000 markers (other objectives). Traits with different heritability (5%, 20%, 40%, 60%, 80% and 99%) and numbers of QTLs (60, 120, 180 and 240 – objectives 2, 3, and 4) were simulated. Phenotypic and genotypic predictive ability, phenotypic and genotypic accuracy, genomic heritability, genetic variance, selection gain, conincidence index, and processing time were estimated for all scenarios. 5-fold cross validation was repeated 50 times. The mainly conclusion in this research were: 1) SG model performed with the significant markers found by GWAS as fixed effect and the remaining SNPs as random effects is a useful strategy to select superior individuals with high accuracy; 2) GS model performed with the QTLs, previously reported for the traits in study, as fixed effect allows the selection of superior individuals more accurate; 3) Genomic selection models should be composed with number of markers ranged from 200 to 900 and number of individuals in the training population beyond 600.

1. INTRODUÇÃO GERAL

A identificação de genótipos superiores requer métodos de seleção capazes de explorar eficientemente o material genético disponível, maximizando o ganho genético em relação às características de interesse (Lorenz,

2013). Diversos métodos de seleção têm sido empregados nos programas de melhoramento, com destaque para a seleção entre e dentro de famílias e seleção combinada (Bhering et al., 2013), seleção via modelos mistos (BLUP) (Piepho et al., 2008) e índice de seleção (Peixoto et al., 2016).

Ganhos genéticos adicionais que possibilitam o aperfeiçoamento de linhagens, híbridos e variedades comerciais têm se tornado cada vez mais difíceis, no contexto de espécies submetidas a longos processos seletivos devido ao decréscimo da variabilidade genética causada principalmente pela deriva genética (Fu, 2015). Assim, recursos extras, além daqueles pertinentes à escolha de delineamentos genéticos, métodos de seleção e boa experimentação agrícola, fazem parte de uma tendência recente. O uso de procedimentos analíticos mais refinados, como o emprego de marcadores moleculares pelas metodologias de seleção assistida por marcadores moleculares (Ashraf & Foolad, 2013; Steele et al., 2013), estudo de associação genômica (Li et al., 2013; Samayoa et al., 2015) e seleção genômica (Poland et al., 2012; Spindel et al., 2015) tem-se tornado ferramenta de uso rotineiro nos programas de melhoramento de plantas.

De maneira geral, a grande questão envolvida no melhoramento genético é o conhecimento do valor genético do indivíduo, para que se possa praticar a seleção com o máximo de acurácia. Para melhor predizer o valor genético, é possível recorrer à informação fenotípica do indivíduo ou de seus aparentados (descendentes ou ancestrais) e correlaciona-las as informações de marcadores moleculares tornando a predição do valor genético mais eficiente (Meuwissen et al., 2001).

O valor genético é baseado no modelo aditivo, e tem desempenhado papel importante no ganho de seleção de características complexas em plantas e animais (Crossa et al., 2010). A partir do modelo aditivo, têm-se utilizado BLUP (Piepho et al., 2008), interações bayesianas (Gianola et al., 2009) e redes neurais artificiais (Peixoto et al., 2015) para predizer de forma mais acurada o valor genético.

Com o advento dos marcadores moleculares é possível obter o valor genético de cada indivíduo a partir do genótipo, ou seja, utilizando os efeitos de cada marcador sobre a característica. A primeira metodologia baseada em marcador desenvolvida foi a seleção assistida por marcadores moleculares (Ashraf & Foolad, 2013; Steele et al., 2013), seguida pelo estudo de associação genômica ampla (Morris et al., 2013; Tian et al., 2011), e por último a seleção genômica ampla (Lorenz et al., 2012; Spindel et al., 2015). Essas três metodologias serão destacadas em tópicos especiais.

1.1. MÉTODOS DE SELEÇÃO NO MELHORAMENTO

Uma das grandes contribuições da genética quantitativa é a indicação de estratégias de melhoramento que proporcionem avanços na direção desejada, em relação às características de interesse (Cruz et al., 2012). Nesse sentido, procura-se, desenvolver métodos de seleção que sejam mais eficientes em aumentar a frequência de genes favoráveis, em comparação com métodos 'clássicos', como a seleção massal e a seleção entre e dentro (Bhering et al., 2013).

A identificação de genótipos superiores requer métodos de seleção capazes de explorar eficientemente o material genético disponível, maximizando o ganho genético em relação às características de interesse (Junqueira et al., 2016). Diversos métodos de seleção tem sido utilizados no melhoramento genético vegetal tais como, seleção entre e dentro de famílias (Gomes Jr et al., 2014), seleção combinada (Bhering et al., 2013), e seleção por modelos mistos pelo método BLUP (*best linear unbiased prediction*) (Baretta et al., 2016), e redes neurais artificiais (Peixoto et al., 2015).

Na seleção massal a eficiência seletiva depende da quantidade de variabilidade existente na população-base a ser explorada, da herdabilidade do caráter a ser melhorado e da extensão do ganho genético deste caráter selecionado (Gomes Jr et al., 2014). Para diminuir um pouco o efeito ambiental sobre a seleção massal, este método foi modificado dividindo a área em faixas com características ambientais homogêneas dentro destas faixas. Este método é conhecido como seleção massal estratificada (Bhering et al., 2013). Outro método clássico é a seleção entre e dentro de famílias, que tem sido muito utilizada e tem apresentado, em geral, bons resultados (Bhering et al., 2013; Daros et al., 2004; de Carvalho & de Souza, 2007; Gomes Jr et al., 2014; Matta & Viana, 2003; Santos et al., 2008).

Um método alternativo a esses métodos clássicos é a seleção combinada, o que é baseada em um índice que leva em consideração, simultaneamente, o comportamento dos indivíduos e de suas famílias (Vencovsky & Barriga, 1992). Inúmeros trabalhos mostraram que a seleção combinada supera os demais métodos citados (Bhering et al., 2013; Gomes Jr et al., 2014). Arnhold et al. (2009) trabalhando com milho pipoca e Bhering et al. (2013) trabalhando com genótipos de *Jatropha curcas* L., visaram comparar a eficiência relativa da seleção massal, seleção entre e dentro e seleção combinada. Os autores concluíram que a seleção combinada proporcionou maiores ganhos e é mais adequada para estas espécies em comparação com os outros métodos.

Atualmente, para o estudo de famílias tem se adotado o método dos modelos mistos REML/BLUP (REML é a máxima verossimilhança restrita, e BLUP é a melhor predição linear não viciada), que permite estimar os parâmetros genéticos e predizer os valores genotípicos das famílias (Resende, 2002b). O BLUP consiste basicamente na predição de valores genéticos dos efeitos aleatórios do modelo estatístico associado às observações fenotípicas, ajustando-se os dados aos efeitos fixos e ao número desigual de informações nas parcelas por meio da metodologia de modelos mistos (Resende, 2002b).

De Oliveira et al. (2011) objetivando comparar a seleção via procedimento BLUP individual simulado (BLUPIS) versus seleção massal em famílias de irmãos-completos de cana-de-açúcar, concluíram que a seleção clonal via procedimento BLUPIS indica maior número de clones promissores para caracteres quantitativos dentro de famílias com elevados efeitos genotípicos. Rosado et al. (2009) trabalharam com famílias de meio irmãos de *Eucalyptus urophylla* com o objetivo de comparar vários métodos de seleção como seleção entre e dentro, seleção combinada e seleção baseada em modelos mistos (REML/BLUP). Os autores concluíram que a seleção combinada e a seleção por modelos mistos (BLUP) proporcionam estimativas de ganhos significativamente maiores às obtidas com a seleção entre e dentro, e maior eficiência na escolha dos melhores indivíduos dentro da população. Rocha et al. (2009) trabalhando com árvores de *Dipteryx alata* concluíram que a seleção combinada e BLUP proporcionaram maiores ganhos com relação aos métodos de seleção massal e seleção entre e dentro. Li & Lindgren (2006), usando dados simulados, compararam a seleção individual e a seleção combinada. Eles concluíram que a utilização de índices é vantajosa em situação de populações grandes e

características com baixa ou média herdabilidade. David et al. (2003) compararam quatro métodos e 10 intensidades de seleção em mudas de *Pinus resinosa*, e concluíram que a seleção combinada foi o método que proporcionou maiores ganhos e maximizou a diversidade genética.

Além das metodologias citadas, as redes neurais artificiais (RNAs) tem se tornado um método promissor para ser utilizado em algumas áreas tais como predição do valor genético (Peixoto et al., 2015), estudo de diversidade genética (Barbosa et al., 2011), identificação de genótipos (Pandolfi et al., 2009), predição de produção em trigo (Alvarez, 2009), entre outros. Peixoto et al. (2015) avaliaram as RNAs para predição do valor genético em dados simulados para características com diferentes herdabilidades, e verificaram que as RNAs são promissoras para predição do valor genético, especialmente em características de baixa herdabilidade.

1.2. PREDIÇÃO DO VALOR GENÉTICO

O sucesso de um programa de melhoramento genético depende da acurácia da predição do valor genético a partir de valores fenotípicos (Heffner et al., 2009). As estimativas de parâmetros genéticos podem orientar o melhorista durante as três fases principais de um programa de melhoramento: i) Aumento da variabilidade genética por meio de cruzamentos (Bhering et al., 2015a); ii) seleção de indivíduos superiores na população segregante (Junqueira et al., 2016); iii) utilização dos indivíduos selecionados no programa (Teodoro et al., 2016). Assim, tem sido usual e indispensável a quantificação dos parâmetros genéticos, especialmente a herdabilidade que é a medida do quadrado da correlação entre o valor genético e a média fenotípica (Bhering et al., 2013; Peixoto et al., 2015).

A seleção de genótipos superiores tem sido uma tarefa de difícil execução, uma vez que os caracteres de importância agronômica, em sua maioria quantitativos, apresentam base genética complexa, além de serem altamente influenciados pelo ambiente (Vieira et al., 2005)

A redução da influência ambiental sobre o valor fenotípico tem sido obtida por meio da condução de experimentos de forma zelosa e a adoção de delineamentos experimentais apropriados observando sempre os princípios básicos da casualização, repetição e controle local (Ramalho, 2005).

Assim, por meio de metodologias mais elaboradas é possível identificar a partir dos valores fenotípicos, os indivíduos de valores genotípicos desejáveis e com maior concentração de alelos favoráveis (Junqueira et al., 2016).

A metodologia de modelos mistos pressupõe que os componentes de variância sejam conhecidos, o que raramente ocorre na prática. Quando tais componentes são desconhecidos, utiliza-se como estratégia de análise a substituição dos valores paramétricos dos componentes de variância por estimativas de máxima verossimilhança restrita (REML) (Resende, 2000).

Estimativas dos componentes de variância e coeficientes de herdabilidade têm-se mostrado heterogêneas de acordo com diferentes níveis de produção e diferentes classes de desvio-padrão genético ou ambiental. Quando a heterogeneidade não é considerada, diferenças de variâncias dentro das subclasses podem resultar na predição de valores genéticos viesados, redução no progresso genético e desproporcional número de indivíduos selecionados de ambientes com diferentes variâncias (Weigel & Gianola, 1992).

Dentro do contexto relativo ao processo de seleção, o conhecimento dos componentes de variância é de fundamental importância para estimar a herdabilidade, predizer o ganho genético e avaliar as potencialidades de uma população e a eficiência relativa dos diferentes métodos de melhoramento (Peixoto et al., 2016). Além disto pode auxiliar a identificar a estratégia de seleção mais adequada (Bhering et al., 2013; Junqueira et al., 2016; Laviola et al., 2010).

Outro conjunto de informações extremamente útil é aquele obtido por meio dos cálculos de correlações. Segundo Ferreira et al. (2003), a correlação fenotípica fornece uma estimativa da influência conjunta de causas genéticas e ambientais na expressão de uma dada característica. Por sua vez, os valores de correlação genotípica (que corresponde à porção genética da correlação fenotípica) têm sido empregados para orientar programas de melhoramento genético, uma vez que eles refletem a fração da expressão fenotípica que é de natureza herdável.

Quando os caracteres são de baixa herdabilidade a correlação fenotípica pode ter pouca aplicabilidade, podendo induzir o melhorista a erros. Assim, é importante distinguir as causas genéticas e de ambiente que, combinadas, resultam na correlação fenotípica (Resende, 2002a).

De acordo com Hallauger & Miranda Filho (1981), a correlação tem importância no melhoramento de plantas, porque mede o grau de associação genética

ou não genética entre dois ou mais caracteres. Cruz et al. (2012) ressaltaram a importância das correlações, afirmando que elas quantificam a possibilidade de ganhos indiretos por seleção e que caracteres de baixa herdabilidade têm a seleção mais eficiente quando realizada sobre caracteres correlacionados.

1.3. MARCADORES MOLECULARES

Por volta dos anos de 1960, os marcadores utilizados nos estudos de genética e melhoramento eram aqueles determinados por locos associados a características morfológicas, sendo estas, geralmente de fácil identificação visual. Essa estratégia de trabalho utilizando marcadores morfológicos foi proposta por Sax (1923).

A revolução no campo dos marcadores iniciou-se com o desenvolvimento dos marcadores isoenzimáticos. A partir daí o número de marcadores foi grandemente ampliado, e aplicabilidade da técnica passou a ser geral (Tanksley, 1993). Contudo, o número de marcadores previsto pelos ensaios isoenzimáticos é limitado, ficando geralmente entre 10 a 20 por espécie. Mesmo sabendo que o número de locos isoenzimáticos que podem ser detectados seja maior do que 100, o nível de resolução dos marcadores isoenzimáticos não permitia a cobertura completa do genoma, o que limitava certos estudos como a construção de mapas genéticos.

Na busca por marcadores mais abundantes, surgiram na década de 80 os marcadores moleculares. Os marcadores moleculares possuem algumas vantagens em relação aos marcadores morfológicos e isoenzimáticos. A primeira vantagem é que são baseados na variação de sequência de nucleotídeos, sendo assim, aparecem em grande número no genoma da espécie. A segunda vantagem é que quando codominantes permitem a identificação de genótipos em qualquer cruzamento. A terceira vantagem é que não são influenciados pelo ambiente. A utilização desses marcadores está em estabelecer relações de ligação entre eles e os locos de interesse. Essa ligação permite ao pesquisador, inferir sobre a presença dos locos, mediante o ensaio de alguns destes marcadores, facilitando trabalhos de transferência e de mapeamento (Tanksley, 1993).

Existem inúmeros tipos de marcadores de DNA que são geralmente conhecidos por suas siglas. Inicialmente, a utilização de enzimas de restrição permitiu a análise do polimorfismo de comprimento de fragmentos de restrição de DNA (Restriction Fragment Lenght Polymosphism - RFLP) (Botstein et al., 1980). RFLP é uma técnica que se baseia no polimorfismo gerado por digestão do DNA genômico

com enzimas de restrição. Após a digestão, ocorre a separação dos fragmentos de DNA por eletroforese em gel de agarose, ou poliacrilamida, que, em seguida, são transferidos por uma membrana de nilon e hibridizados por uma sonda marcada.

O desenvolvimento de um processo de ampliação em cadeia utilizando uma DNA polimerase (Polymerase Chain reaction – PCR) (Saiki et al., 1988) levou a descrição de outras classes de marcadores moleculares como o polimorfismo de DNA amplificado ao acaso (Random Applied Polymorphism DNA – RAPD) (Williams et al., 1990). Este tipo de marcador permite gerar grande número de informações em curto tempo e por custo acessível. A técnica RAPD consiste basicamente na extração de DNA de indivíduos, amplificação de fragmentos desses DNA pela técnica de PCR, separação de fragmentos amplificados de comprimentos diferentes por eletroforese em gel de agarose, e na visualização de bandas correspondentes a regiões amplificadas do genoma por meio da coloração de fragmentos de DNA com brometo de etidio.

Os microssatélites, também conhecidos como marcadores SSR (Sequency Single Repetition), consiste de uma sequência do genoma que varia de 1 a 6 pares de bases e que repetem consecutivamente n vezes. Este tipo de marcador é co-dominante, multi alélico, e apresenta robustez analítica e transferibilidade (Grattapaglia, 2007a). Essas sequências são amplificadas por meio da reação de PCR utilizando primers que flanqueiam a região microssatélite. O resultado pode ser observado em um gel de agarose ou poliacrilamida ou por meio de eletroforese capilar. No entanto, embora microssatélites possuam as vantagens mencionadas acima, uma limitação inerente ao método de detecção é a incapacidade de genotipar um grande número de marcadores em um grande número de indivíduos, sendo este uma limitação de todos os tipos de marcadores citados até agora. Assim quando o objetivo é genotipar um grande número de marcas em um grande número de indivíduos, outros tipos de marcadores precisam ser utilizados como descrito abaixo (Pires et al., 2011).

Mais recentemente surgiram os marcadores Single Nucleotide Polymorphism (SNPs) e Diversity Arrays Technology (DArTs). Devido ao baixo custo por marca, alta abundância no genoma, especificidade de locos, potencial de alta densidade de marcas, baixos níveis de erros (Rafalski, 2002) e facilidade de automação (Van Inghelandt et al., 2010) esses marcadores têm emergido como poderosas ferramentas para várias aplicações dentro da genética, incluindo estudos de diversidade (Bhering

et al., 2015a; Zhu et al., 2003), associação genômica (Morris et al., 2013; Neumann et al., 2011) e seleção genômica (Lorenz et al., 2012; Sansaloni et al., 2010).

Os marcadores DArTs apresentam a vantagem de detecção da variação em milhares de locos gênicos sem a necessidade de sequenciamento prévio (Wenzl et al., 2006). O polimorfismo detectado nos DArTs incluem SNPs, inserções-deleções (InDels) e modificações de metilação herdáveis (Jaccoud et al., 2001). Características essas que têm feito com que esses marcadores se tornem ferramentas interessantes em estudos de diversidade (Bhering et al., 2015a) e de genética de associação (Neumann et al., 2011).

Os marcadores SNPs são caracterizados pela diferença em uma única base entre dois cromossomos homólogos ou entre dois indivíduos da mesma espécie. Um dos principais métodos de genotipagem de SNPs consiste em arranjo que contenha a sequência de DNA a ser analisada. O DNA de vários indivíduos é hibridizado à sonda e um sinal referente a hibridização é capturado. Caso o fragmento possua um polimorfismo de base única, uma hibridização diferencial acontece e o SNP é então identificado (Kisrt, 2007). SNPs são então marcadores codominantes, considerados bi alélicos pois, em geral, apenas duas formas alélicas são observadas, embora em teoria, todas as quatro possíveis combinações possam ser capturadas (Pires et al., 2011). A principal vantagem deste marcador é a capacidade de genotipagem em alta escala de muitos marcadores em muitos indivíduos. Sendo assim, trata-se de um marcador ideal para a integração da seleção genômica ampla em um programa de melhoramento genético (Meuwissen et al., 2001).

1.4. SELEÇAO ASSISTIDA POR MARCADORES MOLECULARES

O melhoramento genético tem, há vários anos, proporcionado com muito sucesso o aumento da produtividade e a melhoria de várias características de interesse na agricultura e na pecuária. Embora muitos métodos tenham surgido nos últimos anos com intuito de otimizar o processo de seleção de indivíduos superiores, a estratégia básica utilizada até hoje é a de predizer o valor genético do indivíduo, baseado em informações fenotípicas e em alguns casos na genealogia. No entanto, com o desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular, existe a expectativa de que informações genotípicas (obtidas por meio dos marcadores moleculares), uma vez correlacionada com características fenotípicas de interesse, possam ser amplamente utilizados na identificação e seleção

de indivíduos com maiores valores genéticos. Adicionalmente, espera-se que a seleção com base em informações genotípicas possam ser realizadas precocemente, o que no caso do melhoramento animal ou de espécies vegetais perenes, tende a elevar os ganhos (Resende, 2008).

Neste sentido, o primeiro método proposto para o uso de marcadores no melhoramento ficou conhecido como seleção assistida por marcador molecular (SAM) (Lande & Thompson, 1990). Essa metodologia se baseia nas análises de populações segregantes (uma ou algumas famílias) para identificação e mapeamento de regiões controladoras de características quantitativas (QTL – Quantitative Trait Loci). A geração de progênies oriundas de indivíduos constrastantes gera alto desequilíbrio de ligação (DL), dentro da família ou cruzamento, o que permite a identificação de marcadores que cossegregam com a característica fenotípica, mesmo quando número reduzido de marcadores é utilizado.

O conceito de DL refere-se a associação não aleatória entre dois genes ou entre um QTL e um marcador. Neste caso, quando as frequências alélicas e genotípicas de dois locos são constantes de uma geração para outra e as frequências genotípicas são determinadas pelas frequências alélicas, diz-se que esses locos se encontram em equilíbrio de Hardy-Weinberg e de ligação (EL). Em razão da ligação gênica, dois genes/marcadores ligados apresentam associação que não se dá ao acaso, e diz-se então que esses genes estão em DL.

Dado ao grande interesse pela técnica de SAM um grande número de QTLs já foram detectados e mapeados nas mais variadas culturas e características tais como tolerância salinidade (Ashraf et al., 2012; Ashraf & Foolad, 2013), características de raiz em arroz (Steele et al., 2013), resistência a fusarium em trigo (Arruda et al., 2016), resistência a ferrugem em trigo (Yaniv et al., 2015), e Potato Virus Y (PVY) em batata (Szajko et al., 2014).

Steele et al. (2013) avaliaram a introgressão de genes via SAM em cultivares de arroz visando resistência a seca e aumento da produtividade. Os autores observaram que a introgressão de genes identificados via SAM aumentou em 1 $ton.ha^{-1}$ e concluíram que a SAM é uma técnica promissora para aumentar a resistência a seca e consequentemente a produtividade em cultivares de arroz. Dong et al. (2014) avaliaram a SAM para identificar o gene *qhir1* que é responsável pela indução de haploide *in vivo* de milho. Os autores verificaram que utilizando a SAM combinado com a taxa de indução de haploides (característica morfológica) foi eficiente para

selecionar plantas com boa performance na indução de haploides. (Mandoulakani et al., 2015) avaliaram 1.256 plantas F_2 buscando encontrar plantas com o gene *Yr15* que é localizado no cromossomo 1BS. Este gene controla a resistência do trigo a ferrugem. Foram avaliadas 7 marcadores SCARs e os autores concluíram que essas marcas são úteis para serem utilizadas na SAM com o objetivo de incorporar o gene *Yr15* em linhagens elite de trigo ou cultivares comerciais. Fallen et al. (2015) avaliaram 875 progênies de uma população de RILs (*reconbinant inbreed lines*) em soja com o objetivo de por meio da SAM encontrar QTLs que tem grande efeito sobre a produtividade. Os autores utilizaram a análise de variância e o mapeamento por intervalo composto para identificar os QTLs e identificaram 44 QTLs variando o efeito individual de cada um deles de 4,5 % a 11,9% sobre a produtividade. Entre os QTLs identificados, cinco não tinham sido reportados anteriormente.

No entanto, grande parte dos QTLs detectados e mapeados em cada espécie não tem sido aplicados de forma prática nos programas de melhoramento (Bernardo, 2008). Uma das causas deste insucesso é a necessidade do estabelecimento de associações entre os marcadores e os QTLs de cada família avaliada. Isso acontece, pois, os níveis populacionais de DL em uma população de melhoramento são muito inferiores quando comparado ao DL analisado na progênie segregante. Além disso, uma segunda razão que limita o uso prático da SAM é o fato de apenas pequenos números de QTLs de grande efeito terem sido detectados e mapeados, os quais, devido a natureza poligênica e alta influência ambiental dos caracteres quantitativos, não explicam suficientemente toda a variação genética (Dekkers, 2004).

1.5. ESTUDO DE ASSOCIAÇÃO GENÔMICA AMPLA

Em razão das limitações inerentes as técnicas de SAM, novas metodologias foram propostas para utilização dos marcadores moleculares na identificação de locos que controlam características de interesse no melhoramento. Entre elas, o estudo de associação genômica ampla (EAGA).

O estabelecimento desta nova metodologia só foi possível graças ao extraordinário avanço das tecnologias de genotipagem. A partir do inicio do século XXI, métodos que permitem a descoberta e genotipagem em larga escala dos Single Nucleotide Polymosphism (SNPs), em plataformas de micro arranjo multiplicaram-se, de modo que a maioria das espécies de interesse econômico dispõe de um número relativamente elevado (na ordem de algumas centenas/milhares) de marcadores

passiveis de uso em programas de melhoramento (Jenkins & Gibson, 2002). Com o advento da tecnologia de genotipagem por sequenciamento (genotyping by sequencing – GBS) (Baird et al., 2008) espere-se que haja incremento ainda maior no número de marcadores disponíveis, simultaneamente com a redução do custo por *data point*, o qual tende a viabilizar o uso rotineiro de marcadores em programas de melhoramento.

O EAGA baseia-se na teoria de que, com grande número de marcadores espalhados pelo genoma, aumenta-se a probabilidade de que QTLs de interesse estejam em forte DL com os marcadores (Hästbacka et al., 1992). O princípio do EAGA é semelhante àqueles utilizados nos métodos de mapeamento de QTLs em famílias segregantes, em que se identificam locos cuja frequência alélica esta correlacionada com a variação fenotípica em uma população. Nos dois métodos, diferenças significativas entre os valores fenotípicos observados nos indivíduos que herdaram alelos distintos sugerem que o loco está em desequilíbrio de ligação com o loco que efetivamente controla a característica fenotípica (Resende et al., 2013).

O mapeamento de associação baseia-se, entretanto, na análise de uma população de indivíduos não relacionados. A razão desta condição está relacionada com o objetivo do mapeamento de associação, em que se deseja detectar e mapear os genes controladores de características fenotípicas. Assim, para que se atinja uma resolução em nível de um gene na determinação da associação entre um marcador e um fenótipo, é necessário que o tamanho do bloco de ligação seja extremamente reduzido. Além disso, o uso de uma população não estruturada permite que os locos amostrados potencialmente capturem toda a variabilidade genética da população em estudo e não apenas a variabilidade dos dois genótipos parentais.

No entanto, o número de marcadores necessários para identificar genes associados ao fenótipo é inversamente proporcional à extensão do DL. Assim, como essa ausência de estrutura das populações leva a um reduzido nível de DL, as análises de EAGA requerem elevadíssima densidade de marcadores (Resende et al., 2013). Por exemplo, em *Eucaliptus grandis*, se a extensão média do DL for de 1.000 pares de base e o genoma tiver 630 Mgb, será necessária a genotipagem de 630.000 marcadores SNPs (Grattapaglia, 2007b). Se o desequilíbrio de ligação se estender por apenas 500 pares de bases, o número de marcadores dobra para 1,26 milhão de SNPs. Em outras palavras, com as tecnologias atuais de genotipagem, essa abordagem não é possível economicamente (Resende et al., 2013).

A metodologia consiste em ajustar um modelo marcador por marcador em grandes populações naturais, não estruturadas, em que os indivíduos supostamente se relacionam entre si, por meio de um ancestral comum em um dado tempo. O principal objetivo desta técnica é identificar genes candidatos para o controle genético de determinada característica.

O EAGA tem sido utilizado com sucesso para várias características tais como arquitetura de folhas em milho (Tian et al., 2011), características agroclimáticas em sorgo (Morris et al., 2013), características relacionadas a produção em arroz (Huang et al., 2010), proteína e óleo em soja (Hwang et al., 2014), entre outros.

Wen et al. (2014) buscando encontrar genes de resistência a síndrome da morte súbita (SDS), avaliaram duas populações de soja contendo 392 e 300 acessos utilizando 52,041 e 5,361 SNPs respectivamente. Por meio do EAGA 20 QTLs foram identificados, sendo 7 já reportados na literatura e 13 como novos QTLs. Os QTLs identificados foram responsáveis por 54,5% da variância fenotípica. Kump et al. (2011) avaliaram 5,000 RILs (*recombinant inbred lines*) provenientes de 25 famílias em milho com o objetivo de encontrar genes de resistência a southern leaf blight (SLB). A metodologia de EAGA encontrou 32 QTLs afetando a expressão desta característica em milho, sendo todos de pequeno efeito e efeito aditivo. Huang et al. (2010) avaliaram 517 genótipos de arroz buscando QTLs com efeito sobre 14 características de interesse econômico. Em média os QTLs encontrados para cada característica explicavam 36% da variação fenotípica. Os autores concluíram que a utilização do EAGA para identificação de QTLs em características agronômicas em arroz é de grande importância para ajudar a entender a estrutura genética de características complexas. Le Gouis et al. (2012) utilizaram EAGA para identificar regiões cromossômicas associadas com a precocidade em trigo. Os autores avaliaram 227 genótipos de trigo durante três anos, e encontraram 62 marcas associadas ao controle genético da precocidade em trigo, correspondendo a 33 diferentes regiões cromossômicas. As marcas identificadas explicaram 34% da variância fenotípica da característica.

Cabe destacar que o tipo de população utilizada tem impacto relevante sobre os padrões de DL e, consequentemente, sobre o número de marcadores necessários para identificar genes que controlam características de interesse no melhoramento e selecionar indivíduos superiores. Por isso, o EAGA não tem obtido muito sucesso na prática. Em razão destas características e de outras que serão detalhadas no próximo

tópico, a seleção genômica ampla tem chamado mais atenção de melhoristas pela possibilidade real de sua operacionalização em programas de melhoramento (Hayes et al., 2009).

1.6. SELEÇÃO GENÔMICA AMPLA

As atuais metodologias de seleção assistida por marcadores moleculares conseguem capturar o efeito apenas de poucos genes de grande efeito (Dekkers & Hospital, 2002). No entanto para o sucesso de uma nova variedade é necessário que se faça seleção para características quantitativas, e essas características são afetadas por inúmeros genes de menor efeito, tornando a SAM inviável para este tipo de característica (Heffner et al., 2009).

Com o recente desenvolvimento das plataformas de sequenciamento de última geração tem se tornado possível a genotipagem de um grande número de indivíduos e um grande número de marcadores. Os tipos de marcadores que tem sido genotipados por esta metodologia são os DArTs e os SNPs. Assim, a aplicação da seleção genômica proposta por Meuwissen et al. (2001) tem se tornado uma ferramenta para resolver as principais limitações mostradas pela SAM (Heffner et al., 2009).

A ideia principal da metodologia de seleção genômica é a estimação do valor genético genômico (VGG) de cada indivíduo baseado em um grande número de marcadores. A metodologia consiste em estimar o efeito de cada marcador a partir de uma população que foi previamente fenotípada e genotípada. Esta população é conhecida como população de treinamento (Meuwissen et al., 2001). A partir dos efeitos de marcadores é possível estimar o VGG de uma população que foi apenas genotipada conhecida como população de validação (Meuwissen et al., 2001). Os valores de VGG são então utilizados para praticar seleção na população de validação. Para maximizar a acurácia de predição do VGG a população de treinamento precisa ser representativa da população de melhoramento (Heffner et al., 2009).

Inúmeros são os fatores que interferem na acurácia de predição de um modelo de seleção genômica, dentre eles podemos destacar o tamanho da população de treinamento, tamanho efetivo populacional, densidade de marcadores utilizada, herdabilidade da característica, número de QTls envolvidos no controle das características-alvo e o modelo estatístico utilizado (Grattapaglia & Resende, 2011). É interessante salientar que os três primeiros fatores podem ser controlados pelo

melhorista. O quarto fator (herdabilidade), embora possa ser mais bem estimado em função de um delineamento experimental bem planejado, não pode ser controlado. O mesmo acontece com o número de QTLs envolvidos no caráter, característica essa inerente a arquitetura genética do caráter em questão (Resende et al., 2013). Esses fatores são a seguir discutidos em mais detalhes, bem como o seu impacto na GWS.

No que diz respeito ao tamanho da população de treinamento, tem-se demonstrado via simulações que populações de treinamento de pequeno tamanho não permitem a estimação adequada dos efeitos dos marcadores (Resende, 2008). Isso provavelmente se deve ao fato de que, nessas condições, a amplitude dos efeitos alélicos amostrados não é adequada, e os efeitos dos genes/alelos importantes para a correta determinação do caráter podem não estar sendo estimados com precisão (Grattapaglia & Resende, 2011). É interessante destacar, no entanto, que esse fator não é o principal determinante de qualidade dos modelos preditivos, uma vez que, embora exista incremento na acurácia dos modelos, a medida que o tamanho da população de treinamento aumenta, esse incremento é relativamente pequeno após o número de 1.000 indivíduos. Usando densidade elevada de marcadores, populações com menos de 1.000 indivíduos permitem atingir acurácias acima de 0,80. No caso de um número reduzido de marcadores ser utilizado, o tamanho populacional deve ser aumentado para cerca de 2.000 (Grattapaglia & Resende, 2011). Porém existe poucos estudos mostrando como o número de indivíduos na população de treinamento afeta a acurácia de predição.

O tamanho efetivo da população (Ne) possui impacto muito maior que o tamanho absoluto da população de treinamento. Isto porque o tamanho efetivo determina ao menos parcialmente a extensão do DL na população (Sved, 1971). Em populações de maior Ne, a extensão do DL tende a ser bastante limitada e o número de alelos em um dado locos, maior. Entretanto, em populações com Ne reduzido a extensão do DL é consideravelmente maior e, de maneira geral, o número de alelos presentes na população para cada locos é menor. Isso implica que, teoricamente, população de alto Ne, maior número de marcadores deverão ser necessariamente genotipados na população de treinamento, a fim de garantir que ao menos um deles esteja em DL com cada um dos QTLs que controlam a característica de interesse (Resende et al., 2013). Como Ne é uma característica da população de melhoramento, pode-se manejar a população de treinamento, de modo a manter o Ne em níveis intermediários. Isto pode ser feito, restringindo-se o número de genitores que são

intercruzados para gerar as famílias segregantes (Resende, 2002a). Embora limitação extrema do número de genitores pode ser interessante do ponto de vista de manter um DL extenso, uma mesma redução pode ter impactos negativos significantes para o programa de melhoramento, uma vez que, ao diminuir o Ne, reduz-se também a variabilidade genética da população, o que consequentemente, pode diminuir o ganho esperado em gerações futuras (White et al., 2007). Portanto, são necessárias pesquisas para definir um Ne ideal de modo a melhorar a acurácia de predição e manter a variabilidade genética da população de melhoramento.

O número de marcadores utilizado nos modelos de seleção genômica devem ser elevado, ao ponto de cobrir todo o genoma da espécie, de forma a maximizar o número de QTLs em DL com pelo menos um marcador. Desta forma a variância genética explicada pelos QTLs é maximizada (Heffner et al., 2009). A densidade de marcas alvo será detectada pelo decréscimo do DL ao longo do genoma. Este valor é mensurado pelo coeficiente de determinação entre os marcadores e a distância genética.

O DL é afetado por inúmeros fatores da própria população tais como o processo de evolução, tamanho da população, taxa de recombinação (crossing over), e efeitos da seleção (Gaut & Long, 2003). Portanto o decréscimo na taxa de DL está altamente relacionado com características da espécie, população ou região genômica. A estimativa do DL pode ser usada para determinar o número de marcas que precisa ser utilizado no modelo de seleção genômica. Por exemplo Calus & Veerkamp (2007) usaram a média do coeficiente de determinação entre marcas adjacentes para mensurar o número de marcas necessárias para haver decréscimo na taxa de desequilíbrio de ligação. Os autores encontraram que para características de alta herdabilidade um coeficiente de determinação de 0,15 é suficiente para manter o marcador na análise. No entanto para características de baixa herdabilidade é necessário manter marcadores com coeficiente de determinação acima de 0,2. Desta forma a acurácia de predição é maximizada.

Como a cada dia o custo de genotipagem tem se tornado mais barato, no futuro será possível genotipar um número maior de marcadores de forma que o genoma da espécie será completamente coberto e todos os QTLs estarão em DL com pelo menos um marcador (Zhu et al., 2008). Porém, as condições de saturação completa do genoma e pelo menos um marcador em DL com o QTL, nem sempre garante aumento na acurácia de predição. Portanto, é necessário pesquisas mostrando qual é o número

necessário de marcas para maximizar a acurácia de predição e também como estas marcas devem ser selecionadas (Heffner et al., 2009).

Com relação ao impacto da herdabilidade das características sobre seleção, estudos tem demonstrado que os modelos preditivos funcionam relativamente bem, mesmo para características de baixa herdabilidade. Grattapaglia & Resende (2011) demonstraram que a acurácia aumentou apenas 10-20% a medida que a herdabilidade aumentou de 0,2 para 0,6, independentemente do tamanho da população. Isso indica que, ao contrário da SAM, a GS é eficiente para selecionar indivíduos superiores, mesmo para características de baixa herdabilidade (Calus et al., 2008).

Ao considerar o fator número de QTLs envolvido no controle genético das características-alvo da GWS, Grattapaglia & Resende (2011) verificaram que esse possui impacto significativo sobre a acurácia dos modelos preditivos apenas quando se faz uso de baixa densidade de marcadores moleculares. Isso, provavelmente, se deve ao fato de que nesta situação nem todos os QTLs estão em DL, com pelo menos um marcador. Neste caso, com baixa densidade de marcadores, a acurácia dos modelos preditivos é maior se a característica for controlada por número menor de QTLs (Resende et al., 2013).

Nos programas de melhoramento, métodos estatísticos utilizados na seleção genômica serão necessários para estimar simultaneamente o efeito de todas as marcas presentes no modelo, a partir de um número limitado de fenótipos. O número maior de marcas em relação ao número de fenótipos faz com que os métodos de quadrados mínimos não possam ser utilizados devido à falta de graus de liberdade para o resíduo (Heffner et al., 2009). Os métodos estatísticos comumente utilizados para aplicação da seleção genômica no melhoramento são aqueles baseados em modelos mistos tais como o RR-BLUP (Meuwissen et al., 2001) e o G-BLUP (Hayes et al., 2009), e os métodos baseados na inferência bayesiana tais como Bayes A e Bayes B (Meuwissen et al., 2001), Bayes Cπ (Jannink et al., 2010), LASSO Bayesiano (De Los Campos et al., 2009b) e Reproducing Kernel Hilbert Space regression – RKHS (De Los Campos et al., 2009a). Bhering et al. (2015b) compararam RR-BLUP, G-BLUP e LASSO bayesiano utilizando dados simulados e verificaram que não houve diferença significativa entre os métodos para computo da acurácia de predição. Azevedo et al. (2015) compararam 10 modelos de seleção genômica baseados nos métodos de LASSO, inferência Bayesiana e modelos mistos, em duas herdabilidades

(0,3 e 0,5) e duas estruturas genéticas (a primeira onde todos os QTLs tinham pequeno efeito sobre a característica e a segunda onde 5 QTLs exerciam grande efeito sobre a característica e os demais eram de pequeno efeito. Os autores concluíram que as metodologias G-BLUP e Bayes B modificado apresentaram melhores resultados e são os mais adequados para predição do valor genético genômico e do valor genotípico total, assim como a estimação dos efeitos aditivos e de dominância no modelo genômico aditivo-dominante.

Em geral, os modelos utilizados atualmente na seleção genômica definem todas as marcas como efeito aleatório, ou seja, não se faz o uso de nenhuma covariável. Pesquisas recentes mostraram que a modelagem de QTLs identificados pelas metodologias de SAM ou EAGA como efeito fixo e os demais marcadores como efeito aleatório no modelo de seleção genômica pode aumentar a acurácia de predição (Arruda et al., 2016; Bernardo, 2014; Spindel et al., 2015). Spindel et al. (2015) sugeriram que o uso das informações obtidas utilizando as análises EAGA dentro dos modelos de SG pode providenciar informações sobre a arquitetura genética da característica em estudo e informações sobre a estrutura da população que está sendo utilizada no programa de melhoramento. Em seu trabalho Spindel et al. (2015) demonstraram que o uso das marcas significativas pelo EAGA utilizada como efeito fixo nos modelos de SG para produção de grãos, altura de plantas e florescimento em arroz podem revelar a presença de QTL de maior efeito segregando na população de melhoramento, na qual podem ser estabelecidos como covariáves nos modelos de SG melhorando a acurácia. Porém quando o número de QTLs é maior que 10 esse efeito pode ser contrário, ou seja, pode haver um decréscimo no valor da acurácia (Bernardo, 2014). Bernardo (2014) observou que um único gene tratado como efeito fixo no modelo de SG usando RRBLUP nunca será desvantajoso, exceto em alguns casos onde a variabilidade explicada pelo QTL for inferior a 10%. No entanto, ainda existe muitas perguntas sem respostas sobre o uso de QTLs como covariáveis no modelo.

1.7. REFERÊNCIAS BIBLIOGRÁFICAS

ALVAREZ, R., Predicting average regional yield and production of wheat in the Argentine Pampas by an artificial neural network approach. **European Journal of Agronomy**, v. 30, p. 70-77, 2009.

ARNHOLD, E., VIANA, J. M. S., SILVA, R. G., MORA, F., Eficiências relativas de métodos de seleção de famílias endogâmicas em milho-pipoca. **Acta Scientiarum. Agronomy**, v. 31, p. 203-207, 2009.

ARRUDA, M. P., LIPKA, A. E., BROWN, P. J., KRILL, A. M., THURBER, C., BROWN-GUEDIRA, G., DONG, Y., FORESMAN, B. J., KOLB, F. L., Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum*). **Molecular Breeding**, v. 36, p. 1-11, 2016.

ASHRAF, M., AKRAM, N. A., FOOLAD, M. R., Marker-assisted selection in plant breeding for salinity tolerance. **Plant Salt Tolerance: Methods and Protocols**, v., p. 305-333, 2012.

ASHRAF, M., FOOLAD, M. R., Crop breeding for salt tolerance in the era of molecular markers and marker-assisted selection. **Plant Breeding**, v. 132, p. 10-20, 2013.

AZEVEDO, C. F., DE RESENDE, M. D. V., E SILVA, F. F., VIANA, J. M. S., VALENTE, M. S. F., RESENDE, M. F. R., MUÑOZ, P., Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC genetics**, v. 16, p. 1, 2015.

BAIRD, N. A., ETTER, P. D., ATWOOD, T. S., CURREY, M. C., SHIVER, A. L., LEWIS, Z. A., SELKER, E. U., CRESKO, W. A., JOHNSON, E. A., Rapid SNP discovery and genetic mapping using sequenced RAD markers. **PloS one**, v. 3, p. e3376, 2008.

BARBOSA, C. D., VIANA, A. P., QUINTAL, S. S. R., PEREIRA, M. G., Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology**, v. 11, p. 224-231, 2011.

BARETTA, D., NARDINO, M., CARVALHO, I. R., DE OLIVEIRA, A. C., DE SOUZA, V. Q., DA MAIA, L. C., Performance of maize genotypes of Rio Grande do Sul using mixed models. **Científica**, v. 44, p. 403-411, 2016.

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

BERNARDO, R., Genomewide selection when major genes are known. **Crop Science**, v. 54, p. 68-75, 2014.

BHERING, L. L., BARRERA, C. F., ORTEGA, D., LAVIOLA, B. G., ALVES, A. A., ROSADO, T. B., CRUZ, C. D., Differential response of Jatropha genotypes to different selection methods indicates that combined selection is more suited than other methods for rapid improvement of the species. **Industrial Crops and Products**, v. 41, p. 260-265, 2013.

BHERING, L. L., DE AZEVEDO PEIXOTO, L., LEITE, N. L. S. F., LAVIOLA, B. G., Molecular analysis reveals new strategy for data collection in order to explore variability in Jatropha. **Industrial Crops and Products**, v. 74, p. 898-902, 2015a.

BHERING, L. L., JUNQUEIRA, V. S., PEIXOTO, L. A., CRUZ, C. D., LAVIOLA, B. G., Comparison of methods used to identify superior individuals in genomic selection in plant breeding. **Genetics and molecular research: GMR**, v. 14, p. 10888, 2015b.

BOTSTEIN, D., WHITE, R. L., SKOLNICK, M., DAVIS, R. W., Construction of a genetic linkage map in man using restriction fragment length polymorphisms. **American journal of human genetics**, v. 32, p. 314, 1980.

CALUS, M. P. L., DE ROOS, A. P. W., VEERKAMP, R. F., Accuracy of genomic selection using different methods to define haplotypes. **Genetics**, v. 178, p. 553-561, 2008.

CALUS, M. P. L., VEERKAMP, R. F., Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. **Journal of Animal Breeding and Genetics**, v. 124, p. 362-368, 2007.

CROSSA, J., DE LOS CAMPOS, G., PÉREZ, P., GIANOLA, D., BURGUEÑO, J., ARAUS, J. L., MAKUMBI, D., SINGH, R. P., DREISIGACKER, S., YAN, J., Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. **Genetics**, v. 186, p. 713-724, 2010.

CRUZ, C. D., REGAZZI, A. J., CARNEIRO, P. C. S., 2012. Modelos biométricos aplicados ao melhoramento genético. UFV, Viçosa.

DAROS, M., AMARAL JR, A. T. D., PEREIRA, M. G., SANTOS, F. S., GABRIEL, A. P. C., SCAPIM, C. A., FREITAS JR, S. D. P., SILVÉRIO, L., Recurrent selection in inbred popcorn families. **Scientia Agricola**, v. 61, p. 609-614, 2004.

DAVID, A., PIKE, C., STINE, R., Comparison of selection methods for optimizing genetic gain and gene diversity in a red pine (*Pinus resinosa* Ait.) seedling seed orchard. **Theoretical and Applied Genetics**, v. 107, p. 843-849, 2003.

DE CARVALHO, H. W. L., DE SOUZA, E. M., Ciclos de seleção de progênies de meios-irmãos do milho BR 5011 Sertanejo. **Pesq. agropec. bras., Brasília**, v. 42, p. 803-809, 2007.

DE LOS CAMPOS, G., GIANOLA, D., ROSA, G., Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, v. 87, p. 1883-1887, 2009a.

DE LOS CAMPOS, G., NAYA, H., GIANOLA, D., CROSSA, J., LEGARRA, A., MANFREDI, E., WEIGEL, K., COTES, J. M., Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, p. 375-385, 2009b.

DE OLIVEIRA, R. A., DAROS, E., DE RESENDE, M. D. V., BESPALHOK-FILHO, J. C., ZAMBON, J. L. C., SOUZA, T. R., LUCIUS, A. S. F., Procedimento Blupis e seleção massal em cana-de-açúcar. **Bragantia**, v. 70, p. 796-800, 2011.

DEKKERS, J. C., Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. **Journal of animal science**, v. 82, p. E313-E328, 2004.

DEKKERS, J. C., HOSPITAL, F., The use of molecular genetics in the improvement of agricultural populations. **Nature Reviews Genetics**, v. 3, p. 22-32, 2002.

DONG, X., XU, X., LI, L., LIU, C., TIAN, X., LI, W., CHEN, S., Marker-assisted selection and evaluation of high oil in vivo haploid inducers in maize. **Molecular Breeding**, v. 34, p. 1147-1158, 2014.

FALLEN, B. D., ALLEN, F. L., KOPSELL, D. A., SAXTON, A. M., MCHALE, L., SHANNON, J. G., KANTARTZI, S. K., CARDINAL, A. J., CREGAN, P. B., HYTEN, D. L., Selective Genotyping for Marker Assisted Selection Strategies for Soybean Yield Improvement. **Plant Genetics, Genomics, and Biotechnology** v. 2, p. 95-119, 2015.

FERREIRA, M., QUEIRÓZ, M. A. D., BRAZ, L. T., VENCOVSKY, R., Correlações genotípicas, fenotípicas e de ambiente entre dez caracteres de melancia e suas implicações para o melhoramento genético. **Horticultura Brasileira**, v. 21, p. 438-442, 2003.

FU, Y.-B., Understanding crop genetic diversity under modern plant breeding. **Theoretical and Applied Genetics**, v. 128, p. 2131-2142, 2015.

GAUT, B. S., LONG, A. D., The lowdown on linkage disequilibrium. **The Plant Cell**, v. 15, p. 1502-1506, 2003.

GIANOLA, D., DE LOS CAMPOS, G., HILL, W. G., MANFREDI, E., FERNANDO, R., Additive genetic variability and the Bayesian alphabet. **Genetics**, v. 183, p. 347-363, 2009.

GOMES JR, R. A., DE LIMA GURGEL, F., DE AZEVEDO PEIXOTO, L., BHERING, L. L., DA CUNHA, R. N. V., LOPES, R., DE ABREU PINA, A. J., VEIGA, A. S., Evaluation of interspecific hybrids of palm oil reveals great genetic variability and potential selection gain. **Industrial Crops and Products**, v. 52, p. 512-518, 2014.

GRATTAPAGLIA, D., 2007a. Aplicações operacionais de marcadores moleculares, in: BORÉM, A. (Ed.), Biotecnologia florestal. Suprema, Visconde do Rio Branco.

GRATTAPAGLIA, D., 2007b. Mapas genéticos e seleção assistida por marcadores moleculares, in: BORÉM, A. (Ed.), Biotecnologia florestal. UFV, Viçosa, MG, pp. 201-230.

GRATTAPAGLIA, D., RESENDE, M. D. V., Genomic selection in forest tree breeding. **Tree Genetics & Genomes**, v. 7, p. 241-255, 2011.

HALLAUGER, A. R., MIRANDA FILHO, J. B., 1981. Quantitative genetics in maize Breeding. Iowa State University, Ames.

HÄSTBACKA, J., DE LA CHAPELLE, A., KAITILA, I., SISTONEN, P., WEAVER, A., LANDER, E., Linkage disequilibrium mapping in isolated founder populations: diastrophic dysplasia in Finland. **Nature genetics**, v. 2, p. 204-211, 1992.

HAYES, B. J., BOWMAN, P. J., CHAMBERLAIN, A. J., GODDARD, M. E., Invited review: Genomic selection in dairy cattle: Progress and challenges. **Journal of dairy science**, v. 92, p. 433-443, 2009.

HEFFNER, E. L., SORRELLS, M. E., JANNINK, J.-L., Genomic selection for crop improvement. **Crop Science**, v. 49, p. 1-12, 2009.

HUANG, X., WEI, X., SANG, T., ZHAO, Q., FENG, Q., ZHAO, Y., LI, C., ZHU, C., LU, T., ZHANG, Z., Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature genetics**, v. 42, p. 961-967, 2010.

HWANG, E.-Y., SONG, Q., JIA, G., SPECHT, J. E., HYTEN, D. L., COSTA, J., CREGAN, P. B., A genome-wide association study of seed protein and oil content in soybean. **BMC genomics**, v. 15, p. 1-12, 2014.

JACCOUD, D., PENG, K., FEINSTEIN, D., KILIAN, A., Diversity arrays: a solid state technology for sequence information independent genotyping. **Nucleic acids research**, v. 29, p. e25-e25, 2001.

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

JENKINS, S., GIBSON, N., High-throughput SNP genotyping. **Comparative and functional genomics**, v. 3, p. 57-66, 2002.

JUNQUEIRA, V. S., PEIXOTO, L. D. A., LAVIOLA, B. G., BHERING, L. L., MENDONÇA, S., COSTA, T. D. S. A., ANTONIASSI, R., Bayesian Multi-Trait Analysis Reveals a Useful Tool to Increase Oil Concentration and to Decrease Toxicity in *Jatropha curcas* L. **PloS one**, v. 11, p. e0157038, 2016.

KISRT, M., 2007. Genômica florestal: Novas abordagens, desafios e perspectivas, in: BORÉM, A. (Ed.), Biotecnologia florestal. Suprema, Visconde do Rio Branco, MG.

KUMP, K. L., BRADBURY, P. J., WISSER, R. J., BUCKLER, E. S., BELCHER, A. R., OROPEZA-ROSAS, M. A., ZWONITZER, J. C., KRESOVICH, S., MCMULLEN, M. D., WARE, D., Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. **Nature genetics**, v. 43, p. 163-168, 2011.

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

LAVIOLA, B. G., ROSADO, T. B., BHERING, L. L., KOBAYASHI, A. K., RESENDE, M. D. V. D., Genetic parameters and variability in physic nut accessions during early developmental stages. **Pesquisa Agropecuária Brasileira**, v. 45, p. 1117-1123, 2010.

LE GOUIS, J., BORDES, J., RAVEL, C., HEUMEZ, E., FAURE, S., PRAUD, S., GALIC, N., REMOUE, C., BALFOURIER, F., ALLARD, V., Genome-wide association analysis to identify chromosomal regions determining components of earliness in wheat. **Theoretical and Applied Genetics**, v. 124, p. 597-611, 2012.

LI, H., LINDGREN, D., Comparison of phenotype and combined index selection at optimal breeding population size considering gain and gene diversity. **Silvae Genetica**, v. 55, p. 13-18, 2006.

LI, H., PENG, Z., YANG, X., WANG, W., FU, J., WANG, J., HAN, Y., CHAI, Y., GUO, T., YANG, N., Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. **Nature Genetics**, v. 45, p. 43-50, 2013.

LORENZ, A. J., Resource allocation for maximizing prediction accuracy and genetic gain of genomic selection in plant breeding: a simulation experiment. **G3: Genes| Genomes| Genetics**, v. 3, p. 481-491, 2013.

LORENZ, A. J., SMITH, K. P., JANNINK, J.-L., Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. **Crop science**, v. 52, p. 1609-1621, 2012.

MANDOULAKANI, B. A., YANIV, E., KALENDAR, R., RAATS, D., BARIANA, H. S., BIHAMTA, M. R., SCHULMAN, A. H., Development of IRAP-and REMAP-derived SCAR markers for marker-assisted selection of the stripe rust resistance gene Yr15 derived from wild emmer wheat. **Theoretical and Applied Genetics**, v. 128, p. 211-219, 2015.

MATTA, F. D. P., VIANA, J. M. S., Eficiências relativas dos processos de seleção entre e dentro de famílias de meios-irmãos em população de milho-pipoca. **Ciência e Agrotecnologia**, v. 27, p. 548-556, 2003.

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

MORRIS, G. P., RAMU, P., DESHPANDE, S. P., HASH, C. T., SHAH, T., UPADHYAYA, H. D., RIERA-LIZARAZU, O., BROWN, P. J., ACHARYA, C. B., MITCHELL, S. E., Population genomic and genome-wide association studies of agroclimatic traits in sorghum. **Proceedings of the National Academy of Sciences**, v. 110, p. 453-458, 2013.

NEUMANN, K., KOBILJSKI, B., DENČIĆ, S., VARSHNEY, R., BÖRNER, A., Genome-wide association mapping: a case study in bread wheat (*Triticum aestivum* L.). **Molecular Breeding**, v. 27, p. 37-58, 2011.

PANDOLFI, C., MUGNAI, S., AZZARELLO, E., BERGAMASCO, S., MASI, E., MANCUSO, S., Artificial neural networks as a tool for plant identification: a case study on Vietnamese tea accessions. **Euphytica**, v. 166, p. 411-421, 2009.

PEIXOTO, L. A., BHERING, L. L., CRUZ, C. D., Artificial neural networks reveal efficiency in genetic value prediction. **Genetics and molecular research: GMR**, v. 14, p. 6796, 2015.

PEIXOTO, L. A., LAVIOLA, B. G., BHERING, L. L., MENDONÇA, S., COSTA, T. D. S. A., ANTONIASSI, R., Oil content increase and toxicity reduction in jatropha seeds through family selection. **Industrial Crops and Products**, v. 80, p. 70-76, 2016.

PIEPHO, H. P., MÖHRING, J., MELCHINGER, A. E., BÜCHSE, A., BLUP for phenotypic selection in plant breeding and variety testing. **Euphytica**, v. 161, p. 209-228, 2008.

PIRES, I. E., RESENDE, M. D. V., SILVA, R. L., RESENDE JÚNIOR, M., 2011. Genética florestal. Arka, Viçosa, MG.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M., Genomic selection in wheat breeding using genotyping-by-sequencing. **The Plant Genome**, v. 5, p. 103-113, 2012.

RAFALSKI, A., Applications of single nucleotide polymorphisms in crop genetics. **Current opinion in plant biology**, v. 5, p. 94-100, 2002.

RAMALHO, M. A. P., 2005. Experimentação Em Genética e Melhoramento de Plantas. UFLA, Lavras.

RESENDE, M. D. V., 2000. Análise estatística de modelos mistos via REML/BLUP na experimentação em melhoramento de plantas perenes. Embrapa Florestas, Colombo.

RESENDE, M. D. V., 2002a. Genética biométrica e estatística no melhoramento de plantas perenes, Brasília.

RESENDE, M. D. V., 2008. Genômica quantitativa e seleção no melhoramento de plantas perenes e animais. EMBRAPA Floresta, Colombo, PR.

RESENDE, M. D. V. D., 2002b. Software SELEGEN - REML/BLUP. EMBRAPA Floresta, Colombo.

RESENDE, M. F. R., ALVES, A. A., SÁNCHEZ, C. F. B., RESENDE, M. D. V., CRUZ, C. D., 2013. Seleção genômica ampla, in: CRUZ, C. D., SALGADO, C. C., BHERING, L. L. (Eds.), Genômica aplicada. Suprema, Visconde do Rio Branco, MG, pp. 375-424.

ROCHA, R. B., ROCHA, M. D. G. B., SANTANA, R. C., VIEIRA, A. H., Estimação de parâmetros genéticos e seleção de procedências e famílias de *Dipteryx alata* Vogel (Baru) utilizando metodologia de Reml/Blup e E (QM). **Cerne**, v. 3, p. 331-338, 2009.

ROSADO, A. M., ROSADO, T. B., RESENDE JÚNIOR, M., BHERING, L. L., CRUZ, C. D., Ganhos genéticos preditos por diferentes métodos de seleção em progênies de *Eucalyptus urophylla*. **Pesquisa Agropecuária Brasileira**, v. 44, p. 1653-1659, 2009.

SAIKI, R., GELFAND, D., STOFFEL, S., SCHARF, S., HIGUCHI, R., HORN, G., MULLIS, K., EHRLICH, H., Primer-directed enzymatic amplification of DNA. **Science**, v. 239, p. 487-491, 1988.

SAMAYOA, L. F., MALVAR, R. A., OLUKOLU, B. A., HOLLAND, J. B., BUTRÓN, A., Genome-wide association study reveals a set of genes associated with resistance to the Mediterranean corn borer (*Sesamia nonagrioides* L.) in a maize diversity panel. **BMC plant biology**, v. 15, p. 1, 2015.

SANSALONI, C. P., PETROLI, C. D., CARLING, J., HUDSON, C. J., STEANE, D. A., MYBURG, A. A., GRATTAPAGLIA, D., VAILLANCOURT, R. E., KILIAN, A., A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus. **Plant Methods**, v. 6, p. 1, 2010.

SANTOS, F. S., AMARAL JÚNIOR, A. T. D., JÚNIOR, F., DE PAIVA, S., RANGEL, R. M., SCAPIM, C. A., MORA, F., Genetic gain prediction of the third recurrent selection cycle in a popcorn population. **Acta Scientiarum. Agronomy**, v. 30, p. 651-655, 2008.

SAX, K., The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. **Genetics**, v. 8, p. 552, 1923.

SPINDEL, J., BEGUM, H., AKDEMIR, D., VIRK, P., COLLARD, B., REDOÑA, E., ATLIN, G., JANNINK, J.-L., MCCOUCH, S. R., Genomic Selection and Association Mapping in rice (*Oryza sativa)*: Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS Genet**, v. 11, p. e1004982, 2015.

STEELE, K. A., PRICE, A. H., WITCOMBE, J. R., SHRESTHA, R., SINGH, B. N., GIBBONS, J. M., VIRK, D. S., QTLs associated with root traits increase yield in upland rice when transferred through marker-assisted selection. **Theoretical and applied genetics**, v. 126, p. 101-108, 2013.

SVED, J. A., Linkage disequilibrium and homozygosity of chromosome segments in finite populations. **Theoretical population biology**, v. 2, p. 125-141, 1971.

SZAJKO, K., STRZELCZYK-ŻYTA, D., MARCZEWSKI, W., Ny-1 and Ny-2 genes conferring hypersensitive response to potato virus Y (PVY) in cultivated potatoes: mapping and marker-assisted selection validation for PVY resistance in potato breeding. **Molecular Breeding**, v. 34, p. 267-271, 2014.

TANKSLEY, S. D., Mapping polygenes. **Annual review of genetics**, v. 27, p. 205-233, 1993.

TEODORO, P. E., COSTA, R. D., ROCHA, R. B., LAVIOLA, B. G., Contribuição de caracteres agronômicos para a produtividade de grãos em pinhão-manso. **Bragantia**, v. 75, p. 51-56, 2016.

TIAN, F., BRADBURY, P. J., BROWN, P. J., HUNG, H., SUN, Q., FLINT-GARCIA, S., ROCHEFORD, T. R., MCMULLEN, M. D., HOLLAND, J. B., BUCKLER, E. S., Genome-wide association study of leaf architecture in the maize nested association mapping population. **Nature genetics**, v. 43, p. 159-162, 2011.

VAN INGHELANDT, D., MELCHINGER, A. E., LEBRETON, C., STICH, B., Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. **Theoretical and Applied Genetics**, v. 120, p. 1289-1299, 2010.

VENCOVSKY, R., BARRIGA, P., 1992. Genética biométrica no fitomelhoramento. Sociedade Brasileira de Genética, Ribeirão Preto.

VIEIRA, E. A., COIMBRA, J. L. M., FLOSS, I. P. V. E. L., DA SILVA, I. B. G. O., Adaptabilidade e estabilidade em aveia em ambientes estratificados. **Ciência Rural**, v. 35, p., 2005.

WEIGEL, K. A., GIANOLA, D., Estimation of heterogeneous within-herd variance components using empirical Bayes methods: a simulation study. **Journal of dairy science**, v. 75, p. 2824-2833, 1992.

WEN, Z., TAN, R., YUAN, J., BALES, C., DU, W., ZHANG, S., CHILVERS, M. I., SCHMIDT, C., SONG, Q., CREGAN, P. B., Genome-wide association mapping of quantitative resistance to sudden death syndrome in soybean. **BMC genomics**, v. 15, p. 1, 2014.

WENZL, P., LI, H., CARLING, J., ZHOU, M., RAMAN, H., PAUL, E., HEARNDEN, P., MAIER, C., XIA, L., CAIG, V., A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. **Bmc Genomics**, v. 7, p. 1, 2006.

WHITE, T. L., ADAMS, W. T., NEALE, D. B., 2007. Forest genetics. CABI publishing, Wallingford.

WILLIAMS, J. G., KUBELIK, A. R., LIVAK, K. J., RAFALSKI, J. A., TINGEY, S. V., DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic acids research**, v. 18, p. 6531-6535, 1990.

YANIV, E., RAATS, D., RONIN, Y., KOROL, A. B., GRAMA, A., BARIANA, H., DUBCOVSKY, J., SCHULMAN, A. H., FAHIMA, T., Evaluation of marker-assisted selection for the stripe rust resistance gene Yr15, introgressed from wild emmer wheat. **Molecular Breeding**, v. 35, p. 1-12, 2015.

ZHU, C., GORE, M., BUCKLER, E. S., YU, J., Status and prospects of association mapping in plants. **The plant genome**, v. 1, p. 5-20, 2008.

ZHU, Y., SONG, Q., HYTEN, D., VAN TASSELL, C., MATUKUMALLI, L., GRIMM, D., HYATT, S., FICKUS, E., YOUNG, N., CREGAN, P., Single-nucleotide polymorphisms in soybean. **Genetics**, v. 163, p. 1123-1134, 2003.

2. OBJETIVO GERAL

Avaliar como os fatores inclusão de QTLs como efeito fixo no modelo, método utilizado, herdabilidade da característica, número de QTLs controlando a característica, número de indivíduos na população de treinamento e número de marcadores influenciam na predição do valor genético genômico pela metodologia de seleção genômica.

3. CAPITULO 1

MODELOS DE SELEÇÃO GENÔMICA COM INCLUSÃO DE COVARIÁVEIS

MODELOS DE SELEÇÃO GENÔMICA COM INCLUSÃO DE COVARIÁVEIS

3.1. RESUMO

A seleção genômica (SG) tem se tornado uma ferramenta de grande utilidade no melhoramento de plantas. Porém a utilização da SG em conjunto com a seleção assistida por marcador molecular (SAM) ou estudo de associação genômica ampla (EAGA) ainda é pouco estudada. Portanto, o objetivo da pesquisa foi avaliar modelos utilizando os SNPs significativos encontrados pelos SAM e EAGA como efeito fixo nos modelos comumente utilizados na SG, onde, no modelo tradicional, todos os SNPs são estabelecidos como de efeito aleatório. Estes modelos foram comparados com o modelo padrão utilizado na SG (RRBLUP bayesiano). As comparações entre modelos foram realizadas avaliando a capacidade de predição fenotípica e genotípica, acurácia fenotípica e genotípica, ganho de seleção, coincidência de seleção e tempo de processamento. Os métodos EAGA e SAM não conseguiram identificar de forma acurada os verdadeiros QTLs mostrando que a seleção baseada apenas nos QTLs identificados por estes métodos pode selecionar indivíduos de baixo valor genético. A utilização de um modelo de SG com as marcas significativas encontradas pelo EAGA como efeito fixo e as demais marcas como efeito aleatório é uma boa estratégia para selecionar indivíduos superiores com alta acurácia. A introdução no modelo de SG de QTLs que já foram descritos previamente para a característica em estudo, como efeito fixo, permite a seleção de indivíduos superiores de forma mais acurada.

Palavras-chave: Predição genômica, seleção assistida por marcadores moleculares, estudo de associação genômica, modelagem

3.2. INTRODUÇÃO

O melhoramento de plantas, desde os seus primórdios, tem se baseado na seleção visual de indivíduos, ou seja, a seleção com base apenas no valor fenotípico (Allard, 1999). Com os avanços da genética molecular e da genômica, outras novas estratégias, e consequentemente novos critérios, têm sido desenvolvidas de forma a tornar mais rápido o ciclo de melhoramento e mais eficiente a seleção de indivíduos. A primeira metodologia baseada em marcadores foi a seleção assistida por marcadores molecular (SAM). A SAM baseia-se na informação de um provável QTL, onde alguns são identificados como responsáveis pela expressão de uma

determinada característica fenotípica (Jena & Mackill, 2008). Estes marcadores são conhecidos como locos de características quantitativas (QTL). A SAM tem se mostrado eficiente para características governadas por poucos genes de grande efeito e inúmeros trabalhos nas principais culturas agrícolas têm demonstrado essa eficiência, como tolerância a salinidade (Ashraf et al., 2012; Ashraf & Foolad, 2013), características de raiz em arroz (Steele et al., 2013), resistência a fusarium em trigo (Arruda et al., 2016), ferrugem em trigo (Yaniv et al., 2015) e Potato Virus Y (PVY) em batata (Szajko et al., 2014). Porém para características governadas por muitos genes de pequeno efeito, este método tem se mostrado inapropriado (Gregorio et al., 2013).

A seleção genômica ampla (SG) descrita por Meuwissen et al. (2001) apresenta-se como alternativa para resolver as limitações encontradas pela SAM para características quantitativas. Os modelos de SG baseiam-se na estimação do valor genético genômico utilizando grande número de marcadores e o valor fenotipico dos indivíduos (Meuwissen et al., 2001). Este valor genético genômico é então utilizado como critério para selecionar indivíduos superiores. Primeiramente os efeitos de marcadores são estimados via população de treinamento, onde os indivíduos são genotipados e fenotipados. Estes efeitos de marcadores são então utilizados para estimar o valor genético genômico nas populações de validação, onde os indivíduos são genotipados e fenotipados. Assim a grande diferença entre SG e SAM é que na seleção genômica não é necessário verificar a significância dos marcadores, ou seja, o valor genético genômico é baseado em todas as marcas disponíveis evitando perda de informações e, também não é necessário a construção de um mapa genético, podendo assim, os modelos de SG serem utilizados em populações não estruturadas (Lorenz et al., 2011; Meuwissen et al., 2001).

As maiores aplicações da SG em plantas têm sido demostradas em milho, trigo, soja e espécies florestais. Acurácias elevadas têm sido encontradas para predição da produção e outras características quantitativas em populações de duplo-haploide em milho por meio da validação cruzada (Guo et al., 2012; Lorenzana & Bernardo, 2009). Moderadas acurácias têm sido obtidas para predição da produção e outras características em coleção de germoplasmas em milho, trigo, aveia e cevada (Asoro et al., 2011; Crossa et al., 2014; Heffner et al., 2011; Lorenz et al., 2012; Ornella et al., 2012; Rutkoski et al., 2012). Alguns trabalhos preliminares utilizando SG também tem sido desenvolvido em mandioca, cana-de-açúcar e beterraba (de Oliveira et al., 2012; Gouy et al., 2013; Ly et al., 2013; Würschum et al., 2013).

Outra limitação da SAM é a necessidade do mapa de ligação que só pode ser criado a partir de populações estruturadas. Uma alternativa que se tornou possível apenas após o desenvolvimento dos marcadores SNPs são os mapas de associação. Os mapas de associação têm o potencial de encontrar e mapear os QTLs dentro do genoma, além de identificar polimorfismo causais dentro dos genes que pode ser responsável pela diferença entre dois fenótipos (Palaisa et al., 2003). Assim, foi desenvolvido uma metodologia capaz de identificar QTLs significativos a partir de um grande número de marcas cobrindo todo o genoma conhecido como estudo de associação genômica ampla (EAGA) (Pritchard et al., 2000). EAGA tem sido utilizado para identificar QTLs em inúmeras características quantitativas em programas de melhoramento de plantas, tais como arquitetura de folhas em milho (Tian et al., 2011), características agroclimáticas em sorgo (Morris et al., 2013), características relacionadas a produção em arroz (Huang et al., 2010), proteína e óleo em soja (Hwang et al., 2014), entre outros.

Apesar de serem metodologias distintas, SAM, SG e EAGA tem o mesmo objetivo final: melhorar a acurácia de seleção. Uma alternativa de utilizar essas metodologias simultaneamente é considerar os QTLs identificados pelo SAM e EAGA como efeito fixo e os demais marcadores como efeito aleatório no modelo original de SG (Hayr, 2013). Porém, poucos trabalhos têm demonstrado a utilização destas metodologias simultaneamente (Arruda et al., 2016; Bernardo, 2014; Spindel et al., 2015) e nenhum deles leva em consideração o efeito da herdabilidade sobre a predição genômica. Portanto, o objetivo da pesquisa foi avaliar modelos utilizando os SNPs significativos detectados pelas metodologias de seleção assistida por marcadores moleculares (SAM) e estudo da associação genômica ampla (EAGA) como efeito fixo nos modelos comumente utilizados na seleção genômica ampla (SG). Além dos modelos descritos acima, outros modelos foram utilizados todos os QTLs simulados ou utilizando apenas os 2 QTLs de maior efeito como efeito fixo no modelo de SG e os demais SNPs como efeito aleatório. Estes modelos foram comparados com o modelo padrão utilizado na seleção genômica (RRBLUP bayesiano).

3.3. MATERIAL E MÉTODOS

3.3.1. Simulação dos dados

A simulação da população F_2 foi realizada utilizando o módulo de simulação do aplicativo computacional GENES (Cruz, 2013), que permitiu gerar informações sobre

o genoma, genótipos dos genitores, populações de cruzamentos controlados e dados de características quantitativas.

3.3.1.1. Simulação do genoma

Foi simulado um genoma constituído de 15 grupos de ligação, similar ao de uma espécie diploide 2n=2x=30. Cada grupo de ligação foi simulado com 150 cM, e constituído por 300 marcas, espaçadas de forma equidistante (0,5 cM), totalizando 4.500 marcas. As marcas foram assumidas como codominantes e bialélicas.

3.3.1.2. Simulação dos genitores

Pais homozigotos contrastantes foram simulados, ou seja, o pai 1 foi codificado como portador de um alelo A_1 (recebeu código 2), e o pai 2 foi codificado como portador do alelo alternativo A_2 (recebeu código 0) para todas as marcas existentes. Desta forma, o cruzamento entre pai 1 e pai 2 gerou a população F_1 com todos as marcas em heterozigose e em fase de aproximação do tipo $A_1B_1//A_2B_2$.

3.3.1.3. Simulação da população de mapeamento

População F_2 foram geradas a partir da autofecundação dos indivíduos da população F_1. Para a formação do primeiro indivíduo da população F_2, cada indivíduo da população F_1 produziu 5.000 gametas e quando 2 destes gametas se encontraram ao acaso, o primeiro indivíduo da população F_2 foi gerado. Este processo se repetiu até formação de todos os indivíduos na população.

Para a formação de cada gameta seguiu-se o seguinte critério: O alelo da primeira marca foi escolhido de forma aleatória (A_1 ou A_2) para começar a constituir o gameta (alelo de inicialização); O alelo da segunda marca foi escolhido levando em consideração a distância para o primeiro gene, ou seja, as frequências de crossing over foram contabilizados e a escolha de qual alelo (B_1 ou B_2) constituiria o gameta foi baseado nas probabilidades de cada gameta $P(A_1B_1)$ e $P(A_2B_2)$ que são gametas parentais, e $P(A_1B_2)$ e $P(A_2B_1)$ que são os gametas recombinantes. Este processo se repetiu até chegar no último gene. Foi considerado interferência nula, ou seja, o crossing over ocorrido entre os genes A e B não interferem em um próximo crossing over entre os genes B e C. Desta forma foi garantido que todos os gametas formados eram diferentes devido a escolha aleatória do alelo no primeiro gene e as probabilidades condicionadas a cada alelo para os próximos genes. Como todos os genes foram simulados de forma equidistante a 0,5 cM, a frequência de recombinação

foi de 0,5% para todos os genes, ou seja, a probabilidade de cada gameta foi: $P(A_1B_1)$ = $P(A_2B_2)$= 0.4975 e $P(A_1B_2)$ = $P(A_2B_1)$ = 0,0025.

A população F_2 simulada foi codificada com 0, 1 e 2, sendo que 0 correspondeu aos indivíduos homozigotos (A_2A_2), 1 aos indivíduos heterozigotos (A_1A_2),) e 2 aos indivíduos homozigotos (A_1A_1), para um determinado loco.

3.3.1.4. Simulação das características quantitativas

Para a simulação das características quantitativas primeiramente foi atribuído, como importância de cada loco, um valor correspondente à probabilidade gerada por uma distribuição binomial, de parâmetro p=q= 0,5 e n = 99 (gerando uma família de probabilidade com 100 elementos). Este valor, denominado proporção da variância genética explicada por cada QTL ($\mathrm{PVG/QTL}$), corresponde a importância do loco para a média genotípica e, consequentemente, para a proporção da variância genética da característica explicada por cada QTL.

Cada característica foi simulada sendo controlada por 100 QTLs distribuídos de forma aleatória no genoma. O efeito de cada QTL foi definido por: $A_1A_1=\mu + a$; $A_1A_2=\mu + d$; $A_2A_2=\mu - a$, onde a é o efeito codificado do homozigoto e d é o efeito codificado do heterozigoto, que foi estabelecido como nulo no processo de simulação.

O valor genotípico (VG) de cada indivíduo foi definido pela equação:

$$\mathrm{VG} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{PVG/QTL_i} \ \ \mathrm{x\ efeito\ do\ QTL_i})$$

O efeito ambiental (EA) foi assumido como não correlacionado com o valor genotípico e foi estimado seguindo uma distribuição $N(0,\sigma^2)$. O valor de σ^2 é calculado a partir da herdabilidade da característica e o valor da variância genética (σ_g^2). Foram simuladas características com herdabilidade de 5%, 20%, 40%, 60%, 80% e 99%. A σ_g^2 foi calculada como sendo a variância do valor genotípico dos indivíduos da população F_2. Sendo assim, o valor fenotípico foi calculado como:

$$\mathrm{VF} = \mathrm{u} + \mathrm{VG} + \mathrm{EA}$$

em que u = 100 é média e VF é o valor fenotípico.

3.3.2. Modelos utilizados

Sete modelos de seleção genômica (SG) foram utilizados nas análises. O método de SG utilizado foi o Bayesian ridge regression best linear unbiased prediction (BRR) que tem como objetivo estimar o efeito para cada uma das covariáveis

(marcadores SNPs) incluídos no modelo. O BRR assume que todos os SNPs apresentam controle na expressão fenotípica dos QTLs e assume variância homogênea. O pacote BGLR (Pérez & de los Campos, 2012) foi utilizado para processar o método BRR. Este método foi escolhido pois no pacote BGLR é possível modelar efeitos fixos e efeitos aleatórios e ele é o método bayesiano que requer menor tempo computacional.

Modelo 1 (SAM): Neste modelo os SNPs significativos para cada característica em análise encontrados pela seleção assistida por marcadores moleculares (SAM) foram modelados como fixo. As análises da SAM foram realizadas usando a função cim no pacote qtl no programa R (R Core Team, 2015). Assim, o modelo 1 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\beta}X$$

Em que μ é a média genotípica, β é o vetor de efeito de cada SNP significativo encontrados pelas análises SAM (efeito fixo), e X é matriz de marcadores composta apenas pelos SNPs significativos.

Modelo 2 (AGA): Neste modelo os SNPs significativos para cada característica em análise encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como fixo. As análises EAGA foram realizadas usando a função gwas no pacote rrBLUP no programa R (R Core Team, 2015). Assim, o modelo 2 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\beta}X$$

Em que μ é a média genotípica, β é o vetor de efeito de cada SNP significativo encontrados pelas análises EAGA (efeito fixo), e X é matriz de marcadores composta apenas pelos SNPs significativos.

Modelo 3 (SGA): Neste modelo todos os SNPs foram modelados como aleatório. Assim, o modelo 3 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\alpha}W$$

Em que μ é a média genotípica, α é o vetor de efeito aditivo de cada SNP (efeito aleatório), e W é matriz de marcadores composta por todos os SNPs.

Modelo 4 (S_S): Neste modelo os SNPs significativos para cada característica em análise encontrados pela SAM foram modelados como efeito fixo e os SNPs restantes foram modelados como efeito aleatório. Assim, o modelo 4 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W$$

Em que μ é a média genotípica, β é o vetor de efeito de cada SNP significativo encontrados pelas análises SAM (efeito fixo), X é matriz de marcadores composta apenas pelos SNPs significativos, α é o vetor de efeito aditivo de cada SNP (efeito aleatório), e W é matriz de marcadores composta por todos os SNPs não significativos nas análises SAM.

Modelo 5 (A_S): Neste modelo os SNPs significativos para cada característica em análise encontrados pelo EAGA foram modelados como efeito fixo e os SNPs restantes foram modelados como efeito aleatório. Assim, o modelo 5 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W$$

Em que μ é a média genotípica, β é o vetor de efeito de cada SNP significativo encontrados pelas análises EAGA (efeito fixo), X é matriz de marcadores composta apenas pelos SNPs significativos, α é o vetor de efeito aditivo de cada SNP (efeito aleatório), e W é matriz de marcadores composta por todos os SNPs não significativos nas análises EAGA.

Modelo 6 (M_S): Neste modelo os dois QTls de maior efeito (Tabela 1) simulados para cada característica em análise foram modelados como efeito fixo e os SNPs restantes foram modelados como efeito aleatório. Assim, o modelo 6 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W$$

Em que μ é a média genotípica, β é o vetor de efeito dos dois QTls mais significativos pelo processo de simulação (efeito fixo), X é matriz de marcadores composta apenas pelos dois QTLs, α é o vetor de efeito aditivo de cada SNP (efeito aleatório), e W é matriz de marcadores composta por todos os SNPs exceto os dois QTLs de maior efeito sobre a característica.

Modelo 7 (Q_S): Neste modelo todos os QTls simulados (Tabela 1) para cada característica em análise foram modelados como efeito fixo e os SNPs restantes foram modelados como efeito aleatório. Assim, o modelo 7 segue a seguinte equação:

$$GEBV = \hat{\mu} + \hat{\beta}X + \hat{\alpha}W$$

Em que μ é a média genotípica, β é o vetor dos efeitos de todos os QTls (efeito fixo), X é matriz de marcadores composta por todos os QTLs, α é o vetor de efeito aditivo de cada SNP (efeito aleatório), e W é matriz de marcadores composta por todos os SNPs exceto os QTLs.

Tabela 1. Posição (Pos) e efeito de cada QTL simulado para cada característica.

h^2=0,05		h^2=0,20		h^2=0,40		h^2=0,60		h^2=0,80		h^2=0,99	
Pos	**Efeito**	**Pos**	**Efeito**	**Pos**	**Efeito**	**Pos**	**Efeito**	**Pos**	**Efeito**	**Pos**	**Efeito**
2.510	0	2.949	0	651	0	1.389	0	67	0	2.291	0
1.922	0	2.589	0	2.513	0	908	0	2.803	0	2.936	0
499	0	936	0	158	0	2.294	0	1.858	0	512	0
202	0	2.787	0	1.695	0	68	0	2.346	0	2.993	0
187	0	671	0	1.949	0	543	0	1.081	0	1.706	0
2.656	0	649	0	95	0	83	0	170	0	2.309	0
1.336	0	2.879	0	281	0	1.104	0	325	0	1.201	0
2.314	0	1.954	0	2.399	0	2.899	0	2.389	0	1.544	0
207	0	1.253	0	2.895	0	2.895	0	2.260	0	1.061	0
1.221	0	1.006	0	818	0	2.939	0	689	0	2.644	0
1.198	0	2.357	0	2.960	0	57	0	1.847	0	2.918	0
962	0	2.469	0	46	0	2.287	0	1.986	0	2.232	0
2.480	0	1.836	0	344	0	798	0	790	0	236	0
1.901	0	779	0	1.494	0	662	0	211	0	1.878	0
2.500	0	1.183	0	488	0	2.172	0	896	0	967	0
2.098	0	2.902	0	2.783	0	1.320	0	2.568	0	61	0
1.481	0	1.105	0	283	0	1.760	0	1.600	0	982	0
2.773	0	2.216	0	1.889	0	2.040	0	2.758	0	1.520	0
642	0	1.740	0	2.394	0	2.648	0	1.985	0	1.223	0
1.140	0	1.804	0	1.304	0	1.388	0	291	0	1.238	0
657	0	1.500	0	2.405	0	2.684	0	269	0	158	0
2.523	0	1.292	0	831	0	2.763	0	392	0	1.498	0
384	1E-08	1.505	1E-08	2.656	1E-08	382	1E-08	1.800	1E-08	2.044	1E-08
1.314	3E-08	2.024	3E-08	2.001	3E-08	1.148	3E-08	873	3E-08	857	3E-08
2.833	1E-07	1.286	1E-07	1.538	1E-07	2.970	1E-07	815	1E-07	1.104	1E-07
255	2,9E-07	1.788	2,9E-07	1.088	2,9E-07	1.283	2,9E-07	1.977	2,9E-07	649	2,9E-07
2.063	8,2E-07	45	8,2E-07	1.326	8,2E-07	933	8,2E-07	919	8,2E-07	2.895	8,2E-07
631	2,21E-06	2.313	2,21E-06	2.609	2,21E-06	1.288	2,21E-06	2.721	2,21E-06	1.580	2,21E-06
2.078	5,68E-06	1.214	5,68E-06	117	5,68E-06	657	5,68E-06	2.344	5,68E-06	1.557	5,68E-06
2.746	1,39E-05	2.194	1,39E-05	2.434	1,39E-05	1.678	1,39E-05	1.097	1,39E-05	2.287	1,39E-05
1.562	3,24E-05	2.961	3,24E-05	372	3,24E-05	1.852	3,24E-05	1..912	3,24E-05	1.925	3,24E-05
2.334	7,22E-05	2.200	7,22E-05	165	7,22E-05	2.453	7,22E-05	116	7,22E-05	2.162	7,22E-05
1.442	0,000153	1.717	0,000153	1.675	0,000153	784	0,000153	1.038	0,000153	1.423	0,000153
480	0,000312	67	0,000312	1.679	0,000312	1.512	0,000312	1.500	0,000312	570	0,000312
790	0,000605	1.303	0,000605	1.106	0,000605	522	0,000605	1.493	0,000605	378	0,000605
351	0,001123	1.858	0,001123	2.087	0,001123	2.832	0,001123	1.261	0,001123	541	0,001123
2.165	0,001996	893	0,001996	693	0,001996	1.163	0,001996	1.429	0,001996	398	0,001996
213	0,003399	1.831	0,003399	763	0,003399	2.943	0,003399	862	0,003399	2.701	0,003399
245	0,005546	920	0,005546	2.552	0,005546	2.475	0,005546	2.440	0,005546	388	0,005546
2.661	0,008675	2.011	0,008675	1.470	0,008675	1.086	0,008675	1.611	0,008675	1.592	0,008675
2.747	0,013013	935	0,013013	2.531	0,013013	1.266	0,013013	1.134	0,013013	1.132	0,013013
1.224	0,018726	1.685	0,018726	1.345	0,018726	2.059	0,018726	2.664	0,018726	2.648	0,018726
2.751	0,025859	1.860	0,025859	93	0,025859	2.997	0,025859	1.203	0,025859	720	0,025859
2.427	0,034278	394	0,034278	2.698	0,034278	586	0,034278	1.364	0,034278	1.987	0,034278
1.125	0,043627	111	0,043627	290	0,043627	856	0,043627	1.205	0,043627	841	0,043627
1.394	0,053322	40	0,053322	1.927	0,053322	1.305	0,053322	1.158	0,053322	914	0,053322
2.651	0,062595	210	0,062595	171	0,062595	1.653	0,062595	1.890	0,062595	1.407	0,062595
185	0,070586	1.692	0,070586	2.681	0,070586	939	0,070586	540	0,070586	836	0,070586
302	0,076468	1.865	0,076468	2.590	0,076468	1.996	0,076468	2.687	0,076468	1.993	0,076468
1.800*	**0,079589**	**99**	**0,079589**	**1.572**	**0,079589**	**480**	**0,079589**	**314**	**0,079589**	**953**	**0,079589**
2.997	**0,079589**	**1.305**	**0,079589**	**2.074**	**0,079589**	**890**	**0,079589**	**1.478**	**0,079589**	**1.300**	**0,079589**
2.494	0,076468	297	0,076468	1.267	0,076468	1.937	0,076468	582	0,076468	389	0,076468
697	0,070586	1.088	0,070586	865	0,070586	1.100	0,070586	653	0,070586	522	0,070586
798	.0,062595	1.065	0,062595	1.967	0,062595	2.773	0,062595	2.463	0,062595	2.915	0,062595

956	0,053322	1.188	0,053322	1.354	0,053322	2.882	0,053322	508	0,053322	413	0,053322
386	0,043627	1.050	0,043627	754	0,043627	2.471	0,043627	2.184	0,043627	2.133	0,043627
2.499	0,034278	1.670	0,034278	1.865	0,034278	80	0,034278	1.576	0,034278	975	0,034278
571	0,025859	815	0,025859	2.410	0,025859	496	0,025859	553	0,025859	1.789	0,025859
1.348	0,018726	523	0,018726	1.126	0,018726	1.270	0,018726	1.420	0,018726	2.766	0,018726
115	0,013013	2.606	0,013013	136	0,013013	500	0,013013	1.863	0,013013	1.309	0,013013
320	0,008675	2.392	0,008675	784	0,008675	2.216	0,008675	2.127	0,008675	2.247	0,008675
1.305	0,005546	2.344	0,005546	1.076	0,005546	1.875	0,005546	1.595	0,005546	1.336	0,005546
2.823	0,003399	2.496	0,003399	1.063	0,003399	245	0,003399	2.518	0,003399	1.560	0,003399
892	0,001996	2.709	0,001996	1.652	0,001996	892	0,001996	1.267	0,001996	555	0,001996
1.054	0,001123	912	0,001123	699	0,001123	794	0,001123	924	0,001123	1.473	0,001123
1.786	0,000605	1.859	0,000605	2.371	0,000605	2.457	0,000605	1.277	0,000605	1.409	0,000605
613	0,000312	2.296	0,000312	2.482	0,000312	1.752	0,000312	194	0,000312	810	0,000312
1.580	0,000153	2.243	0,000153	1.769	0,000153	2.559	0,000153	2.704	0,000153	2.730	0,000153
229	7,22E-05	1.307	7,22E-05	759	7,22E-05	2.491	7,22E-05	1.980	7,22E-05	1.956	7,22E-05
2.142	3,24E-05	1.146	3,24E-05	2.413	3,24E-05	695	3,24E-05	669	3,24E-05	1.539	3,24E-05
2.489	1,39E-05	2.440	1,39E-05	213	1,39E-05	751	1,39E-05	2.699	1,39E-05	1.100	1,39E-05
1.730	5,68E-06	1.141	5,68E-06	870	5,68E-06	1.705	5,68E-06	1.231	5,68E-06	2.773	5,68E-06
2.475	2,21E-06	2.678	2,21E-06	2.869	2,21E-06	409	2,21E-06	1.007	2,21E-06	586	2,21E-06
2.171	8,2E-07	461	8,2E-07	1.589	8,2E-07	1.090	8,2E-07	143	8,2E-07	1.721	8,2E-07
1.918	2,9E-07	1.799	2,9E-07	2.775	2,9E-07	788	2,9E-07	2.104	2,9E-07	830	2,9E-07
2.789	1E-07	1.772	1E-07	1.812	1E-07	2.074	1E-07	583	1E-07	496	1E-07
2.776	3E-08	2.704	3E-08	2.973	3E-08	1.848	3E-08	378	3E-08	1.270	3E-08
1.218	1E-08	2.330	1E-08	1.374	1E-08	220	1E-08	2.098	1E-08	500	1E-08
1.466	0	437	0	1.492	0	866	0	2.997	0	1.466	0
9	0	1.290	0	722	0	666	0	1.066	0	1.875	0
358	0	483	0	1.764	0	2.377	0	1.274	0	2.003	0
1.553	0	1.110	0	2.097	0	1.835	0	2.212	0	143	0
2.989	0	2.274	0	2.178	0	2.487	0	2.004	0	935	0
1.564	0	1.332	0	1.959	0	2.236	0	2.060	0	2.598	0
2.581	0	2.462	0	1.766	0	2.891	0	520	0	253	0
1.379	0	508	0	2.679	0	893	0	1.654	0	78	0
2.888	0	1.434	0	2.548	0	2.479	0	2.710	0	1.695	0
219	0	1.151	0	1.886	0	998	0	2.275	0	2.899	0
960	0	1.022	0	1.588	0	1.069	0	492	0	1	0
1.312	0	1.419	0	261	0	1.871	0	1.715	0	1.705	0
218	0	2.613	0	1.887	0	1.507	0	777	0	2.659	0
2.046	0	767	0	1.177	0	174	0	1.276	0	294	0
2.432	0	2.391	0	1.475	0	1.803	0	2.949	0	1.492	0
1.109	0	2.002	0	2.484	0	793	0	1.793	0	2.824	0
1.567	0	2.906	0	1.478	0	1.391	0	1.686	0	348	0
1.588	0	1.674	0	1.691	0	2.349	0	2.694	0	1.487	0
409	0	1.135	0	1.367	0	1.575	0	672	0	2.649	0
1.560	0	990	0	2.879	0	832	0	649	0	666	0
1.069	0	2.000	0	2.490	0	2.313	0	2.082	0	742	0
68	0	1.698	0	888	0	330	0	177	0	2.963	0

*As duas linhas em negrito significam os dois QTLs de maior efeito simulados que foram utilizados no modelo 6.

3.3.3. Análise dos dados

Após a geração da população, seguiram-se as etapas do processo de mapeamento, iniciando pela análise de segregação de locos individuais. Foram aplicados testes de qui-quadrado (χ^2) para verificar se as marcas segregavam de acordo com o esperado em uma população F_2. Verificou-se ainda se todos os Grupos

de Ligação foram restaurados, com tamanho, distância e ordem dos marcadores, podendo concluir assim que se tratava de uma população F_2 com as propriedades de simulação desejadas.

Para verificar a acurácia das metodologias, SAM e EAGA, em encontrar SNPs significativos foi realizado a comparação entre os QTLs simulados e os QTLs encontrados por estes métodos.

Para comparar os modelos propostos neste trabalho alguns parâmetros foram estimados como segue abaixo:

A capacidade preditiva fenotípica (CPF) e genotípica (CPG) foram estimadas sendo a correlação de Pearson entre o valor genético genômico (GEBV) estimado pelos modelos e o valor fenotípico e genotípico, respectivamente.

A acurácia fenotípica (AF) e genotípica (AG) foram calculadas pelas equações abaixo:

$$AF = \frac{CPF}{\sqrt{h^2}}$$

$$AG = \frac{CPG}{\sqrt{h^2}}$$

Onde h^2 é a herdabilidade da característica.

O ganho de seleção (GS) foi estimado pela seguinte equação:

$$GS = DS * h^2$$

Onde DS é o diferencial de seleção que foi estimado por:

$$DS = m_s - m_o$$

Onde m_s é a média dos indivíduos selecionados e m_o é a média da população inicial. Foi considerado uma porcentagem de seleção de 20%.

A coincidência de seleção (CS) foi calculado da seguinte forma:

$$CS = \frac{NIS}{NT} * 100$$

Onde NIS é o número de indivíduos selecionados com base no valor fenotípico que foram os mesmos selecionados com base no GEBV, e NT é o número total de indivíduos selecionados.

Além dos parâmetros mencionados acima, o tempo de processamento também foi computado. Todos esses parâmetros foram plotados em gráficos de barras com seus respectivos erros, de forma a facilitar a comparação entre os métodos.

O ganho de seleção máximo ($GS_{máx}$) foi estimado pela seguinte expressão:

$$GS_{máx} = \bar{X}_S - \bar{X}_O$$

Onde $\bar{X}_S$ é a média genotípica dos indivíduos selecionados baseados nos valores genéticos simulados (valores verdadeiros) e $\bar{X}_O$ é a média genotípica da população original.

3.4. RESULTADO

Foi observado que o número de QTLs encontrado pelo EAGA foi muito abaixo do total simulado (100 QTLs onde 66 tiveram efeito significativo sobre a característica), enquanto que o número de QTLs encontrado pela SAM foi muito superior (Tabela 2). Poucos QTLs identificados pelo EAGA eram realmente QTLs, ou seja, a maioria dos QTLs era falso positivo. A SAM conseguiu identificar quase 50% dos QTLs simulados, porém apresentou número muito alto de falso positivo.

Tabela 2. Número de QTLs detectados (NQE), número de QTLs detectados na posição correta (NQEPC) e número de falso positivo (NFP) pelos métodos de seleção assistida por marcadores (SAM) e estudo de associação genômica ampla (EAGA) para diferentes herdabilidades.

Modelo + Herdabilidade	NQD	NQDPC	NFP
SAM – h^2 = 5%	1.687	46	1.641
EAGA – h^2 = 5%	12	0	12
SAM – h^2 = 20%	2.221	52	2.169
EAGA – h^2 = 20%	7	1	6
SAM – h^2 = 40%	1.807	47	1.760
EAGA – h^2 = 40%	11	3	8
SAM – h^2 = 60%	1.957	51	1.906
EAGA – h^2 = 60%	24	5	19
SAM – h^2 = 80%	2.096	52	2.044
EAGA – h^2 = 80%	27	7	20
SAM – h^2 = 99%	1.697	46	1.651
EAGA – h^2 = 99%	32	7	25

Os modelos A_S, SGA, Q_S e M_S foram superiores aos demais métodos para estimar a capacidade preditiva fenotípica para todas as herdabilidades avaliadas, exceto na herdabilidade de 5%, onde os modelos SAM e AGA foram superiores (Figura 1). Para herdabilidades de 60, 80 e 99% os modelos SAM e S_S apresentaram a capacidade preditiva fenotípica próxima de zero, sendo muito inferior aos demais modelos que apresentaram valores acima de 0,4.

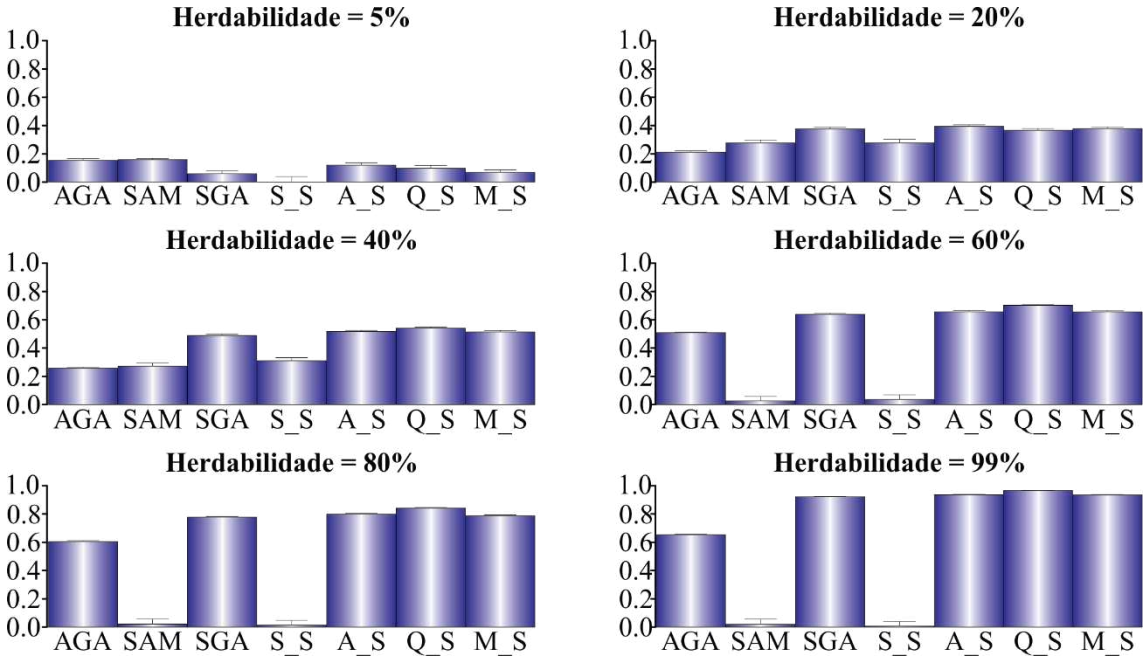

Figura 1. Comparação dos modelos via capacidade preditiva fenotípica para características com diferentes herdabilidades. A capacidade preditiva fenotípica é estimada pela correlação de Pearson entre o valor fenotípico e o valor genético estimado (GEBV) pelos métodos de seleção genômica.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

A capacidade preditiva genotípica estimada pelos modelos SGA, A_S, Q_S e M_S foram superiores aos demais (Figura 2) para todas as herdabilidades avaliadas. O modelo AGA foi inferior aos modelos SAM e S_S para baixa herdabilidade (5, 20 e 40%) e superior para alta herdabilidade (60, 80 e 99%). A capacidade preditiva

genotípica aumentou a medida que a herdabilidade aumentou para todos os modelos avaliados, exceto os modelos SAM e S_S.

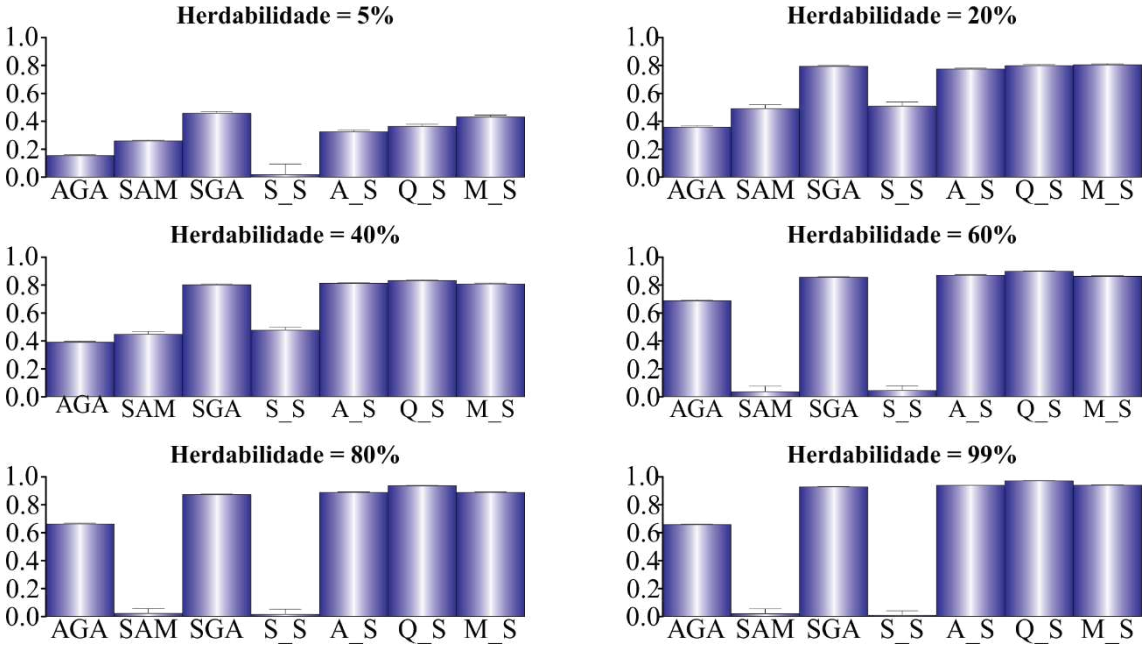

Figura 2. Comparação dos modelos via capacidade preditiva genotípica para características com diferentes herdabilidades. A capacidade preditiva genotípica é estimada pela correlação de Pearson entre o valor genotípico verdadeiro e o valor genético estimado (GEBV) pelos métodos de seleção genômca.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

Seguindo o mesmo padrão de resposta observado para a capacidade preditiva fenotípica, os modelos SGA, A_S, Q_S, e M_S foram superiores para a acurácia

fenotípica para todas as herdabilidades avaliadas, exceto na herdabilidade de 5% onde os modelos SAM e AGA foram superiores (Figura 3). Para herdabilidades de 60, 80 e 99% os modelos SAM e S_S apresentaram a acurácia fenotípica próxima de zero, sendo muito inferior aos demais que apresentaram valores acima de 0.6. O modelo Q_S apresentou os valores mais altos de acurácia fenotípica para herdabilidades igual ou superior a 40%.

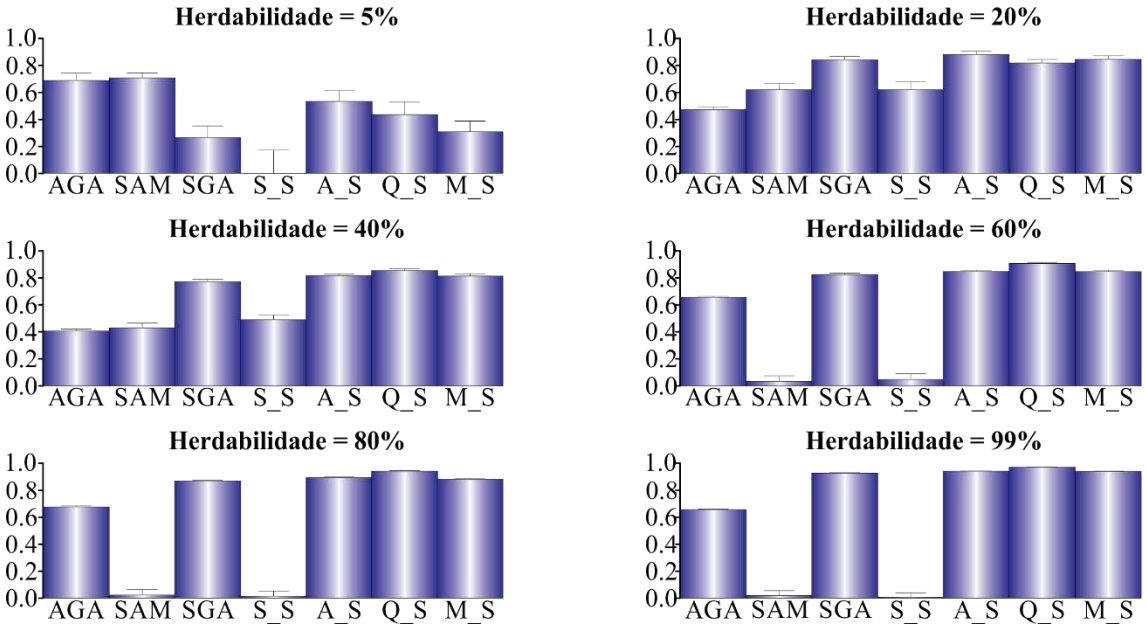

Figura 3. Comparação dos modelos via acurácia fenotípica para características com diferentes herdabilidades. A acurácia fenotípica é estimada dividindo a capacidade preditiva fenotípica pela raiz quadrada da herdabilidade.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

Os modelos SGA, A_S, Q_S e M_S apresentaram acurácia genotípica maior que os demais modelos para herdabilidades igual ou superior a 20% (Figura 4). Para herdabilidade de 5% o modelo SGA apresentou maior acurácia genotípica seguido pelo modelo M_S. O modelo AGA foi superior aos modelos SAM e S_S para as herdabilidades de 60, 80 e 99%.

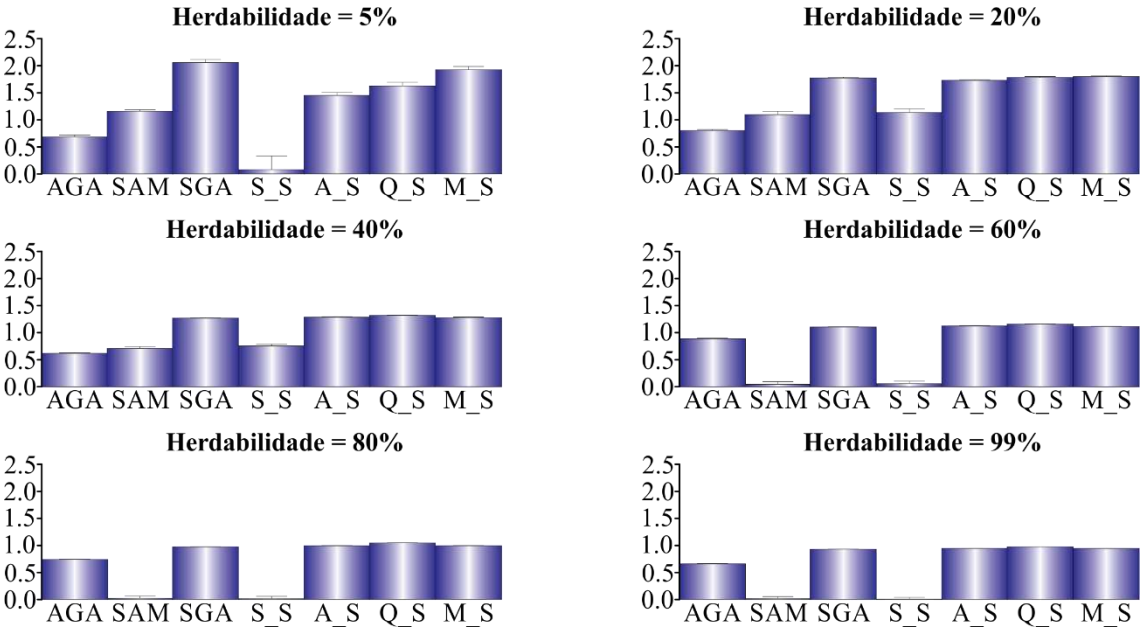

Figura 4. Comparação dos modelos via acurácia genotipica para características com diferentes herdabilidades. A acurácia genotípica é estimada dividindo a capacidade preditiva genotípica pela raiz quadrada da herdabilidade.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

O modelo AGA apresentou maior coincidência de seleção para características de baixa herdabilidade (5 e 20%) (Figura 5). No entanto para características com herdabilidades de 60%, 80% e 99% os modelos SGA, A_S, Q_S e M_S foram superiores. A medida que a herdabilidade da característica aumentou a coincidência de seleção também aumentou para todos os modelos avaliados, exceto os modelos SAM e S_S que se mantiveram praticamente constante (variando a coincidência de seleção de 20 a 40%).

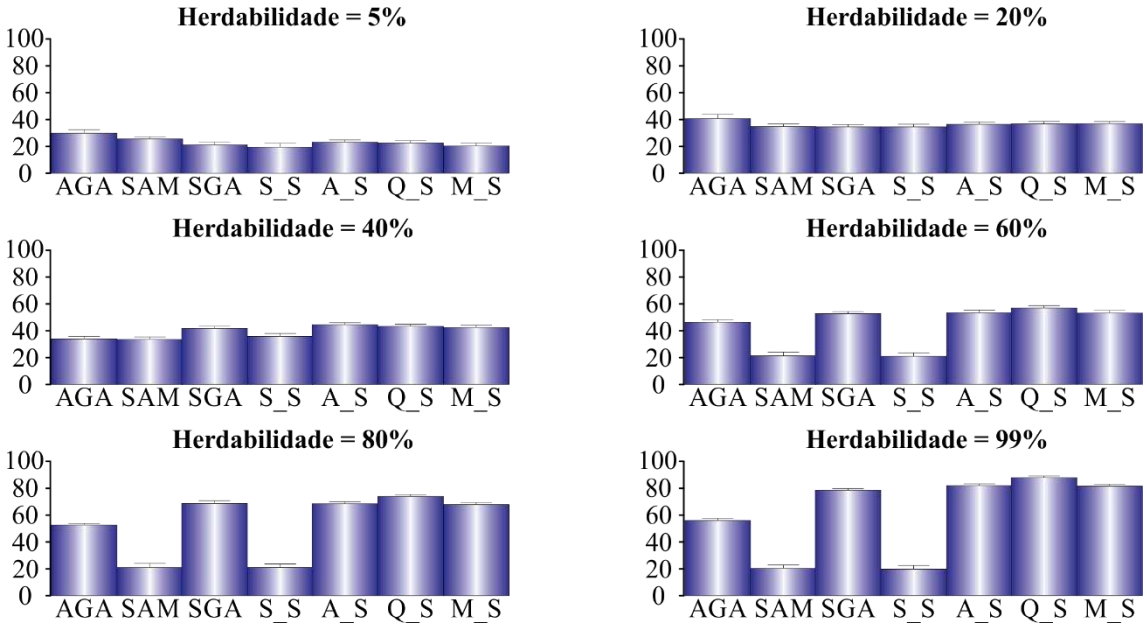

Figura 5. Comparação dos modelos via análise de coincidência de seleção para características com diferentes herdabilidades. A coincidência de seleção foi calculada dividindo o número de indivíduos selecionados com base no valor fenotípico que foram os mesmos selecionados com base no GEBV, pelo número total de indivíduos selecionados.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

O ganho de seleção estimado pelas modelos AGA, SAM e A_S foram maiores que a estimativa pelos demais modelos para herdabilidade de 5% (Figura 6). No entanto a partir da herdabilidade de 20%, os modelos SAM e S_S apresentaram maiores estimativas do ganho de seleção. Porém o ganho de seleção foi

superestimado pelos modelos SAM e S_S para as herdabilidades de 60%, 80% e 99% quando comparado ao ganho de seleção máximo mostrado na tabela 3.

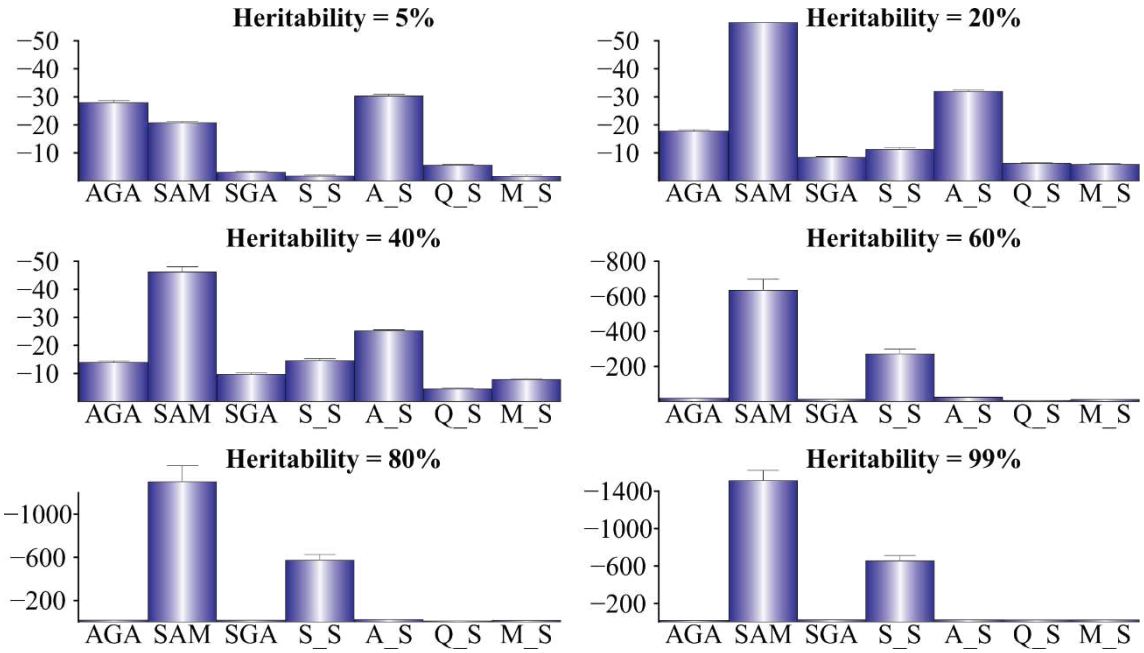

Figura 6. Comparação dos modelos via ganho de seleção para características com diferentes herdabilidades. O ganho de seleção foi estimado pela multiplicação da herdabilidade pelo diferencial de seleção. O diferencial de seleção foi calculado subtraindo a média dos selecionados da média original.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

O ganho de seleção máximo é maior quanto menor é a herdabilidade da característica (Tabela 3). Os ganhos de seleção foram negativos pois trabalhou-se

com características onde o objetivo era reduzir a média na população de melhoramento.

Tabela 3. Ganho de seleção máximo para as características avaliadas com diferentes herdabilidades.

Parâmetros	h^2 = 5%	h^2 = 20%	h^2 = 40%	h^2 = 60%	h^2 = 80%	h^2 = 99%
$\bar{X}_O$	149,35	151,27	147,71	151,05	153,38	156,03
$\bar{X}_S$	30,89	77,88	101,43	115,42	124,41	128,76
$GS_{máx}$	-118,45	-73,40	-46,28	-35,63	-28,97	-27,27

h^2 – herdabilidade, $\bar{X}_O$ – média genotípica da população, $\bar{X}_S$ – média genotípica dos indivíduos selecionados, $GS_{máx}$ – ganho de seleção máximo calculado subtraindo a média genotípica original pela média genotípica dos selecionados.

Foi observado que os modelos AGA e SAM foram mais rápidos que os demais para todas as características avaliadas (Figura 7). Porém a medida que a herdabilidade aumentou o tempo de processamento no modelo SAM também aumentou, tornando-o mais demorado que o modelo AGA. Os demais modelos apresentaram tempo de processamento bem parecidos, exceto o modelo S_S que foi o mais demorado para herdabilidade igual ou superior a 20%.

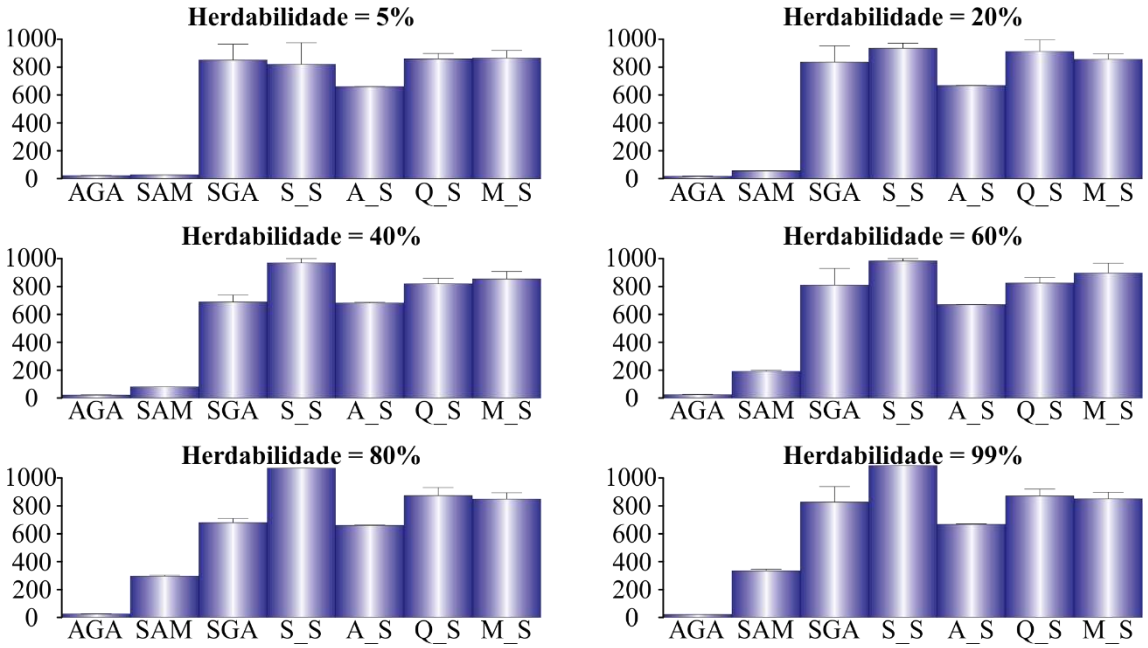

Figura 7. Comparação dos modelos via tempo de processamento em segundos para características com diferentes herdabilidades.

AGA – Todos os SNPs significativos encontrados pelo estudo de associação genômica ampla (EAGA) foram modelados como efeito fixo no modelo de seleção genômica (SG); SAM - Todos os SNPs significativos encontrados pela análise de seleção assistida por marcadores moleculares (SAM) foram modelados como efeito fixo no modelo de SG; SGA – Todos os SNPs foram modelados como efeito aleatório (modelo padrão utilizado na seleção genômica); S_S - Todos os SNPs significativos encontrados pela SAM foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; A_S - Todos os SNPs significativos encontrados pelo EAGA foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; Q_S – Todos os QTLs simulados foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG; M_S – Os QTLs de maior efeito foram modelados como efeito fixo e os demais SNPs como efeito aleatório no modelo de SG.

3.5. DISCUSSÃO

Spindel et al. (2015) sugeriram que o uso das informações obtidas utilizando as análises EAGA dentro dos modelos de SG pode providenciar informações sobre a arquitetura genética da característica em estudo e informações sobre a estrutura da população que esta sendo utilizada no programa de melhoramento. Em seu trabalho Spindel et al. (2015) demonstraram que o uso das marcas significativas pelo EAGA utilizada como efeito fixo nos modelos de SG para produção de grãos, altura de

plantas e florescimento em arroz podem revelar a presença de QTL de maior efeito segregando na população de melhoramento, na qual podem ser estabelecidos como covariáves nos modelos de SG melhorando a acurácia. Porém quando o número de QTLs é maior que 10 esse efeito pode ser contrário, ou seja haver um decréscimo no valor da acurácia (Bernardo, 2014). Esse fato explica o motivo pelo qual a capacidade preditiva (Figura 1 e 2) e a acurácia (Figura 3 e 4) dos modelos SAM e S_S foram inferiores aos demais, pois nestes modelos foram utilizadas mais de 10 marcas como efeito fixo (Tabela 2). Nos modelos AGA e A_S a acurácia e a capacidade preditiva foram igual ou ás vezes superior ao modelo padrão de SG (Figuras 1, 2, 3 e 4). Isto ocorreu pois nestes modelos o número de marcadores utilizados como efeito fixo foi de no máximo 32 marcadores. O mesmo foi observado no modelo M_S onde apenas os dois QTLs de maior efeito (Tabela 1) foram utilizados como efeito fixo no modelo (Figuras 1, 2, 3 e 4).

Tratar um QTL como efeito fixo pode aumentar a acurácia de predição dos modelos de SG, porém se esse QTL for um falso positivo, ele vai na verdade diminuir a acurácia do modelo. No presente estudo foi observado um número de falso positivo muito grande para os modelos SAM e S_S, evidenciando mais um fator que fez com que esses modelos fossem inferiores aos demais. Assim é preferível que os falsos positivo sejam tratados como aleatório, e consequentemente, tenham sua variância bem próxima de zero, que tratado como fixo e influenciar muito na predição do valor genético (Arruda et al., 2016). O modelo Q_S evidencia bem esse fato, pois neste modelo 100 QTLs foram utilizados como efeito fixo, onde todos os QTLs tiveram efeito sobre a característica, ou seja, nenhum falso positivo foi utilizado como efeito fixo, e foi observado valores de capacidade preditiva e acurácia igual ou superior ao modelo tradicional de SG (modelo com todos os SNPs como efeito aleatório) (Figuras 1, 2, 3 e 4).

Bernardo (2014) observou que um único gene tratado como efeito fixo no modelo de SG usando RRBLUP nunca será desvantajoso, exceto em alguns casos onde a variabilidade explicada pelo QTL for inferior a 10%. Desta forma, o uso dos QTLs significativos pelas metodologias de EAGA e SAM são mais influentes na predição quanto maior for o efeito destes QTLs (Spindel et al., 2015). No modelo M_S os dois QTLs utilizados como efeito fixo tinham 7,96% de efeito sobre a característica em estudo (Tabela 1), mostrando que a utilização de apenas dois QTLs de maior efeito sobre a característica pode estimar valores de capacidade de predição e acurácia

similares à modelos utilizando um número maior de marcadores como efeito fixo, como visto nos modelos Q_S, A_S e S_A (Figuras 1, 2, 3 e 4). Este fato é importante, pois para muitas características já são conhecidos alguns QTLs de grande efeito, e desta forma esses QTLs podem ser introduzidos como efeito fixo nos modelos de SG. Por exemplo em soja já foram identificados 18 QTLs para tolerância a alumínio (Bianchi-Hall et al., 2000; Korir et al., 2011; Sharma et al., 2011), um QTL para ferrugem asiática da soja (Kim et al., 2012), e quatro QTLs para tolerância a seca (Carpentieri-Pipolo et al., 2012). Assim, todos esses QTLs identificados anteriormente podem ser utilizados como efeito fixo no modelo de SG.

Foi observado que o grande número de marcadores considerados como efeito fixo no modelo devido aos falso positivos também afetou o ganho de seleção (Figura 6 e Tabela 3), que foi superestimado nos modelos SAM e S_S, enquanto que nos modelos AGA e A_S, onde o número de marcadores considerados como efeito fixo não excedeu 32, o ganho genético não foi superestimado (Figura 6 e Tabela 3) ocorrendo o mesmo para os modelos Q_S e M_S. A coincidência de seleção foi outro parâmetro que foi afetado pelo número de QTLs falso negativos considerados como efeito fixo no modelo. Para os modelos SAM e S_S a coincidência de seleção foi inferior aos demais modelos e foi diminuindo a medida que a herdabilidade da característica aumentou, pois o número de falso positivo também aumentou (Figura 5). Portanto, conhecer a arquitetura genética da característica em estudo pode ser de suma importância na aplicação do modelo correto de seleção genômica, e consequentemente, aumento da acurácia.

A consideração da arquitetura genética da caraterística em estudo por meio das análises EAGA e SAM pode melhorar muito a acurácia dos modelos de SG, visto que os QTLs de maior efeito serão tratados de forma independente dos marcadores de pequeno ou nenhum efeito. Bernardo (2014) utilizando dados simulados verificou em características com herdabilidade de moderada a alta, considerar QTLs com efeito maior que 30% como efeito fixo no modelo de seleção genômica pode aumentar a eficiência relativa baseada no ganho de seleção entre 07 a 21%. No entanto, quando QTLs com efeito menor que 5% foi utilizado no modelo houve decréscimo na eficiência relativa mostrando que marcadores que explicam uma pequena fração da variância genética devem ser tratados como efeito aleatório no modelo de seleção genômica. Como no nosso trabalho, quase todos os genes que foram considerados como efeito fixo explicavam uma fração muito pequena da variância genética (menor que 5%) não

foi verificado aumento significativo na capacidade preditiva, acurácia e ganho com a seleção, e em muitos casos houve decréscimo do valor destas estimativas (Figura 1, 2, 3, 4 e 6). Desta forma, a diferença na arquitetura genética entre as diferentes espécies, assim como a diferença de arquitetura genética entre as características de importância econômica dentro das principais espécies vão influenciar na acurácia dos modelos de SG (Spindel et al., 2015). Este fato é importante para a maioria das principais culturas agrícolas. Em milho, resultados provenientes das análises EAGA tem encontrado inúmeros genes de menor efeito controlando as principais características agrícolas desta espécie (Briggs et al., 2007; McMullen et al., 2009). Enquanto que em arroz, muitos QTLs de grande efeito tem sido encontrados pelo EAGA e SAM para as características de produção de grãos, florescimento, altura de plantas e tolerância a alumínio (Ashikari et al., 2002; Chen et al., 2014; Famoso et al., 2011; Li et al., 1997; Venuprasad et al., 2012).

Para características governadas por menor número de genes de maior efeito os modelos SAM e S_S podem ser superiores aos demais modelos, pois eles conseguem capturar grande parte da variância genética total presente em apenas algumas marcas (QTLs) (Arruda et al., 2016). Spindel et al. (2015) verificou que SAM foi superior a SG para a característica tempo de florescimento em arroz, característica esta governada por poucos genes de grande efeito, enquanto que SG foi superior á SAM para produção de grãos, característica governada por um grande número de genes de pequeno efeito. Arruda et al. (2016) compararam os modelos SAM, SG e S_S em seis características associadas a resistência a fusarium em trigo, e verificaram que a SG apresentou acuracia preditiva (0,4-0,9) superior á SAM (<0,3). Porém quando eles utilizaram os QTLs encontrados na SAM como efeito fixo no modelo de SG (modelo S_S) a acurácia preditiva foi superior ao modelo SG. Portanto podemos verificar que a performance de cada metodologia depende muito da característica e de sua estrutura genética, e assim devemos conhecer profundamente a característica em estudo para escolher o modelo mais apropriado de forma a aumentar a capacidade preditiva e a acurácia.

Uma importante característica do RRBLUP bayesiano, modelo utilizado como padrão neste estudo, é que todas as marcas possuem a mesma variância genética. Porém, nós sabemos que isso é praticamente impossível acontecer nas características de importância agronômica, e desta forma, marcas que exercem efeito sobre a característica podem estar sendo subestimadas e marcas que não tem

nenhum efeito estão sendo superestimadas. Pensando neste contexto, colocar os QTLs de grande efeito indicados pela SAM e EAGA como efeito fixo no modelo de SG, garante que estes QTLs sejam estimados de forma mais realística aumentando assim a capacidade preditiva e a acurácia do modelo (Spindel et al., 2015). Esta subestimação dos QTLs de maior efeito pode afetar a resposta na seleção por vários ciclos no programa de melhoramento (Combs & Bernardo, 2013). Rutkoski et al. (2014) verificaram que a acurácia do modelo de SG aumenta quando os QTLs são tratados como efeito fixo para resistência a ferrugem em trigo. Heffner et al. (2011) compararam os modelos SG e SAM para 13 características agronômicas em trigo e verificaram que a capacidade preditiva fenotípica e genotípica foram 28% maior na SG. Porém alguns trabalhos mostram que dependendo da característica pode não ocorrer aumento da acurácia quando efeitos fixos são colocados no modelo, como verificado por Rutkoski et al. (2012) avaliando a resistência de trigo a fusarium. Além disso os autores verificaram que a acurácia na SAM foi superior a SG. Zhao et al. (2014) compararam SG e SAM para altura de plantas em trigo e verificaram que a capacidade preditiva foi a mesma para ambos.

Portanto, com tudo que foi exposto neste trabalho, verificamos a importância em conhecer a característica em estudo com relação a herdabilidade, estrutura genética, e número de genes controlando esta característica de forma a escolher o modelo mais apropriado para maximizar a capacidade preditiva e a acurácia, e consequentemente, aumentar o ganho com a seleção.

3.6. CONCLUSÃO

Os métodos AGA e SAM não conseguiram identificar de forma acurada os verdadeiros QTLs mostrando que a seleção baseada nestes apenas nos QTLs identificados por estes métodos, pode selecionar indivíduos de baixo valor genético;

A utilização de um modelo de seleção genômica com as marcas significativas encontradas pelo EAGA como efeito fixo e as demais marcas como efeito aleatório é uma boa estratégia para selecionar indivíduos superiores com alta acurácia.

A introdução no modelo de seleção genômica de QTLs, que já foram descritos previamente para a característica em estudo, como efeito fixo permite a seleção de indivíduos superiores de forma mais acurada.

3.7. REFERÊNCIAS BIBLIOGRÁFICAS

ALLARD, R. W., 1999. Principles of plant breeding. John Wiley & Sons, New York.

ARRUDA, M. P., LIPKA, A. E., BROWN, P. J., KRILL, A. M., THURBER, C., BROWN-GUEDIRA, G., DONG, Y., FORESMAN, B. J., KOLB, F. L., Comparing genomic selection and marker-assisted selection for Fusarium head blight resistance in wheat (*Triticum aestivum*). **Molecular Breeding**, v. 36, p. 1-11, 2016.

ASHIKARI, M., SASAKI, A., UEGUCHI-TANAKA, M., ITOH, H., NISHIMURA, A., DATTA, S., ISHIYAMA, K., SAITO, T., KOBAYASHI, M., KHUSH, G. S., Loss-of-function of a Rice Gibberellin Biosynthetic Gene, GA20 oxidase (GA20ox-2), Led to the Rice'Green Revolution'. **Breeding Science**, v. 52, p. 143-150, 2002.

ASHRAF, M., AKRAM, N. A., FOOLAD, M. R., Marker-assisted selection in plant breeding for salinity tolerance. **Plant Salt Tolerance: Methods and Protocols**, v., p. 305-333, 2012.

ASHRAF, M., FOOLAD, M. R., Crop breeding for salt tolerance in the era of molecular markers and marker-assisted selection. **Plant Breeding**, v. 132, p. 10-20, 2013.

ASORO, F. G., NEWELL, M. A., BEAVIS, W. D., SCOTT, M. P., JANNINK, J.-L., Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. **The Plant Genome**, v. 4, p. 132-144, 2011.

BERNARDO, R., Genomewide selection when major genes are known. **Crop Science**, v. 54, p. 68-75, 2014.

BIANCHI-HALL, C. M., CARTER, T. E., BAILEY, M. A., MIAN, M. A. R., RUFTY, T. W., ASHLEY, D. A., BOERMA, H. R., ARELLANO, C., HUSSEY, R. S., PARROTT, W. A., Aluminum tolerance associated with quantitative trait loci derived from soybean PI 416937 in hydroponics. **Crop Science**, v. 40, p. 538-545, 2000.

BRIGGS, W. H., MCMULLEN, M. D., GAUT, B. S., DOEBLEY, J., Linkage mapping of domestication loci in a large maize–teosinte backcross resource. **Genetics**, v. 177, p. 1915-1928, 2007.

CARPENTIERI-PIPOLO, V., PIPOLO, A., ABDEL-HALEEM, H., BOERMA, H., SINCLAIR, T., Identification of QTLs associated with limited leaf hydraulic conductance in soybean. **Euphytica**, v. 186, p. 679-686, 2012.

CHEN, J., LI, X., CHENG, C., WANG, Y., QIN, M., ZHU, H., ZENG, R., FU, X., LIU, Z., ZHANG, G., Characterization of epistatic interaction of QTLs LH8 and EH3 controlling heading date in rice. **Scientific reports**, v. 4, p. 4263, 2014.

COMBS, E., BERNARDO, R., Genomewide selection to introgress semidwarf maize germplasm into US Corn Belt inbreds. **Crop Science**, v. 53, p. 1427-1436, 2013.

CROSSA, J., PÉREZ, P., HICKEY, J., BURGUEÑO, J., ORNELLA, L., CERÓN-ROJAS, J., ZHANG, X., DREISIGACKER, S., BABU, R., LI, Y., Genomic prediction in CIMMYT maize and wheat breeding programs. **Heredity**, v. 112, p. 48-60, 2014.

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

DE OLIVEIRA, E. J., DE RESENDE, M. D. V., DA SILVA SANTOS, V., FERREIRA, C. F., OLIVEIRA, G. A. F., DA SILVA, M. S., DE OLIVEIRA, L. A., AGUILAR-VILDOSO, C. I., Genome-wide selection in cassava. **Euphytica**, v. 187, p. 263-276, 2012.

FAMOSO, A. N., ZHAO, K., CLARK, R. T., TUNG, C.-W., WRIGHT, M. H., BUSTAMANTE, C., KOCHIAN, L. V., MCCOUCH, S. R., Genetic architecture of aluminum tolerance in rice (*Oryza sativa*) determined through genome-wide association analysis and QTL mapping. **PLoS Genet**, v. 7, p. e1002221, 2011.

GOUY, M., ROUSSELLE, Y., BASTIANELLI, D., LECOMTE, P., BONNAL, L., ROQUES, D., EFILE, J.-C., ROCHER, S., DAUGROIS, J., TOUBI, L., Experimental assessment of the accuracy of genomic selection in sugarcane. **Theoretical and applied genetics**, v. 126, p. 2575-2586, 2013.

GREGORIO, G. B., ISLAM, M. R., VERGARA, G. V., THIRUMENI, S., Recent advances in rice science to design salinity and other abiotic stress tolerant rice varieties. **SABRAO J. Breed. Genet**, v. 45, p. 31-41, 2013.

GUO, Z., TUCKER, D. M., LU, J., KISHORE, V., GAY, G., Evaluation of genome-wide selection efficiency in maize nested association mapping populations. **Theoretical and Applied Genetics**, v. 124, p. 261-275, 2012.

HAYR, M., 2013. Increasing the accuracy of genomic estimated breeding values of milk traits in New Zealand dairy cattle using DGAT genotypes, Plant and Animal Genome XXI Conference. Plant and Animal Genome.

HEFFNER, E. L., JANNINK, J.-L., IWATA, H., SOUZA, E., SORRELLS, M. E., Genomic selection accuracy for grain quality traits in biparental wheat populations. **Crop Science**, v. 51, p. 2597-2606, 2011.

HUANG, X., WEI, X., SANG, T., ZHAO, Q., FENG, Q., ZHAO, Y., LI, C., ZHU, C., LU, T., ZHANG, Z., Genome-wide association studies of 14 agronomic traits in rice landraces. **Nature genetics**, v. 42, p. 961-967, 2010.

HWANG, E.-Y., SONG, Q., JIA, G., SPECHT, J. E., HYTEN, D. L., COSTA, J., CREGAN, P. B., A genome-wide association study of seed protein and oil content in soybean. **BMC genomics**, v. 15, p. 1-12, 2014.

JENA, K. K., MACKILL, D. J., Molecular markers and their use in marker-assisted selection in rice. **Crop Science**, v. 48, p. 1266-1276, 2008.

KIM, K.-S., UNFRIED, J. R., HYTEN, D. L., FREDERICK, R. D., HARTMAN, G. L., NELSON, R. L., SONG, Q., DIERS, B. W., Molecular mapping of soybean rust resistance in soybean accession PI 561356 and SNP haplotype analysis of the Rpp1 region in diverse germplasm. **Theoretical and Applied Genetics**, v. 125, p. 1339-1352, 2012.

KORIR, P. C., QI, B., WANG, Y., ZHAO, T., YU, D., CHEN, S., GAI, J., A study on relative importance of additive, epistasis and unmapped QTL for aluminium tolerance at seedling stage in soybean. **Plant breeding**, v. 130, p. 551-562, 2011.

LI, Z., PINSON, S. R. M., PARK, W. D., PATERSON, A. H., STANSEL, J. W., Epistasis for three grain yield components in rice (*Oryxa sativa* L.). **Genetics**, v. 145, p. 453-465, 1997.

LORENZ, A. J., CHAO, S., ASORO, F. G., HEFFNER, E. L., HAYASHI, T., IWATA, H., SMITH, K. P., SORRELLS, M. E., JANNINK, J.-L., 2 Genomic Selection in Plant Breeding: Knowledge and Prospects. **Advances in agronomy**, v. 110, p. 77, 2011.

LORENZ, A. J., SMITH, K. P., JANNINK, J.-L., Potential and optimization of genomic selection for Fusarium head blight resistance in six-row barley. **Crop science**, v. 52, p. 1609-1621, 2012.

LORENZANA, R. E., BERNARDO, R., Accuracy of genotypic value predictions for marker-based selection in biparental plant populations. **Theoretical and applied genetics**, v. 120, p. 151-161, 2009.

LY, D., HAMBLIN, M., RABBI, I., MELAKU, G., BAKARE, M., GAUCH, H. G., OKECHUKWU, R., DIXON, A. G., KULAKOW, P., JANNINK, J.-L., Relatedness and genotype× environment interaction affect prediction accuracies in genomic selection: a study in cassava. **Crop Science**, v. 53, p. 1312-1325, 2013.

MCMULLEN, M. D., KRESOVICH, S., VILLEDA, H. S., BRADBURY, P., LI, H., SUN, Q., FLINT-GARCIA, S., THORNSBERRY, J., ACHARYA, C., BOTTOMS, C., Genetic properties of the maize nested association mapping population. **Science**, v. 325, p. 737-740, 2009.

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

MORRIS, G. P., RAMU, P., DESHPANDE, S. P., HASH, C. T., SHAH, T., UPADHYAYA, H. D., RIERA-LIZARAZU, O., BROWN, P. J., ACHARYA, C. B.,

MITCHELL, S. E., Population genomic and genome-wide association studies of agroclimatic traits in sorghum. **Proceedings of the National Academy of Sciences**, v. 110, p. 453-458, 2013.

ORNELLA, L., SINGH, S., PEREZ, P., BURGUEÑO, J., SINGH, R., TAPIA, E., BHAVANI, S., DREISIGACKER, S., BRAUN, H.-J., MATHEWS, K., Genomic prediction of genetic values for resistance to wheat rusts. **The Plant Genome**, v. 5, p. 136-148, 2012.

PALAISA, K. A., MORGANTE, M., WILLIAMS, M., RAFALSKI, A., Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. **The Plant Cell**, v. 15, p. 1795-1806, 2003.

PÉREZ, P., DE LOS CAMPOS, G., 2012. BGLR: A Statistical Package for Whole-Genome Regression.

PRITCHARD, J. K., STEPHENS, M., ROSENBERG, N. A., DONNELLY, P., Association mapping in structured populations. **The American Journal of Human Genetics**, v. 67, p. 170-181, 2000.

R CORE TEAM, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

RUTKOSKI, J., BENSON, J., JIA, Y., BROWN-GUEDIRA, G., JANNINK, J.-L., SORRELLS, M., Evaluation of genomic prediction methods for Fusarium head blight resistance in wheat. **The Plant Genome**, v. 5, p. 51-61, 2012.

RUTKOSKI, J. E., POLAND, J. A., SINGH, R. P., HUERTA-ESPINO, J., BHAVANI, S., BARBIER, H., ROUSE, M. N., JANNINK, J.-L., SORRELLS, M. E., Genomic selection for quantitative adult plant stem rust resistance in wheat. **The plant genome**, v. 7, p., 2014.

SHARMA, A. D., SHARMA, H., LIGHTFOOT, D. A., The genetic control of tolerance to aluminum toxicity in the ‘Essex’by ‘Forrest’recombinant inbred line population. **Theoretical and applied genetics**, v. 122, p. 687-694, 2011.

SPINDEL, J., BEGUM, H., AKDEMIR, D., VIRK, P., COLLARD, B., REDOÑA, E., ATLIN, G., JANNINK, J.-L., MCCOUCH, S. R., Genomic Selection and Association Mapping in rice (*Oryza sativa)*: Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. **PLoS Genet**, v. 11, p. e1004982, 2015.

STEELE, K. A., PRICE, A. H., WITCOMBE, J. R., SHRESTHA, R., SINGH, B. N., GIBBONS, J. M., VIRK, D. S., QTLs associated with root traits increase yield in upland

rice when transferred through marker-assisted selection. **Theoretical and applied genetics**, v. 126, p. 101-108, 2013.

SZAJKO, K., STRZELCZYK-ŻYTA, D., MARCZEWSKI, W., Ny-1 and Ny-2 genes conferring hypersensitive response to potato virus Y (PVY) in cultivated potatoes: mapping and marker-assisted selection validation for PVY resistance in potato breeding. **Molecular Breeding**, v. 34, p. 267-271, 2014.

TIAN, F., BRADBURY, P. J., BROWN, P. J., HUNG, H., SUN, Q., FLINT-GARCIA, S., ROCHEFORD, T. R., MCMULLEN, M. D., HOLLAND, J. B., BUCKLER, E. S., Genome-wide association study of leaf architecture in the maize nested association mapping population. **Nature genetics**, v. 43, p. 159-162, 2011.

VENUPRASAD, R., BOOL, M. E., QUIATCHON, L., CRUZ, M. T. S., AMANTE, M., ATLIN, G. N., A large-effect QTL for rice grain yield under upland drought stress on chromosome 1. **Molecular Breeding**, v. 30, p. 535-547, 2012.

WÜRSCHUM, T., REIF, J. C., KRAFT, T., JANSSEN, G., ZHAO, Y., Genomic selection in sugar beet breeding populations. **Bmc Genetics**, v. 14, p. 85, 2013.

YANIV, E., RAATS, D., RONIN, Y., KOROL, A. B., GRAMA, A., BARIANA, H., DUBCOVSKY, J., SCHULMAN, A. H., FAHIMA, T., Evaluation of marker-assisted selection for the stripe rust resistance gene Yr15, introgressed from wild emmer wheat. **Molecular Breeding**, v. 35, p. 1-12, 2015.

ZHAO, Y., METTE, M. F., GOWDA, M., LONGIN, C. F. H., REIF, J. C., Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. **Heredity**, v. 112, p. 638-645, 2014.

hybrid wheat. **Heredity**, v. 112, p. 638-645, 2014.

4. CAPITULO 2

FATORES QUE AFETAM A PREDIÇÃO GENÔMICA E RELAÇÃO ENTRE ACURÁCIA FENOTÍPICA E GENOTÍPICA

FATORES QUE AFETAM A PREDIÇÃO GENÔMICA E RELAÇÃO ENTRE ACURÁCIA FENOTÍPICA E GENOTÍPICA

4.1. RESUMO

Os objetivos deste trabalho foram, a partir de uma população F_2, comparar os métodos de seleção genômica, verificar como a herdabilidade e o número de QTLs que controlam a característica podem influenciar na predição do valor genético, e estabelecer equação de predição da correlação genética em função da correlação fenotípica. Uma população F_2 com 1.000 indivíduos foi simulada em diferentes cenários. Em cada cenário variou-se a herdabilidade (5, 20, 40, 60, 80 e 99%) e o número de QTL (60, 120, 180 e 240). Quatro métodos de seleção genômica foram utilizados nas análises (RRBLUP, GBLUP, Bayes B, e RKHS). As capacidades preditivas fenotípicas e genotípicas foram calculadas para cada método, e foi utilizado o teste de Tukey para a comparação das médias. O efeito da herdabilidade e do número de QTLs controlando a característica foi avaliado pela análise de regressão. Foi observado diferença entre os métodos pelo teste de Tukey em que os métodos Bayes B e RRBLUP foram superiores aos demais em quase todos os cenários. A herdabilidade apresentou uma relação linear positiva com a capacidade preditiva fenotípica e quadrática com a capacidade preditiva genotípica. O número de QTL que controlam a característica não apresentou nenhum tipo de relação com a capacidade preditiva fenotípica e genotípica.

Palavras-chave: Seleção genômica, herdabilidade, QTL, bayesiana, acurácia

4.2. INTRODUÇÃO

A seleção de genótipos superiores, na maioria dos programas de melhoramento de plantas e animais até cerca de 30 anos atrás, era realizada baseando-se apenas na seleção visual dos indivíduos. Com o advento dos marcadores moleculares foi possível incorporar estas informações para melhorar a acurácia de predição e de seleção (Stuber et al., 1982; Tanksley et al., 1989). A primeira metodologia baseada em marcadores utilizada foi a seleção assistida por marcadores moleculares (SAM) (Xu & Crouch, 2008). No entanto essa metodologia era útil apenas para características com QTLs de grande efeito, sendo ineficiente para características governadas por genes de menor efeito (Jena & Mackill, 2008; Zhong et al., 2006).

Desta forma era necessário apresentar metodologia capaz de identificar esses genes de pequeno efeito e predizer o valor genético de cada indivíduo baseado nestes genes. Com a evolução dos marcadores moleculares e a introdução de marcadores abundantes no genoma como os SNPs e os DarTs foi possível estabelecer novos modelos estatísticos para capturar a influência destes genes de pequeno efeito conhecidos como modelos de seleção genômica (Meuwissen et al., 2001). Esses modelos baseiam se na utilização do efeito de todas as marcas disponíveis para estimar o valor genético genômico (VGG) do indivíduo (Lorenz et al., 2011). Assim foi possível aumentar, de forma considerável, a acurácia de predição e/ou diminuir o tempo por ciclo de seleção, e, consequentemente, acelerar o ganho com a seleção e aumentar o lucro no programa de melhoramento (Heffner et al., 2009).

A partir do trabalho pioneiro de Meuwissen et al. (2001) inúmeros métodos foram desenvolvidos tentando capturar da melhor forma possível a variância genética e reduzir a variância residual de forma a aumentar a acurácia de predição (De Los Campos et al., 2010; De Los Campos et al., 2009; Jannink et al., 2010).

Porém, além da influência dos diversos métodos estatísticos de seleção genômica na predição da acurácia, muitos outros fatores influenciam a acurácia, entre eles podemos destacar a herdabilidade da característica e o número de genes que controlam esta característica.

A herdabilidade é um fator importante para predição do valor genético e os resultados encontrados para este fator tem sido controverso. Heffner et al. (2011) avaliaram duas característica em trigo e verificaram que característica com menor herdabilidade apresentou maior acurácia. No entanto Ornella et al. (2012) verificaram em trigo que características de alta herdabilidade apresentaram maior valor de acurácia quando comparada com uma característica de herdabilidade menor. Provavelmente, esta incoerência nos resultados ocorre devido a outros fatores estarem atuando conjuntamente com a herdabilidade de forma a tornar mais difícil o entendimento das características e assim a escolha do melhor método a ser utilizado nas análises.

A acurácia nos métodos de seleção genômica parece estar inversamente relacionada ao número de QTL (Zhong et al., 2009). A acurácia estimada pelos métodos bayesianos é maior para características governadas por poucos genes de maior efeito. Já para características governadas por muitos genes de menor efeito os

modelos baseados no BLUP tem apresentado melhor performance (Daetwyler et al., 2010; Meuwissen et al., 2001; Zhong et al., 2009).

Apesar de existirem na literatura inúmeros trabalhos comparando os métodos de seleção genômica (Bhering et al., 2015; Daetwyler et al., 2010; Heslot et al., 2012; Jannink et al., 2010), poucos trabalhos levam em consideração a herdabilidade (Heffner et al., 2009) e o número de QTLs controlando a característica (Desta & Ortiz, 2014), e nenhum deles levou em consideração todos estes fatores simultaneamente.

Para a melhor avaliação os métodos de seleção genômica é necessário que pelo menos um marcador esteja em desequilíbrio de ligação com o QTL (Heffner et al., 2009). Pensando neste contexto, a população F_2 é ideal para avaliação das metodologias de seleção genômica, pois é nesta fase que a população apresenta maior variabilidade genética, e consequentemente, maior desequilíbrio de ligação (Falconer & Mackay, 1996).

Portanto, os objetivos deste trabalho, utilizando uma população F_2 como população base, foram comparar diversos métodos estatísticos utilizados para seleção genômica, verificar como a herdabilidade e o número de QTLs que controlam a característica podem influenciar na predição do valor genético, e estabelecer uma equação de predição da correlação genética em função da correlação fenotípica.

4.3. MATERIAL E MÉTODOS

4.3.1. Simulação dos dados

A simulação da população F_2 foi realizada utilizando o módulo de simulação do aplicativo computacional GENES (Cruz, 2013), que permitiu gerar informações sobre o genoma, genótipos dos genitores, populações de cruzamentos controlados e dados de características quantitativas.

4.3.1.1. Simulação do genoma

Foi simulado um genoma constituído de 15 Grupos de ligação, similar ao de uma espécie diploide 2n=2x=30. Cada grupo de ligação foi simulado com 200 cM, com 200 marcas por grupo de ligação, espaçadas de forma equidistante (1 cM), totalizando 3.000 marcas. As marcas foram assumidas como codominantes e bialélicas.

4.3.1.2. Simulação dos genitores

Pais homozigotos contrastantes foram simulados, ou seja, o pai 1 foi codificado como dominante (2), e o pai 2 foi codificado como recessivo (0) para todas as marcas existentes. Desta forma o cruzamento entre pai 1 e pai 2 gerou a população F_1 com todos os genes em heterozigose.

4.3.1.3. Simulação da populaçõe de mapeamento

População F_2 compostas 1.000 indivíduos foram geradas a partir da autofecundação dos indivíduos da população F_1. Neste processo cada indivíduo da população F_1 produziu 5.000 gametas e quando 2 destes gametas se encontraram ao acaso, o primeiro indivíduo da população F_2 foi gerado. Este processo se repetiu até formação da todos os indivíduos em cada população.

Para a formação de cada gameta seguiu-se o seguinte critério: O alelo do primeiro gene foi escolhido de forma aleatória (A ou a) para começar a constituir o gameta; O alelo do segundo gene foi escolhido levando em consideração a distância para o primeiro gene, ou seja, as frequências de crossing over foram contabilizados e a escolha de qual alelo (B ou b) que constituiria o gameta foi baseado nas probabilidades; este processo se repetiu até chegar no último gene. Foi considerado interferência nula. Desta forma foi garantido que todos os gametas formados eram diferentes devido a escolha aleatória do alelo no primeiro gene e as probabilidades condicionadas a cada alelo para os próximos genes.

A população F_2 simulada foi codificada com 0, 1 e 2, sendo que 0 correspondeu aos indivíduos homozigotos recessivos, 1 aos indivíduos heterozigotos e 2 aos indivíduos homozigotos para um determinado loco.

Com o objetivo de verificar como o número de QTLs controlando a característica poderiam influenciar a predição pelos métodos de seleção genômica, características governadas por diferentes números de QTLs (60, 120, 180 e 240) foram simuladas.

4.3.1.4. Simulação das características quantitativas

Primeiramente foi atribuído a distribuição binomial para a importância de cada QTL seguindo a equação abaixo:

$$Importância\ do\ QTL = \frac{n!}{k!\,(n-k\,)!} p^k q^{(n-k)}$$

Onde:

p=q=0.5;

N=n-1, sendo n o número de QTLs.

$$\sum_{i=1}^{n} Importância\ do\ QTL_i = 1$$

A expressão de cada QTL foi definida por: AA=μ + a; Aa=μ + d; aa=μ – a. Como o valor de d foi definido como nulo o grau médio de dominância (d/a) foi igual a zero para todos os locos.

O valor genotípico (VG) de cada indivíduo foi definido pela equação:

$$VG = \sum_{i=1}^{n} (importância\ do\ QTL_i * expressão\ do\ QTL_i)$$

Onde:

VG – valor genotípico.

O efeito ambiental foi definido como um vetor independente do valor genotípico, e foi estimado seguindo uma distribuição N(0,σ^2). O valor de σ^2 é calculado a partir da herdabilidade da característica e o valor da variância genética (σ_g^2). O valor da herdabilidade foi definido previamente. Para este trabalho foram simuladas características com herdabilidade de 5%, 20%, 40%, 60%, 80% e 99%. A σ_g^2 é calculada como sendo a variância do valor genotípico dos indivíduos da população F_2.

Sendo assim o valor fenotípico é calculado da seguinte forma:

$$VF = u + VG + VA$$

Onde:

u – média definida pelo usuário. No presente trabalho foi definido que u = 100.

VF – valor fenotípico;

VA – valor ambiental.

4.3.2. Análise dos dados

Após a geração da população, seguiram-se as etapas do processo de mapeamento, iniciando pela análise de segregação de locos individuais. Foram aplicados testes de qui-quadrado (χ^2) para verificar se as marcas geradas segregavam de acordo com uma população F_2. Verificou-se ainda se todos os Grupos de Ligação foram restaurados, com tamanho, distância e ordem dos marcadores, podendo concluir assim que se tratava de uma população F_2 com as propriedades de simulação desejadas.

Quatro métodos de seleção genômica largamente utilizados no melhoramento de plantas e animais foram testados neste trabalho.

RR-BLUP e Bayes B foram descritos por Meuwissen et al. (2001). RR-BLUP assume que cada marcador possui uma variância igual a V_G/M, onde V_G é a variância genética e M é o número de marcas.

No método Bayes B, a priori da proporção de marcas associadas com a variância fenotípica igual a zero, π, assumiu uma distribuição de qui quadrado invertida. Os outros hiperparâmetros para os componentes de variâncias atribuídas a cada marca no método Bayes B foram descritos por Meuwissen et al. (2001).

No reproducing kernel Hilbert spaces regression (RKHS), valores genéticos são estimados pelo processo gaussiano. Quando marcas e pedigree estão disponíveis o valor genético pode ser estimado como a soma dos dois componentes:

$$g_i = u_i + f_i$$

onde u_i é a média geral e f_i é um processo gaussiano com a função de covariância proporcional para avaliação dos reproducing kernel, $K(x_i, x_j)$, avaliados nos marcadores; aqui x_i e x_j são vetores contendo as marcas codificadas dos indivíduos i_{th} e j_{th} respectivamente. Todos os parâmetros das prioris são descritos por De Los Campos et al. (2010).

Para comparação dos métodos de seleção genômica foi calculado a capacidade preditiva fenotípica definida como a correlação de Pearson entre o valor fenotípico e valor genético estimado (GEBV), e a capacidade preditiva genotípica definido como a correlação de Pearson entre o valor genético verdadeiro e o GEBV. Os métodos foram comparados utilizando teste de Tukey a 5% de probabilidade.

Com o objetivo de verificar a influência da herdabilidade e do número de QTLs que controlam a característica na predição da acurácia pelos métodos de seleção genômica foi utilizado análise de regressão avaliando cada método nas 6 herdabilidades (5, 20, 40, 60, 80 e 99%) e nos diferentes números de QTLs simulados (60, 120, 180 e 240). Os modelos de regressão linear, quadrática e cúbica foram utilizados.

Com o objetivo de predizer a correlação genética (correlação de Pearson entre o valor genético verdadeiro e o valor genético genômico) a partir da correlação fenotípica (correlação de Pearson entre o valor fenotípico e o valor genético genômico), os modelos de regressão linear, quadrática e cúbica foram testados, regredindo-se a correlação genética em função da correlação fenotípica. As

regressões foram realizadas para cada método de seleção genômica, em todos os cenários (6 herdabilidades e 4 números de QTLs).

4.3.3. Informações de software e hardware

Todas as análises foram executadas no programa estatístico R (Team, 2014). RR-BLUP e G-BLUP foram analisados utilizando as funções mixed.solve e kin.BLUP (parte do pacote rrBLUP (Endelman, 2011)). BayesB, e RKHS foram analisados usando a função BGLR (parte do pacote BGLR (Pérez & de los Campos, 2012)),

Um total de 20,000 burn-ins e 100,000 iterações da Markov-Chain Monte Carlo (MCMC) foram usadas nas análises bayesianas. A convergência dos modelos bayesianos foi analisada por meio dos parâmetros de variância pelo trace plot.

Dois computadores de alta performance foram utilizados para rodar as análises de seleção genômica. As especificações deles foram as seguintes: Intel Xeon com processador E5-26 12° geração 3.30 GHz, com memória RAM de 64 e 96 GB respectivamente, e hard drive de 1024 GB.

4.4. RESULTADO

4.4.1. Comparação entre os métodos de seleção genômica

Foi verificado diferenças significativas entre os métodos de seleção genômica para todas as herdabilidades avaliadas independentemente do número de QTL para a capacidade preditiva fenotípica e genotipica (Tabelas 4 e 5).

Os métodos GBLUP e RKHS mostraram-se inferiores em quase todos os cenários avaliados tanto para a capacidade preditiva fenotipcia quanto para a capacidade preditiva genotipica (Tabelas 4 e 5).

Os métodos RRBLUP e Bayes B foram significativamente superiores aos demais métodos em quase todos os cenários avaliados (Tabelas 4 e 5) para a capacidade preditiva fenotípica e genotipica. Para herdabilidades acima de 40% o Bayes B foi superior ao RRBLUP.

Tabela 4. Estimativa da capacidade preditiva fenotípica (correlação de Pearson entre o valor fenotípico e o valor genético estimado) para as diferentes herdabilidade (5, 20, 40, 60, 80 e 99%) e números de QTL (60, 120, 180 e 240).

Métodos	h^2 = 5%	h^2 = 20%	h^2 = 40%	h^2 = 60%	h^2 = 80%	h^2 = 99%
			Número de QTLs = 60			
RRBLUP	0.04[a]	0.37[a]	0.56[b]	0.71[b]	0.84[b]	0.96[b]
Bayes B	0.05[a]	0.37[a]	0.58[a]	0.73[a]	0.86[a]	0.99[a]
RKHS	0.05[a]	0.35[b]	0.55[c]	0.71[b]	0.83[c]	0.95[c]
GBLUP	0.02[b]	0.35[b]	0.55[c]	0.71[b]	0.82[c]	0.95[c]
			Número de QTLs = 120			
RRBLUP	0.16[b]	0.34[a]	0.56[a]	0.69[b]	0.80[b]	0.96[b]
Bayes B	0.16[b]	0.35[a]	0.57[a]	0.70[a]	0.85[a]	0.99[a]
RKHS	0.16[b]	0.34[b]	0.56[a]	0.69[b]	0.78[c]	0.95[c]
GBLUP	0.17[a]	0.34[b]	0.56[a]	0.69[b]	0.79[c]	0.96[b]
			Número ae QTLs = 180			
RRBLUP	0.19[a]	0.39[a]	0.52[a]	0.69[b]	0.84[b]	0.97[b]
Bayes B	0.19[a]	0.39[a]	0.52[a]	0.70[a]	0.86[a]	0.99[a]
RKHS	0.18[b]	0.38[b]	0.51[b]	0.69[b]	0.84[b]	0.96[c]
GBLUP	0.18[b]	0.38[b]	0.51[b]	0.69[b]	0.84[b]	0.96[c]
			Número de QTLs = 240			
RRBLUP	0.14[a]	0.32[a]	0.54[b]	0.69[a]	0.79[c]	0.97[b]
Bayes B	0.14[a]	0.32[a]	0.55[a]	0.69[a]	0.86[a]	0.99[a]
RKHS	0.13[b]	0.31[b]	0.53[c]	0.68[b]	0.85[b]	0.96[c]
GBLUP	0.13[b]	0.31[b]	0.53[c]	0.68[b]	0.85[b]	0.97[b]

Também foi observado que a medida que a herdabilidade aumenta o valor da capacidade preditiva fenotípica e genotipica também aumentou independentemente do método utilizado ou do número de QTLs governando a característica (Tabelas 4 e 5).

Tabela 5. Estimativa da capacidade preditiva genotípica (correlação de Pearson entre o valor genotípico verdadeiro e o valor genético estimado) para as diferentes herdabilidade (5, 20, 40, 60, 80 e 99%) e números de QTL (60, 120, 180 e 240).

Métodos	h^2 = 5%	h^2 = 20%	h^2 = 40%	h^2 = 60%	h^2 = 80%	h^2 = 99%
			Número de QTLs = 60			
RRBLUP	0.56[a]	0.77[b]	0.89[b]	0.91[b]	0.94[b]	0.97[b]
Bayes B	0.55[b]	0.78[a]	0.92[a]	0.93[a]	0.97[a]	0,99[a]
RKHS	0.51[d]	0.76[c]	0.88[c]	0.91[c]	0.93[c]	0.95[c]
GBLUP	0.53[c]	0.76[c]	0.88[c]	0.91[c]	0.93[c]	0.96[c]
			Número de QTLs = 120			
RRBLUP	0.64[a]	0.75[a]	0.86[a]	0.91[b]	0.90[b]	0.97[b]
Bayes B	0.64[a]	0.75[a]	0.86[a]	0.92[a]	0.96[a]	0,99[a]
RKHS	0.60[c]	0.74[b]	0.85[b]	0.91[b]	0.89[c]	0.96[c]
GBLUP	0.63[b]	0.74[b]	0.85[b]	0.91[b]	0.89[c]	0.96[c]
			Número de QTLs = 180			
RRBLUP	0.64[b]	0.80[a]	0.85[b]	0.89[b]	0.94[b]	0.97[b]
Bayes B	0.64[b]	0.80[a]	0.86[a]	0.91[a]	0.96[a]	0,99[a]
RKHS	0.63[c]	0.79[b]	0.84[c]	0.88[c]	0.94[b]	0.96[c]
GBLUP	0.65[a]	0.79[b]	0.84[c]	0.88[c]	0.94[b]	0.96[c]
			Número de QTLs = 240			
RRBLUP	0.56[a]	0.77[a]	0.86[b]	0.90[b]	0.95[b]	0.98[b]
Bayes B	0.56[a]	0.77[a]	0.87[a]	0.91[a]	0.96[a]	0,99[a]
RKHS	0.52[b]	0.75[b]	0.85[c]	0.89[c]	0.94[c]	0.97[c]
GBLUP	0.52[b]	0.74[c]	0.85[c]	0.89[c]	0.94[c]	0.97[c]

Apesar de ter sido observados diferenças entre os métodos para as capacidades fenotípicas e genotípicas, essa diferença em termos de ganhos em porcentagem foi muito pequena, tornando os métodos RRBLUP e GBLUP mais promissores na prática, pois sua performance computacional (tempo de processamento) é muito superior aos métodos bayesianos.

4.4.2. Influência da herdabilidade na predição do valor genético

Com o objetivo de avaliar a influência da herdabilidade na predição do valor genético pelos métodos de seleção genômica foram simulados diferentes cenários variando a herdabilidade (5, 20, 40, 60, 80 e 99%).

Foi observado uma relação linear positiva entre a capacidade preditiva fenotípica e herdabilidade em todos os cenários (número diferente de QTL controlando a característica) (Figura 8). Para todos os métodos de seleção genômica utilizados o valor de R^2 da regressão linear foi superior a 0.94.

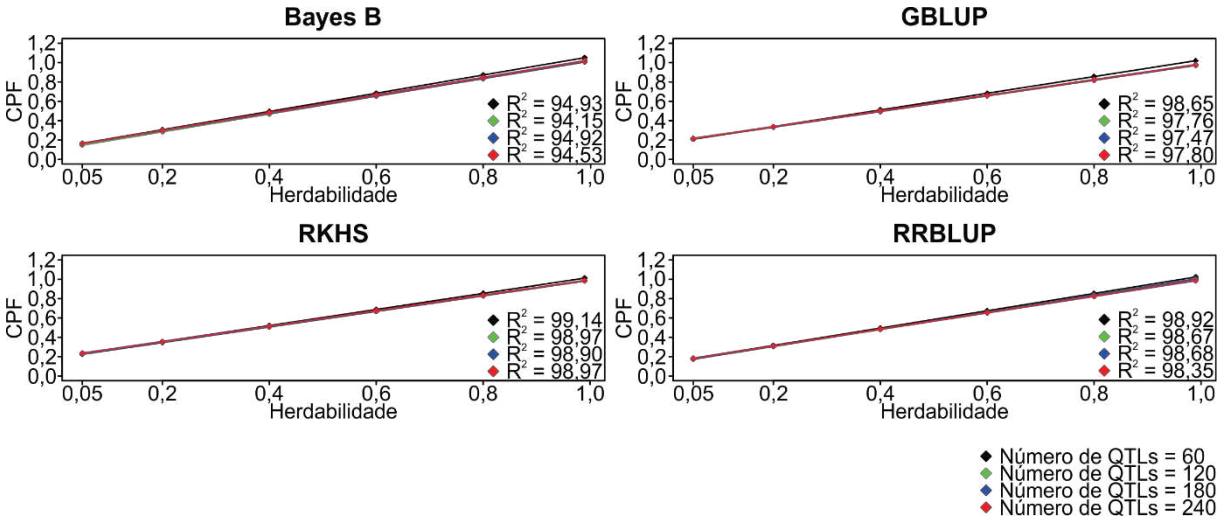

Figura 8. Capacidade preditiva fenotípica (CPF) (correlação de Pearson entre o valor fenotípico e o valor genético estimado) predita pelos quatro métodos de seleção genômica em função da herdabilidade, com diferentes números de QTLs controlando a característica (60, 120, 180 e 240). Os pontos são os valores reais estimados pelos métodos de seleção genômica e as linhas são as regressões para cada método.

Apesar da capacidade preditiva fenotípica ter apresentado resposta linear com a herdabilidade, a capacidade preditiva genotípica teve uma relação quadrática com a herdabilidade em todos os cenários (número diferente de QTLs controlando a característica) avaliados, onde um platô foi alcançado quando a herdabilidade da característica atingiu 60% (Figura 9). Para todos os métodos de seleção genômica utilizados o valor de R^2 da regressão quadrática foi superior a 0.93.

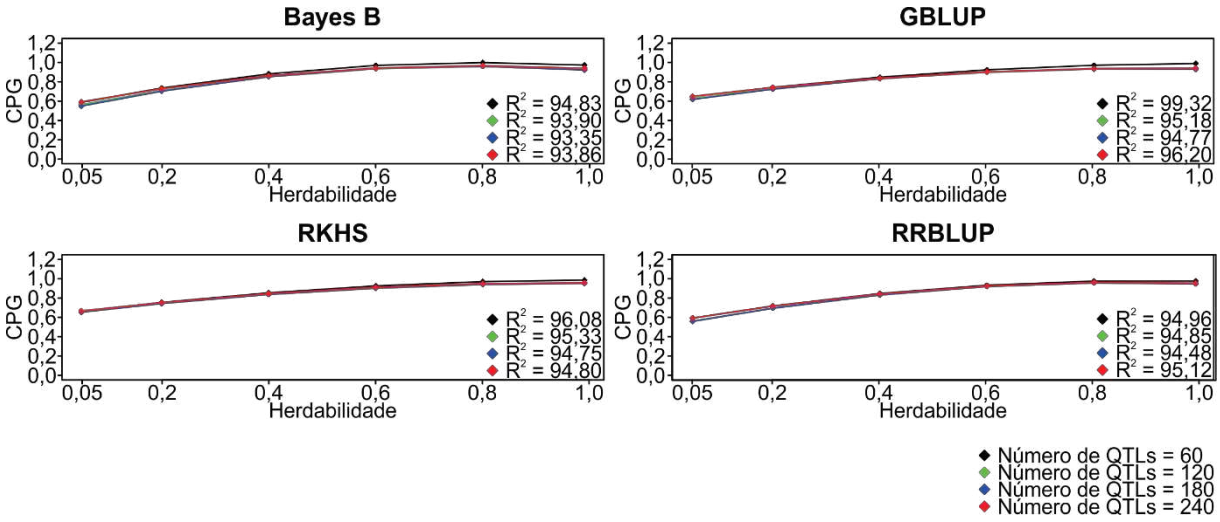

Figura 9. Capacidade preditiva genotípica (CPG) (correlação de Pearson entre o valor genotípico verdadeiro e o valor genético estimado) predita pelos quatro métodos de seleção genômica em função da herdabilidade, variando o número de QTL controlando a característica (60, 120, 180 e 240). Os pontos são os valores reais estimados pelos métodos de seleção genômica e as linhas são as regressões para cada método.

4.4.3. Influência do número de QTL na predição do valor genético

Com o objetivo de avaliar a influência do número de QTLs que controlam a característica na predição do valor genético pelos métodos de seleção genômica foram simulados diferentes cenários variando o número de QTL (60, 120, 180 e 240).

Foi observado que os valores de R^2 das regressões cúbicas variaram de 0.58 a 0.97 (Figura 10). Foi utilizado regressão cúbica pois esta foi a que apresentou melhores resultados. Desta forma não foi possível observar nenhum tipo de relação entre o número de QTL que controla a característica e a capacidade preditiva fenotípica independente da herdabilidade da característica.

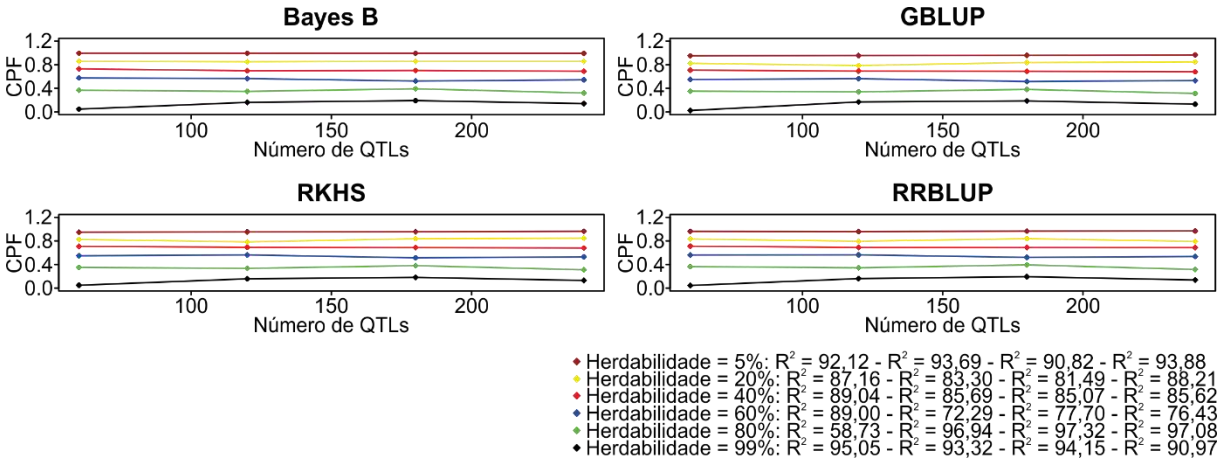

Figura 10. Capacidade preditiva fenotípica (CPF) (correlação de Pearson entre o valor fenotípico e o valor genético estimado) predita pelos quatro métodos de seleção genômica em função do número de QTL que controlam a característica, avaliado em diferentes herdabilidade (5, 20, 40, 60, 80 e 99%). Os pontos são os valores reais estimados pelos métodos de seleção genômica e as linhas são as regressões para cada método.

Para a capacidade preditiva genotípica também não foi possível encontrar nenhum tipo de relação entre os valores preditos pelos métodos de seleção genômica e o número de QTL controlando a característica quantitativa. A regressão cúbica foi a que apresentou melhores resultados, com R^2 variando de 85,29% a 98,58 (Figura 11).

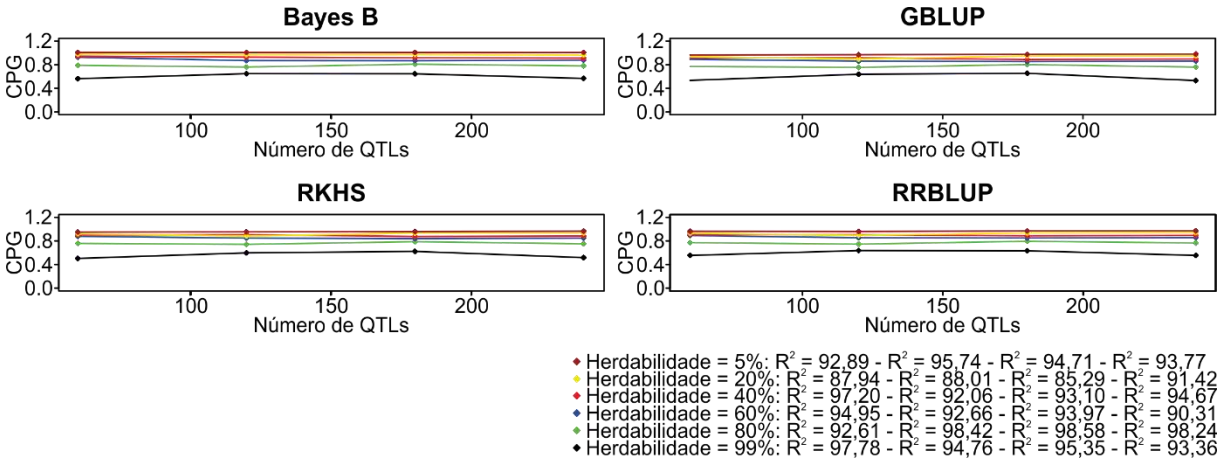

Figura 11. Capacidade preditiva genotípica (CPG) (correlação de Pearson entre o valor genotípico verdadeiro e o valor genético estimado) predita pelos quatro métodos de seleção genômica em função do número de QTL que controlam a característica, avaliado em diferentes herdabilidade (5, 20, 40, 60, 80 e 99%). Os pontos são os valores reais estimados pelos métodos de seleção genômica e as linhas são as regressões para cada método.

4.4.4. Predição da correlação genética via correlação fenotípica

Os modelos de regressão linear, quadrática e cúbica foram testados com o objetivo de encontrar uma relação entre a correlação genética e a correlação fenotípica.

Foi verificado baixo valor de R^2 para todos os modelos de regressão avaliados independentemente do método de seleção genômica utilizado para estimação da correlação fenotípica e genotípica, exceto para a herdabilidade de 99% onde o valor de R^2 foi superior a 89% para RRBLUP, RKHS e GBLUP (Tabelas 6, 7, 8 e 9).

Tabela 6. Coeficiente de determinação para predição da acurácia genética via acurácia fenotípica para características governadas por 60 QTLs.

Modelos	RRBLUP	GBLUP	Bayes B	RKHS
		Herdabilidade = 5%		
Linear	32,20	30,53	30,68	35,27
Quadrática	33,16	30,69	31,17	35,56
Cúbica	34,10	36,09	31,20	35,70
		Herdabilidade = 20%		
Linear	48,31	10,57	32,67	38,97
Quadrática	48,67	11,20	32,78	38,97
Cúbica	49,80	13,97	33,68	40,29
		Herdabilidade = 40%		
Linear	29,00	35,41	61,54	54,33
Quadrática	29,03	37,49	61,69	55,73
Cúbica	29,68	37,62	63,35	57,10
		Herdabilidade = 60%		
Linear	51,08	61,65	40,73	34,34
Quadrática	51,15	61,67	41,17	34,37
Cúbica	51,34	62,22	41,17	34,68
		Herdabilidade = 80%		
Linear	50,63	69,96	54,71	60,70
Quadrática	51,11	69,98	54,72	61,74
Cúbica	51,11	70,08	54,72	61,79
		Herdabilidade = 99%		
Linear	94,30	96,00	44,52	97,30
Quadrática	94,44	96,14	44,52	97,31
Cúbica	94,44	96,14	44,52	97,31

Foi verificado diferenças entre os métodos de seleção genômica para predição da correlação genética a partir da correlação fenotípica (Tabelas 6, 7, 8 e 9). Porém não foi possível detectar nenhum padrão entre os métodos, sendo completamente aleatório os resultados de R^2.

Tabela 7. Coeficiente de determinação para predição da acurácia genética via acurácia fenotípica para características governadas por 120 QTLs.

Modelos	RRBLUP	GBLUP	Bayes B	RKHS
		Herdabilidade = 5%		
Linear	10,46	41,52	18,32	24,00
Quadrática	13,15	41,56	19,30	30,45
Cúbica	13,23	47,76	19,71	35,70
		Herdabilidade = 20%		
Linear	25,92	44,02	27,74	46,04
Quadrática	27,28	44,55	30,25	46,04
Cúbica	27,46	44,75	30,25	46,85
		Herdabilidade = 40%		
Linear	21,10	17,80	52,50	11,62
Quadrática	23,67	25,27	57,68	12,09
Cúbica	28,49	25,72	58,29	12,23
		Herdabilidade = 60%		
Linear	36,73	43,84	34,77	52,93
Quadrática	37,29	49,81	35,52	52,97
Cúbica	37,79	50,49	45,11	53,08
		Herdabilidade = 80%		
Linear	68,36	72,07	58,31	54,47
Quadrática	68,47	72,56	58,77	54,56
Cúbica	69,01	72,63	59,44	59,04
		Herdabilidade = 99%		
Linear	94,63	96,88	10,23	95,43
Quadrática	94,64	96,91	10,23	95,89
Cúbica	94,64	96,91	10,23	95,89

Foi verificado que a medida que a herdabilidade aumentou os valores de R^2 das regressões também aumentaram para quase todos os cenários avaliados (Número de QTLs controlando a característica) independente do método de seleção genômica utilizado para estimação da correlação fenotípica e genética (Tabelas 6, 7, 8 e 9).

Tabela 8. Coeficiente de determinação para predição da acurácia genética via acurácia fenotípica para características governadas por 180 QTLs.

Modelos	RRBLUP	GBLUP	Bayes B	RKHS
		Herdabilidade = 5%		
Linear	51,23	21,41	24,94	27,04
Quadrática	51,31	21,58	24,97	27,13
Cúbica	51,61	21,94	26,09	30,31
		Herdabilidade = 20%		
Linear	32,00	44,90	35,76	30,18
Quadrática	33,06	44,94	36,49	38,49
Cúbica	35,96	45,09	36,59	41,45
		Herdabilidade = 40%		
Linear	39,73	35,04	36,26	19,38
Quadrática	39,84	35,41	36,26	19,56
Cúbica	40,03	36,84	36,27	20,15
		Herdabilidade = 60%		
Linear	45,74	49,68	44,41	51,75
Quadrática	46,41	53,45	56,03	51,75
Cúbica	47,30	53,59	56,09	52,25
		Herdabilidade = 80%		
Linear	45,01	52,35	26,63	63,75
Quadrática	45,01	52,52	26,69	65,35
Cúbica	45,03	52,70	26,69	65,48
		Herdabilidade = 99%		
Linear	95,96	95,00	40,83	95,24
Quadrática	95,97	95,00	40,83	95,29
Cúbica	95,97	95,00	40,83	95,29

Tabela 9. Coeficiente de determinação para predição da acurácia genética via acurácia fenotípica para características governadas por 240 QTLs.

Modelos	RRBLUP	GBLUP	Bayes B	RKHS
		Herdabilidade = 5%		
Linear	10,43	38,68	33,45	42,81
Quadrática	14,06	38,84	34,80	43,03
Cúbica	14,43	40,94	37,03	43,94
		Herdabilidade = 20%		
Linear	25,02	42,59	35,19	25,70
Quadrática	26,67	42,60	35,19	26,26
Cúbica	28,06	43,27	38,07	26,65
		Herdabilidade = 40%		
Linear	26,49	39,15	42,28	26,06
Quadrática	30,60	39,60	43,64	26,07
Cúbica	33,41	42,57	43,65	26,12
		Herdabilidade = 60%		
Linear	52,81	55,62	42,87	56,80
Quadrática	53,38	58,53	42,93	56,80
Cúbica	53,54	63,81	43,37	58,06
		Herdabilidade = 80%		
Linear	61,50	49,10	35,61	67,03
Quadrática	61,92	49,14	35,78	69,03
Cúbica	61,92	49,14	35,78	69,09
		Herdabilidade = 99%		
Linear	89,23	92,79	58,35	93,37
Quadrática	89,24	92,83	58,35	93,83
Cúbica	89,24	92,83	59,08	93,83

4.5. DISCUSSÃO

4.5.1. Comparação entre os métodos de seleção genômica

Quatro métodos de seleção genômica foram utilizados para estimar a capacidade preditiva em diversos cenários (herdabilidade versus número de QTLs) e foram comparados pelo teste de Tukey. Inúmeros fatores influenciam a predição da acurácia pelos métodos de seleção genômica, tais como, performance do modelo,

tamanho da população de treinamento, estrutura da população, herdabilidade, estrutura genética e distribuição dos QTLs ao longo do genoma (Desta and Ortiz 2014). Estes fatores fazem com que os métodos utilizados na seleção genômica se comportem de forma diferenciada de acordo com as mudanças inerentes a estes fatores.

A performance do modelo é muito influenciado pela interação interalélica. Ornella *et al.* (2012) verificaram que Bayesian Lasso e Bayesian Ridge Regression apresentaram resultados superiores a support vector regression (método não paramétrico assim como o RKHS utilizado no presente estudo) para a resistência de trigo a ferrugem. Os autores concluíram que a superioridade dos métodos paramétricos (RR-BLUP, Bayes B e GBLUP por exemplo) é devido a esta característica ser governada por genes de efeito aditivo. Por outro lado, os métodos não paramétricos (RKHS) conseguem capturar os efeitos não aditivos como dominância e epistasia, porém quando a característica tem controle gênico aditivo a utilização destes métodos pode até diminuir a acurácia como encontrado por Zhao *et al.* (2013) e no presente trabalho onde o método RKHS apresentou resultados inferiores na maioria dos cenários avaliados. Como as características foram todas simuladas com apenas efeito aditivo este fato pode ter influenciado para que todos os métodos apresentasse resultados semelhantes com exceção do RKHS (não paramétrico) (Tabela 4 e 5). Heslot *et al.* (2012), trabalhando com milho e cevada, compararam 11 métodos de seleção genômica e conseguiram separar os métodos em dois grupos, o grupo dos métodos paramétricos e o grupo dos métodos não paramétricos. No presente trabalho foi verificado que o método RKHS foi classificado em diferente grupo dos demais métodos.

Outro fator que influenciou para que os métodos tradicionais (GBLUP e RRBLUP) apresentassem resultados semelhantes aos métodos bayesianos foi a utilização de *prioris* não tão informativa, pois foi utilizado o *default* do pacote BGLR. Quando utilizamos *prioris* não informativas, a *posteriori* fica baseada apenas na função de verossimilhança, ou seja, apesar de ser métodos bayesianos os resultados apenas transcreveram a função de verossimilhança da mesma forma que os métodos tradicionais fazem (De Los Campos *et al.* 2009; Bhering *et al.* 2015). Todos os métodos bayesianos de seleção genômica utilizam o mesmo modelo original e o único fato que os diferenciam são as diferenças entre os hiperparametros nas *prioris*. Como

as *prioris* foram não informativas, os métodos acabam sendo muito semelhantes e consequentemente não apresentaram diferenças significativas.

Portanto, se não temos informações a priori sobre a característica em estudo, os métodos tradicionais de seleção genômica, RRBLUP e GBLUP, podem ser utilizados na predição do valor genético. Em caso contrário a utilização dos métodos bayesianos quando se tem acesso a informações a priori apresentarão melhores resultados (Meuwissen *et al.* 2001b). E se além das informações a priori também estamos estimando o efeito de dominância e/ou epistático o método RKHS é mais apropriado.

4.5.2. Influência da herdabilidade na predição do valor genético

Para avaliar o efeito da herdabilidade sobre a predição do valor genético pelos métodos de seleção genômica foram simuladas características com diferentes valores de herdabilidade (5, 20, 40, 60, 80 e 99%).

A herdabilidade tem correlação positiva com a acurácia como verificado em trigo para ferrugem amarela e ferrugem do colmo (Ornella *et al.* 2012), e em milho para produção de grãos e umidade dos grãos (Zhao *et al.* 2013). No entanto, a herdabilidade e número de QTLs que controlam a característica são fatores correlacionados e as vezes características com menor valor de herdabilidade e maior número de QTLs apresentam maior acurácia que características com maior herdabilidade e menor número de QTL como mostrado por Heffner *et al.* (2011). Este fato foi verificado no presente trabalho onde todas as características governadas por 240 QTLs apresentaram maior acurácia que as características governadas por 60 QTLs independentemente da herdabilidade, apesar desta relação não ser linear (Figura 10 e 11). No entanto para características governadas pelo mesmo número de QTL foi observado que, quanto maior o valor de herdabilidade, maior foi o valor da acurácia fenotípica (relação linear – Figura 8) e genotípica (relação quadrática – Figura 9).

A utilização da seleção genômica nos programas de melhoramento pode melhorar muito a precisão na seleção principalmente de características com alto custo para avaliação fenotípica (teor de proteínas e óleo) ou caraterísticas muito complexas (resistência a doenças), pois essas características normalmente possuem baixa herdabilidade (menor que 30%) e assim a seleção apenas com base no fenótipo torna-se muito difícil. Portanto como observado no presente trabalho e por Villumsen *et al.*

(2008) se compararmos a confiabilidade (quadrado da acurácia preditiva) com a herdabilidade da característica a diferença entre elas é maior para herdabilidade de menor valor, ou seja, a seleção com base no valor genético genômico predito pelos métodos de seleção genômica será muito mais acurada quando comparada a seleção com base nos valores fenotípicos para características de baixa herdabilidade.

4.5.3. Influência do número de QTL na predição do valor genético

Com o objetivo de avaliar o efeito do número de QTLs que controlam a característica foram simulados cenários variando o número de QTL (60, 120, 180 e 240).

Não foi possível verificar algum tipo de relação entre número de QTL e acurácia fenotípica ou genotípica. Resultados semelhantes foram encontrados em espécies florestais (Grattapaglia and Resende 2011; Iwata *et al.* 2011) e em milho (Riedelsheimer *et al.* 2012).

No entanto, quando a característica é governada por um número pequeno de QTL de maior efeito, os métodos bayesianos apresentam melhor performance que os métodos tradicionais (GBLUP e RRBLUP) e o oposto ocorre quando o número de QTL é alto (Meuwissen *et al.* 2001a; Zhong *et al.* 2009a; Daetwyler *et al.* 2010). Todavia esta diferença pode esta sendo mais influenciada por outras características tais como herdabilidade, tamanho da população de treinamento, estrutura da população do que propriamente o número de QTL (Desta and Ortiz 2014).

Pelos resultados vistos aqui podemos concluir que a verificação de um fator separadamente pode super ou subestimar os valores estimados pelos métodos de seleção genômica e, portanto, é necessário estudos criteriosos como este, onde vários fatores são estudados em conjunto, de forma a estabelecer o melhor modelo de seleção genômica para cada estrutura de população, que no presente trabalho foi F_2.

4.5.4. Predição da correlação genética via correlação fenotípica

Segundo Dekkers (2007) a acurácia, também conhecida como correlação genética entre o valor genético verdadeiro e o valor genético genômico, é estimado pela relação entre a correlação entre o valor fenotípico e o valor genético genômico e a herdabilidade da característica. Com o objetivo de predizer a correlação genética m função da correlação fenotípica os modelos de regressão linear, quadrática e cúbica

foram utilizados. No entanto foi verificado pela avaliação do R^2 dos modelos de regressão que a relação entre a correlação genética e fenotípica não pode ser explicado via os modelos simples de regressão (Tabelas 5, 6, 7 e 8).

Para herdabilidades mais próximas de 1, ou seja, quando o efeito ambiental é pequeno os modelos de regressão conseguiram explicar de forma mais acurada a correlação genética a partir da correlação fenotípica (Tabelas 5, 6, 7 e 8). Este fato é explicado devido a relação entre herdabilidade e correlação, onde a correlação do valor fenotípico com o valor genético é a raiz quadrada da herdabilidade.

Porém para valores baixos de herdabilidade essa relação não foi observada. Assim podemos concluir que não existe uma relação linear entre as correlações genética e fenotípica quando a herdabilidade da característica é inferior a 80%. Desta forma é necessário a utilização de modelos não lineares como as redes neurais artificiais para estimar a correlação genética em função da correlação fenotípica de forma mais acurada.

4.6. CONCLUSÃO

Todos os métodos de seleção genômica avaliados no presente estudo podem ser utilizados para predição do valor genético em uma população F_2, no entanto os métodos Bayes B apresenta resultados um pouco superiores aos demais;

A herdabilidade possui uma relação linear positiva com a capacidade preditiva fenotípica e quadrática com a capacidade preditiva genotípica;

O número de QTL que controlam a característica não possui nenhum tipo de relação com a capacidade preditiva fenotípica e genotípica.

4.7. REFERÊNCIAS BIBLIOGRÁFICAS

BHERING, L. L., JUNQUEIRA, V. S., PEIXOTO, L. A., CRUZ, C. D., LAVIOLA, B. G., Comparison of methods used to identify superior individuals in genomic selection in plant breeding. **Genetics and Molecular Research:**, v. 14, p. 10888-10896, 2015.

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

DAETWYLER, H. D., PONG-WONG, R., VILLANUEVA, B., WOOLLIAMS, J. A., The impact of genetic architecture on genome-wide evaluation methods. **Genetics**, v. 185, p. 1021-1031, 2010.

DE LOS CAMPOS, G., GIANOLA, D., ROSA, G. J. M., WEIGEL, K. A., CROSSA, J., Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. **Genetics Research**, v. 92, p. 295-308, 2010.

DE LOS CAMPOS, G., NAYA, H., GIANOLA, D., CROSSA, J., LEGARRA, A., MANFREDI, E., WEIGEL, K., COTES, J. M., Predicting quantitative traits with regression models for dense molecular markers and pedigree. **Genetics**, v. 182, p. 375-385, 2009.

DEKKERS, J., Prediction of response to marker-assisted and genomic selection using selection index theory. **Journal of Animal Breeding and Genetics**, v. 124, p. 331-341, 2007.

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

ENDELMAN, J. B., Ridge regression and other kernels for genomic selection with R package rrBLUP. **The Plant Genome**, v. 4, p. 250-255, 2011.

FALCONER, D., MACKAY, T., 1996. Introduction to Quantitative Genetics. Longman Scientific & Technical, Harlow, UK.

HEFFNER, E. L., JANNINK, J.-L., IWATA, H., SOUZA, E., SORRELLS, M. E., Genomic selection accuracy for grain quality traits in biparental wheat populations. **Crop Science**, v. 51, p. 2597-2606, 2011.

HEFFNER, E. L., SORRELLS, M. E., JANNINK, J.-L., Genomic selection for crop improvement. **Crop Science**, v. 49, p. 1-12, 2009.

HESLOT, N., YANG, H.-P., SORRELLS, M. E., JANNINK, J.-L., Genomic selection in plant breeding: a comparison of models. **Crop Science**, v. 52, p. 146-160, 2012.

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

JENA, K. K., MACKILL, D. J., Molecular markers and their use in marker-assisted selection in rice. **Crop Science**, v. 48, p. 1266-1276, 2008.

LORENZ, A. J., CHAO, S., ASORO, F. G., HEFFNER, E. L., HAYASHI, T., IWATA, H., SMITH, K. P., SORRELLS, M. E., JANNINK, J.-L., 2 Genomic Selection in Plant Breeding: Knowledge and Prospects. **Advances in agronomy**, v. 110, p. 77, 2011.

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

ORNELLA, L., SINGH, S., PEREZ, P., BURGUEÑO, J., SINGH, R., TAPIA, E., BHAVANI, S., DREISIGACKER, S., BRAUN, H.-J., MATHEWS, K., Genomic

prediction of genetic values for resistance to wheat rusts. **The Plant Genome**, v. 5, p. 136-148, 2012.

PÉREZ, P., DE LOS CAMPOS, G., 2012. BGLR: A Statistical Package for Whole-Genome Regression.

STUBER, C. W., GOODMAN, M. M., MOLL, R. H., Improvement of yield and ear number resulting from selection at allozyme loci in a maize population. **Crop Science**, v. 22, p. 737-740, 1982.

TANKSLEY, S. D., YOUNG, N. D., PATERSON, A. H., BONIERBALE, M. W., RFLP mapping in plant breeding: new tools for an old science. **Nature Biotechnology**, v. 7, p. 257-264, 1989.

TEAM, R. C., 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria, Vienna.

VILLUMSEN, T. M., JANSS, L., LUND, M. S., The importance of haplotype length and heritability using genomic selection in dairy cattle. **Animal breeding and genetics**, v. 126, p. 3-13, 2008.

XU, Y., CROUCH, J. H., Marker-assisted selection in plant breeding: from publications to practice. **Crop Science**, v. 48, p. 391-407, 2008.

ZHONG, S., DEKKERS, J. C. M., FERNANDO, R. L., JANNINK, J.-L., Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. **Genetics**, v. 182, p. 355-364, 2009.

ZHONG, S., TOUBIA-RAHME, H., STEFFENSON, B. J., SMITH, K. P., Molecular mapping and marker-assisted selection of genes for septoria speckled leaf blotch resistance in barley. **Phytopathology**, v. 96, p. 993-999, 2006.

5. CAPITULO 3

NÚMERO DE MARCADORES E DE INDIVÍDUOS NA POPULAÇÃO DE TREINAMENTO NECESSÁRIOS PARA OBTER MÁXIMA ACURÁCIA DE PREDIÇÃO EM POPULAÇÕES F_2 VIA MODELOS DE SELEÇÃO GENÔMICA

NÚMERO DE MARCADORES E DE INDIVÍDUOS NA POPULAÇÃO DE TREINAMENTO NECESSÁRIOS PARA OBTER MÁXIMA ACURÁCIA DE PREDIÇÃO EM POPULAÇÕES F_2 VIA MODELOS DE SELEÇÃO GENÔMICA

5.1. RESUMO

A seleção genômica tem se tornado uma técnica muito útil para ajudar os melhoristas a selecionar os melhores genótipos de forma mais acurada. Como a seleção fenotípica em F_2 possui baixa acurácia devido ao fato de cada genótipo ser representado apenas por um indivíduo, a seleção genômica pode aumentar a acurácia de seleção nesta etapa do programa de melhoramento. Desta forma, os objetivos do trabalho foram estabelecer o número ideal de indivíduos para compor a população de treinamento e estabelecer a quantidade necessária de marcadores para obter máxima acurácia pelos métodos de seleção genômica em populações F_2. Uma população F_2 com 1.000 indivíduos foram simuladas, e seis características foram simuladas variando a herdabilidade (5%, 20%, 40%, 60%, 80% e 99%). O método de seleção genômica RR-BLUP foi utilizado em todas as análises. Os modelos de seleção genômica foram montados variando o número de indivíduos na população de treinamento (2 a 1.000 indivíduos) e o número de marcadores (2 a 3.060 marcadores). Os parâmetros avaliados foram a acurácia fenotípica, acurácia genotípica, variância genética, variância residual e herdabilidade. Foi verificado que quanto maior o número de indivíduos nas populações de treinamento maior é o valor da acurácia e valores das variâncias genotípicas e residuais e herdabilidade ficam mais próximos do valor real. Quanto maior a herdabilidade da característica maior foi o número de marcadores necessários para obter máxima acurácia, variando de 300 para característica com herdabilidade de 5% a 800 para características com herdabilidade de 99%. Portanto, os modelos de seleção genômica para predição em populações F_2 devem ser compostos por 300 a 800 marcadores de maior efeito sobre a característica e mais de 600 indivíduos na população de treinamento.

Palavras-chave: Predição genômica, herdabilidade, capacidade preditiva, melhoramento, genética quantitativa

5.2. INTRODUÇÃO

A seleção de plantas vem sendo realizada pela humanidade desde os primórdios da história. Porém, esta seleção foi intensificada no início do último século, onde os programas de melhoramento das principais culturas foram formados (Allard, 1999). Durante os últimos 100 anos houve evolução exponencial nos métodos e nas tecnologias que auxiliam a seleção de plantas (Borém & Miranda, 2013). Os métodos de seleção evoluíram da seleção massal (Borém & Miranda, 2013), que consiste na seleção visual dos indivíduos, passando pela seleção combinada (Bhering et al., 2013), que leva em consideração as informações entre e dentro de famílias, chegando a seleção recorrente reciproca (Ordas et al., 2012), que consiste no aumento gradativo da frequência de alelos favoráveis, por meio de repetidos ciclos de seleção, sem reduzir a variabilidade genética da população.

Com o advento dos marcadores moleculares na década de 80 foi possível melhorar a acurácia de seleção por meio da seleção assistida por marcadores (SAM) moleculares (He et al., 2014). Apesar da SAM ter sido um avanço significativo no melhoramento de plantas, esta técnica só é efetiva para características qualitativas, ou governadas por poucos genes, como morte súbita na soja (Lightfoot, 2015), ferrugem em trigo (Yaniv et al., 2015), tolerância a salinidade (Ashraf et al., 2012) e ferrugem bacteriana em arroz (Pandey et al., 2013). Em 30 anos a partir do seu surgimento, as técnicas de marcadores moleculares evoluíram de forma efetiva, passando desde isoenzimas (Dirlewanger et al., 1998), RAPD (Lynch & Milligan, 1994), RFLP (Langer & Maixner, 2015), AFLP (Frascaroli et al., 2013), microssatélites (Soldati et al., 2013), chegando aos marcadores de base única conhecidos como SNP (Belaj et al., 2012).

Dentre as variações encontradas no genoma, as variações do tipo SNP são as mais amplamente distribuídas e abundantes no genoma. Com o desenvolvimento das plataformas de genotipagem de marcadores SNPs, e o aprimoramento dos métodos estatísticos, Meuwissen et al. (2001) apresentaram uma nova abordagem baseada em regressão múltipla utilizando os marcadores como covariáveis, conhecida como seleção genômica. O objetivo da seleção genômica é identificar possíveis marcadores em desiquilíbrio de ligação com as regiões gênicas de interesse. A partir do trabalho pioneiro de Meuwissen et al. (2001), inúmeros autores começaram a utilizar desta técnica para predizer o valor genético em diversas espécies vegetais, tais como milho (Beyene et al., 2015), soja (Zhang et al., 2016), trigo (Bassi et al., 2016), espécies

florestais (Cros et al., 2015), cana de açúcar (Gouy et al., 2013), e arroz (Spindel et al., 2015).

Apesar dos inúmeros trabalhos existentes, poucos deles mostram o efeito dos diferentes fatores que afetam a acurácia de predição na seleção genômica, tais como o número de marcadores e de indivíduos na população de treinamento. Isidro et al. (2015) avaliaram cinco critérios para determinar o número ótimo de indivíduos para compor a população de treinamento em cinco características em trigo. Os autores observaram que quanto maior o número de indivíduos na população de treinamento maior o valor da acurácia predita, porém outros efeitos também devem ser considerados como a arquitetura da característica e a estrutura da população. de los Campos et al. (2013) realizaram a seleção de marcadores baseados na sua importância para a característica pelos resultados provenientes das análises GWAS via meta-análises e verificaram que a seleção de marcadores foi eficiente em humanos, pois a acurácia foi 7.5% maior para os modelos de seleção genômica utilizando 5k SNPs quando comparado com modelos usando 400.000 SNPs.

A geração filial F_2 é uma das fases mais importantes em um programa de melhoramento de plantas porque maior variabilidade genética e heterose são encontradas nesta fase (Tang et al., 1993). Além disso, na geração filial F_2 é possível estimar a frequência alélica para cada gene via segregação mendeliana, avaliar possíveis distorções no equilíbrio de Hardy-Weinberg (Falconer & Mackay, 1996), a variância genotípica e ambiental, e consequentemente, a herdabilidade (Tang et al., 1996). A utilização da seleção genômica em populações F_2 ainda é restrita a poucos trabalhos (Ren et al., 2015), e pouco se sabe sobre como os fatores que afetam a acurácia preditiva podem afetar a estimativa dos parâmetros genéticos em populações F_2. Portanto, os objetivos do trabalho foram estabelecer o número ideal de indivíduos e marcadores para compor a população de treinamento nos modelos de seleção genômica, de modo a capturar máxima variância genética, e consequentemente obter maior acurácia em populações F_2.

5.3. MATERIAL E MÉTODOS

5.3.1. Simulação dos dados

A simulação da população F_2 foi realizada utilizando o módulo de simulação do aplicativo computacional GENES (Cruz, 2013), que permitiu gerar informações

sobre o genoma, genótipos dos genitores, populações de cruzamentos controlados e dados de características quantitativas.

5.3.1.1. Simulação do genoma

Foi simulado um genoma constituído de 15 grupos de ligação, similar ao de uma espécie diploide 2n=2x=30. Cada grupo de ligação foi simulado com 200 cM, e constituído por 200 marcas, espaçadas de forma equidistante (1 cM), totalizando 3.060 marcas. As marcas foram assumidas como codominantes e bialélicas. Além disso, foi considerado 4 marcas por grupo de ligação como responsável pelo controle da expressão fenotípica das características quantitativas, que foram inseridas de forma aleatória no genoma.

5.3.1.2. Simulação dos genitores

Pais homozigotos contrastantes foram simulados, ou seja, o pai 1 foi codificado como portador de um alelo A_1 (recebeu código 2), e o pai 2 foi codificado como portador do alelo alternativo A_2 (recebeu código 0) para todas as marcas existentes. Desta forma, o cruzamento entre pai 1 e pai 2 gerou a população F_1 com todos as marcas em heterozigose e em fase de aproximação do tipo $A_1B_1//A_2B_2$.

5.3.1.3. Simulação da população de mapeamento

A população F_2 foram geradas a partir da autofecundação dos indivíduos da população F_1. Para a formação do primeiro individuo da população F_2, cada indivíduo da população F_1 produziu 5.000 gametas e quando 2 destes gametas se encontraram ao acaso, o primeiro indivíduo da população F_2 foi gerado. Este processo se repetiu até formação de todos os indivíduos em cada população.

Para a formação de cada gameta seguiu-se o seguinte critério: O alelo da primeira marca foi escolhido de forma aleatória (A_1 ou A_2) para começar a constituir o gameta (alelo de inicialização); O alelo da segunda marca foi escolhido levando em consideração a distância para o primeiro gene, ou seja, as frequências de crossing over foram contabilizados e a escolha de qual alelo (B_1 ou B_2) que constituiria o gameta foi baseado nas probabilidades de cada gameta $P(A_1B_1)$ e $P(A_2B_2)$ que são gametas parentais, e $P(A_1B_2)$ e $P(A_2B_1)$ que são os gametas recombinantes. Este processo se repetiu até chegar no último gene. Foi considerado interferência nula, ou seja, o crossing over ocorrido entre os genes A e B não interferem em um próximo

crossing over entre os genes B e C. Desta forma foi garantido que todos os gametas formados eram diferentes devido a escolha aleatória do alelo no primeiro gene e as probabilidades condicionadas a cada alelo para os próximos genes. Como todos os genes foram simulados de forma equidistante a 1 cM, a frequência de recombinação foi de 1% para todos os genes, ou seja, a probabilidade de cada gameta foi: $P(A_1B_1) = P(A_2B_2) = 0.49$ e $P(A_1B_2) = P(A_2B_1) = 0.005$.

A população F_2 simulada foi codificada com 0, 1 e 2, sendo que 0 correspondeu aos indivíduos homozigotos (A_2A_2), 1 aos indivíduos heterozigotos (A_1A_2),) e 2 aos indivíduos homozigotos (A_1A_1), para um determinado loco.

5.3.1.4. Simulação das características quantitativas

Para a simulação das características quantitativas primeiramente foi atribuído, como importância de cada loco, um valor correspondente à probabilidade gerada por uma distribuição binomial, de parâmetro p=q= 0,5 e n = 59 (gerando uma família de probabilidade com 60 elementos). Este valor, denominado proporção da variância genética explicada por cada QTL ($\mathrm{PVG/QTL}$), corresponde a importância do loco para a média genotípica e, consequentemente, para a proporção da variância genética da característica explicada por cada QTL.

Cada característica foi simulada sendo controlada por 60 QTLs distribuídos de forma equidistante no genoma (4 QTls por grupo de ligação). O efeito de cada QTL foi definido por: $A_1A_1=\mu + a$; $A_1A_2=\mu + d$; $A_2A_2=\mu - a$, onde *a* é o efeito aditivo de cada gene e *d* é o efeito de dominância de cada gene. Como o valor de *d* foi definido como nulo o grau médio de dominância (*d/a*) foi igual a zero para todos os locos, ou seja, todos os locos possuíam apenas efeito aditivo.

O valor genotípico (VG) de cada indivíduo foi definido pela equação:

$$\mathrm{VG} = \sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{PVG/QTL_i} \ \ \mathrm{x\ efeito\ do\ QTL_i})$$

O efeito ambiental (EA) foi assumido como não correlacionado com o valor genotípico e foi estimado seguindo uma distribuição $N(0,\sigma^2)$. O valor de σ^2 é calculado a partir da herdabilidade da característica e o valor da variância genética (σ_g^2). Foram simuladas características com herdabilidade de 5%, 20%, 40%, 60%, 80% e 99%. A σ_g^2 foi calculada como sendo a variância do valor genotípico dos indivíduos da população F_2. Sendo assim, o valor fenotípico foi calculado como:

$$\mathrm{VF} = \mathrm{u} + \mathrm{VG} + \mathrm{EA}$$

em que u = 100 é média e VF é o valor fenotípico.

5.3.2. Análise dos dados

Após a geração da população, seguiram-se as etapas do processo de mapeamento, iniciando pela análise de segregação de locos individuais. Foram aplicados testes de qui-quadrado (χ^2) para verificar se as marcas segregavam de acordo com o esperado em uma população F_2. Verificou-se ainda se todos os grupos de ligação foram restaurados, com tamanho, distância e ordem dos marcadores, podendo concluir assim que se tratava de uma população F_2 com as propriedades de simulação desejadas.

O método de seleção genômica utilizado nas análises foi o ridge regression best linear unbiased prediction (RR-BLUP) que tem como objetivo estimar o efeito para cada uma das covariáveis (marcadores SNP) incluídos no modelo. O RR-BLUP assume que todos os SNPs apresentam controle na expressão fenotípica dos QTLs e assume variância homogênea.

Para verificar o número ideal de indivíduos para compor a população de treinamento (PT), foi utilizado modelos de seleção genômica variando de 2 a 1000 indivíduos na PT e todos os marcadores disponíveis, ou seja, 3060 marcas. A população de validação foi sempre composta por 200 indivíduos tomados aleatoriamente na população. Gráficos de tendência foram plotados para a variância genética, variância residual, herdabilidade, acurácia genotípica e fenotípica. Foram realizados 50 análises para cada número de indivíduos testados na população de treinamento, de forma que em cada análise os indivíduos tomados aleatoriamente eram diferentes.

Com o objetivo de avaliar o número de marcadores necessário para capturar toda a variância genética e, consequentemente, obter maior acurácia, foram utilizados modelos de seleção genômica variando o número de marcas de 2 a 3060. A PT foi formada por 800 indivíduos e a população de validação por 200 indivíduos. Foi realizada seleção de marcas, ou seja, o marcador com menor efeito foi deletado da matriz original de marcadores e não foi utilizado na próxima análise, ou seja, o próximo modelo teria um marcador a menos. Essa seleção aconteceu até que que o modelo fosse composto apenas por 2 marcadores. Gráficos de tendência foram plotados para a variância genética, variância residual, herdabilidade, e acurácia fenotípica.

As acurácias fenotípica e genotípica foram estimadas como a correlação de Pearson entre o valor fenotípico e valor genético genômico (GEBV), e entre o valor genotípico verdadeiro e o GEBV respectivamente.

A variância genética (σ_g^2) foi estimada segundo Falconer & Mackay (1996):

$$\sigma_g^2 = 2\sum_{i=1}^{n} p * q * \alpha_i^2$$

em que α_i^2 é o efeito de substituição alélica para cada loco.

A herdabilidade (h^2) foi estimada como:

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma^2}$$

5.3.3. Informações de software e hardware

As simulações foram realizadas no software GENES (Cruz, 2013), enquanto que as análises de teste de segregação e seleção genômica foram executadas no programa estatístico R (Team, 2014). O pacote rrBLUP (Endelman, 2011) foi utilizado para rodar o modelo RR-BLUP. Dois computadores de alta performance (Intel Xeon, processador E5-26 12° geração 3.30 GHz, memória RAM de 64 e 96 GB e hard dri ve de 1024 GB) foram utilizados para rodar as análises de seleção genômica.

5.4. RESULTADO

Com o objetivo de avaliar como o tamanho da população de treinamento (PT) e o número de marcadores influenciariam na predição do valor genético genômico com o método RR-BLUP, variou-se o tamanho da PT de 2 a 1000 indivíduos e de 2 a 3060 marcadores SNPs.

5.4.1. Teste de segregação dos marcadores

O teste de segregação foi realizado com o objetivo de verificar se o processo de simulação conseguiu estabelecer uma população com características genéticas de uma população F_2 como proposto por Falconer & Mackay (1996).

Foi observado que as frequências alélica e genotípica foram próximos do valor esperado para uma população F_2 para todas as marcas que controlam a característica (QTL – Tabela 10).

Tabela 10. Teste de segregação, frequência do alelo menos frequente (MAF), p-valor associado ao teste χ^2 da avaliação do equilíbrio de Hardy-Weinberg (hwe.p.valor) e efeito dos marcadores associados ao quantitative trait loci (QTL).

QTL	AA	Aa	aa	Total	maf	hwe.p.valor	PVG/QTL	EA(u+a)
M42	248	503	249	1000	0.4995	0.849491	0.00000000	0.000000
M83	220	525	255	1000	0.4825	0.104833	0.00000000	0.000000
M123	249	488	263	1000	0.4930	0.451512	0.00000000	0.000000
M165	254	502	244	1000	0.4950	0.896830	0.00000000	0.000000
M246	253	478	269	1000	0.4920	0.166462	0.00000000	0.000000
M287	275	480	245	1000	0.4850	0.215878	0.00000000	0.000000
M327	290	487	223	1000	0.4665	0.494414	0.00000000	0.000000
M369	255	506	239	1000	0.4920	0.698262	0.00000000	0.000001
M450	234	491	275	1000	0.4795	0.605211	0.00000000	0.000004
M491	248	499	253	1000	0.4975	0.950199	0.00000002	0.000022
M531	245	499	256	1000	0.4945	0.952612	0.00000011	0.000109
M573	231	507	262	1000	0.4845	0.635811	0.00000049	0.000486
M654	255	495	250	1000	0.4975	0.752424	0.00000194	0.001942
M695	238	503	259	1000	0.4895	0.838532	0.00000702	0.007021
M735	257	502	241	1000	0.4920	0.892912	0.00002310	0.023069
M777	255	502	243	1000	0.4940	0.895725	0.00006920	0.069208
M858	234	529	237	1000	0.4985	0.066591	0.00019000	0.190321
M899	247	520	233	1000	0.4930	0.203601	0.00048100	0.481401
M939	263	489	248	1000	0.4925	0.490986	0.00112300	1.123269
M981	253	509	238	1000	0.4925	0.564308	0.00242400	2.423897
M1062	251	490	259	1000	0.4960	0.528386	0.00484800	4.847794
M1103	264	494	242	1000	0.4890	0.715601	0.00900300	9.003046
M1143	232	508	260	1000	0.4860	0.595299	0.01555100	15.55072
M1185	265	517	218	1000	0.4765	0.251149	0.02501600	25.01637
M1266	246	491	263	1000	0.4915	0.575321	0.03752500	37.52455
M1307	269	493	238	1000	0.4845	0.679807	0.05253400	52.53438
M1347	273	487	240	1000	0.4835	0.430338	0.06869900	68.69880
M1389	254	498	248	1000	0.4970	0.900241	0.08396500	83.96520
M1470	248	497	255	1000	0.4965	0.850723	0.09596000	95.96023
M1511	254	481	265	1000	0.4945	0.230923	0.10257800	102.5782

M1551	249	501	250	1000	0.4995	0.949546	0.10257800	102.5782
M1593	275	470	255	1000	0.4900	0.059366	0.09596000	95.96023
M1674	234	497	269	1000	0.4825	0.879831	0.08396500	83.96520
M1715	260	499	241	1000	0.4905	0.958649	0.06869900	68.69880
M1755	239	509	252	1000	0.4935	0.565527	0.05253400	52.53438
M1797	252	505	243	1000	0.4955	0.749867	0.03752500	37.52455
M1878	241	501	258	1000	0.4915	0.942279	0.02501600	25.01637
M1919	239	495	266	1000	0.4865	0.769225	0.01555100	15.55072
M1959	253	519	228	1000	0.4875	0.221634	0.00900300	9.003046
M2001	252	508	240	1000	0.4940	0.609637	0.00484800	4.847794
M2082	238	505	257	1000	0.4905	0.743092	0.00242400	2.423897
M2123	255	480	265	1000	0.4950	0.206994	0.00112300	1.123269
M2163	254	504	242	1000	0.4940	0.796736	0.00048100	0.481401
M2205	229	499	272	1000	0.4785	0.996183	0.00019000	0.190321
M2286	254	495	251	1000	0.4985	0.752043	0.00006920	0.069208
M2327	280	501	219	1000	0.4695	0.855905	0.00002310	0.023069
M2367	268	492	240	1000	0.4860	0.630126	0.00000702	0.007021
M2409	256	504	240	1000	0.4920	0.793981	0.00000194	0.001942
M2490	255	508	237	1000	0.4910	0.605591	0.00000049	0.000486
M2531	244	502	254	1000	0.4950	0.896830	0.00000011	0.000109
M2571	244	492	264	1000	0.4900	0.621650	0.00000002	0.000022
M2613	256	502	242	1000	0.4930	0.894419	0.00000000	0.000004
M2694	242	496	262	1000	0.4900	0.809997	0.00000000	0.000001
M2735	257	504	239	1000	0.4910	0.792309	0.00000000	0.000000
M2775	260	501	239	1000	0.4895	0.938444	0.00000000	0.000000
M2817	261	491	248	1000	0.4935	0.572781	0.00000000	0.000000
M2898	241	539	220	1000	0.4895	0.013079	0.00000000	0.000000
M2939	255	499	246	1000	0.4955	0.951607	0.00000000	0.000000
M2979	232	505	263	1000	0.4845	0.728628	0.00000000	0.000000
M3021	236	510	254	1000	0.4910	0.520283	0.00000000	0.000000

M – marcador. O número na frente da letra M significa o número do marcador na tabela original; PVG/QTL – proporção da variância genética de cada característica explicada por cada QTL; EA – efeito aditivo.

A avaliação do equilíbrio de Hardy-Weinberg foi realizada pelo teste de qui-quadrado (χ^2), o qual indicou segregação esperada de acordo com uma população F_2 (p-value é apresentado na Tabela 10 e no material suplementar 1).

A proporção da variância genética da característica explicada por cada QTL seguiu distribuição binomial como esperado via o processo de simulação (Tabela 10). Os valores dos efeitos aditivos variaram de QTL para QTL, sendo maior nos QTLs localizados nos grupos de ligação 5, 6, 7, 8, 9 e 10 e menor nos QTLs localizados nos grupos de ligação 1, 2, 3, 4, 11, 12, 13, 14 e 15 (Tabela 10).

5.4.2. Avaliação do tamanho da população de treinamento

As variâncias genotípicas e residuais mostraram comportamento semelhante independente da herdabilidade simulada (Figura 12). Foi observado que quanto maior a quantidade de indivíduos utilizados na população referência, maior foi a habilidade do modelo em estimar componentes de variância semelhantes aos parâmetros simulados. Também foi observado que a herdabilidade influencia no número de indivíduos para compor a PT para estimar componentes de variância (genotípico e residual) acurados.

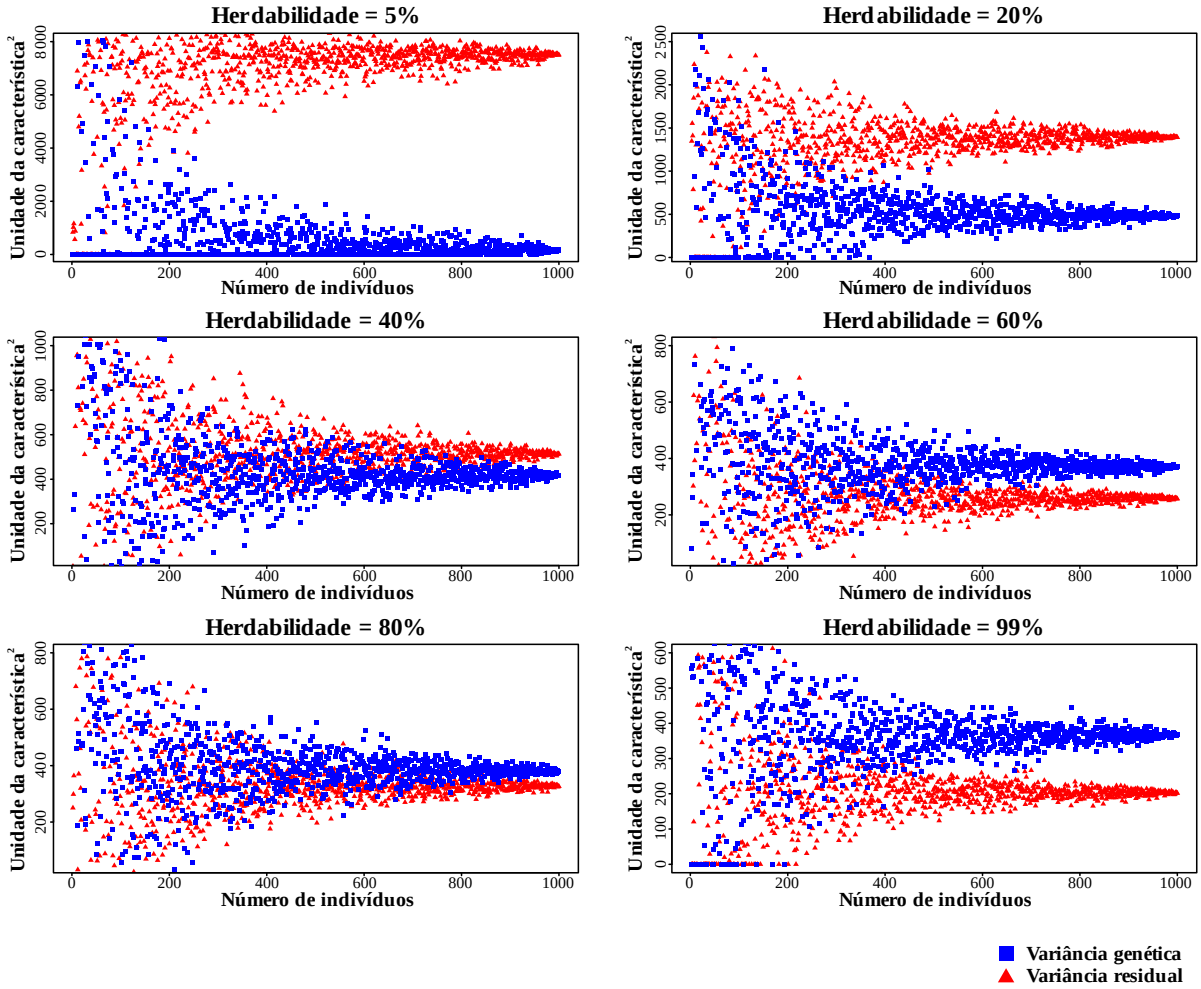

Figura 12. Tendência da variância genética e residual em função do número de indivíduos na população de treinamento para características com diferentes herdabilidades: a) 5%; b) 20%; c) 40%; d) 60%; e) 80%; f) 99%.

As acurácias genotípica e fenotípica apresentaram maiores estimativas em cenários de PT com maior número de indivíduos (Figura 13). Também foi observado que, quanto maior a herdabilidade da característica, maiores foram as estimativas de acurácias fenotípica e genotípica (Figura 13). Porém, PT com mais de 600 indivíduos proporcionaram pequeno ganho nas acurácias fenotípica e genotípica. A acurácia genotípica apresentou valores inferiores a acurácia fenotípica para todas as herdabilidade avaliadas, porém quanto maior o valor da herdabilidade simulada mais as estimativas de acurácia fenotípica e genotípica se aproximam (Figura 13).

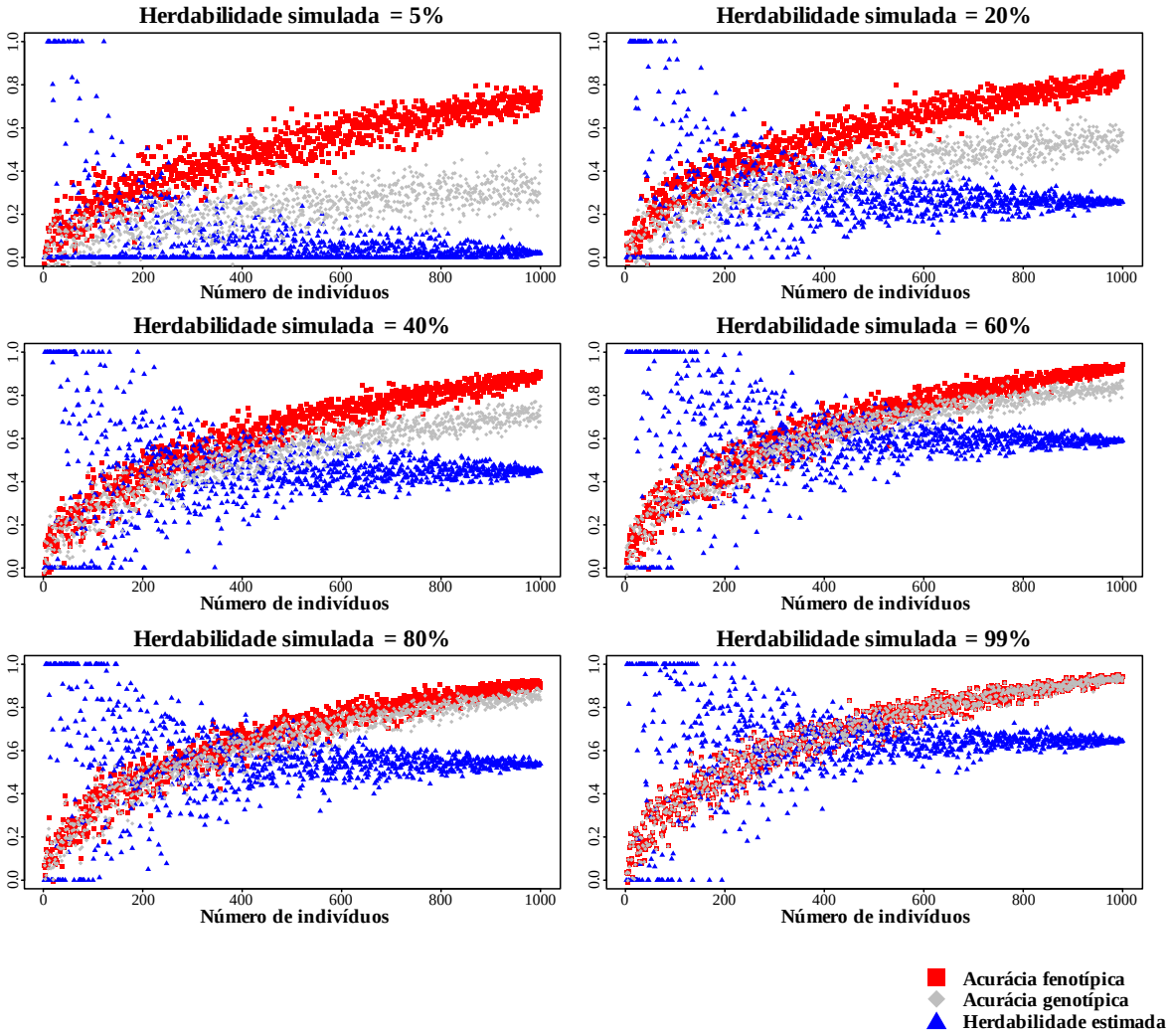

Figure 13. Tendência da acurácia fenotípica, acurácia genotípica e herdabilidade em função do número de indivíduos na população de treinamento para características com diferentes herdabilidades: a) 5%; b) 20%; c) 40%; d) 60%; e) 80%; f) 99%.

A herdabilidade estimada mostrou comportamento semelhante às variâncias genotípicas e residuais, ou seja, apresentou valores variáveis quando poucos indivíduos foram utilizados na PT (Figura 13). À medida que o número de indivíduos na PT aumentou, valores mais estáveis para a herdabilidade estimada foram observados e, estes valores foram mais próximos dos valores de herdabilidade simulada, com exceção das características com herdabilidade de 80% e 99%, onde os valores da herdabilidade estimada foram inferiores aos valores simulados (Figura 13).

5.4.3. Avaliação do número de marcadores necessários para predição genômica em uma população F_2

A variância genética apresentou tendência quadrática, ou seja, ela aumentou até um determinado número de marcadores, e depois foi decrescendo à medida que aumentou o número de marcas (Figura 14). O número ótimo de marcadores variou de acordo com a herdabilidade da característica (Tabela 11). Verificou-se também que quanto maior a herdabilidade da característica maior o número de marcadores necessário para obter a melhor estimativa da variância genética.

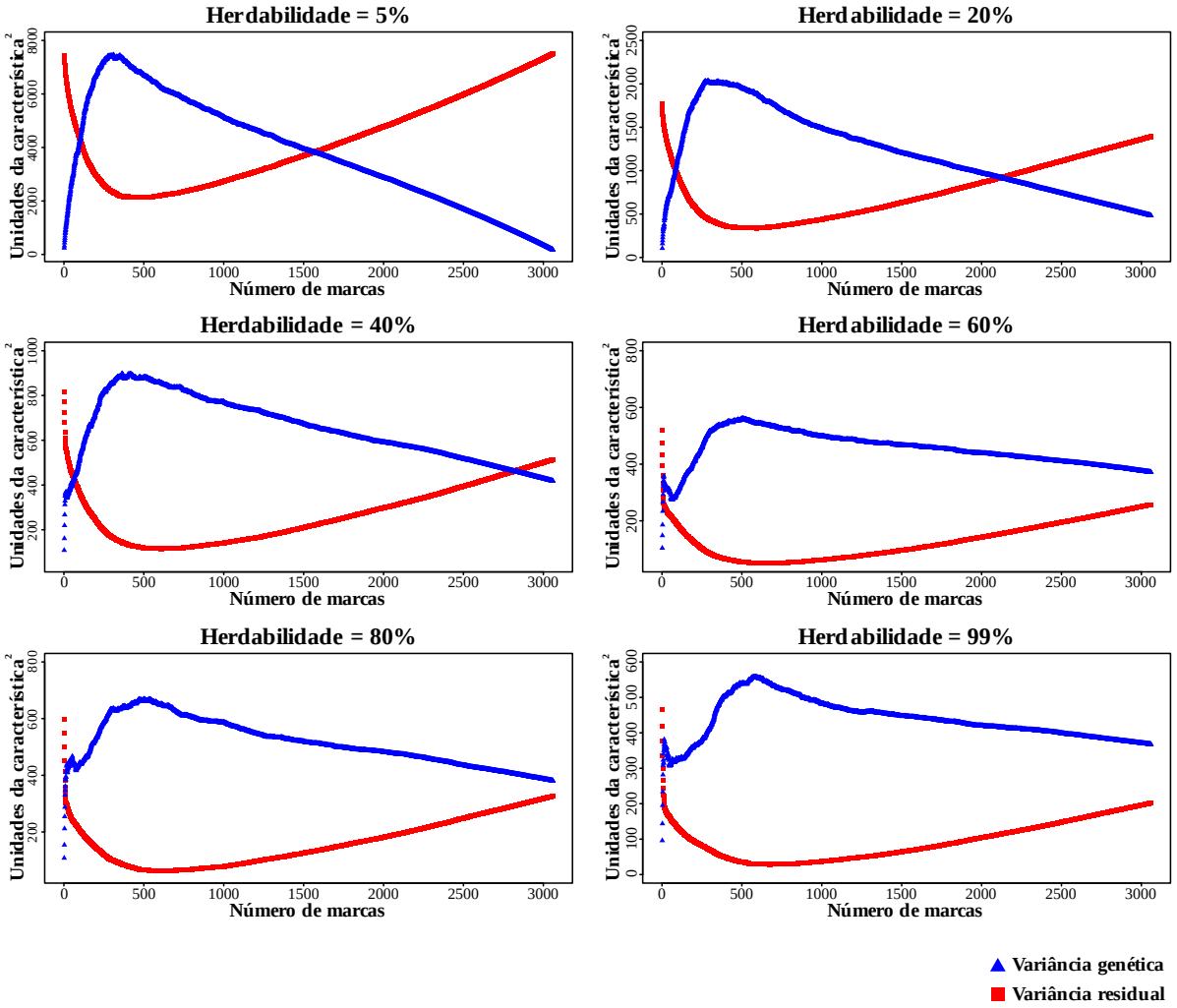

Figura 14. Tendência da variância genética e residual em função do número de marcadores utilizado para treinamento do modelo de seleção genômica para características com diferentes herdabilidades: a) 5%; b) 20%; c) 40%; d) 60%; e) 80%; f) 99%.

A variância residual apresentou uma tendência quadrática com concavidade positiva, ou seja, ela diminuiu até um determinado número de marcadores, e depois aumentou à medida que aumentou o número de marcas (Figura 14). O número ótimo de marcadores para a variância residual aumentou à medida que a herdabilidade da característica aumentou (Tabela 11).

Tabela 11. Número de marcadores (NM) para obtenção do valor ótimo (VO) dos parâmetros variância genética ($\boldsymbol{\sigma_g^2}$), variância residual ($\boldsymbol{\sigma^2}$), herdabilidade (h^2) e acurácia em uma população F_2**.**

		$\boldsymbol{\sigma_g^2}$	$\boldsymbol{\sigma^2}$	$\mathbf{h^2}$	**acurácia**
	C1	7.466	2.126	0,77	0,91
	C2	2.035	337	0,85	0,94
VO	**C3**	897	115	0,88	0,96
	C4	560	50	0,91	0,98
	C5	669	62	0,91	0,98
	C6	557	28	0,95	0,99
	C1	240-397	372-554	307-487	545-739
	C2	258-415	479-634	408-579	589-741
NM*	**C3**	332-525	518-709	489-647	693-864
	C4	410-573	602-757	522-711	704-870
	C5	428-588	532-685	500-666	639-827
	C6	486-674	585-746	557-710	729-884

*O intervalo mostrado corresponde aos 5% melhores valores para cada parâmetro avaliado. C1 a C6 – corresponde a cada característica simulada variando o valor da herdabilidade (5%, 20%, 40%, 60%, 80% e 99%).

O valor da herdabilidade estimada aumentou de forma exponencial até um número ótimo de marcadores, e depois diminuiu de forma linear (Figura 15). O número de marcadores para o ponto de máxima herdabilidade aumentou a medida que a herdabilidade simulada para cada característica aumentou (Tabela 11). O decréscimo da herdabilidade estimada com o aumento do número de marcadores foi menor para características de maior herdabilidade simulada.

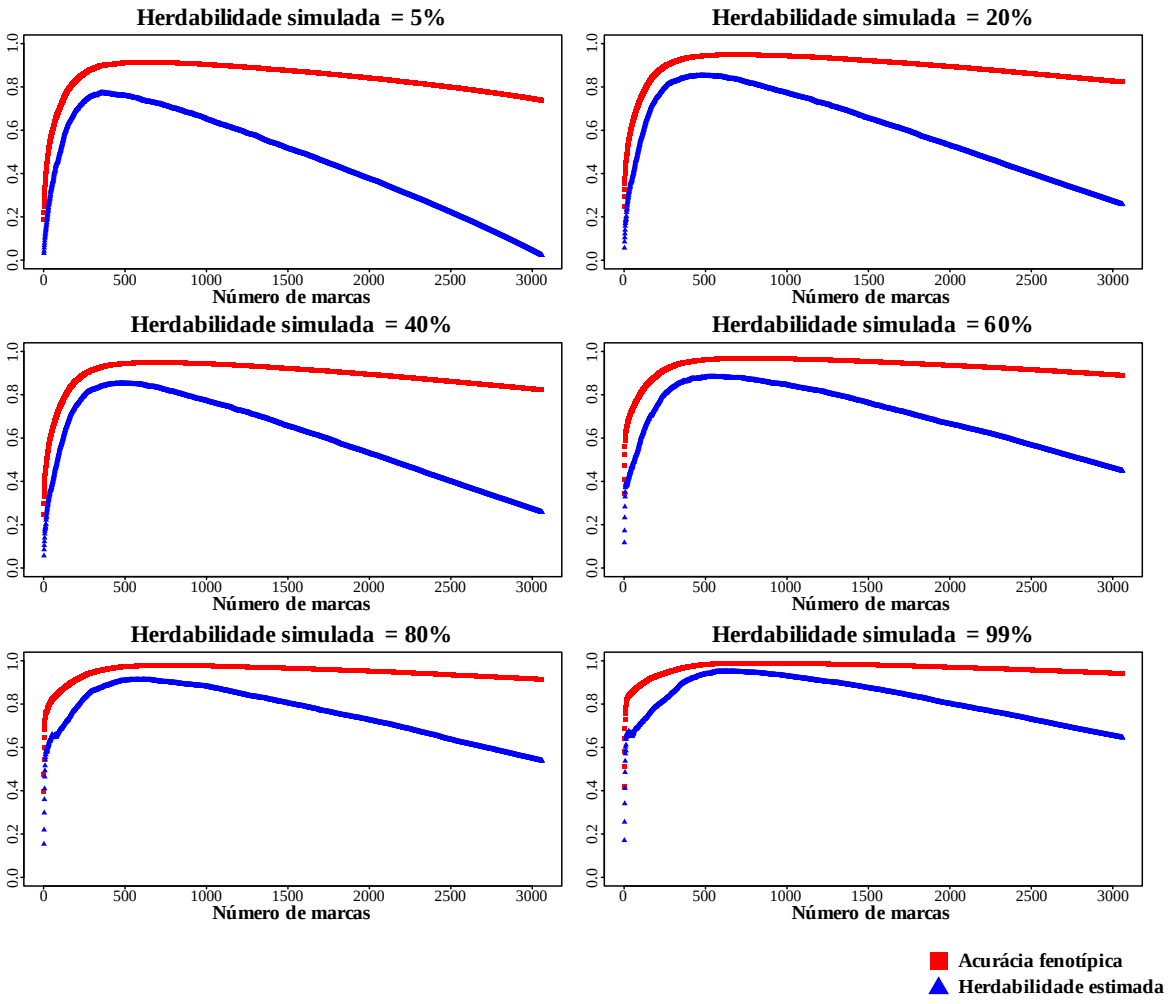

Figura 15. Tendência da acurácia fenotípica e herdabilidade em função do número de marcadores utilizado para treinamento do modelo de seleção genômica para características com diferentes herdabilidades: a) 5%; b) 20%; c) 40%; d) 60%; e) 80%; f) 99%.

A acurácia de predição da população de treinamento apresentou um incremento exponencial até um ponto de máximo e depois uma leve queda de forma linear (Figura 15). Essa queda foi menor para características de alta herdabilidade simulada. O número ótimo de marcadores para obter maior acurácia foi maior à medida que se aumentou o valor da herdabilidade da característica (Tabela 11).

5.5. DISCUSSÃO

5.5.1. Teste de segregação de marcadores

Verificou-se pela frequência alélica, frequência genica e equilíbrio de Hardy-Weinberg que a população simulada realmente representou uma população com todas as características de uma população F_2, ou seja, (A)p=(a)q=0,5, (AA)p^2 = (aa)q^2 = 0,25, e (Aa)2pq = 0,5.

A grande importância de conseguir recuperar todas as informações de uma população F_2 via o processo de simulação é que a variância genética, variância ambiental e a herdabilidade são facilmente estimadas para este tipo de população. Segundo Falconer & Mackay (1996) a variância genética em população F_2 é estimada sendo:

$$\sigma_g^2 = 2pq\alpha^2 + (2pqd)^2$$

onde o valor de d foi simulado como 0 para todos os locos, e desta forma a variância genética é igual a variância aditiva, sendo esta facilmente calculada, pois α^2 é a variância dos marcadores via o método RR-BLUP, e as frequências p e q já foram estimadas (material suplementar 1). Desta forma a herdabilidade pode ser estimada seguindo a equação proposta por Falconer & Mackay (1996) para uma população F_2:

$$h^2 = \frac{\sigma_g^2}{\left(\sigma_g^2 + \sigma^2\right)}$$

e σ^2 é a variância residual dos marcadores estimada pelo RR-BLUP.

Assim todos os parâmetros genéticos e ambientais foram calculados de forma acurada, e desta forma estes parâmetros foram critérios na escolha do melhor modelo de seleção genômica, ou seja, o modelo composto pelo número ideal de indivíduos na população de treinamento e o número de marcadores necessários para treinar o modelo de forma acurada.

5.5.2. Tamanho da população de treinamento versus valor genético estimado

Geralmente o aumento do número de indivíduos na PT aumenta a acurácia da predição do valor genético (Desta & Ortiz, 2014). Porém, apesar do aumento da acurácia, a partir de 600 indivíduos na PT esse incremento foi muito pequeno tornando-se quase nulo para características com herdabilidade de 80 e 99% no presente trabalho.

Além do número de indivíduos na PT, a estrutura da população pode influenciar na predição pelos métodos de seleção genômica. Trabalhos realizados em aveia

(Asoro et al., 2011), milho (Ogutu et al., 2012) e beterraba (Würschum et al., 2013) mostraram que a utilização de uma população estruturada em conjunto com uma PT suficientemente grande aumenta a acurácia de predição consideravelmente. Desta forma todos os resultados apresentados aqui têm validade para uma população F_2, necessitando de outros trabalhos como este avaliando os outros tipos de populações como retrocruzamentos, RILs, FMI e FIC, pois cada tipo de população tem uma estrutura diferente influenciando na predição pelos métodos de seleção genômica.

Isidro et al. (2015) avaliando várias metodologias para otimizar a escolha dos indivíduos para compor a PT verificaram que a estrutura de população e a arquitetura da característica são os fatores que mais influenciam na performance da PT. Desta forma torna-se difícil verificar um tamanho padrão da PT para as diversas herdabilidades possíveis e os diferentes tipos de população. Pelo presente trabalho realizado em F_2 podemos concluir que 600 indivíduos são suficientes independente da arquitetura da característica, porém para característica com baixa herdabilidade os valores de acurácia são mais instáveis, ou seja, dependendo dos indivíduos que estão na PT a acurácia é maior ou menor, e quando a herdabilidade vai aumentando o valor de acurácia fica constante independente dos indivíduos que constituem A PT. Este fato foi verificado neste trabalho pois as análises foram repetidas 50 vezes para cada tamanho de PT. Desta forma o conhecimento prévio da característica em estudo pode ajudar o pesquisador a delinear o experimento de forma a obter resultados acurados via seleção genômica, e consequentemente diminuir os custos com o programa de melhoramento.

5.5.3. Densidade de marcas versus valor genético estimado

Foi verificado que um número de marcas variando de 300 a 800 foi suficiente para capturar toda a variância genética de uma população F_2, e consequentemente alcançar máxima acurácia. Este valor é variável dependendo da herdabilidade da característica, pois quanto maior a herdabilidade da característica maior é o número de marcadores necessários para obter máxima acurácia. Este fato pode ser explicado devido ao efeito de cada QTL e a sua influência nos métodos de seleção genômica. Todas as características foram simuladas com 60 QTLs, porém quanto maior a herdabilidade da característica maior é o efeito de cada QTL. Uma das características dos métodos de seleção genômica é a captura de genes de menor efeito, principalmente pelo fato do método RR-BLUP utilizar mesma variância para todos os

marcadores. Isso significa que o RR-BLUP não consegue capturar todo o efeito dos QTLs de maior efeito, precisando desta forma de mais marcadores para explicar toda a variância genética da característica. Uma alternativa para melhorar a captura da variância dos QTLs de maior efeito é a utilização dos métodos bayesianos que pressupõe variância especifica para cada marcador, como o Bayes A e o Bayes B (Gianola et al., 2009).

Erbe et al. (2013) avaliaram Brown Swiss em bovinos, onde estes animais foram genotipados com chips de 777k, e observaram que a variância genética estimada pelos modelos de seleção genômica aumentou até 20k, tornando-se constante a partir deste número de marcadores. Eles concluíram que mesmo com uma população infinita de indivíduos para treinamento e um grande número de marcadores não seria possível aumentar a acurácia para esta população. Poland et al. (2012) verificaram que 1.827 SNPs foram suficientes para capturar toda variância genética em populações de trigo. Como no presente trabalho foi utilizado seleção de marcadores, ou seja, o marcador de menor efeito foi excluído a cada iteração, não houve necessidade de muitos marcadores para explicar a variância genética da característica, pois pelo processo de simulação apenas 60 marcas explicavam toda a variação da característica. Desta forma, o conhecimento prévio da característica pode ser importante para o desenvolvimento de chips de baixa densidade específicos para determinadas características ou espécies. O desenvolvimento deste tipo de chip é importante para diminuir o custo com a genotipagem. No melhoramento animal tem sido desenvolvido chips de baixa densidade para bovinos (Boichard et al., 2012; Heaton et al., 2002) e suínos (Wellmann et al., 2013). Além do mais, o custo com genotipagem de um chip de baixa densidade é muito inferior quando comparado aos chips de alta densidade (Habier et al., 2009).

Além disso, a alta acurácia verificada no presente trabalho para modelos utilizando número baixo de marcadores (200 a 900 SNPs) pode ser explicado pelo fato dos indivíduos da população de treinamento serem altamente correlacionados, devido a todos serem descendentes dos mesmos parentais, ou seja os indivíduos da população F_2 compartilham alelos idênticos por descendência (Poland & Rife, 2012). No entanto, apesar do número reduzido de marcadores para cada característica em estudo, estes marcadores são diferentes para cada uma delas, tornando a construção de um chip de baixa densidade multicaracterístico muito difícil (Habier et al., 2009). Portanto, em estudos futuros é necessário buscar estratégias que consigam fazer a

seleção de marcadores para várias características simultaneamente, e desta forma, montar um chip de baixa densidade multicaracterístico.

5.6. CONCLUSÃO

O número ideal de indivíduos para compor a população de treinamento está fortemente correlacionado com a herdabilidade da característica, porém uma população de treinamento composta por mais de 600 indivíduos garante máxima acurácia independente da herdabilidade para uma população F_2.

O modelo de seleção genômica que utiliza entre 300 e 800 marcas é suficiente para capturar toda a variância genética e diminuir a variância residual de forma a obter máxima acurácia de predição em uma população F_2.

5.7. REFERÊNCIAS BIBLIOGRÁFICAS

ALLARD, R. W., 1999. Principles of plant breeding. John Wiley & Sons, New York.

ASHRAF, M., AKRAM, N., FOOLAD, M., Marker-assisted selection in plant breeding for salinity tolerance. **Plant Salt Tolerance: Methods and Protocols**, v., p. 305-333, 2012.

ASORO, F. G., NEWELL, M. A., BEAVIS, W. D., SCOTT, M. P., JANNINK, J.-L., Accuracy and training population design for genomic selection on quantitative traits in elite North American oats. **The Plant Genome**, v. 4, p. 132-144, 2011.

BASSI, F. M., BENTLEY, A. R., CHARMET, G., ORTIZ, R., CROSSA, J., Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). **Plant Science**, v. 242, p. 23-36, 2016.

BELAJ, A., DEL CARMEN DOMINGUEZ-GARCÍA, M., ATIENZA, S. G., URDÍROZ, N. M., DE LA ROSA, R., SATOVIC, Z., MARTÍN, A., KILIAN, A., TRUJILLO, I., VALPUESTA, V., Developing a core collection of olive (*Olea europaea* L.) based on molecular markers (DArTs, SSRs, SNPs) and agronomic traits. **Tree genetics & genomes**, v. 8, p. 365-378, 2012.

BEYENE, Y., SEMAGN, K., MUGO, S., TAREKEGNE, A., BABU, R., MEISEL, B., SEHABIAGUE, P., MAKUMBI, D., MAGOROKOSHO, C., OIKEH, S., Genetic gains in grain yield through genomic selection in eight bi-parental maize populations under drought stress. **Crop Science**, v. 55, p. 154-163, 2015.

BHERING, L. L., BARRERA, C. F., ORTEGA, D., LAVIOLA, B. G., ALVES, A. A., ROSADO, T. B., CRUZ, C. D., Differential response of Jatropha genotypes to different

selection methods indicates that combined selection is more suited than other methods for rapid improvement of the species. **Industrial Crops and Products**, v. 41, p. 260-265, 2013.

BOICHARD, D., CHUNG, H., DASSONNEVILLE, R., DAVID, X., EGGEN, A., FRITZ, S., GIETZEN, K. J., HAYES, B. J., LAWLEY, C. T., SONSTEGARD, T. S., Design of a bovine low-density SNP array optimized for imputation. **PloS one**, v. 7, p. e34130, 2012.

BORÉM, A., MIRANDA, G. V., 2013. Melhoramento de Plantas. UFV, Viçosa.

CROS, D., DENIS, M., SÁNCHEZ, L., COCHARD, B., FLORI, A., DURAND-GASSELIN, T., NOUY, B., OMORÉ, A., POMIÈS, V., RIOU, V., Genomic selection prediction accuracy in a perennial crop: case study of oil palm (*Elaeis guineensis* Jacq.). **Theoretical and Applied Genetics**, v. 128, p. 397-410, 2015.

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

DE LOS CAMPOS, G., VAZQUEZ, A. I., FERNANDO, R., KLIMENTIDIS, Y. C., SORENSEN, D., Prediction of complex human traits using the genomic best linear unbiased predictor. **PLoS Genet**, v. 9, p. e1003608, 2013.

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

DIRLEWANGER, E., PRONIER, V., PARVERY, C., ROTHAN, C., GUYE, A., MONET, R., Genetic linkage map of peach [*Prunus persica* (L.) Batsch] using morphological and molecular markers. **Theoretical and Applied Genetics**, v. 97, p. 888-895, 1998.

ENDELMAN, J. B., Ridge regression and other kernels for genomic selection with R package rrBLUP. **The Plant Genome**, v. 4, p. 250-255, 2011.

ERBE, M., GREDLER, B., SEEFRIED, F. R., BAPST, B., SIMIANER, H., A function accounting for training set size and marker density to model the average accuracy of genomic prediction. **PloS one**, v. 8, p. e81046, 2013.

FALCONER, D., MACKAY, T., 1996. Introduction to Quantitative Genetics. Longman Scientific & Technical, Harlow, UK.

FRASCAROLI, E., SCHRAG, T. A., MELCHINGER, A. E., Genetic diversity analysis of elite European maize (*Zea mays* L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs. **Theoretical and applied genetics**, v. 126, p. 133-141, 2013.

GIANOLA, D., DE LOS CAMPOS, G., HILL, W. G., MANFREDI, E., FERNANDO, R., Additive genetic variability and the Bayesian alphabet. **Genetics**, v. 183, p. 347-363, 2009.

GOUY, M., ROUSSELLE, Y., BASTIANELLI, D., LECOMTE, P., BONNAL, L., ROQUES, D., EFILE, J.-C., ROCHER, S., DAUGROIS, J., TOUBI, L., Experimental assessment of the accuracy of genomic selection in sugarcane. **Theoretical and applied genetics**, v. 126, p. 2575-2586, 2013.

HABIER, D., FERNANDO, R. L., DEKKERS, J. C., Genomic selection using low-density marker panels. **Genetics**, v. 182, p. 343-353, 2009.

HE, J., ZHAO, X., LAROCHE, A., LU, Z.-X., LIU, H., LI, Z., Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. **Front Plant Sci**, v. 5, p. 484, 2014.

HEATON, M. P., HARHAY, G. P., BENNETT, G. L., STONE, R. T., GROSSE, W. M., CASAS, E., KEELE, J. W., SMITH, T. P., CHITKO-MCKOWN, C. G., LAEGREID, W. W., Selection and use of SNP markers for animal identification and paternity analysis in US beef cattle. **Mammalian Genome**, v. 13, p. 272-281, 2002.

ISIDRO, J., JANNINK, J.-L., AKDEMIR, D., POLAND, J., HESLOT, N., SORRELLS, M. E., Training set optimization under population structure in genomic selection. **Theoretical and Applied Genetics**, v. 128, p. 145-158, 2015.

LANGER, M., MAIXNER, M., Molecular characterisation of grapevine yellows associated phytoplasmas of the stolbur-group based on RFLP-analysis of non-ribosomal DNA. **VITIS-Journal of Grapevine Research**, v. 43, p. 191, 2015.

LIGHTFOOT, D. A., Two Decades of Molecular Marker-Assisted Breeding for Resistance to Soybean Sudden Death Syndrome. **Crop Science**, v. 55, p. 1460-1484, 2015.

LYNCH, M., MILLIGAN, B. G., Analysis of population genetic structure with RAPD markers. **Molecular ecology**, v. 3, p. 91-99, 1994.

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

OGUTU, J. O., SCHULZ-STREECK, T., PIEPHO, H.-P., 2012. Genomic selection using regularized linear regression models: ridge regression, lasso, elastic net and their extensions, BMC proceedings. BioMed Central Ltd, p. S10.

ORDAS, B., BUTRON, A., ALVAREZ, A., REVILLA, P., MALVAR, R., Comparison of two methods of reciprocal recurrent selection in maize (*Zea mays* L.). **Theoretical and Applied Genetics**, v. 124, p. 1183-1191, 2012.

PANDEY, M. K., RANI, N. S., SUNDARAM, R. M., LAHA, G. S., MADHAV, M. S., RAO, K. S., SUDHARSHAN, I., HARI, Y., VARAPRASAD, G. S., RAO, L. V. S., Improvement of two traditional Basmati rice varieties for bacterial blight resistance and plant stature through morphological and marker-assisted selection. **Molecular breeding**, v. 31, p. 239-246, 2013.

POLAND, J., ENDELMAN, J., DAWSON, J., RUTKOSKI, J., WU, S., MANES, Y., DREISIGACKER, S., CROSSA, J., SÁNCHEZ-VILLEDA, H., SORRELLS, M., Genomic selection in wheat breeding using genotyping-by-sequencing. **The Plant Genome**, v. 5, p. 103-113, 2012.

POLAND, J. A., RIFE, T. W., Genotyping-by-sequencing for plant breeding and genetics. **The Plant Genome**, v. 5, p. 92-102, 2012.

REN, R., RAY, R., LI, P., XU, J., ZHANG, M., LIU, G., YAO, X., KILIAN, A., YANG, X., Construction of a high-density DArTseq SNP-based genetic map and identification of genomic regions with segregation distortion in a genetic population derived from a cross between feral and cultivated-type watermelon. **Molecular Genetics and Genomics**, v. 290, p. 1457-1470, 2015.

SOLDATI, M. C., FORNES, L., VAN ZONNEVELD, M., THOMAS, E., ZELENER, N., An assessment of the genetic diversity of *Cedrela balansae* C. DC.(Meliaceae) in Northwestern Argentina by means of combined use of SSR and AFLP molecular markers. **Biochemical Systematics and Ecology**, v. 47, p. 45-55, 2013.

SPINDEL, J., BEGUM, H., AKDEMIR, D., VIRK, P., COLLARD, B., REDOÑA, E., ATLIN, G., JANNINK, J.-L., MCCOUCH, S. R., Correction: Genomic Selection and Association Mapping in Rice (*Oryza sativa*): Effect of Trait Genetic Architecture, Training Population Composition, Marker Number and Statistical Model on Accuracy of Rice Genomic Selection in Elite, Tropical Rice Breeding Lines. **PLoS Genet**, v. 11, p. e1005350, 2015.

TANG, B., JENKINS, J., WATSON, C., MCCARTY, J., CREECH, R., Evaluation of genetic variances, heritabilities, and correlations for yield and fiber traits among cotton F2 hybrid populations. **Euphytica**, v. 91, p. 315-322, 1996.

TANG, B., JENKINS, J. N., MCCARTY, J., WATSON, C., F2 hybrids of host plant germplasm and cotton cultivars: II. Heterosis and combining ability for fiber properties. **Crop science**, v. 33, p. 706-710, 1993.

TEAM, R. C., 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Austria, Vienna.

WELLMANN, R., PREUß, S., THOLEN, E., HEINKEL, J., WIMMERS, K., BENNEWITZ, J., Genomic selection using low density marker panels with application to a sire line in pigs. **Genet Sel Evol**, v. 45, p. 28, 2013.

WÜRSCHUM, T., REIF, J. C., KRAFT, T., JANSSEN, G., ZHAO, Y., Genomic selection in sugar beet breeding populations. **Bmc Genetics**, v. 14, p. 85, 2013.

YANIV, E., RAATS, D., RONIN, Y., KOROL, A. B., GRAMA, A., BARIANA, H., DUBCOVSKY, J., SCHULMAN, A. H., FAHIMA, T., Evaluation of marker-assisted selection for the stripe rust resistance gene Yr15, introgressed from wild emmer wheat. **Molecular Breeding**, v. 35, p. 1-12, 2015.

ZHANG, J., SONG, Q., CREGAN, P. B., JIANG, G.-L., Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). **Theoretical and Applied Genetics**, v. 129, p. 117-130, 2016.

6. CONCLUSÕES GERAIS

1) A utilização de um modelo de SG com as marcas significativas encontradas pelo EAGA como efeito fixo e as demais marcas como efeito aleatório é uma boa estratégia para selecionar indivíduos superiores com alta acurácia;

2) A introdução no modelo de SG de QTLs que já foram descritos previamente para a característica em estudo, como efeito fixo, permite a seleção de indivíduos superiores de forma mais acurada;

3) Os modelos de seleção genômica para predição em populações F_2 devem ser compostos por 300 a 800 marcadores de maior efeito sobre a característica e mais de 600 indivíduos na população de treinamento.

ANEXO 1. Códigos do R utilizados no capítulo 1 da tese.

```
## Códigos para rodar o modelo 1
setwd("G:\\doutorado\\tese\\chapter1\\Dados")

##Lendo os arquivos de genótipos e fenótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}

gen<-read.table("gen.txt", head=T)
pheno<-data.frame(fen)

##carregando pacotes
library(qtl)
library(BGLR)

##Lendo arquivos para serem utilizados no pacote qtl
dat1 <- read.cross("csvr", dir="F:\\doutorado\\tese\\chapter1\\Dados",
                   file="MASh_5.csv")

data<-calc.genoprob(dat1)
###Rodando mapeamento por intervalo composto
result<-cim(data, pheno.col=2, n.marcovar=3, window=10,
    method=c("em"),error.prob=0.0001,
    map.function=c("morgan"))

##Selecionando SNPs de maior efeito (LOD>3)
selSNP<-result[result$lod>3, c(1:3)]
result1<-data.frame(selSNP)
names<-as.matrix(rownames(selSNP))
```

```
colnames(names)<-c("snp")
selSNP1<-cbind(names, result1)
selSNP1$snp<-as.character(selSNP1$snp)

###mapeamento por interval composto sobre o valor genético
result_1<-cim(data, pheno.col=1, n.marcovar=3, window=10,
        method=c("em"),error.prob=0.0001,
        map.function=c("morgan"))

##Selecionando SNPs de alto efeito baseado no valor genético (LOD>3)
selSNP<-result_1[result_1$lod>3, c(1:3)]
result_2<-data.frame(selSNP)
names<-as.matrix(rownames(selSNP))
colnames(names)<-c("snp")
selSNP_1<-cbind(names, result_2)
selSNP_1$snp<-as.character(selSNP_1$snp)

##Lendo arquivo de cromossomo
chro<-read.table("pos.txt", h=FALSE)

gen1<-data.frame(cbind(chro,gen))
colnames(gen1)<-c("snp", "chro", "pos", colnames(gen))

##Criando a matriz de efeito fixo
fixedeffect<-merge(x=selSNP1,y=gen1, by="snp")
marker<-t(fixedeffect[,7:ncol(fixedeffect)])
colnames(marker)<-selSNP1$snp

##Calculando a frequência alélica
library(HapEstXXR)

##Criando uma matriz codificada para o pacote HapEstXXR
geno = matrix(0,nrow=nrow(marker),ncol=ncol(marker))
idx = marker == 1
```

```
geno[idx] = 2
idx = marker == 0
geno[idx] = 3
idx = marker == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)

##Rodando a função maf que calcula menor frequ~encia alélica, call rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin, marker.label=colnames(marker)))

freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

### modelo: y = média + SNPs significativos (efeito fixo) + erro

##Rodando o modelo de seleção genômica
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
  sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
```

```
for (i in 1:folds)
{
  timei<- proc.time() #initial time

  fixedf1<-marker[sample,]
  phen=data.frame(pheno[sample,3])
  pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
  pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                          pop1, pop2, pop3, pop4, pop5))

  phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
  phentest<-as.matrix(popfin[,i+4])

  GEN=data.frame(pheno[sample,2])
  popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
  popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
                           popg1, popg2, popg3, popg4, popg5))

  gentest<-as.matrix(popgfin[,i+4])

  ##criando a matriz de efeito fixo
  fe1<-fixedf1[1:((1/5)*nrow(pheno)),]
  fe2<-fixedf1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
  fe3<-fixedf1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
```

```
fe4<-fixedf1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
fe5<-fixedf1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  fixedeffect3=rbind(fe1, fe2, fe3, fe4)# markers matrix
  fixedeffectval=fe5
}
if (i==2){
  fixedeffect3=rbind(fe2, fe3, fe4, fe5)# markers matrix
  fixedeffectval=fe1
}
if (i==3){
  fixedeffect3=rbind(fe3, fe4, fe5, fe1)# markers matrix
  fixedeffectval=fe2
}
if (i==4){
  fixedeffect3=rbind(fe4, fe5, fe1, fe2)# markers matrix
  fixedeffectval=fe3
}
if (i==5){
  fixedeffect3=rbind(fe5, fe1, fe2, fe3)# markers matrix
  fixedeffectval=fe4
}

ETA=list(list(X=fixedeffect3, model="FIXED"))
BRR = BGLR(y=phentrain, response_type= "gaussian",
           ETA=ETA,
           nIter=100000, burnIn=20000, thin=10,
           saveAt = "BRR")

## Estimando o valor genético genômico
F_effect<-as.matrix(BRR$ETA[[1]]$b)
M<-as.matrix(fixedeffectval)
GEBVval<-M%*%F_effect
```

```
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(fixedeffectval)
GEBV1<-mean(phentest) + GEBVval

##Estimando a herdabilidade
sig2a=mean(BRR$ETA[[1]]$varB) #marker variance
sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
sig2e=mean(BRR$varE) #residual variance
h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability

## Estimando a acurácia e o ganho com a seleção
result<-cbind(phentest, GEBV1, gentest)
rownames(result)<-seq(1,200,1)

q1<-quantile(result[,2], probs =.20)
top20<-result[result[,2]<q1, c(1:2)]

q3<-quantile(result[,1], probs =.20)
top20phen<-result[result[,1]<q3, c(1:2)]

# Calculando o coeficiente de coincidência entre os indivíduos selecionados
coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
mp<-mean(result[,2]) #mean population
ms20<-mean(top20[,2]) #mean selected individual
GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain

phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
genpredictionability[j,i]<-cor(result[,3], result[,2])
phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
timef<- proc.time() #final time
timecal<-as.matrix(timef-timei)
t<-as.numeric(timecal[3,])
```

```
    time[j,i]<-t
  }
}

MAS_final_results<-rbind(mean(phenaccuracy), mean(genaccuracy), mean(time),
                         mean(sig2uval), mean(h2_rval), mean(GS20),
                         mean(co20), mean(phenpredictionability),
                         mean(genpredictionability))
colnames(MAS_final_results)<-cbind("Genetic parameters")
rownames(MAS_final_results)<-rbind("Phenotipic Pearson Correlation",
                                   "Genotipic Pearson Correlation", "Time",
                                   "Genetic variance", "Heritability", "Genetic Gain 20%",
                                   "Coincidence 20%", "phenpredictionability",
                                   "genpredictionability")

write.table(sig2uval,"h5_MAS_sig2uval.txt", row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_MAS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_MAS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_MAS_h2_rval.txt", row.names=TRUE,quote=FALSE)
write.table(GS20,"h5_MAS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_MAS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_MAS_time.txt", row.names=TRUE,quote=FALSE)
write.table(selSNP1,"h5_MAS_selSNP1.txt", row.names=TRUE,quote=FALSE)
write.table(selSNP_1,"h5_MAS_selSNP_1.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_MAS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(genpredictionability,"h5_MAS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(MAS_final_results,"h5_MAS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)

## Códigos para rodar o modelo 2
```

```
setwd("C:\\Users\\Leonardo\\Google Drive\\tese\\chapter1\\MAS+GWS")
##Carregando pacotes
library(rrBLUP)
library(BGLR)

##Lendo os arquivos de genótipos e fenótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
#function to read all files together
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}

gen<-read.table("gen.txt", head=T)

##Calculando a frequência alélica
library(HapEstXXR)

geno = matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx = gen == 1
geno[idx] = 2
idx = gen == 0
geno[idx] = 3
idx = gen == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)

##Rodando a função maf que calcula a frequência alélica, Call-rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin))
```

```
freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

##Trocando os códigos para rodar a função GWAS
##2=1
##1=0
##0=-1
geno4 = matrix(0,nrow=nrow(genfin),ncol=ncol(genfin))
idx = genfin == 0
geno4[idx] = -1
idx = genfin == 1
geno4[idx] = 0
idx = genfin == 2
geno4[idx] = 1

##Lendo o arquivo de cromossomos
chro<-read.table("pos.txt", h=FALSE)

##Criando o arquivo de genótipos
geno5<-data.frame(cbind(chro,t(geno4)))
colnames(geno5)<-c("snp", "chro", "pos", pheno[,1])

##Rodando GWAS
pheno<-data.frame(fen)
gwas<-GWAS(pheno, geno5, min.MAF=0.05)

selSNP<-gwas[gwas$phen>3, c(1:5)]
selSNP_1<-gwas[gwas$gen>3, c(1:5)]

#Criando uma matriz com os SNPs selecioandos pelo GWAS
geno3<-merge(x=selSNP, y=geno5, by = "snp")
geno4<-t(geno3[,8:1007])
colnames(geno4)<-geno3[,1]
```

```
rownames(geno4)<-rownames (genfin)

### modelo: y = média + SNPs significativos (efeito fixo) + erro

##Rodando o modelo de seleção genômica
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
 sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
 for (i in 1:folds)
 {
  timei<- proc.time() #initial time

  fixedf1<-geno4[sanmple,]
  phen=data.frame(pheno[sanmple,3])
  pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
  pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                          pop1, pop2, pop3, pop4, pop5))
```

```
phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
phentest<-as.matrix(popfin[,i+4])

GEN=data.frame(pheno[sample,2])
popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
                         popg1, popg2, popg3, popg4, popg5))

gentest<-as.matrix(popgfin[,i+4])

##Criando a matriz de efeito fixos
fe1<-fixedf1[1:((1/5)*nrow(pheno)),]
fe2<-fixedf1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
fe3<-fixedf1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
fe4<-fixedf1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
fe5<-fixedf1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  fixedeffect3=rbind(fe1, fe2, fe3, fe4)# markers matrix
  fixedeffectval=fe5
}
if (i==2){
  fixedeffect3=rbind(fe2, fe3, fe4, fe5)# markers matrix
  fixedeffectval=fe1
}
if (i==3){
  fixedeffect3=rbind(fe3, fe4, fe5, fe1)# markers matrix
  fixedeffectval=fe2
```

```
}
if (i==4){
  fixedeffect3=rbind(fe4, fe5, fe1, fe2)# markers matrix
  fixedeffectval=fe3
}
if (i==5){
  fixedeffect3=rbind(fe5, fe1, fe2, fe3)# markers matrix
  fixedeffectval=fe4
}

ETA=list(list(X=fixedeffect3, model="FIXED"))
BRR = BGLR(y=phentrain, response_type= "gaussian",
           ETA=ETA,
           nIter=100000, burnIn=20000, thin=10,
           saveAt = "BRR")

## Calculando o valor genético genômico
F_effect<-as.matrix(BRR$ETA[[1]]$b)
M<-as.matrix(fixedeffectval)
GEBVval<-M%*%F_effect
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(fixedeffectval)
GEBV1<-mean(phentest) + GEBVval

##Estimando a herdabilidade genômica
sig2a=mean(BRR$ETA[[1]]$varB) #marker variance
sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
sig2e=mean(BRR$varE) #residual variance
h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability

## Estimando o ganho de seleção e a acurácia
result<-cbind(phentest, GEBV1, gentest)
rownames(result)<-seq(1,200,1)
```

```
    q1<-quantile(result[,2], probs =.20)
    top20<-result[result[,2]<q1, c(1:2)]

    q3<-quantile(result[,1], probs =.20)
    top20phen<-result[result[,1]<q3, c(1:2)]

    #Calculando o índice de coincidência entre os indivíduos selecioandos
    coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
    co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
    mp<-mean(result[,2]) #mean population
    ms20<-mean(top20[,2]) #mean selected individual
    GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain

    phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
    genpredictionability[j,i]<-cor(result[,3], result[,2])
    phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
    genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
    timef<- proc.time() #final time
    timecal<-as.matrix(timef-timei)
    t<-as.numeric(timecal[3,])
    time[j,i]<-t
  }
}

GWAS_final_results<-rbind(mean(phenaccuracy), mean(genaccuracy), mean(time),
                          mean(sig2uval), mean(h2_rval), mean(GS20),
                          mean(co20), mean(phenpredictionability),
                          mean(genpredictionability))
colnames(GWAS_final_results)<-cbind("Genetic parameters")
rownames(GWAS_final_results)<-rbind("Phenotipic Pearson Correlation",
                                    "Genotipic Pearson Correlation", "Time",
                                    "Genetic variance", "Heritability", "Genetic Gain 20%",
                                    "Coincidence 20%", "phenpredictionability",
                                    "genpredictionability")
```

```
write.table(sig2uval,"h5_GWAS_sig2uval.txt", row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_GWAS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_GWAS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_GWAS_h2_rval.txt", row.names=TRUE,quote=FALSE)
write.table(GS20,"h5_GWAS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_GWAS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_GWAS_time.txt", row.names=TRUE,quote=FALSE)
write.table(selSNP1,"h5_GWAS_selSNP.txt", row.names=TRUE,quote=FALSE)
write.table(selSNP_1,"h5_GWAS_selSNP_1.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_GWAS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(genpredictionability,"h5_GWAS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(GWAS_final_results,"h5_GWAS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)

## Códigos para rodar o modelo 3
setwd("F:\\doutorado\\tese\\chapter1\\Dados")

##Carregando pacotes
library(BGLR)

##Lendo arquivos de fenótipos e genótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}
```

```
gen<-read.table("gen.txt", head=T)
pheno<-data.frame(fen)

## Calculando a frequência alélica
library(HapEstXXR)

## Criando uma matriz codificada para o pacote HapEstXXR
geno = matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx = gen == 1
geno[idx] = 2
idx = gen == 0
geno[idx] = 3
idx = gen == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)

##Rodando a função maf que calcula menor frequ~encia alélica, call rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin))

freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

### modelo: y = média + SNPs (efeito aleatório) + erro

##Rodando o modelo de seleção genômica
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
```

```
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
 sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
 for (i in 1:folds)
 {
  timei<- proc.time() #initial time
  markers1<-gen[sample,]
  phen=data.frame(pheno[sample,3])
  pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
  pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                          pop1, pop2, pop3, pop4, pop5))

  phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
  phentest<-as.matrix(popfin[,i+4])

  GEN=data.frame(pheno[sample,2])
  popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
  popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
```

```
                popg1, popg2, popg3, popg4, popg5))

gentest<-as.matrix(popgfin[,i+4])

##Criando a matriz de efeitos aleatórios
M1<-markers1[1:((1/5)*nrow(pheno)),]
M2<-markers1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
M3<-markers1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
M4<-markers1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
M5<-markers1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  markerstrain=rbind(M1, M2, M3, M4)# markers matrix
  markersval=M5
}
if (i==2){
  markerstrain=rbind(M2, M3, M4, M5)# markers matrix
  markersval=M1
}
if (i==3){
  markerstrain=rbind(M3, M4, M5, M1)# markers matrix
  markersval=M2
}
if (i==4){
  markerstrain=rbind(M4, M5, M1, M2)# markers matrix
  markersval=M3
}
if (i==5){
  markerstrain=rbind(M5, M1, M2, M3)# markers matrix
  markersval=M4
}
ETA=list(list(X=markerstrain, model = "BRR"))
BRR = BGLR(y=phentrain, response_type= "gaussian",
           ETA=ETA,
```

```
        nIter=100000, burnIn=20000, thin=10,
        saveAt = "BRR")

## Estimando o valor genético genômico
R_effect<-as.matrix(BRR$ETA[[1]]$b)
M1<-as.matrix(markersval)
GEBVval<-M1%*%R_effect
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(markersval)
GEBV1<-mean(phentest) + GEBVval

## Estimando a herdabilidade
sig2a=mean(BRR$ETA[[1]]$varB) #marker variance
sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
sig2e=mean(BRR$varE) #residual variance
h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability

## Estimando a acurácia e o ganho com a seleção
result<-cbind(phentest, GEBV1, gentest)
rownames(result)<-seq(1,200,1)

q1<-quantile(result[,2], probs =.20)
top20<-result[result[,2]<q1, c(1:2)]

q3<-quantile(result[,1], probs =.20)
top20phen<-result[result[,1]<q3, c(1:2)]

# Calculando o coeficiente de coincidência entre os indivíduos selecionados
coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
mp<-mean(result[,2]) #mean population
ms20<-mean(top20[,2]) #mean selected individual
GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain
```

```
    phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
    genpredictionability[j,i]<-cor(result[,3], result[,2])
    phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
    genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
    timef<- proc.time() #final time
    timecal<-as.matrix(timef-timei)
    t<-as.numeric(timecal[3,])
    time[j,i]<-t
  }
}

GWS_final_results<-rbind(mean(phenaccuracy), mean(genaccuracy), mean(time),
                         mean(sig2uval), mean(h2_rval), mean(GS20),
                         mean(co20), mean(phenpredictionability),
                         mean(genpredictionability))
colnames(GWS_final_results)<-cbind("Genetic parameters")
rownames(GWS_final_results)<-rbind("Phenotipic Pearson Correlation",
                                   "Genotipic Pearson Correlation", "Time",
                                   "Genetic variance", "Heritability", "Genetic Gain 20%",
                                   "Coincidence 20%", "phenpredictionability",
                                   "genpredictionability")

write.table(sig2uval,"h5_GWS_sig2uval.txt", row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_GWS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_GWS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_GWS_h2_rval.txt", row.names=TRUE,quote=FALSE)
write.table(GS20,"h5_GWS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_GWS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_GWS_time.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_GWS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
```

```
write.table(genpredictionability,"h5_GWS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(GWS_final_results,"h5_GWS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)

## Códigos para rodar o modelo 4
setwd("D:\\Leonardo\\Capitulo 1")

##Lendo os arquivos de genótipos e fenótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
#function to read all files together
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}

gen<-read.table("gen.txt", head=T)
pheno<-data.frame(fen)

##carregando pacotes
library(qtl)
library(BGLR)

##Lendo arquivos para serem utilizados no pacote qtl
dat1 <- read.cross("csvr", dir="D:\\Leonardo\\Capitulo 1",
                   file="MASh_5.csv")

data<-calc.genoprob(dat1)
###Rodando mapeamento por intervalo composto
result<-cim(data, pheno.col=2, n.marcovar=3, window=10,
            method=c("em"),error.prob=0.0001,
            map.function=c("morgan"))
```

```
##Selecionando SNPs de maior efeito (LOD>3)
selSNP<-result[result$lod>3, c(1:3)]
result1<-data.frame(selSNP)
names<-as.matrix(rownames(selSNP))
colnames(names)<-c("snp")
selSNP1<-cbind(names, result1)
selSNP1$snp<-as.character(selSNP1$snp)

##Lendo arquivo de cromossomo
chro<-read.table("pos.txt", h=F)
gen1<-data.frame(cbind(chro,t(gen)))
colnames(gen1)<-c("snp", "chro", "pos", rownames(gen))

##Criando a matriz de efeito fixo
fixedeffect<-merge(x=selSNP1,y=gen1, by="snp")
fixedeffect1<-t(fixedeffect[,7:ncol(fixedeffect)])
colnames(fixedeffect1)<-selSNP1$snp
snpnames<-as.integer(colnames(fixedeffect1))
gennew<-gen[,-snpnames]

##Calculando a frequência alélica
library(HapEstXXR)

##Criando uma matriz codificada para o pacote HapEstXXR
geno = matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx = gen == 1
geno[idx] = 2
idx = gen == 0
geno[idx] = 3
idx = gen == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)
```

```
##Rodando a função maf que calcula menor frequ~encia alélica, call rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin))

freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

### modelo: y = média + SNPs significativos (efeito fixo) + SNPs não significativos (efeito aleatório) + erro

##Rodando o modelo de seleção genômica
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
 sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
 for (i in 1:folds)
 {
  timei<- proc.time() #initial time

  fixedf1<-fixedeffect1[sample,]
  markers1<-gennew[sample,]
```

```
phen=data.frame(pheno[sanmple,3])
pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                        pop1, pop2, pop3, pop4, pop5))

phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
phentest<-as.matrix(popfin[,i+4])

GEN=data.frame(pheno[sample,2])
popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
                         popg1, popg2, popg3, popg4, popg5))

gentest<-as.matrix(popgfin[,i+4])

##Criando a matriz de efeito fixo
fe1<-fixedf1[1:((1/5)*nrow(pheno)),]
fe2<-fixedf1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
fe3<-fixedf1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
fe4<-fixedf1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
fe5<-fixedf1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  fixedeffect3=rbind(fe1, fe2, fe3, fe4)# markers matrix
```

```
  fixedeffectval=fe5
}
if (i==2){
  fixedeffect3=rbind(fe2, fe3, fe4, fe5)# markers matrix
  fixedeffectval=fe1
}
if (i==3){
  fixedeffect3=rbind(fe3, fe4, fe5, fe1)# markers matrix
  fixedeffectval=fe2
}
if (i==4){
  fixedeffect3=rbind(fe4, fe5, fe1, fe2)# markers matrix
  fixedeffectval=fe3
}
if (i==5){
  fixedeffect3=rbind(fe5, fe1, fe2, fe3)# markers matrix
  fixedeffectval=fe4
}

##Criando a matriz de efeito aleatório
M1<-markers1[1:((1/5)*nrow(pheno)),]
M2<-markers1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
M3<-markers1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
M4<-markers1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
M5<-markers1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  markerstrain=rbind(M1, M2, M3, M4)# markers matrix
  markersval=M5
}
if (i==2){
  markerstrain=rbind(M2, M3, M4, M5)# markers matrix
  markersval=M1
}
```

```
if (i==3){
  markerstrain=rbind(M3, M4, M5, M1)# markers matrix
  markersval=M2
}
if (i==4){
  markerstrain=rbind(M4, M5, M1, M2)# markers matrix
  markersval=M3
}
if (i==5){
  markerstrain=rbind(M5, M1, M2, M3)# markers matrix
  markersval=M4
}
ETA=list(list(X=fixedeffect3, model="FIXED"),
         list(X=markerstrain, model = "BRR"))
BRR = BGLR(y=phentrain, response_type= "gaussian",
           ETA=ETA,
           nIter=100000, burnIn=20000, thin=10,
           saveAt = "BRR")

## Estimando o valor genético genômico
F_effect<-as.matrix(BRR$ETA[[1]]$b)
R_effect<-as.matrix(BRR$ETA[[2]]$b)
M<-as.matrix(fixedeffectval)
M1<-as.matrix(markersval)
GEBVval<-M%*%F_effect + M1%*%R_effect
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(fixedeffectval)
GEBV1<-mean(phentest) + GEBVval

## Estimando a herdabilidade
sig2a=mean(BRR$ETA[[2]]$varB) #marker variance
sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
sig2e=mean(BRR$varE) #residual variance
h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability
```

```
  ## Estimando a acurácia e o ganho com a seleção
  result<-cbind(phentest, GEBV1, gentest)
  rownames(result)<-seq(1,200,1)

  q1<-quantile(result[,2], probs =.20)
  top20<-result[result[,2]<q1, c(1:2)]

  q3<-quantile(result[,1], probs =.20)
  top20phen<-result[result[,1]<q3, c(1:2)]

  # Calculando o coeficiente de coincidência entre os indivíduos selecionados
  coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
  co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
  mp<-mean(result[,2]) #mean population
  ms20<-mean(top20[,2]) #mean selected individual
  GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain

  phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
  genpredictionability[j,i]<-cor(result[,3], result[,2])
  phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
  genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
  timef<- proc.time() #final time
  timecal<-as.matrix(timef-timei)
  t<-as.numeric(timecal[3,])
  time[j,i]<-t
 }
}

MAS+GWS_final_results<-rbind(mean(phenaccuracy),          mean(genaccuracy),
mean(time),
                  mean(sig2uval), mean(h2_rval), mean(GS20),
                  mean(co20), mean(phenpredictionability),
                  mean(genpredictionability))
```

```
colnames(MAS+GWS_final_results)<-cbind("Genetic parameters")
rownames(MAS+GWS_final_results)<-rbind("Phenotipic Pearson Correlation",
                        "Genotipic Pearson Correlation", "Time",
                        "Genetic variance", "Heritability", "Genetic Gain 20%",
                        "Coincidence 20%", "phenpredictionability",
                        "genpredictionability")

write.table(sig2uval,"h5_MAS+GWS_sig2uval.txt", row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_MAS+GWS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_MAS+GWS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_MAS+GWS_h2_rval.txt", row.names=TRUE,quote=FALSE)
write.table(GS20,"h5_MAS+GWS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_MAS+GWS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_MAS+GWS_time.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_MAS+GWS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(genpredictionability,"h5_MAS+GWS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(MAS+GWS_final_results,"h5_MAS+GWS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)

## Códigos para rodar o modelo 5
setwd("F:\\doutorado\\tese\\chapter1\\Dados")

##Carregando pacotes
library(rrBLUP)
library(BGLR)

##Lendo os arquivos de genótipos e fenótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
```

```
#function to read all files together
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}

gen<-read.table("gen.txt", head=T)
pheno<-data.frame(fen)

##Calculando a frequência alélica
library(HapEstXXR)

geno = matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx = gen == 1
geno[idx] = 2
idx = gen == 0
geno[idx] = 3
idx = gen == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)

##Rodando a função maf que calcula a frequência alélica, Call-rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin))

## Calculate de frequency to allele A, allele a
## and A x a
## This file will be used to estimate genetic variance
freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

##Trocando os códigos para rodar a função GWAS
```

```
##2=1
##1=0
##0=-1
geno4 = matrix(0,nrow=nrow(genfin),ncol=ncol(genfin))
idx = genfin == 0
geno4[idx] = -1
idx = genfin == 1
geno4[idx] = 0
idx = genfin == 2
geno4[idx] = 1

##Lendo o arquivo de cromossomos
chro<-read.table("pos.txt", h=FALSE)

##Criando o arquivo de genótipos
geno5<-data.frame(cbind(chro,t(geno4)))
colnames(geno5)<-c("snp", "chro", "pos", pheno[,1])

##Rodando GWAS
gwas<-GWAS(pheno, geno5, min.MAF=0.05)

selSNP<-gwas[gwas$vfen>3, c(1:5)]

#Criando uma matriz com os SNPs selecioandos pelo GWAS
geno3<-merge(x=selSNP, y=geno5, by = "snp")
geno4<-t(geno3[,8:1007])
colnames(geno4)<-geno3[,1]
rownames(geno4)<-rownames (genfin)
snpnames<-as.integer(colnames(geno4))
gennew<-gen[,-snpnames]
### modelo: y = média + SNPs significativos (efeito fixo) + SNPs não significaticos (efeito aleatório) + erro

##Rodando o modelo de seleção genômica
```

```
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
 sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
 for (i in 1:folds)
 {
  timei<- proc.time() #initial time
  markers1<-gennew[sample,]
  fixedf1<-geno4[sample,]
  phen=data.frame(pheno[sample,3])
  pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
  pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                          pop1, pop2, pop3, pop4, pop5))

  phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
  phentest<-as.matrix(popfin[,i+4])

  GEN=data.frame(pheno[sample,2])
```

```
popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
                         popg1, popg2, popg3, popg4, popg5))

gentest<-as.matrix(popgfin[,i+4])

##Criando a matriz de efeito fixos
fe1<-fixedf1[1:((1/5)*nrow(pheno)),]
fe2<-fixedf1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
fe3<-fixedf1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
fe4<-fixedf1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
fe5<-fixedf1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  fixedeffect3=rbind(fe1, fe2, fe3, fe4)# markers matrix
  fixedeffectval=fe5
}
if (i==2){
  fixedeffect3=rbind(fe2, fe3, fe4, fe5)# markers matrix
  fixedeffectval=fe1
}
if (i==3){
  fixedeffect3=rbind(fe3, fe4, fe5, fe1)# markers matrix
  fixedeffectval=fe2
}
if (i==4){
  fixedeffect3=rbind(fe4, fe5, fe1, fe2)# markers matrix
  fixedeffectval=fe3
}
```

```
if (i==5){
  fixedeffect3=rbind(fe5, fe1, fe2, fe3)# markers matrix
  fixedeffectval=fe4
}

##Criando a matriz de efeito aleatórios
M1<-markers1[1:((1/5)*nrow(pheno)),]
M2<-markers1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
M3<-markers1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
M4<-markers1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
M5<-markers1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  markerstrain=rbind(M1, M2, M3, M4)# markers matrix
  markersval=M5
}
if (i==2){
  markerstrain=rbind(M2, M3, M4, M5)# markers matrix
  markersval=M1
}
if (i==3){
  markerstrain=rbind(M3, M4, M5, M1)# markers matrix
  markersval=M2
}
if (i==4){
  markerstrain=rbind(M4, M5, M1, M2)# markers matrix
  markersval=M3
}
if (i==5){
  markerstrain=rbind(M5, M1, M2, M3)# markers matrix
  markersval=M4
}
ETA=list(list(X=fixedeffect3, model="FIXED"),
         list(X=markerstrain, model = "BRR"))
```

```
BRR = BGLR(y=phentrain, response_type= "gaussian",
        ETA=ETA,
        nIter=100000, burnIn=20000, thin=10,
        saveAt = "BRR")

## Calculando o valor genético genômico
F_effect<-as.matrix(BRR$ETA[[1]]$b)
R_effect<-as.matrix(BRR$ETA[[2]]$b)
M<-as.matrix(fixedeffectval)
M1<-as.matrix(markersval)
GEBVval<-M%*%F_effect + M1%*%R_effect
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(fixedeffectval)
GEBV1<-mean(phentest) + GEBVval

##Estimando a herdabilidade genômica
sig2a=mean(BRR$ETA[[2]]$varB) #marker variance
sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
sig2e=mean(BRR$varE) #residual variance
h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability

## Estimando o ganho de seleção e a acurácia
result<-cbind(phentest, GEBV1, gentest)
rownames(result)<-seq(1,200,1)

q1<-quantile(result[,2], probs =.20)
top20<-result[result[,2]<q1, c(1:2)]

q3<-quantile(result[,1], probs =.20)
top20phen<-result[result[,1]<q3, c(1:2)]

#Calculando o índice de coincidência entre os indivíduos selecionados
coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
```

```
   mp<-mean(result[,2]) #mean population
   ms20<-mean(top20[,2]) #mean selected individual
   GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain

   phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
   genpredictionability[j,i]<-cor(result[,3], result[,2])
   phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
   genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
   timef<- proc.time() #final time
   timecal<-as.matrix(timef-timei)
   t<-as.numeric(timecal[3,])
   time[j,i]<-t
  }
}

GWAS_GWS_final_results<-rbind(mean(phenaccuracy), mean(genaccuracy),
mean(time),
                  mean(sig2uval), mean(h2_rval), mean(GS20),
                  mean(co20), mean(phenpredictionability),
                  mean(genpredictionability))
colnames(GWAS_GWS_final_results)<-cbind("Genetic parameters")
rownames(GWAS_GWS_final_results)<-rbind("Phenotipic Pearson Correlation",
                        "Genotipic Pearson Correlation", "Time",
                        "Genetic variance", "Heritability", "Genetic Gain 20%",
                        "Coincidence 20%", "phenpredictionability",
                        "genpredictionability")

write.table(sig2uval,"h5_GWAS_GWS_sig2uval.txt",
row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_GWAS_GWS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_GWAS_GWS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_GWAS_GWS_h2_rval.txt", row.names=TRUE,quote=FALSE)
```

```
write.table(GS20,"h5_GWAS_GWS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_GWAS_GWS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_GWAS_GWS_time.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_GWAS_GWS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(genpredictionability,"h5_GWAS_GWS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(GWAS_GWS_final_results,"h5_GWAS_GWS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)

## Códigos para rodar o modelo 6
setwd("F:\\doutorado\\tese\\chapter1\\Dados")

##Carregando pacotes
library(BGLR)

##Lendo arquivos de fenótipos e genótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}

gen<-read.table("gen.txt", head=T)
pheno<-data.frame(fen)

## Calculando a frequência alélica
library(HapEstXXR)

## Criando uma matriz codificada para o pacote HapEstXXR
geno = matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx = gen == 1
```

```
geno[idx] = 2
idx = gen == 0
geno[idx] = 3
idx = gen == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)

##Rodando a função maf que calcula menor frequ~encia alélica, call rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin))

freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

##Criando as matrizes de efeitos fixos e aleatórios
a<-c(1800,2997)
gennew<-genfin[,-a]
geno4<-genfin[,a]

### modelo: y = média + 2 SNPs de maior efeito (efeito fixo) + SNPs restantes (efeito aleatório) + erro

##Rodando o modelo de seleção genômica
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
```

```
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
 sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
 for (i in 1:folds)
 {
  timei<- proc.time() #initial time
  markers1<-gennew[sample,]
  fixedf1<-geno4[sample,]
  phen=data.frame(pheno[sample,3])
  pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
  pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                          pop1, pop2, pop3, pop4, pop5))

  phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
  phentest<-as.matrix(popfin[,i+4])

  GEN=data.frame(pheno[sample,2])
  popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
  popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
                           popg1, popg2, popg3, popg4, popg5))
```

```
gentest<-as.matrix(popgfin[,i+4])

##Criando a matriz de efeitos fixo
fe1<-fixedf1[1:((1/5)*nrow(pheno)),]
fe2<-fixedf1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
fe3<-fixedf1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
fe4<-fixedf1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
fe5<-fixedf1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  fixedeffect3=rbind(fe1, fe2, fe3, fe4)# markers matrix
  fixedeffectval=fe5
}
if (i==2){
  fixedeffect3=rbind(fe2, fe3, fe4, fe5)# markers matrix
  fixedeffectval=fe1
}
if (i==3){
  fixedeffect3=rbind(fe3, fe4, fe5, fe1)# markers matrix
  fixedeffectval=fe2
}
if (i==4){
  fixedeffect3=rbind(fe4, fe5, fe1, fe2)# markers matrix
  fixedeffectval=fe3
}
if (i==5){
  fixedeffect3=rbind(fe5, fe1, fe2, fe3)# markers matrix
  fixedeffectval=fe4
}

##Criando a matriz de efeitos aleatórios
M1<-markers1[1:((1/5)*nrow(pheno)),]
M2<-markers1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
M3<-markers1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
```

```
M4<-markers1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
M5<-markers1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  markerstrain=rbind(M1, M2, M3, M4)# markers matrix
  markersval=M5
}
if (i==2){
  markerstrain=rbind(M2, M3, M4, M5)# markers matrix
  markersval=M1
}
if (i==3){
  markerstrain=rbind(M3, M4, M5, M1)# markers matrix
  markersval=M2
}
if (i==4){
  markerstrain=rbind(M4, M5, M1, M2)# markers matrix
  markersval=M3
}
if (i==5){
  markerstrain=rbind(M5, M1, M2, M3)# markers matrix
  markersval=M4
}
ETA=list(list(X=fixedeffect3, model="FIXED"),
         list(X=markerstrain, model = "BRR"))
BRR = BGLR(y=phentrain, response_type= "gaussian",
           ETA=ETA,
           nIter=100000, burnIn=20000, thin=10,
           saveAt = "BRR")

## Estimando o valor genético genômico
F_effect<-as.matrix(BRR$ETA[[1]]$b)
R_effect<-as.matrix(BRR$ETA[[2]]$b)
M<-as.matrix(fixedeffectval)
```

```
M1<-as.matrix(markersval)
GEBVval<-M%*%F_effect + M1%*%R_effect
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(fixedeffectval)
GEBV1<-mean(phentest) + GEBVval

## Estimando a herdabilidade
sig2a=mean(BRR$ETA[[2]]$varB) #marker variance
sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
sig2e=mean(BRR$varE) #residual variance
h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability

## Estimando a acurácia e o ganho com a seleção
result<-cbind(phentest, GEBV1, gentest)
rownames(result)<-seq(1,200,1)

q1<-quantile(result[,2], probs =.20)
top20<-result[result[,2]<q1, c(1:2)]

q3<-quantile(result[,1], probs =.20)
top20phen<-result[result[,1]<q3, c(1:2)]

# Calculando o coeficiente de coincidência entre os indivíduos selecionados
coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
mp<-mean(result[,2]) #mean population
ms20<-mean(top20[,2]) #mean selected individual
GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain

phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
genpredictionability[j,i]<-cor(result[,3], result[,2])
phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
timef<- proc.time() #final time
```

```
    timecal<-as.matrix(timef-timei)
    t<-as.numeric(timecal[3,])
    time[j,i]<-t
  }
}

MG_GWS_final_results<-rbind(mean(phenaccuracy), mean(genaccuracy),
mean(time),
                    mean(sig2uval), mean(h2_rval), mean(GS20),
                    mean(co20), mean(phenpredictionability),
                    mean(genpredictionability))
colnames(MG_GWS_final_results)<-cbind("Genetic parameters")
rownames(MG_GWS_final_results)<-rbind("Phenotipic Pearson Correlation",
                          "Genotipic Pearson Correlation", "Time",
                          "Genetic variance", "Heritability", "Genetic Gain 20%",
                          "Coincidence 20%", "phenpredictionability",
                          "genpredictionability")

write.table(sig2uval,"h5_MG_GWS_sig2uval.txt", row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_MG_GWS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_MG_GWS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_MG_GWS_h2_rval.txt", row.names=TRUE,quote=FALSE)
write.table(GS20,"h5_MG_GWS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_MG_GWS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_MG_GWS_time.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_MG_GWS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(genpredictionability,"h5_MG_GWS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(MG_GWS_final_results,"h5_MG_GWS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)
```

```
## Códigos para rodar o modelo 7
setwd("F:\\doutorado\\tese\\chapter1\\Dados")

##carregando pacotes
library(BGLR)

##Lendo os arquivos de genótipos e fenótipos
files = dir()
idx = grepl(pattern="fen", x=files)
files = files[idx]; files
for(i in 1:length(files)){
  assign(x=substr(x=files[i], start=1, stop=(nchar(files[i])-4)),
         value=read.table(file=files[i], header=T))
}

##Abrindo o arquivo com os QTLs simulados
qtl<-as.matrix(read.table("QTLs.txt", h=T))

gen<-read.table("gen.txt", head=T)
pheno<-data.frame(fen)

##Calculando a frequência alélica
library(HapEstXXR)

##Criando uma matriz codificada para o pacote HapEstXXR
geno = matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx = gen == 1
geno[idx] = 2
idx = gen == 0
geno[idx] = 3
idx = gen == -1
geno[idx] = 1
genfin<-geno
colnames(genfin)<-colnames(geno)
```

```
##Rodando a função maf que calcula menor frequ~encia alélica, call rate e equilíbrio de Hardy-Weinberg
newgen=data.frame(maf(genfin))

freq<-cbind(((newgen[,1]+(newgen[,2]/2))/newgen[,4]),
((newgen[,3]+(newgen[,2]/2))/newgen[,4]),
((newgen[,1]+(newgen[,2]/2))/newgen[,4])*((newgen[,3]+(newgen[,2]/2))/newgen[,4]))

##Criando as matrizes de efeito fixo e aleatório
a<-as.matrix(qtl[,1])
gennew<-genfin[,-a]
geno4<-genfin[,a]

### modelo: y = média + QTLs (efeito fixo) + SNPs não significativos (efeito aleatório) + erro

##Rodando o modelo de seleção genômica
iteraction=50
folds=5
phenaccuracy=matrix(nrow=iteraction, ncol=folds)
genaccuracy<-matrix(nrow=iteraction, ncol=folds)
phenpredictionability<-matrix(nrow=iteraction, ncol=folds)
genpredictionability<-matrix(nrow=iteraction, ncol=folds)
sig2uval=matrix(nrow=iteraction, ncol=folds)
h2_rval=matrix(nrow=iteraction, ncol=folds)
GS20=matrix(nrow=iteraction, ncol=folds)
co20=matrix(nrow=iteraction, ncol=folds)
time=matrix(nrow=iteraction, ncol=folds)

for (j in 1:iteraction)
{
  sample<-as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
  for (i in 1:folds)
```

```
{
  timei<- proc.time() #initial time
  markers1<-gennew[sample,]
  fixedf1<-geno4[sample,]
  phen=data.frame(pheno[sample,3])
  pop1<-as.matrix(phen[1:((1/5)*nrow(pheno)),])
  pop2<-as.matrix(phen[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  pop3<-as.matrix(phen[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  pop4<-as.matrix(phen[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  pop5<-as.matrix(phen[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popfin<-as.matrix(cbind(pop1, pop2, pop3, pop4, pop5,
                          pop1, pop2, pop3, pop4, pop5))

  phentrain<-as.matrix(c(popfin[,i], popfin[,i+1], popfin[,i+2], popfin[,i+3]))
  phentest<-as.matrix(popfin[,i+4])

  GEN=data.frame(pheno[sample,2])
  popg1<-as.matrix(GEN[1:((1/5)*nrow(pheno)),])
  popg2<-as.matrix(GEN[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),])
  popg3<-as.matrix(GEN[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),])
  popg4<-as.matrix(GEN[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),])
  popg5<-as.matrix(GEN[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),])

  popgfin<-as.matrix(cbind(popg1, popg2, popg3, popg4, popg5,
                           popg1, popg2, popg3, popg4, popg5))

  gentest<-as.matrix(popgfin[,i+4])

  ##Criando a matriz de efeito fixo
  fe1<-fixedf1[1:((1/5)*nrow(pheno)),]
  fe2<-fixedf1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
  fe3<-fixedf1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
  fe4<-fixedf1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
```

```
fe5<-fixedf1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  fixedeffect3=rbind(fe1, fe2, fe3, fe4)# markers matrix
  fixedeffectval=fe5
}
if (i==2){
  fixedeffect3=rbind(fe2, fe3, fe4, fe5)# markers matrix
  fixedeffectval=fe1
}
if (i==3){
  fixedeffect3=rbind(fe3, fe4, fe5, fe1)# markers matrix
  fixedeffectval=fe2
}
if (i==4){
  fixedeffect3=rbind(fe4, fe5, fe1, fe2)# markers matrix
  fixedeffectval=fe3
}
if (i==5){
  fixedeffect3=rbind(fe5, fe1, fe2, fe3)# markers matrix
  fixedeffectval=fe4
}

##Criando a matriz de efeito aleatório
M1<-markers1[1:((1/5)*nrow(pheno)),]
M2<-markers1[(1+((1/5)*nrow(pheno))):((2/5)*nrow(pheno)),]
M3<-markers1[(1+((2/5)*nrow(pheno))):((3/5)*nrow(pheno)),]
M4<-markers1[(1+((3/5)*nrow(pheno))):((4/5)*nrow(pheno)),]
M5<-markers1[(1+((4/5)*nrow(pheno))):((5/5)*nrow(pheno)),]

if (i==1){
  markerstrain=rbind(M1, M2, M3, M4)# markers matrix
  markersval=M5
}
```

```
if (i==2){
  markerstrain=rbind(M2, M3, M4, M5)# markers matrix
  markersval=M1
}
if (i==3){
  markerstrain=rbind(M3, M4, M5, M1)# markers matrix
  markersval=M2
}
if (i==4){
  markerstrain=rbind(M4, M5, M1, M2)# markers matrix
  markersval=M3
}
if (i==5){
  markerstrain=rbind(M5, M1, M2, M3)# markers matrix
  markersval=M4
}
ETA=list(list(X=fixedeffect3, model="FIXED"),
         list(X=markerstrain, model = "BRR"))
BRR = BGLR(y=phentrain, response_type= "gaussian",
           ETA=ETA,
           nIter=100000, burnIn=20000, thin=10,
           saveAt = "BRR")

## Estimando o valor genético genômico
F_effect<-as.matrix(BRR$ETA[[1]]$b)
R_effect<-as.matrix(BRR$ETA[[2]]$b)
M<-as.matrix(fixedeffectval)
M1<-as.matrix(markersval)
GEBVval<-M%*%F_effect + M1%*%R_effect
colnames(GEBVval)<-"GEBV"
rownames(GEBVval)<-rownames(fixedeffectval)
GEBV1<-mean(phentest) + GEBVval

## Estimando a herdabilidade
```

```
  sig2a=mean(BRR$ETA[[2]]$varB) #marker variance
  sig2uval[j,i]=2*sum(freq[,3]*sig2a) #additive genetic variance
  sig2e=mean(BRR$varE) #residual variance
  h2_rval[j,i]=sig2uval[j,i]/(sig2uval[j,i]+sig2e) #heritability

  ## Estimando a acurácia e o ganho com a seleção
  result<-cbind(phentest, GEBV1, gentest)
  rownames(result)<-seq(1,200,1)

  q1<-quantile(result[,2], probs =.20)
  top20<-result[result[,2]<q1, c(1:2)]

  q3<-quantile(result[,1], probs =.20)
  top20phen<-result[result[,1]<q3, c(1:2)]

  # Calculando o coeficiente de coincidência entre os indivíduos selecionados
  coi20 <- as.matrix(union(row.names(top20), row.names(top20phen)))
  co20[j,i] <- ((2*nrow(top20phen)-nrow(coi20))/nrow(top20phen))*100
  mp<-mean(result[,2]) #mean population
  ms20<-mean(top20[,2]) #mean selected individual
  GS20[j,i]<-h2_rval[j,i]*(ms20-mp) #genetic gain

  phenpredictionability[j,i]<-cor(result[,1], result[,2], method=c("pearson"))
  genpredictionability[j,i]<-cor(result[,3], result[,2])
  phenaccuracy[j,i]<-phenpredictionability[j,i]/sqrt(0.99)
  genaccuracy[j,i]<-genpredictionability[j,i]/sqrt(0.99)
  timef<- proc.time() #final time
  timecal<-as.matrix(timef-timei)
  t<-as.numeric(timecal[3,])
  time[j,i]<-t
 }
}
```

```
GWAS_GWS_final_results<-rbind(mean(phenaccuracy), mean(genaccuracy),
mean(time),
                    mean(sig2uval), mean(h2_rval), mean(GS20),
                    mean(co20), mean(phenpredictionability),
                    mean(genpredictionability))
colnames(GWAS_GWS_final_results)<-cbind("Genetic parameters")
rownames(GWAS_GWS_final_results)<-rbind("Phenotipic Pearson Correlation",
                        "Genotipic Pearson Correlation", "Time",
                        "Genetic variance", "Heritability", "Genetic Gain 20%",
                        "Coincidence 20%", "phenpredictionability",
                        "genpredictionability")

write.table(sig2uval,"h5_QTL_GWS_sig2uval.txt", row.names=TRUE,quote=FALSE)
write.table(phenaccuracy,"h5_QTL_GWS_phenotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(genaccuracy,"h5_QTL_GWS_genotypic_accuracy.txt",
row.names=TRUE,quote=FALSE)
write.table(h2_rval,"h5_QTL_GWS_h2_rval.txt", row.names=TRUE,quote=FALSE)
write.table(GS20,"h5_QTL_GWS_GS20.txt", row.names=TRUE,quote=FALSE)
write.table(co20,"h5_QTL_GWS_co20.txt", row.names=TRUE,quote=FALSE)
write.table(time,"h5_QTL_GWS_time.txt", row.names=TRUE,quote=FALSE)
write.table(phenpredictionability,"h5_QTL_GWS_phenpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(genpredictionability,"h5_QTL_GWS_genpredictionability.txt",
row.names=TRUE,quote=FALSE)
write.table(GWAS_GWS_final_results,"h5_QTL_GWS_final_results.txt",
row.names=TRUE,col.names=TRUE, quote=FALSE)
```

ANEXO 2. Códigos do R utilizados no capítulo 2 da tese.

```
setwd("D:\\Leonardo\\Capitulo2")

pheno<-read.table("pop_f1.txt",header=T)
gen<-read.table("pop_g.txt",header=T)
folds =5
replications =50

## Criando uma nova matriz para rodar os métodos de seleção genômica
geno= matrix(0,nrow=nrow(gen),ncol=ncol(gen))
idx= gen == -1
geno[idx] = 0
idx= gen == 0
geno[idx] = 1
idx= gen == 1
geno[idx] = 2

subset= cut(seq(1,nrow(geno)),breaks=folds,labels=FALSE)
subset= cut(seq(1,nrow(pheno)),breaks=folds,labels=FALSE)

#Rodando RRBLUP
library(rrBLUP)

correlation= matrix(nrow=folds, ncol=replications)
genetic_correlation= matrix(nrow=folds, ncol=replications)

for (j in 1:replications)
 {
 train= as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
 pheno1= data.frame(pheno[train,])
 geno1<- data.frame(geno[train,])
 for (i in 1:folds)
 {
```

```
    testIndexes= which(subset==i,arr.ind=TRUE)

    markerstrain= geno1[-testIndexes, ] ## matriz de marcadores de treinamento
    markersval= geno1[testIndexes, ] ## matriz d emarcadores de validação

    phentrain= pheno1[-testIndexes,] ## vetor fenotipico de treinamento
    phentest= pheno1[testIndexes,]  ## vetor fenotipico de validação

    rrblup= mixed.solve(y = phentrain[,3], Z = markerstrain)

    ## Calculando o valor genético genômico
    M6= as.matrix(markersval)
    GEBVval= M6%*%as.matrix(rrblup$u)  ## Calculate genetic estimate breeding
value
    colnames(GEBVval)= "GEBV"
    rownames(GEBVval)= rownames(markersval)
    GEBV1= mean(phentest[,3]) + GEBVval

    ## Estimando a acurácia
    result= cbind(phentest, GEBV1)
    rownames(result)= rownames(phentest)

    correlation[i,j]= cor(result[,3], result[,4], method=c("pearson"))
    genetic_correlation[i,j]= cor(result[,2], result[,4], method=c("pearson"))
  }
}
h_5_RRBLUP<-data.frame((colMeans(correlation)/sqrt(0.05)),
colMeans(genetic_correlation))
colnames(h_5_RRBLUP)<-c("pcor", "gcor")

##Rodando os modelos de regressão linear, quadrática e cúbica
lr<-lm(gcor~pcor, data= h_5_RRBLUP)
lr1<-lm(gcor~pcor + I(pcor^2), data= h_5_RRBLUP)
```

```
lr2<-lm(gcor~pcor + I(pcor^2) + I(pcor^3), data= h_5_RRBLUP)

write.table(h_5_RRBLUP, "RRBLUP_h_5.txt",
            quote=FALSE, row.names = FALSE, col.names = FALSE)

#Rodando Bayes B
library(BGLR)

correlation= matrix(nrow=folds, ncol=replications)
genetic_correlation= matrix(nrow=folds, ncol=replications)

for (j in 1:replications)
{
  train= as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
  pheno1= data.frame(pheno[train,])
  geno1<- data.frame(geno[train,])
  for (i in 1:folds)
  {

    testIndexes= which(subset==i,arr.ind=TRUE)

    markerstrain= geno1[-testIndexes, ] ## matriz de marcadores de treinamento
    markersval= geno1[testIndexes, ] ## matriz d emarcadores de validação

    phentrain= pheno1[-testIndexes,] ## vetor fenotipico de treinamento
    phentest= pheno1[testIndexes,]  ## vetor fenotipico de validação

    BayesB = BGLR(y=phentrain, response_type= "gaussian",
                  ETA=list(list(X=markerstrain, model="BayesB")),
                  nIter=100000, burnIn=20000, thin=100, saveAt = "BayesB")

    ## Calculando o valor genético genômico
    M6= as.matrix(markersval)
    GEBVval= M6%*%as.matrix(BayesB$ETA[[1]]$b)
```

```
    colnames(GEBVval)= "GEBV"
    rownames(GEBVval)= rownames(markersval)
    GEBV1= mean(phentest[,3]) + GEBVval

    ## Estimando a acurácia
    result= cbind(phentest, GEBV1)
    rownames(result)= rownames(phentest)

    correlation[i,j]= cor(result[,3], result[,4], method=c("pearson"))
    genetic_correlation[i,j]= cor(result[,2], result[,4], method=c("pearson"))
  }
}
final<-data.frame((colMeans(correlation)/sqrt(0.05)), colMeans(genetic_correlation))
colnames(final)<-c("pcor", "gcor")

##Rodando os modelos de regressão linear, quadrática e cúbica
lr<-lm(gcor~pcor, data= final)
lr1<-lm(gcor~pcor + I(pcor^2), data= final)
lr2<-lm(gcor~pcor + I(pcor^2) + I(pcor^3), data= final)

write.table(final, "BayesB_h_5.txt",
            quote=FALSE, row.names = FALSE, col.names = FALSE)

#Rodando RKHS
library(BGLR)

correlation= matrix(nrow=folds, ncol=replications)
genetic_correlation= matrix(nrow=folds, ncol=replications)

for (j in 1:replications)
{
  train= as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
  pheno1= data.frame(pheno[train,])
  geno1<- data.frame(geno[train,])
```

```
  for (i in 1:folds)
  {

    testIndexes= which(subset==i,arr.ind=TRUE)

    markerstrain= geno1[-testIndexes, ] ## matriz de marcadores de treinamento
    markersval= geno1[testIndexes, ] ## matriz d emarcadores de validação

    phentrain= pheno1[-testIndexes,] ## vetor fenotipico de treinamento
    phentest= pheno1[testIndexes,]  ## vetor fenotipico de validação

    RKHS = BGLR(y=phentrain, response_type= "gaussian",
                ETA=list(list(X=markerstrain, model="RKHS")),
                nIter=100000, burnIn=20000, thin=100, saveAt = "RKHS")
    gebv<-as.matrix(RKHS$yHat)
    snp<-mixed.solve(Z = markerstrain, y = gebv)
    ## Calculando o valor genético genômico
    M6= as.matrix(markersval)
    GEBVval= M6%*%as.matrix(snp$u)
    colnames(GEBVval)= "GEBV"
    rownames(GEBVval)= rownames(markersval)
    GEBV1= mean(phentest[,3]) + GEBVval

    ## Estimando a acurácia
    result= cbind(phentest, GEBV1)
    rownames(result)= rownames(phentest)

    correlation[i,j]= cor(result[,3], result[,4], method=c("pearson"))
    genetic_correlation[i,j]= cor(result[,2], result[,4], method=c("pearson"))
  }
}
final<-data.frame((colMeans(correlation)/sqrt(0.05)), colMeans(genetic_correlation))
colnames(final)<-c("pcor", "gcor")
```

```
##Rodando os modelos de regressão linear, quadrática e cúbica
lr<-lm(gcor~pcor, data= final)
lr1<-lm(gcor~pcor + I(pcor^2), data= final)
lr2<-lm(gcor~pcor + I(pcor^2) + I(pcor^3), data= final)

write.table(final, "RKHS_h_5.txt",
            quote=FALSE, row.names = FALSE, col.names = FALSE)

#Rodando GBLUP
library(rrBLUP)

correlation= matrix(nrow=folds, ncol=replications)
genetic_correlation= matrix(nrow=folds, ncol=replications)

for (j in 1:replications)
{
  train= as.matrix(sample(1:nrow(pheno), nrow(pheno), replace=FALSE))
  pheno1= data.frame(pheno[train,])
  geno1<- data.frame(geno[train,])
  for (i in 1:folds)
  {

    testIndexes= which(subset==i,arr.ind=TRUE)

    markerstrain= geno1[-testIndexes, ] ## matriz de marcadores de treinamento
    markersval= geno1[testIndexes, ] ## matriz d emarcadores de validação

    phentrain= pheno1[-testIndexes,] ## vetor fenotipico de treinamento
    phentest= pheno1[testIndexes,]  ## vetor fenotipico de validação

    MM<-A.mat(markerstrain)
    rownames(MM)<-seq(1,nrow(MM),1)
    phentrain<-data.frame(cbind(seq(1,nrow(MM),1), phentrain))
    colnames(phentrain)<-c("id", "vfen")
```

```
    Gblup = kin.blup(data=phentrain, geno="id", pheno="vfen", K=MM)

    gebv<-as.matrix(Gblup$g)
    snp<-mixed.solve(Z = markerstrain, y = gebv)
    ## Calculando o valor genético genômico
    M6= as.matrix(markersval)
    GEBVval= M6%*%as.matrix(snp$u)
    colnames(GEBVval)= "GEBV"
    rownames(GEBVval)= rownames(markersval)
    GEBV1= mean(phentest[,3]) + GEBVval

    ## Estimando a acurácia
    result= cbind(phentest, GEBV1)
    rownames(result)= rownames(phentest)

    correlation[i,j]= cor(result[,3], result[,4], method=c("pearson"))
    genetic_correlation[i,j]= cor(result[,2], result[,4], method=c("pearson"))
  }
}
final<-data.frame((colMeans(correlation)/sqrt(0.05)), colMeans(genetic_correlation))
colnames(final)<-c("pcor", "gcor")

##Rodando os modelos de regressão linear, quadrática e cúbica
lr<-lm(gcor~pcor, data= final)
lr1<-lm(gcor~pcor + I(pcor^2), data= final)
lr2<-lm(gcor~pcor + I(pcor^2) + I(pcor^3), data= final)

write.table(final, "GBLUP_h_5.txt",
            quote=FALSE, row.names = FALSE, col.names = FALSE)
```

ANEXO 3. Códigos do R utilizados no capítulo 3 da tese.

Estimação do número ótimo de marcadores

```
library(rrBLUP)
setwd("C:\\TPS")

## Lendo arquivo de marcadores
snp=read.table("gen.txt",h=T)
M=as.matrix(snp)
M1=as.matrix(snp)

## Lendo arquivo de fenótipos
data=read.table("fen.txt",h=T)
phe=as.matrix(data)

iteraction <- (ncol(M1)-1)
residual_variance = matrix(nrow = iteraction,ncol = 2)
additive_variance = matrix(nrow = iteraction,ncol = 2)
accuracy = matrix(nrow = iteraction,ncol = 3)

for (j in 1:iteraction){
  var<-(phe[,3])

  ##Rodando RRBLUP
  fit2=mixed.solve(var, Z=M, K=diag(ncol(M)))

  ##Estimando o valor genético genômico
  GEBV<-M%*%fit2$u

  ## Selecionando marcadores com base no efeito
  eff<- abs(fit2$u)
  eff1<- sort(eff, decreasing=TRUE) # Marker effect in decreasing order

  ## identificando o marcador com o menor efeito
  del<-eff1[nrow(fit2$u)]
```

```
  del

  ## Identificando SNPs abaixo do valor crítico em eff
  SNPdel<- eff == del ## Identify the SNP that will be elimite "abs(fit2$u)". Este parecerá como TRUE
  snpM<-as.matrix(SNPdel) # transforma em matriz

  ##Identificando o número de marcas que serão descartadas
  desc<- which(snpM[,1] == "TRUE")
  desc

  residual_variance[j,1] <- ncol(M)
  residual_variance[j,2] <- fit2$Ve
  additive_variance[j,1] <- ncol(M)
  additive_variance[j,2] <- fit2$Vu
  accuracy[j,2] <- ncol(M)
  accuracy[j,2] <-cor(GEBV, phe[,3])
  accuracy[j,3] <-cor(GEBV, phe[,2])

  M<- M[ ,-c(desc)] # deleting the SNP of the original marker matrix
}

write.table(residual_variance,"sig2e_h5.txt")
write.table(additive_variance,"sig2a_h5.txt")
write.table(accuracy, "accuracy_h5.txt")

Estimação do número ótimo de indivíduos na população de treinamento
library(rrBLUP)
setwd("C:\\TPS")

## Lendo arquivo de marcadores
snp=read.table("gen.txt",h=T)
M=as.matrix(snp)
M1=as.matrix(snp)
```

```
## Lendo arquivo de fenótipos
data=read.table("fen.txt",h=T)
phe=as.matrix(data)

iteraction <- (nrow(M1)-1)

residual_variance = matrix(nrow = iteraction,ncol = 2)
additive_variance = matrix(nrow = iteraction,ncol = 2)
accuracy = matrix(nrow = iteraction,ncol = 3)
for (j in 1:iteraction){

  var<-(phe[,3])

  ##Rodando RRBLUP
  fit2=mixed.solve(var, Z=M, K=diag(ncol(M)))

  ##Estimando o valor genético genômico
  GEBV<-M%*%fit2$u

  ## ## Selecionando indivíduos com base no valor genético genômico
  gebv<- abs(GEBV)
  gebv1<- sort(gebv, decreasing=TRUE) # GEBV in decreasing order

  ## identificando o indivíduo com o menor valor genético genômico
  del<-gebv1[nrow(GEBV)]
  del

  ## Identificando indivíduos abaixo do valor crítico em eff
  inddel<- gebv == del ## Identify the individual that will be elimite "abs(GEBV)". That
appears as TRUE
  indM<-as.matrix(inddel) # transforma em matriz

  ##Identificando o número de indivíduos que serão descartados
```

```
  desc<- which(indM[,1] == "TRUE")
  desc

  residual_variance[j,1] <- nrow(M)
  residual_variance[j,2] <- fit2$Ve
  additive_variance[j,1] <- nrow(M)
  additive_variance[j,2] <- fit2$Vu
  accuracy[j,1] <- nrow(M)
  accuracy[j,2] <-cor(GEBV, phe[,3])
  accuracy[j,3] <-cor(GEBV, phe[,2])

  M<- M[-c(desc) ,] # deleting the individual of the original marker matrix
  phe <- phe[-c(desc), ] # deleting the individual of the original phenotypic file
}

write.table(residual_variance,"sig2e_h5.txt")
write.table(additive_variance,"sig2a_h5.txt")
write.table(accuracy, "accuracy_h5.txt")
```