

ÍSIS FERNANDA DE ALMEIDA

MÉTODOS DE ESTIMAÇÃO E VALIDAÇÃO NA SELEÇÃO
GENÔMICA NA PRESENÇA OU AUSÊNCIA DE CORREÇÃO DE
FENÓTIPOS

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*.

VICOSA
MINAS GERAIS – BRASIL
2013

**Ficha catalográfica preparada pela Seção de Catalogação e
Classificação da Biblioteca Central da UFV**

T

A447m
2013 Almeida, Isis Fernanda de, 1984-
Métodos de estimação e validação na seleção genômica na
presença ou ausência de correção de fenótipos / Isis Fernanda de
Almeida. – Viçosa, MG, 2013.
xi, 55f. : il. (algumas color.) ; 29 cm.

Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Genômica. 2. Marcadores genéticos. 3. BLASSO.
4. RR-BLUP. I. Universidade Federal de Viçosa. Departamento
de Biologia Geral. Programa de Pós-Graduação em Genética e
Melhoramento. II. Título.

CDD 22. ed. 576.5

ÍSIS FERNANDA DE ALMEIDA

MÉTODOS DE ESTIMAÇÃO E VALIDAÇÃO NA SELEÇÃO
GENÔMICA NA PRESENÇA OU AUSÊNCIA DE CORREÇÃO DE
FENÓTIPOS

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: 01 de agosto de 2013.

Prof. Dr. Marcos Deon Vilela de
Resende
(Coorientador)

Prof. Dr. Leonardo Lopes Bhering

Prof. Dr. Ramon Vinícius de Almeida

Dra. Marciane da Silva Oliveira

Prof. Dr. Cosme Damião Cruz
(Orientador)

Aos meus pais Iracy e Cely
Ao meu irmão Ramon.

AGRADECIMENTOS

A Deus por conceder o dom da vida e inspirar-me em todos instantes.

À Universidade Federal de Viçosa, pela oportunidade de realizar este curso.

Ao CNPq e CAPES pelo auxílio financeiro.

Ao meu orientador, professor Cosme Damião Cruz pela orientação e pelo apoio.

Ao meu coorientador Marcos Deon Vilela de Resende pelas valiosas sugestões, disponibilidade e atenção.

Aos amigos, Caio, Camila, André, Gustavo e Gislayne pela colaboração e pela agradável convivência.

Às ex-companheiras de república, Daniela, Paulinha e Mila pela amizade e pelos momentos de alegria compartilhados.

Aos professores e colegas do curso de pós-graduação em genética.

Às amigas Pollyana, Adalgisa, Pâmela, Taís e Stela pelo carinho, apoio e companheirismo.

Ao amigo Felipe por sua colaboração, disponibilidade e sugestões.

Às funcionárias da secretaria, Edna e Rita, pela atenção e pela amizade.

Ao meu Irmão, Ramon, que me trouxe para Viçosa e que sem o qual, seria muito mais difícil estar onde estou hoje.

Aos membros da banca, Drs. Marcos Deon Vilela de Resende, Leonardo Lopes Bhering, Ramon Vinícuís de Almeida e Marciane da Silva Oliveira pela disponibilidade e colaboração.

A todos os que direta ou indiretamente colaboraram para realização deste trabalho, minha sincera gratidão.

BIOGRAFIA

ÍISIS FERNANDA DE ALMEIDA, filha de Iracy de Almeida e Cely Aparecida de Almeida, nasceu na cidade de Volta Redonda - RJ, em 09 de novembro de 1984.

Realizou os estudos do primeiro e segundo graus, na Escola Estadual Getúlio Vargas e na Escola Técnica Pandiá Calógeras, respectivamente, até o ano de 2001.

Ingressou no curso de Agronomia no ano de 2002 na Universidade Federal de Lavras.

Em março de 2007, obteve o diploma de Engenheira Agrônomo pela Universidade Federal de Lavras.

Em 2007, iniciou no Programa de Mestrado em Genética e Melhoramento, pela Universidade Federal de Viçosa, com ênfase em Genética Quantitativa, obtendo o respectivo título em fevereiro de 2009.

Neste mesmo ano, ingressou no Programa de Doutorado em Genética e Melhoramento, também pela Universidade Federal de Viçosa, com ênfase em Genética Quantitativa, obtendo o respectivo título em agosto de 2013.

Em 2012 ingressou na Universidade Estadual de Goiás como professora efetiva na área de Genética e Melhoramento de Plantas.

SUMÁRIO

RESUMO	viii
ABSTRACT	x
1 INTRODUÇÃO	1
2 REVISÃO DE LITERATURA	5
2.1 Desequilíbrio de Ligação.....	5
2.2 Seleção genômica ampla.....	8
2.3 Validação na seleção genômica ampla.....	10
2.4 Correção de fenótipos para estrutura de população	12
2.5 Metodologias de análise na seleção genômica ampla	12
3 MATERIAL E MÉTODOS	16
3.1 Descrição das populações	16
3.2 Simulação dos valores genotípicos e fenotípicos	17
3.3 Caracterização genética das populações simuladas	18
3.4 Metodologias de análise	19
3.4.1 Ridge Regression BLUP (RR-BLUP)	19

3.4.2 LASSO Bayesiano(BLASSO)	21
3.5 Validação	22
3.6 Correção dos fenótipos.....	23
3.7 Análises estatísticas	23
4 RESULTADOS E DISCUSSÃO	24
4.1 Simulação	24
4.2 Efeito de marcadores	29
4.3 Valores genéticos, genômicos e fenotípicos	32
4.4 Acurácia na seleção genômica ampla sem correção de fenótipos	37
4.5 Acurácia na seleção genômica ampla com correção de fenótipos	41
5 REFERÊNCIAS	46

RESUMO

ALMEIDA, Ísis Fernanda de D.Sc., Universidade Federal de Viçosa, Agosto de 2013. **Métodos de estimação e validação na seleção genômica na presença ou ausência de correção de fenótipos.** Orientador: Cosme Damião Cruz. Coorientadores: Fabyano Fonseca e Silva e Marcos Deon Vilela de Resende.

Com os avanços de tecnologias de genotipagem em larga escala, tornou-se possível uma cobertura completa do genoma, o que levou os pesquisadores a criar uma nova forma de utilização dessa informação genética gerada em benefício do melhoramento de plantas. Essa nova metodologia foi denominada de seleção genômica ampla (GWS), e consiste na utilização simultânea de centenas ou milhares de marcadores, os quais cobrem o genoma de maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Com isso, o trabalho teve por objetivo avaliar a acurácia da seleção genômica em diferentes cenários de distribuição de efeitos genéticos e populações de validação, além de analisar o efeito da correção de fenótipos na estrutura das populações. Uma vez que os dados foram gerados por simulação, observou-se também a eficácia do processo de simulação em gerar populações cujos princípios genéticos sejam preservados. Para isso, foram simuladas dez repetições de duas estruturas populacionais, com o número de 500 indivíduos. Para os dados genotípicos considerou-se 1000 locos marcadores, sendo 100 efetivamente ligados a QTL. Distintos efeitos de

QTL foram simulados, um de acordo com uma distribuição uniforme e outro de acordo com uma distribuição exponencial com o objetivo de retratar diferentes arquiteturas genéticas. Para a validação 1 (V1) uma amostra de 100 indivíduos constituiu a população de validação; para a validação 2 (V2), aplicou-se a validação cruzada de Jackknife; e na validação 3 (V3), uma nova população de validação foi gerada. As metodologias RR-BLUP e BLASSO foram utilizadas a fim de se computar a acurácia da seleção genômica nos diferentes cenários estabelecidos. A simulação das populações foi eficiente ao retratar a estrutura genética das populações. Sem correção de fenótipos para estrutura de população a distribuição de efeitos de QTL exponencial conduziu a maiores acurácia, sendo o método BLASSO o mais acurado para este tipo de distribuição de efeitos genéticos. Neste cenário, as validações V1 e V3 foram mais acuradas. Com correção de fenótipos os QTL com efeitos genéticos de distribuição exponencial e uniforme conduziram a acurácias similares e o método BLASSO mostrou-se mais robusto, sendo o mais acurado para ambas as distribuições de efeitos genéticos; nesta situação, as validações V1 e V2 apresentaram-se mais acuradas. A correção de fenótipos proporcionou maiores valores de acurácia devido a uma maior eficiência em capturar o desequilíbrio de ligação. De maneiral, o método RR-BLUP foi mais preciso e o BLASSO menos viciado.

ABSTRACT

ALMEIDA, Ísis Fernanda de D. Sc., Universidade Federal de Viçosa, August of 2013.
Methods of estimation and validation in genomic selection in the presence or absence of phenotypic correction. Adviser: Cosme Damião Cruz. Co-advisers: Fabyano Fonseca e Silva and Marcos Deon Vilela de Resende.

With new breakthroughs in large-scale genotyping technologies, it became possible to achieve a genome complete coverage, which led the researchers to create a new way of using this new genetic information acquired to the benefit of plant breeding. This new methodology was called genome-wide selection (GWS), and consists of a simultaneous use of hundreds or thousands of markers, which have a wide and dense coverage of the genome then all genes of a quantitative trait are in linkage disequilibrium with at least one of those markers. Thus, that study aimed to evaluate the accuracy of genomic selection in different scenarios of genetic distribution effects and validation populations, and analyze the effect of phenotype adjustment in the population structure. Once the data were generated by simulation, it was also observed the effectiveness of that simulation process to generate populations whose genetic principles are preserved. For this, ten replicates of two population structures were simulated, with 500 individuals each. For genotypic data was considered 1000 marker loci with 100 effectively linked to QTL. Distinct QTL effects were simulated, one according to a uniform distribution and another according to an exponential distribution

with the goal of represent different genetic architectures. For validation 1 (V1) a sample of 100 individuals composed the validation population, for validation 2 (V2), we applied the Jackknife cross-validation, and validation 3 (V3), a new validation population was generated. The RR-BLUP and BLASSO methodologies were used in order to compute the genomic selection accuracy on the different established scenarios. The simulation of populations was effective in represent the genetic structure of populations. Without phenotypic correction for population structure the QTL exponential distribution effects led to higher accuracy, the BLASSO method showed more accuracy for this distribution of genetic effects. In this scenario, the validations V1 and V3 were more accurated. With phenotypes adjustment the QTL genetic effects in exponential and uniform distributions led to similar accuracies and BLASSO method proved more robust and more accurate for both distributions of genetic effects. In this situation, validations V1 and V2 showed up more accurated. The phenotypes adjustment led to higher accuracy due to greater efficiency in capturing linkage disequilibrium. Generally the RR-BLUP method was more accurate and the BLASSO method less biased.

1INTRODUÇÃO

O melhoramento genético tem, há vários anos, proporcionado, com muito sucesso, o aumento de produtividade e a melhoria de várias características de interesse na agricultura e na pecuária. A eficiência dos programas de melhoramento genético depende basicamente de duas ações do geneticista: a criação e a identificação de genótipos superiores. Em ambas as ações, a seleção desempenha papel fundamental na definição dos cruzamentos a serem realizados, visando à criação de novos genótipos, e na indicação dos indivíduos superiores a serem usados comercialmente ou em novos ciclos de seleção (RESENDE et al., 2008). Em outras palavras, a seleção consiste em escolher os indivíduos que, pela união de seus gametas, formarão a próxima geração, alterando de forma favorável a frequência dos alelos nas populações.

A complexidade da seleção reside no fato de que o desempenho dos indivíduos é resultado do patrimônio genético que eles possuem, e, ainda, dos efeitos de meio ambiente. Embora muitos métodos tenham surgido ao longo dos anos, a estratégia básica utilizada foi a de prever o valor genético do indivíduo, com base em informações fenotípicas, e em alguns casos de genealogia, como indicador do potencial ou da predisposição genética em se expressar sob influência de determinado ambiente.

Com o desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular, criou-se a expectativa de que as informações genotípicas dos marcadores moleculares, uma vez correlacionados com características fenotípicas de interesse, pudessem ser amplamente utilizadas na obtenção e seleção de indivíduos com

maior valor genético. Desta forma, antes do advento dos marcadores moleculares, os caracteres quantitativos foram estudados, principalmente, utilizando-se modelos genético-estatísticos, que consideravam o efeito médio de todos os locos afetando os caracteres.

Os marcadores moleculares aproveitam o polimorfismo da própria molécula de DNA, ocorrendo, portanto, em número suficiente para que seja possível a marcação de qualquer alelo presente em uma dada espécie (RAMALHO; SANTOS; PINTO, 2000). Com isso, foi possível a elaboração de mapas genéticos que apresentam grande número desses marcadores, etapa fundamental para o início dos trabalhos de mapeamento de QTL, ou seja, o mapeamento de regiões contendo locos que controlam caracteres quantitativos, permitindo obter estimativas de efeito e posição para cada um dos locos que afetam esses caracteres.

O principal atrativo da genética molecular em benefício do melhoramento genético aplicado é a utilização direta das informações de DNA na seleção, de modo a permitir alta eficiência seletiva, rapidez na obtenção de ganhos genéticos com a seleção e baixo custo (RESENDE et al., 2008). Inicialmente, os trabalhos de mapeamento de QTL (*Quantitative Trait Loci*) foram realizados para estudos de herança de caracteres de importância agrônoma/econômica. Entretanto, a aplicação mais promissora dos QTL é a utilização dessas informações para realização de seleção, o que é conhecido como seleção assistida por marcadores moleculares (SAM). A SAM pode aumentar a eficiência de programas de melhoramento, pois genótipos contendo alelos favoráveis para o caráter podem ser identificados antes do florescimento, diminuindo o tempo de cada ciclo de seleção.

Com a perspectiva de um aumento nos ganhos de seleção e redução nos ciclos de melhoramento via seleção assistida por marcadores, muitas pesquisas foram feitas e QTL foram detectados e mapeados nas mais variadas culturas (FRARY, et. al. 2000; YANO et. al. 2000; LIU et. al, 2002). No entanto, grande parte destes QTL detectados e mapeados em cada espécie, não foram aplicados de forma prática nos seus programas de melhoramento (BERNARDO, 2008).

As principais causas deste insucesso foram a necessidade do estabelecimento de associações entre os marcadores e os QTL para cada família avaliada e o fato de serem feitas apenas a detecção de um pequeno número de QTL de grande efeito, os quais, devido à natureza poligênica e à alta influência ambiental dos caracteres quantitativos, não explicam suficientemente toda a variação genética (DEKKERS, 2004). Além disso,

pode se destacar, também, que a seleção baseada em marcadores moleculares só é superior em relação à seleção fenotípica quando esta é aplicada em uma família com tamanho superior a 500 indivíduos (RESENDE, 2007).

Com os avanços de tecnologias de genotipagem em larga escala, tornou-se possível uma cobertura completa do genoma, o que levou os pesquisadores a criar uma nova forma de utilização dessa informação genética gerada. A seleção genômica ampla (GWS), proposta por Meuwissen et al. (2001) consiste na utilização simultânea de centenas ou milhares de marcadores, os quais cobrem o genoma de maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores. Por atuar em todo o genoma é considerada ampla, capturando todos os genes que afetam um caráter quantitativo sem a necessidade de identificar previamente os marcadores com efeitos significativos (sem o uso de testes de significância para marcas individuais) e de mapear QTLs, como no caso daSAM.

A diferença básica na predição de valores genéticos tradicionais e na predição de valores genéticos genômicos é que, nos primeiros, as informações fenotípicas são utilizadas visando a inferências sobre efeitos dos genótipos dos indivíduos e, nos últimos, as informações genotípicas (genótipos para os alelos marcadores) são usadas visando à inferência sobre valores fenotípicos futuros (ou valores genéticos genômicos preditos) dos indivíduos. Em outras palavras, os métodos tradicionais usam o fenótipo para inferir sobre o efeito do genótipo e a GWS usa o genótipo, com efeito genético pré-estimado em uma amostra da população, para inferir sobre o fenótipo a ser expresso nos candidatos à seleção (RESENDE, 2008).

Na seleção genômica ampla obtém-se primeiramente o polimorfismo dos marcadores, e, logo após, seus efeitos são estimados baseados em dados fenotípicos de uma população conhecida como população de estimação ou treinamento. Uma vez que seus efeitos são modelados e estimados, estes são testados em uma população de validação e, então, seleciona-se os marcadores que explicam parte da variância genética do caráter em estudo para que sua informação seja efetivamente incorporada à etapa de seleção do programa de melhoramento.

Na população de validação, utiliza-se um conjunto de dados menor do que aquele da população de estimação e contempla indivíduos genotipados e fenotipados para a característica de interesse. Esta amostra independente é utilizada para testar e verificar as acurácias das equações de predição de valores genéticos genômicos.

A implementação da seleção genômica impõe vários desafios estatísticos e computacionais como a dimensionalidade do modelo, colinearidade entre marcas e a complexidade das características quantitativas. Para isso, vários métodos têm sido propostos, que diferem entre si pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo. Entre eles, o método RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*) estima simultaneamente os efeitos de todas as marcas (MEUWISSEN et al., 2001; WHITTAKER et al., 2000), sendo estas consideradas efeitos aleatórios com variância comum, ou seja, assumem que todos os marcadores contribuem igualmente para a variação genética (ausência de genes de efeitos maiores). No entanto, assumir que as marcas individuais possuem a mesma variância pode não ser apropriado para algumas marcas localizadas em regiões não associadas à variância genética, enquanto outras estão localizadas em regiões associadas à QTL (GODDARD; HAYES, 2007). Para contornar esta questão, muitos autores propuseram metodologias que utilizam efeito shrinkage específico para cada marcador. No contexto bayesiano, isto pode ser implementado usando prioris para os efeitos dos marcadores, a exemplo do LASSO Bayesiano de Park e Casella (2008).

O estimador LASSO Bayesiano (BLASSO), além de ajustar uma variância separada para cada marca, possui a vantagem de forçar os estimadores para zero, como no caso do RR-BLUP. No entanto, ele efetivamente permite que alguns estimadores sejam identicamente iguais a zero, realizando simultaneamente o procedimento de shrinkage e seleção de modelos.

Visto que a superioridade de cada método pode estar relacionada com o tipo de distribuição de efeitos genéticos da característica quantitativa envolvida, um dos objetivos do trabalho foi verificar a influência da distribuição dos efeitos de QTL com diferentes formas de distribuição na acurácia de algumas metodologias de análise genômica, entre elas o RR-BLUP e o BLASSO. Adicionalmente, procurou-se avaliar a importância da correção dos fenótipos e o tipo de população de validação na acurácia da estimação. Uma vez que o estudo foi baseado em dados simulados, também se objetivou analisar a eficácia da simulação em reproduzir os princípios genéticos da genética quantitativa e da genética de populações nas populações geradas.

2 REVISÃO DE LITERATURA

2.1 Desequilíbrio de ligação

O desequilíbrio de ligação (LD) ou desequilíbrio de fase gamética é uma medida da dependência ou não entre alelos de dois ou mais locos. Em um grupo de indivíduos, se dois alelos são encontrados juntos com frequência maior do que aquela esperada com base no produto de suas frequências infere-se que tais alelos estão em desequilíbrio de ligação (RESENDE, 2008).

Vários fatores influenciam o LD. Alguns são responsáveis pelo aumento no LD, incluindo as autofecundações, pequenos tamanhos de populações, isolamento genético entre linhagens, subdivisão populacional, baixa taxa de recombinação, mistura populacional, seleção artificial e natural, dentre outros. Alguns outros fatores são responsáveis pela queda ou quebra do LD, incluindo, acasalamento ao acaso, elevadas taxas de recombinação, elevadas taxas de mutações, dentre outros. Existem fatores que podem aumentar ou quebrar o LD, ou podem aumentar o LD entre determinado par de alelos e diminuir-lo entre outros pares de alelos. Por exemplo, a mutação pode romper o LD entre pares de alelos envolvendo alelos selvagens, e promover LD entre os pares de alelos dos mutantes envolvidos. Outros fatores que afetam o LD incluindo estrutura populacional e epistasia não tem recebido atenção desejada nas revisões realizadas (GUPTA et al., 2005).

De forma resumida, os fatores explicados por Garcia et al. (2003) que afetam o LD são apresentados a seguir.

A frequência alélica e a recombinação entre sítios afetam o LD, assim como a maior parte dos processos observados na genética de populações. Mutações fornecem o material para produzir polimorfismos que podem afetá-lo. A recombinação é o principal fenômeno que diminui o LD intracromossômico, enquanto que o intercromossômico diminui com a distribuição independente. O tamanho populacional também tem um importante papel. Em pequenas populações, os efeitos da deriva genética resultam em uma perda consistente de combinações alélicas raras, o que aumenta os níveis de LD.

Os sistemas de acasalamento e misturas também podem fortemente influenciar o LD. Mistura é a circulação gênica entre indivíduos geneticamente distintos seguidos por intercruzamentos. Geralmente, o LD diminui mais rapidamente em espécies alógamas comparada com espécies autógamas. Isto é porque a recombinação é menos efetiva em espécies que se autofecundam, onde os indivíduos são mais semelhantes por serem homozigotos, do que em espécies de fecundação cruzada. O LD pode também ser criado em populações que tiveram recentemente seu tamanho populacional reduzido devido ao afunilamento genético (*bottleneck*). Durante o afunilamento, apenas poucos alelos combinados são passados para gerações futuras, o que pode gerar um LD considerável.

A seleção, que causa afunilamento genético para locos específicos, também pode criar LD entre o alelo selecionado e o loco ligado. Além disso, a seleção a favor ou contra um fenótipo controlado por dois genes não ligados (epistasia) pode criar um LD, embora os locos não estejam fisicamente ligados. Finalmente, o fluxo gênico entre indivíduos de diferentes populações pode introduzir novas combinações cromossômicas e diferentes frequências alélicas, resultando em LD.

Estes fatores mencionados tem sido discutidos com grandes detalhes em outras literaturas (GUPTA et al., 2005; ARDLIE et al., 2002; JANNINK; WALSH, 2002; WEISS; CLARK, 2002; FLINT-GARCIA et al., 2003; GAUT; LONG, 2003; RAFALSKI; MORGANTE, 2004).

O desequilíbrio gamético é dado por $D = \text{prob}(AB) \text{prob}(ab) - \text{prob}(Ab) \text{prob}(aB)$, em que prob denota probabilidade ou frequência dos respectivos haplótipos. Assim, o desequilíbrio existe (D diferente de zero) quando os gametas em associação e repulsão diferem em frequência. Valores de D positivos revelam que os gametas em associação estão em excesso, quando negativos, revelam excesso de gametas em repulsão. Esta estatística é muito dependente das frequências de alelos individuais e, portanto, não é

útil para comparação do LD entre múltiplos pares de locos envolvendo diferentes pontos ao longo do genoma. A estatística r^2 desenvolvida por Hill e Robertson (1968) é a mais adequada, pois é a menos dependente das frequências alélicas. Esta é dada por: $r^2 = D^2 / \text{prob}(A) \text{prob}(a) \text{prob}(B) \text{prob}(b)$. Os valores de r^2 variam de zero (pares de locos com nenhum desequilíbrio entre eles) a 1 (pares de locos com desequilíbrio completo).

As estatísticas anteriores consideram apenas dois locos de cada vez. Para computar a extensão do desequilíbrio de ligação através de um segmento cromossômico que contém múltiplos marcadores, Hayes et al. (2003) propuseram uma definição alternativa de LD multilocos denominada homozigose do segmento cromossômico (CSH). Essa medida é mais informativa quanto às causas do desequilíbrio de ligação. O CSH refere-se à probabilidade de que dois segmentos cromossômicos, tomados ao acaso da população, pertençam a um ancestral comum, ou seja, sejam idênticos por descendência (IBD), significando que não houve quebra por recombinação. A partir de uma geração ancestral, o cromossomo é quebrado em cada geração, e na geração atual apenas pequenas regiões cromossômicas são idênticas por descendência ao que era na geração ancestral. Para inferir sobre o CSH é necessário construir os haplótipos marcadores para os segmentos cromossômicos a partir dos dados marcadores. Desta forma, que tem-se: $HH = CSH + [(Hom_a - CSH)(Hom_b - CSH)] / (1 - CSH)$, em que Hom_a e Hom_b são as homozigosidades dos locos marcadores individuais e dos haplótipos (HH). A medida LD multilocos denominada homozigose do segmento cromossômico (CSH) é mais precisa do que aquelas envolvendo dois locos. (RESENDE, 2008).

Historicamente, a análise de desequilíbrio de ligação tem sido usada para medir a proximidade entre um loco e outro, para mapear características qualitativas e mais recentemente para mapear QTL (FLINT-GARCIA et al., 2003).

O mapeamento de QTL, a seleção auxiliada por marcadores (MAS) proposta por Lande e Thompson (1990) e a seleção genômica ampla (GWS) proposta por Meuwissen et al. (2001), são fundamentadas na ocorrência de desequilíbrio de ligação na população estudada. Especialmente relevante é a extensão desse desequilíbrio no cromossomo em uma população de seleção que se um marcador e um QTL estão em equilíbrio na população, o conhecimento do genótipo do marcador em um indivíduo não apresenta qualquer valor para a seleção. Os alelos dos marcadores informam sobre a presença e efeitos dos locos que governam os caracteres quantitativos, fornecendo meios para estimação dos efeitos dos locos dos QTL e para seu eficiente uso na seleção genética.

O significado genético entre um marcador e um QTL não observado é que ele mede a proporção da variação causada por alelos do QTL que é explicada pelos marcadores. Assim, o decréscimo do desequilíbrio de ligação, dado por r^2 , com o aumento da distância indica quantos marcadores e fenótipos são necessários para a acurada predição no contexto da seleção genômica ampla e da detecção de QTL usando LD em nível populacional (RESENDE, 2008).

2.2 Seleção genômica ampla

A complexidade da seleção reside no fato de que o desempenho dos indivíduos é resultado do patrimônio genético que eles possuem, e, ainda, dos efeitos de meio ambiente. Com isso, fica evidente a preocupação que o melhorista deve ter ao tratar de caracteres de herança quantitativa, em que estão envolvidos grande número de genes, sob influência considerável do ambiente. A natureza da variação manifestada numa população deve ser avaliada, pois aquela proporcionada pelo ambiente é temporária, não sendo explorada pela seleção. Adicionalmente, deve haver instrumentos eficazes para realizar comparações entre indivíduos, de forma que, ao se selecionar o indivíduo com base no seu bom desempenho fenotípico, se obtenha maior sucesso em estar selecionando aquele geneticamente superior.

Uma primeira proposição realizada para aumentar a eficiência do procedimento de seleção baseado em dados fenotípicos foi descrita por Lande e Thompson (1990), por meio da seleção auxiliada por marcadores (SAM) moleculares. A SAM utiliza simultaneamente dados fenotípicos e dados de marcadores moleculares em ligação gênica próxima com alguns locos controladores de características quantitativas (QTL). Desta forma, a utilização de marcadores moleculares permite alta eficiência seletiva, grande rapidez na obtenção de ganhos genéticos com a seleção e baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos (RESENDE, 2008).

A seleção baseada na SAM apresenta as seguintes características: (i) requer o estabelecimento (análise de ligação) de associações marcadores-QTL para cada família em avaliação, ou seja, essas associações apresentam utilidade para seleção apenas dentro de cada família mapeada em espécies alógamas; (ii) para ser útil, precisa explicar grande parte da variação genética de uma característica quantitativa, que é governada por muitos locos de pequenos efeitos. Isto não tem sido observado na prática, exatamente em função

da natureza poligênica e alta influência ambiental nos caracteres quantitativos, fato que conduz à detecção apenas de um pequeno número de QTL de grandes efeitos, os quais não explicam suficientemente toda a variação genética; (iii) só apresenta superioridade considerável em relação à seleção baseada em dados fenotípicos, quando o tamanho de família avaliado e genotipado é muito grande (da ordem de 500 ou mais) (RESENDE, 2008). Em função desses aspectos, a implementação da MAS tem sido limitada e os ganhos em eficiência muito reduzidos (DEKKERS, 2004).

Uma nova abordagem para utilização de marcadores moleculares para seleção assistida, conhecida como *Genome Wide Selection* (GWS) ou Seleção Genômica Ampla, foi proposta por Meuwissen et al. (2001). Devido ao desenvolvimento de marcadores cada vez mais abundantes no genoma e a custos cada vez menores, essa metodologia tornou-se muito atrativa, sendo utilizada por muitos pesquisadores no melhoramento animal. Na seleção genômica, consideram-se informações de todos os marcadores do genoma e não apenas aqueles associados ao caráter por meio da análise de QTL, eliminando-se a etapa de construção de mapa genético e mapeamento de QTL, o que torna o processo mais rápido e barato. Uma vez que a seleção assistida por marcadores baseia-se na detecção, mapeamento e uso de QTL na seleção, a seleção genômica, por outro lado, atua com a seleção simultânea para centenas ou milhares de marcadores, os quais cobrem o genoma de uma maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores.

Meuwissen et al. (2001) verificaram, em estudos de simulação, que a seleção genômica apresenta 41 e 33% a mais de acurácia em relação à SAM e ao BLUP tradicional, respectivamente. Mesmo com altos valores de significância na SAM, Habier et al. (2007) relataram que a seleção genômica resultou em 4 e 11% de aumento na acurácia comparado à SAM e ao BLUP tradicional, respectivamente. Adicionalmente, estudos de simulação em milho indicam que a seleção genômica tem 18-43% de superioridade seletiva em relação à SAM (BERNARDO e YU, 2007). Resultados similares foram observados por Piyasatian et al. (2007), que verificaram que, no primeiro ciclo de seleção no cruzamento entre duas linhagens, o ganho com a seleção genômica foi 109 e 32% maior que o BLUP tradicional e a SAM, respectivamente.

Os marcadores em desequilíbrio de ligação com os QTL, tanto de grandes efeitos quanto de pequenos efeitos, explicarão quase a totalidade da variação genética de um caráter quantitativo. O número de marcadores é de tal magnitude que a

probabilidade de se encontrar um QTL em desequilíbrio de ligação com pelo menos um marcador é muito alta. Este aspecto é importante uma vez que somente os marcadores em desequilíbrio de ligação serão úteis na determinação dos fenótipos e na explicação da variação genética. Os efeitos dos marcadores são estimados em uma amostra de indivíduos pertencentes a várias famílias. Assim, o impacto de determinadas famílias específicas (com específicos padrões de desequilíbrio de ligação) nas estimativas dos efeitos dos marcadores será minimizado. O efeito de ambiente pode ser minimizado por meio do uso de um número de repetições muito alto, que é a mesma filosofia da avaliação e seleção de características quantitativas com base em fenótipos em experimentos de campo. Predições são então obtidas para os efeitos dos haplótipos marcadores ou dos alelos em cada marcador.

Essas predições derivadas de dados fenotípicos e de genótipos em alta densidade em uma geração são então usadas para a obtenção dos valores genéticos genômicos (VGG) dos indivíduos de qualquer geração subsequente, tendo por base os seus próprios genótipos marcadores. Desta forma, como ressaltado anteriormente, na predição de valores genéticos tradicionais informações fenotípicas são utilizadas visando a inferência sobre os efeitos dos genótipos dos indivíduos e, na predição de valores genéticos genômicos informações genotípicas (genótipos para os alelos marcadores) são usadas visando a inferência sobre os valores fenotípicos futuros (ou valores genéticos genômicos preditos) dos indivíduos (RESENDE, 2008).

Apesar dos excelentes resultados obtidos na área de melhoramento animal (GODDARD; HAYES, 2007; KOLBEHDARI et al., 2007; LEGARRA; MISZTAL, 2008; LONG et al., 2007), essa técnica ainda não tem sido muito utilizada na área de melhoramento vegetal, sendo a quase totalidade dos trabalhos existentes resultantes de estudos de simulação (BERNARDO, 2008; BERNARDO, 2009; BERNARDO e YU, 2007; HEFFNER et al., 2009; LIU et al., 2008; RESENDE et al., 2008).

2.3 Validação na seleção genômica ampla

A seleção genômica ampla requer o uso de uma população de estimação ou treinamento para estimar os efeitos dos marcadores e uma população de validação, para analisar a eficiência da estimação destes efeitos na recuperação do valor genômico em uma população independente. Após a validação e estimação de um conjunto de

marcadores, estes serão utilizados em uma população que dispõe apenas de genotipagem (Figura 1).

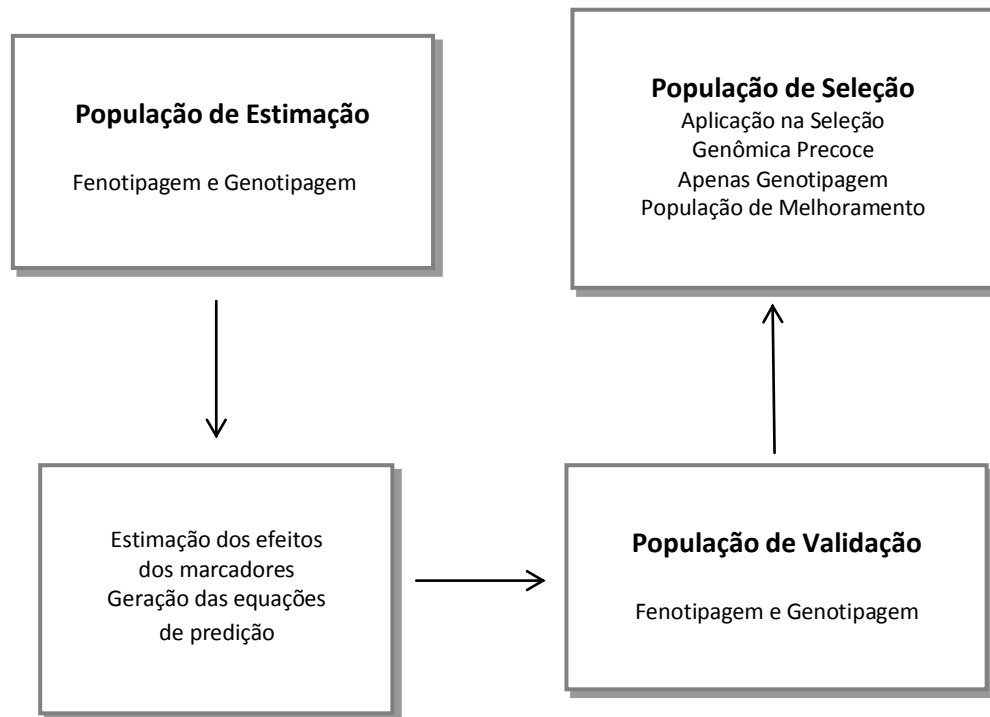


Figura 1-Esquema de aplicação da seleção genômica ampla em um programa de melhoramento genético.

Com isso, durante o processo três populações podem ser definidas: população de estimação, validação e seleção. Essas podem ser fisicamente distintas (três populações diferentes); exercer duas funções ao mesmo tempo (uma só população usada para estimação e validação) ou exercer três funções ao mesmo tempo (uma só população usada para estimação, validação e seleção).

A população de estimação também é denominada população de descoberta, de treinamento ou de referência. Nesta população um grande número de marcadores são avaliados em um número moderado de indivíduos, ou seja, em um número menor que a população de seleção, os quais devem ter seus fenótipos avaliados para os caracteres de interesse. Nessa primeira etapa são descobertos, via genotipagem, os marcadores que explicam os locos que controlam os caracteres, bem como são estimados os seus efeitos. A população de validação, por sua vez, quando fisicamente disjunta da população de estimação, possui um conjunto de dados menor do que aquele da população de descoberta e contém indivíduos avaliados para os marcadores e para os vários caracteres de interesse. Para computar a acurácia, os valores genéticos genômicos são preditos

(usando os efeitos estimados na população de estimação) e submetidos à análise de correlação com os valores fenotípicos observados. Como a amostra de validação não foi envolvida na predição dos efeitos dos marcadores, os erros dos valores genéticos genômicos e dos valores fenotípicos são independentes e a correlação entre esses valores é predominantemente de natureza genética.

Por fim, na população de seleção, há apenas a genotipagem dos candidatos à seleção e as equações de predição derivadas na população de descoberta são então usadas na predição dos valores genéticos genômicos (VGG) ou fenótipos futuros dos candidatos à seleção. Desta forma, nesta fase não é necessária a avaliação dos fenótipos, pois estes serão preditos com base nos valores estimados na população de referência.

2.4 Correção de fenótipos para estrutura de população

Os fenótipos devem ser corrigidos para os efeitos genéticos dos genitores, trabalhando-se basicamente com o efeito da segregação mendeliana desregressada, já que o dado ideal para a população de estimação deve ser o mérito genético verdadeiro de indivíduos não aparentados. O uso da segregação mendeliana capta a associação de alelos de marcas e de QTL, ou seja, captura efeitos genéticos explicados pelo desequilíbrio de ligação e não pelo parentesco ou genealogia (RESENDE, 2008). Selecionar com base somente nesta variação sustenta o progresso genético pela redução na seleção de aparentados e consequentemente, diminui a endogamia (WOOLLIAMS et al., 1999). Entretanto, cabe ressaltar que a ênfase da seleção genômica na segregação mendeliana não nega completamente a contribuição genética de cada indivíduo a longo termo, com consequente aumento na endogamia. Particularmente, seres superiores carregarão alelos superiores, e a seleção com base nesses alelos fará que indivíduos com esses genótipos deixem mais descendentes (DAETWYLER et al., 2007). Com isso, é aconselhável, mesmo no contexto da seleção genômica, a restrição da taxa de endogamia.

2.5 Metodologias de análise para a seleção genômica ampla

O desafio da análise de QTL é a seleção do modelo estatístico apropriado para identificar QTL e estimar os seus efeitos (BROMAN; SPEED, 2002). Em programas de

melhoramento, métodos estatísticos para a seleção genômica precisam estimar simultaneamente o efeito de muitos marcadores através de um número limitado de genótipos. Um maior número de variáveis explicativas (marcadores) do que observações (fenótipos) leva a uma falta de graus de liberdade que deve ser contornada através da seleção e do uso de modelos estatísticos apropriados, que resultem em uma maior acurácia dos valores genéticos preditos, levando em consideração a complexidade do modelo e a demanda computacional (HEFFNER et al., 2009).

Para isso, vários métodos de predição de valores genéticos genômicos foram propostos: quadrados mínimos (LS), BLUP/GWS, BayesA e BayesB (MEUWISSEN et al., 2001), regressão kernel não paramétrica via modelos aditivos generalizados (GIANOLA et al., 2006), aprendizado de máquina (LONG et al., 2007), regressão RKHS (*Reproducing Kernel Hilbert Spaces*) (GIANOLA; VAN KAAM, 2008), LASSO Bayesiano (DE LOS CAMPOS et al., 2009; PARK; CASELLA, 2008), Bayes B Acelerado (MEUWISSEN et al., 2009), Bayes C, BayesC π , BayesD, BayesD π (HABIER et al., 2011), Regressão via Quadrados Mínimos Parciais (PLSR) e Regressão via Componentes Principais (PCR) (SOLBERG et al., 2009). Os métodos BLUP (regressão aleatória), LASSO (*Least Absolute Shrinkage and Selection Operator*) e Bayes A e B pertencem à classe de regressão explícita. Por outro lado, o método RKHS pertence à classe de regressão implícita e é um método semiparamétrico (GIANOLA et al., 2009; GIANOLA; VAN KAAM, 2008).

Conforme Resende et al (2008) essas abordagens diferem na suposição sobre o modelo genético associado ao caráter quantitativo. O método quadrados mínimos (regressão fixa) é ineficiente devido à impossibilidade de estimar todos os efeitos simultaneamente, pois o número de efeitos a estimar é maior do que o número de dados; além disso, o método estima um efeito de cada vez e verifica a sua significância, o que conduz a superestimativas dos efeitos significativos, e, conseqüentemente, a baixos valores de acurácia. Uma alternativa comum é estimar os efeitos das marcas conjuntamente através de métodos de penalização como “*Ridge Regression*” (HOERL; KENNARD, 1970) ou métodos bayesianos (CROSSA et al., 2010).

O BLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos; o AM assume que existe um número limitado de genes a serem ajustados; o método BayesB é intermediário entre esses dois, assumindo poucos genes de grandes efeitos e muitos genes com pequenos efeitos. No método BayesB muitos efeitos de marcadores são assumidos como zero, a priori. Isso reduz o tamanho do genoma por

meioda concentração nas partes do mesmo onde existem QTL. O melhor método é aquele que reflete melhor a natureza biológica do caráter poligênico em questão, em termos de efeitos gênicos.

O BLUP/GWS presuppõe os efeitos de QTL com distribuição normal com variância constante para todos os segmentos cromossômicos. Esse método contorna, por meio da estimação simultânea dos efeitos de todos os marcadores, a questão da necessidade de estimação de um grande número de efeitos a partir de um tamanho amostral restrito e, adicionalmente, a questão do fato de que muitos efeitos mostram colinearidade advinda do desequilíbrio de ligação entre os próprios marcadores. O RR-BLUP (*Ridge Regression* - BLUP), assume que todas as marcas são aleatórias com variância comum (MEUWISSEN et al., 2001). Igual variância não assume que todos os efeitos são iguais (BERNARDO e YU, 2007), mas que todas as marcas possuem o mesmo efeito de shrinkage. Entretanto, assumir que as marcas individuais possuem a mesma variância não é um fato real, com isso, o RR-BLUP incorretamente trata todos os efeitos igualmente (XU, 2003).

Para contornar esta limitação alguns autores propuseram métodos que formulam um fator de shrinkage específico para cada marca, como Bayes A e Bayes B de Meuwissen et al. (2001) e o Bayesian LASSO de Park e Casella (2008). Desta forma, os valores genéticos preditos combinarão informações sobre a variância da distribuição a priori e dos dados utilizados (MEUWISSEN, 2001).

Através de estudos de simulação Meuwissen et al. (2001) verificaram que o Bayes A foi superior aos métodos de regressão de stepwise e o RR-BLUP, proporcionando uma acurácia de 0,798. Usai et al. (2009) comparando métodos como LASSO-LARS, BLUP e Bayes A observaram que a acurácia seletiva de cada metodologia correspondeu, respectivamente, a 0,89; 0,75; 0,84, apontando para a superioridade do método bayesiano sobre o BLUP. De acordo com os autores, o método BLUP localizou corretamente os QTL de maior efeito, entretanto, quando o efeito dos QTL era pequeno a predição de seus efeitos não diferenciava daqueles onde não havia QTL. Os efeitos dos marcadores preditos pelo Bayes A e LASSO-LARS mostraram uma melhor correspondência com os verdadeiros efeitos dos QTL. Contudo, o Bayes A não localizou a posição de alguns QTL, em particular alguns de menor efeito. Moser et al. (2009) por sua vez, ao compararem cinco metodologias de seleção genômica, mínimos quadrados (FR-LS), regressão bayesiana utilizando Bayes A (Bayes-R), melhor predição aleatória não viesada (RR-BLUP), regressão de quadrados mínimos

parcial (PLSR) e regressão suporte não paramétrica (SVR), em touros, verificaram a superioridade desses quatro últimos métodos em relação a predição de mínimos quadrados, observando ainda que a análise bayesiana e o RR-BLUP produziram o mesmo valor de acurácia para as características avaliadas, entretanto, o RR-BLUP gerou predições não viesadas.

O LASSObayesiano (BLASSO) permite que alguns estimadores sejam identicamente iguais a zero, realizando simultaneamente o procedimento de shrinkage e seleção de modelos. Na seleção genômica é esperado que muitos segmentos assumam o valor zero por não possuírem QTL. Desta forma, um parâmetro que tenha valor posterior esperado zero é normalmente especificado através de uma distribuição a priori que coloque um peso maior no valor zero do que as prioris padrões, que no caso do LASSO Bayesiano é a exponencial dupla (DE).

De acordo com de Los Campos et al. (2009), em relação à regressão bayesiana, o estimador LASSO, produz um maior efeito de shrinkage nos coeficientes de regressão próximos a zero e menor naqueles que possuem um maior valor absoluto. Com isso, se é possível que muitos marcadores não estejam associados a QTL que efetivamente contribuam para a característica de interesse, uma melhor escolha seria utilizar como priori a distribuição de exponencial dupla.

No geral, métodos bayesianos são superiores quando a distribuição dos efeitos dos QTL é leptocúrtica (curtose positiva), devido à presença de genes de grandes efeitos. Com distribuição normal dos efeitos dos QTL, o método RR-BLUP é igualmente eficiente (RESENDE, 2008). Crossa et al. (2010) verificaram que o BLASSO apresentou performance superior ao BLUP ao praticarem a seleção genômica em milho e trigo. Estes autores asseveram que esta resposta poderia ser atribuída ao fato de que assim como outros métodos de seleção genômica como o Bayes A e o Bayes B de Meuwissen et al. (2001), o BLASSO aplica um fator de shrinkage específico para cada marca, enquanto o BLUP aplica a mesma regularização em todas as marcas; e também ao fato de que no BLASSO tanto as variâncias quanto os efeitos de marcas são estimados em conjunto, evento que no BLUP é realizado em dois passos distintos (primeiro os componentes de variância e depois cada marca é estimada).

3 MATERIAL E MÉTODOS

3.1 Descrição das populações

A fim de se avaliar os efeitos da seleção genômica ampla foram realizadas 10 simulações das estruturas populacionais aplicadas no estudo. A estrutura de cada simulação foi criada para apresentar dois conjuntos de dados: fenotípicos e genotípicos de cada indivíduo dentro de uma população, imitando alguns dos cenários em que a seleção genômica foi avaliada.

Os dados de simulação foram obtidos por meio do aplicativo GENES (Cruz, 2013) onde se considerou uma população de vinte e cinco famílias de irmãos completos com vinte indivíduos em cada família. Desta forma, o tamanho das populações simuladas foi de 500.

O primeiro tipo de população gerada, constituída pelas famílias de irmãos completos, foi denominada em nosso estudo de geração um (G1). Esta geração foi derivada do cruzamento, planta a planta, entre duas populações genitoras P1 e P2, em equilíbrio de Hardy-Weinberg. A geração um, foi obtida com o intuito de constituir a população de estimação e de validação nas formas de validação V1 e V2 (descritas no item 3.5) e a população de treinamento da validação V3. A fim de gerar a população de validação para a validação V3, obteve-se um segundo tipo de população, ou geração dois

(G2) que, por sua vez, foi confeccionada através do acasalamento ao acaso dos irmãos completos dentro de cada família.

Em cada simulação de dados foram considerados 1000 locos gênicos que expressavam dois alelos codominantes. Dentro dos 1000 locos gênicos simulados, 100 foram caracterizados como associados à característica quantitativa (QTL).

3.2 Simulação dos valores genotípicos e fenotípicos

Para proceder às análises de seleção genômica ampla, foram simulados dados genotípicos e fenotípicos considerando ausência de dominância, diferentes herdabilidades (h^2), e diferentes formas de distribuição dos efeitos dos genes.

Foram simulados caracteres quantitativos com três herdabilidades, que expressam o quadrado da correlação entre o valor do fenótipo e do genótipo, sendo estas: 0,20; 0,40 e 0,60. Para estudar diferentes distribuições de efeitos genéticos, simulou-se dois cenários de efeitos de QTL: (i) genes com efeitos de distribuição uniforme, com o intuito de reproduzir características controladas por vários genes de pequenos efeitos; (ii) genes com efeitos de distribuição exponencial, com o objetivo de retratar caracteres controlados por poucos genes de grandes efeitos, além de vários genes de pequenos efeitos.

A função de densidade da distribuição uniforme e a da distribuição exponencial estão descritas abaixo, respectivamente:

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{se } a \leq x \leq b; \\ 0, & \text{caso contrário.} \end{cases}$$

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{se } x \geq 0; \\ 0, & \text{se } x < 0. \end{cases}$$

Os genótipos dos indivíduos foram gerados segundo o modelo $F_i = G_i + E_i$, em que G_i é o efeito genético dado pelo somatório dos efeitos genéticos em cada loco e E_i o efeito ambiental, gerado segundo uma distribuição normal com média e variância compatível com a herdabilidade do caráter simulado.

O valor genético total expresso por um determinado indivíduo foi estimado a partir da expressão:

$$G_i = \mu + a_i + d_i$$

Em que:

$$a_i = \sum_{j=1}^{1000} p_j \alpha_j$$

$$d_i = 0$$

Sendo α_j é o efeito do alelo favorável no loco j , considerado igual a 1, 0 ou -1 para as classes genóticas AA, Aa e aa, respectivamente, e p_j é a contribuição do loco j para a manifestação da característica considerada, no trabalho, como tendo distribuição uniforme ou exponencial (Figura 2). Foi estabelecido que os 100 primeiros locos genotipados foram os controladores da característica.

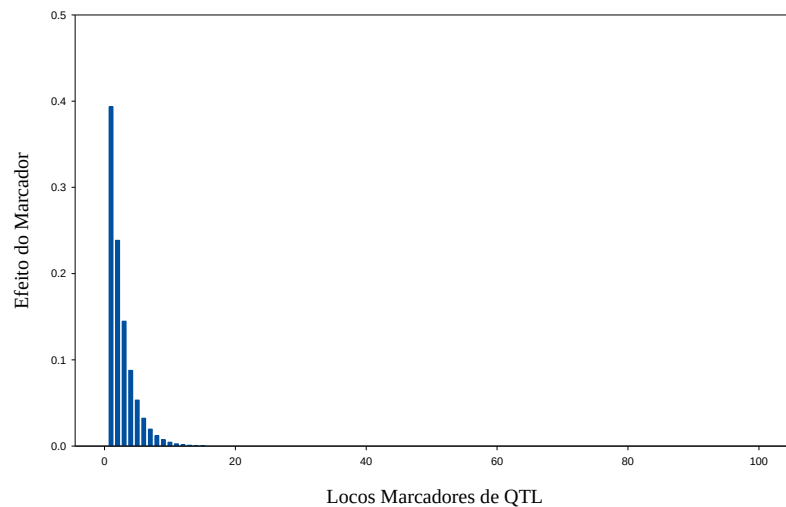


Figura 2- Distribuição exponencial do efeito dos 100 primeiros locos marcadores sobre uma característica quantitativa.

3.3 Caracterização genética das populações simuladas

Para entender o comportamento das populações é importante o conhecimento dos parâmetros genéticos e estatísticos das consequências do uso de métodos de melhoramento genético. Uma vez que as características quantitativas sofrem grande

influência do ambiente, seu estudo é dado em nível de populações, e algumas medidas utilizadas para definir uma população são média, variância e endogamia.

Por essa razão, com base nos valores genotípicos e fenotípicos das populações em todas as gerações, foram estimados os parâmetros: média (μ), variância (σ^2), e a endogamia (F).

Adicionalmente, as populações foram caracterizadas quanto a distribuição de seus fenótipos a fim de se avaliar a normalidade dos dados. Para isso, foi aplicado o teste de Lilliefors, além de se obter os coeficientes de simetria (s) e curtose (k) das populações geradas.

3.4 Metodologias de análise

A partir dos dados fenotípicos gerados, foram estimados os efeitos de cada um dos locos marcadores que somados, compõem o valor genético genômico predito de cada indivíduo. Para isso, foram utilizadas as metodologias RR-BLUP e BLASSO.

3.4.1 Ridge Regression BLUP (RR-BLUP)

Esse método é denominado regressão aleatória ou regressão de cumeieira (*Ridge Regression-Best Linear Unbiased Prediction- RR-BLUP*). Os coeficientes de regressão *ridge* são definidos como aqueles que minimizam a soma de quadrados dada por:

$$(1/N) \sum_j^N (y_j - \sum_{i=1}^n x_{ij} \beta_i)^2 + \lambda \sum_{i=1}^n \beta_i^2$$

em que λ é o parâmetro de penalização (ou shrinkage) ou parâmetro *ridge*, n é o número de marcadores e N é o número de indivíduos. O primeiro termo da equação é a soma de quadrados dos resíduos da regressão e o segundo termo é a penalização, a qual depende da magnitude dos coeficientes de regressão via $\sum_{i=1}^n \beta_i^2$.

Quando λ não é conhecido, a escolha arbitrária do mesmo leva ao método de regressão *Ridge Regression* (RR). Se o parâmetro de regressão for associado a $\lambda = \sigma_e^2 / \sigma_{gi}^2 = \sigma_e^2 / (\sigma_g^2 / n)$ tem-se a regressão aleatória BLUP para o efeito do segmento cromossômico i , em que σ_{gi}^2 é a variância genética associada ao loco ou segmento i e σ_g^2 e σ_e^2 são a variância genética do caráter e variância residual, respectivamente. A quantidade n é desconhecida a priori, mas pode ser inferida por via iterativa ou sintonia,

escolhendo-se aquele que maximiza a correlação entre valor fenotípico e valor genético predito na validação cruzada.

O seguinte modelo linear misto geral será usado para estimar os efeitos dos marcadores, conforme Resende et al. (2008).

$$y = Xb + Za + e,$$

em que y é o vetor de observações fenotípicas,

b é o vetor de efeitos fixos,

a é o vetor dos efeitos aleatórios dos marcadores e

e refere-se ao vetor de resíduos aleatórios.

X e Z são as matrizes de incidência para b e a .

A estrutura de médias e variâncias no modelo em questão é definida como:

$$a \sim N(0, G)$$

$$E(y) = Xb$$

$$e \sim N(0, R = I\sigma_e^2)$$

$$Var(y) = V = ZGZ' + R$$

$$G = I\sigma_g^2/n$$

As equações de modelo misto genômicas para a predição de avia o método BLUP/GWS equivalem a:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + I \frac{\sigma_e^2}{(\sigma_g^2/n)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

em que σ_g^2 refere-se à variância genética total do caráter e σ_e^2 é a variância residual. O valor genético genômico global do indivíduo j é dado por $VGG = \hat{y}_j = \sum_i Z_i \hat{a}_i$, em que Z_i equivale a -1, 0 ou 1 para os genótipos A1A1, A1A2 e A2A2, respectivamente, para marcadores bialélicos e codominantes.

As equações de predição apresentadas acima foram modeladas com σ_g^2 comum. Assim, a variação genética explicada por cada loco é dada por σ_g^2/n , em que σ_g^2 é a variação genética total e n é o número de marcadores utilizados em cada um dos testados. Essa estratégia foi adotada por Meuwissen et al. (2001), Muir (2007), Bernardo (2007) e Kolbehdari et al. (2007).

Na predição dos efeitos aleatórios via BLUP/GWS, não há necessidade de uso da matriz de parentesco (SCHAEFFER, 2006). A matriz de parentesco baseada em *pedigree* usada no BLUP tradicional foi substituída pela própria matriz $Z'Z$ que é uma matriz de parentesco estimada pelos marcadores.

3.4.2 LASSO Bayesiano

O método LASSO (*Least Absolute Shrinkage and Selection Operator*) combina shrinkage (regularização) com seleção de variáveis e envolve o seguinte problema de otimização, via minimização de:

$$(1/N) \sum_j^N (y_j - \sum_{i=1}^n x_{ij} \beta_i)^2 + \lambda \sum_{i=1}^n |\beta_i|$$

é que $\sum_{i=1}^n |\beta_i|$ é a soma dos valores absolutos dos coeficientes de regressão. As soluções em que os coeficientes de regressão se distanciam de zero sofrem penalização. O componente $\lambda \sum_{i=1}^n |\beta_i|$ regulariza a regressão sem penalizar muito. O parâmetro de suavização λ controla a intensidade da regularização. Quando este último parâmetro é igual a zero, não há regularização. No LASSO Bayesiano, esse parâmetro controla a precisão da distribuição a priori atribuída aos coeficientes de regressão. A implementação desse tipo de regularização envolve encurtamento mais forte no sentido de que alguns coeficientes de regressão tenham valores iguais a zero. Para isso, pode ser implementado via análise Bayesiana do LASSO (DE LOS CAMPOS et al., 2009). Esta implementação impõe como distribuição a priori dos p coeficientes de regressão um produto de densidades exponenciais duplas: $p(\beta|\lambda) = \prod_{j=1}^p \frac{\lambda}{2} \exp(-\lambda|\beta_j|)$.

Para construção da distribuição conjunta a priori dos parâmetros, os autores exploram o fato de a distribuição exponencial dupla poder ser representada como uma mistura de densidades normais com parâmetro de escala e com o processo de mistura de variâncias controlado por distribuição exponencial. Na distribuição a priori conjunta construída, a densidade atribuída aos coeficientes de regressão regularizados por LASSO será: $\prod_{j=1}^p N(\beta_{j1}|0, \sigma_e^2 \tau_j^2)$, resultando em variância específicas para cada coeficiente de regressão. Por sua vez, a distribuição a priori para o parâmetro de escala τ_j^2 será representada por: $\prod_{j=1}^p \exp(\tau_j^2|\lambda)$, pela qual o parâmetro de suavização λ influencia o ajuste dos coeficientes de regressão e $G(\lambda^2|\alpha_1, \alpha_2)$ é uma função gama com hiperparâmetros conhecidos. Na implementação de De Los Campos et al., (2009) a priori λ é dada por: $p(\lambda) \propto \text{Beta}(\lambda/100)|\alpha_3 = 1.4, \alpha_4 = 1.4$.

3.5 Validação

A validação cruzada foi realizada pela reamostragem de um grupo de indivíduos via procedimento Jackknife (HELTSHE; FORRESTER, 1983). A metodologia generalizada do Jackknife baseia-se na divisão do conjunto de C dados amostrais em g grupos de tamanho igual a k , de forma que $C = gk$. Em cada um dos g grupos, k indivíduos são retirados para a formação da população de validação. No caso, tomou-se $k=100$ e, portanto, a população de validação usou todos os $C=500$ indivíduos, e a população de estimação, 400 indivíduos e 5 repetições. Em cada repetição, cem indivíduos foram removidos da população e utilizados para a formação da população de validação, e os outros indivíduos restantes foram utilizados na população de estimação dos efeitos dos marcadores. Este tipo de validação recebeu a denominação de validação tipo dois (V2) e foi executada através do software SELEGEN GENÔMICA (RESENDE, 2007).

Além da V2, utilizou-se a validação na qual 100 indivíduos, constituídos pelas últimas quatro famílias simuladas, foram retirados da amostra para servir como população de validação. Assim, a população de estimação foi composta por 400 indivíduos e uma repetição e a de validação por 100. Esta forma de validação independente foi denominada de validação tipo um (V1). Em outro cenário, avaliou-se a utilização de uma nova população de validação (G2), gerada através do acasalamento entre os indivíduos da população de estimação; esta foi nomeada de validação tipo três (V3), desta forma, nesta validação foram utilizados 500 indivíduos para a população de estimação e 500 para a população de validação. Ambas metodologias foram efetuadas através do software R. A tabela 1 resume o número de indivíduos que constituíram tanto população de estimação, quanto população de validação de cada metodologia.

Como na validação dos resultados o valor fenotípico é conhecido, a eficiência da seleção genômica em cada subgrupo analisado foi avaliada ao calcular a correlação do valor genético predito com o fenótipo observado nos indivíduos. Esta correlação é conhecida como capacidade preditiva ($r_{y\hat{y}}$) da seleção genômica em estimar os fenótipos, sendo dada teoricamente pela acurácia de seleção ($r_{g\hat{g}}$)

Tabela 1. Número de indivíduos que constituíram a população de estimação e a população de validação nas diferentes formas de validação.

Forma de Validação	População de Estimação	População de Validação
V1	400	100
V2	400	500
V3	500	500

multiplicada pela raiz quadrada da herdabilidade individual (h) ou, em outras palavras $r_{y\hat{y}} = r_{g\hat{g}} \cdot h$ (RESENDE, 2008).

Outra estatística usada na avaliação da eficiência da GWS foi a acurácia seletiva, dada pelos coeficientes de correlação envolvendo valores genéticos genômicos verdadeiros (GBV fixados na simulação) e preditos ($G\hat{BV}$).

3.6 Correção dos fenótipos

Como ressaltado anteriormente, os fenótipos devem ser corrigidos para os efeitos genéticos dos genitores, trabalhando-se basicamente com o efeito da segregação mendeliana desregressada, já que o dado ideal para a população de estimação deve ser o mérito genético verdadeiro de indivíduos não aparentados. Para a correção dos fenótipos utilizou-se o software SELEGEN-REML/BLUP, que lança mão da análise de modelos mistos para identificar o efeito estimado de cada indivíduo dentro da família. Após a correção destes dados para o efeito de genitores, eles foram utilizados nos estudos de cada cenário analisado.

Em caso de testes de progênie de irmãos germanos o modelo ajustado é dado por $y = Xb + Zg + Wc + e$. O valor individual corrigido ($\hat{f}_{ijk}^*$) para o valor genético médio de seus genitores ($\hat{g}_j$ e $\hat{g}_k$) com matriz de incidência Z , são dados pela expressão:

$$\hat{f}_{ijk}^* = (y_i - 0,5\hat{g}_j - 0,5\hat{g}_k - \hat{b}_i - \hat{c}_i)$$

em que $\hat{b}_i$ e $\hat{c}_i$ são os efeitos estimados de blocos (se houverem) e de capacidade específica (se o modelo for com dominância), com respectivas matrizes de incidência X e W . Neste estudo não foram simulados efeitos de bloco e de capacidade específica.

3.7 Análises estatísticas

As simulações foram implementadas através do software GENES (CRUZ, 2006) e as análises estatísticas através dos softwares SELEGEN GENÔMICA (RESENDE, 2007), SELEGEN-REMLBLUP (RESENDE, 2002), R via o pacote rr-BLUP (ENDELMAN et al., 2012) e via pacote BLR (De los Campos et al., 2010).

4 RESULTADOS E DISCUSSÃO

4.1 Simulação

Acompanhando o comportamento da estrutura das populações da geração um (G1) para a geração dois (G2) verifica-se que houve um acréscimo no coeficiente de endogamia no decorrer dos dois ciclos de estudo (Tabela 2). A endogamia é o fenômeno que ocorre em decorrência do acasalamento entre indivíduos aparentados. O coeficiente de endogamia refere-se à probabilidade de que os alelos de um loco de um indivíduo sejam idênticos por ascendência. Esses alelos são idênticos quando derivam ou são cópias de um alelo comum encontrado nos ancestrais daquele indivíduo (FALCONER, 1989).

Tabela 2. Tabela de heterozigosidade, endogamia e PIC para as gerações um e dois.

	Heterozigosidade Esperada	Heterozigosidade Observada	Endogamia (F)	PIC
G1	0,491	0,502	-0,023	0,370
G2	0,491	0,407	0,170	0,370

Segundo Wright, endogamia ocorre em consequência da identidade dos gametas que se unem, sendo expressa pela correlação entre os valores dos gametas que formam a progênie derivada de uma população. Em uma população panmítica, onde ocorre o acasalamento ao acaso, essa correlação é nula, porém, em uma população endogâmica, a correlação não é nula e é proporcional ao coeficiente de endogamia (CRUZ, 2005).

Segundo Cruz (2005) o coeficiente de endogamia médio de uma progênie é igual ao coeficiente de parentesco de seus genitores. Os valores de endogamia foram dados por: $F = He/Ho$, sendo He a heterozigosidade esperada e Ho a heterozigosidade observada. Uma vez que a geração um foi obtida através da hibridação planta a planta de duas populações genitoras em equilíbrio de Hardy-Weinberg e, a geração dois, por sua vez, foi obtida através do acasalamento ao acaso dentro de cada família de irmãos completos, constata-se que o comportamento da endogamia está como o previsto para essas estruturas de populações. De acordo com Cruz (2006) valores negativos de endogamia podem ser considerados valores nulos, desta forma, observa-se que não houve endogamia na geração um, pois não havia relação de parentesco entre os progenitores P1 e P2 da geração em questão. Por outro lado, em consequência do acasalamento entre os indivíduos aparentados da geração um (irmãos completos) verifica-se um aumento no grau de endogamia com o decorrer dos ciclos. Desta forma, nota-se que a simulação foi fiel ao retratar o que seria esperado na estrutura dessas duas gerações ao longo dos cruzamentos, pois os valores de endogamia passaram de -0,023 ($F=0$) para 0,170.

A heterozigosidade observada é dada por $Ho = \sum_{j=1, j \neq i}^a n_{ij} / N$, sendo o número de heterozigotos representado por n_{ij} . A frequência de heterozigotos esperada (He) representa, por outro lado, uma população supostamente em equilíbrio de Hardy-Weinberg quando se consideram dois alelos por loco. Desta forma tem-se $He = 1 - \sum_{i=1}^a p_i^2$.

Com o aumento da endogamia espera-se, como consequência, um menor número de indivíduos heterozigotos, ou seja, uma menor heterozigosidade. Os valores de heterozigosidade decrescem numa relação de $2pqd-2pqdF$ com a endogamia (Quadro 1). Esse comportamento também foi preservado durante a simulação dos caracteres, de forma que os valores de heterozigosidade observada transitaram de 0,502 para 0,407.

Quadro 1-Valores genotípicos e frequência esperada em uma população endogâmica.

Genótipo	Frequência	Valor	Frequência x Valor
A^1A^1	p^2+pqF	$+ a$	$p^2a+ pqaF$
A^1A^2	$2pq-2pqF$	d	$2pqd-2pqdF$
A^2A^2	q^2+pqF	$- a$	$-q^2a- pqaF$

a = efeito aditivo; d = desvios devido à dominância; p e q = frequência dos alelos A^1 e A^2 , respectivamente e F = endogamia.Fonte: Falconer, 1989.

Os marcadores foram polimórficos, apresentando valores de PIC (conteúdo de informação de polimorfismo)de 0,37. De acordo com Cruz (2006), quando o $PIC \geq 0,1$ o loco será polimórfico, sendo que, neste caso, o alelo mais frequente tem frequência menor que 0,95.

O PIC é usado para quantificar o polimorfismo da população. O valor de PIC fornece uma estimativa do poder discriminatório do marcador por considerar não somente o número de alelos por loco, mas também a frequência relativa desses alelos.

$$PIC = 1 - \sum_{i=1}^a p_i^2 - \sum_{i,j=1}^a \sum_{(i \neq j)} p_i^2 p_j^2$$

Os dados referentes à média foram simulados com o valor 100, e foram devidamente retratados (Tabela 2). A média genotípica da população é dada por $M_O = p^2(\mu+a) + 2pq(\mu+d) + q^2(\mu-a)$, desta forma a obtem-se $M_O = \mu + a(p - q) + 2pqd$. A endogamia altera a média da população com as sucessivas gerações. Através do Quadro 1 verifica-se que a média de uma população endogâmica pode ser obtida da seguinte maneira:

$$M_F = a(p - q) + 2pqd - 2pqdF$$

Considerando-se o efeito combinado de todos os locos que influenciam o caráter e que os valores genotípicos dos locos combinam-se aditivamente, a média da população será dada pela soma da contribuição dos locos separados como descrito abaixo:

$$M_F = M_O - 2F \sum dpq$$

M_O = média da população original.

Estas expressões mostram as circunstâncias sob as quais um caráter métrico mostrará mudança no valor médio. A mudança na média, resultante da endogamia, é, portanto, $- 2dpqF$. Entretanto, observa-se que a média não decaiu da geração um para a

geração dois, ou seja, manteve-se constante independente do tipo de herdabilidade e distribuição de efeitos de QTL (Tabela 3 e 4). Esse fato mostra que um loco contribuirá para a mudança do valor médio sob endogamia, só quando d não for zero, ou seja, se o valor do heterozigoto diferir do valor médio dos homozigotos. Uma vez que neste estudo o valor simulado de $d=0$ (ausência de dominância), depreende-se que o comportamento da média está de acordo com o esperado. Se não há dominância, não haverá influência na média da população endogâmica, ou seja, na geração dois.

Observou-se que quanto maior o valor da endogamia, maior o valor da variância genética total da população. A variância genética total na população, como um todo, é a soma dos componentes dentro de linhas com aqueles entre linhas e é igual a $(1+F)$ vezes a variância genética original $[(1+F)V_G]$, o que também é correto para endogamia estreita (Quadro 2). Quando a endogamia é completa, a variância genética na população, como um todo, é dobrada e toda variância aparece como o componente entre linhas (FALCONER, 1989). A decomposição apresentada no Quadro 2 é característica de populações com ausência de dominância, se encaixando com os dados em estudo. Nesta situação, de ausência de dominância, a variância genética da população é dada por $V_G = 2pqa^2$.

Tabela 3- Tabela de média e variância dos valores fenotípicos nas gerações um e dois, para as características de distribuição uniforme e exponencial.

Distribuição dos efeitos de QTL		Valores Fenotípicos					
		Herdabilidades					
		0,20		0,40		0,60	
		G1	G2	G1	G2	G1	G2
Uniforme	Média	100,050	100,080	99,993	99,994	99,993	99,995
	Variância	0,450	0,570	0,202	0,267	0,138	0,174
Exponencial	Média	100,050	100,082	99,993	99,995	99,993	99,984
	Variância	0,399	0,620	0,202	0,267	0,138	0,175

G1= geração 1; G2=geração 2

Tabela 4- Tabela de média e variância dos valores genotípicos nas gerações um e dois, para as características de distribuição uniforme e exponencial.

Distribuição dos efeitos de QTL		Valores Genotípicos					
		Herdabilidades					
		0,20		0,40		0,60	
		G1	G2	G1	G2	G1	G2
Uniforme	Média	100,050	100,080	99,993	99,994	99,993	99,995
	Variância	0,089	0,121	0,081	0,106	0,081	0,106
Exponencial	Média	100,050	100,081	99,993	99,995	99,993	99,984
	Variância	0,088	0,162	0,081	0,106	0,081	0,105

G1= geração 1; G2=geração 2

Como a endogamia da geração dois apresentou valores mais elevados que na geração um, espera-se que a variância também se eleve, comportamento que também foi mantido durante a simulação.

Quadro 2- Decomposição da variância por causa de genes aditivos, numa população com coeficiente de endogamia F , quando a variância, em virtude de genes aditivos na população base for V_G .

Variância	Valor
Entre Linhas	$2FV_G$
Dentro de Linhas	$(1-F)V_G$
Total	$(1+F)V_G$

Fonte: Falconer, 1989.

Ao se avaliar a normalidade dos fenótipos, através do teste de Lilliefors, observou-se que estes possuem uma aproximação razoável da distribuição normal. A simetria foi estatisticamente igual a zero, consistente com a simetria da distribuição normal (Tabela 5). Valores acima ou abaixo de zero indicam assimetria da curva. Os valores de curtose variaram entre 2,87 e 2,97. O teste fornece o valor de curtose $c=k+3$; portanto, este intervalo numérico aponta para uma curva do tipo mesocúrtica (k estatisticamente igual a zero), em consonância com a distribuição normal.

Tabela 5-Tabela de simetria e curtose dos dados fenotípicos

Distribuição dos efeitos de QTL		0,20		0,40		0,60	
		G1	G2	G1	G2	G1	G2
Uniforme	Simetria	-0,038 ^{NS}	0,045 ^{NS}	-0,075 ^{NS}	-0,023 ^{NS}	-0,056 ^{NS}	-0,012 ^{NS}
	Curtose	2,974 ^{NS}	2,874 ^{NS}	2,905 ^{NS}	2,925 ^{NS}	2,873 ^{NS}	2,909 ^{NS}
Exponencial	Simetria	-0,060 ^{NS}	-0,001 ^{NS}	-0,075 ^{NS}	-0,023 ^{NS}	-0,056 ^{NS}	-0,013 ^{NS}
	Curtose	2,950 ^{NS}	2,873 ^{NS}	2,905 ^{NS}	2,925 ^{NS}	2,873 ^{NS}	2,913 ^{NS}

^{NS} Não significativo a 5% de probabilidade pelo teste t.

O quadrado da correlação entre o valor fenotípico e genotípico expressa a herdabilidade das características, valores esses que permaneceram próximos aos estabelecidos para a simulação (dados não apresentados), garantindo assim, a adequação das variâncias fenotípicas, genotípicas e ambientais simuladas para cada situação ($h^2=0,2$; $h^2=0,4$ e $h^2=0,6$). Em situações de baixa herdabilidade pode-se prever que o melhorista terá certa dificuldade em selecionar genótipos que, de fato, apresentam desempenho superior ou, ainda, de descartar aqueles de desempenho genético não favorável uma vez que a associação entre o fenótipo e genótipo é de baixa magnitude em razão da ação dos efeitos perturbadores do ambiente.

4.2 Efeito de Marcadores

Em todos os programas de melhoramento genético, a predição dos valores genéticos é parte essencial para que obtenha ganhos com a seleção. A maneira mais simples de se deduzir o genótipo do QTL é tratando as marcas como se fossem o QTL e, assim, estimar os efeitos dos alelos marcadores ou genótipos. O parâmetro chave é a proporção da variância do QTL explicada pelas marcas (Goodard e Hayes, 2007). Como enfatizado anteriormente, foi estabelecido durante a simulação que os primeiros 100 locos marcadores seriam associados aos QTL.

As figuras a seguir (Figuras 3 a 6) retratam a eficiência de ambas as metodologias de análise em estabelecer os 100 primeiros marcadores como associados à característica quantitativa, conforme simulação. As estimativas dos efeitos de marcadores foram semelhantes independente do tipo de distribuição dos efeitos genéticos (uniforme ou exponencial), para ambas as metodologias de análise, RR-BLUP

e BLASSO. Com isso, verifica-se que a análise realizada através de RR-BLUP foi mais precisa em retratar o comportamento da característica uniforme, devido à menor variação entre os efeitos de marcadores estimados e, por outro lado, o BLASSO foi mais consistente em descrever as variações de genes com distribuição exponencial. Este resultado está de acordo com o esperado dentro das suposições sobre o modelo genético associado ao caráter quantitativo de cada metodologia. No método RR-BLUP os efeitos genéticos assumem um modelo infinitesimal com muitos locos de pequenos efeitos, comportamento simulado pela distribuição uniforme dos efeitos genéticos; e no método BLASSO, pressupõe-se poucos genes de maior efeito controlando o caráter, o que, no estudo em questão, foi retratado pela distribuição exponencial dos efeitos genéticos. Desta forma, depreende-se que a suposição do modelo genético de cada metodologia influencia diretamente nos valores estimados dos efeitos de marcas.

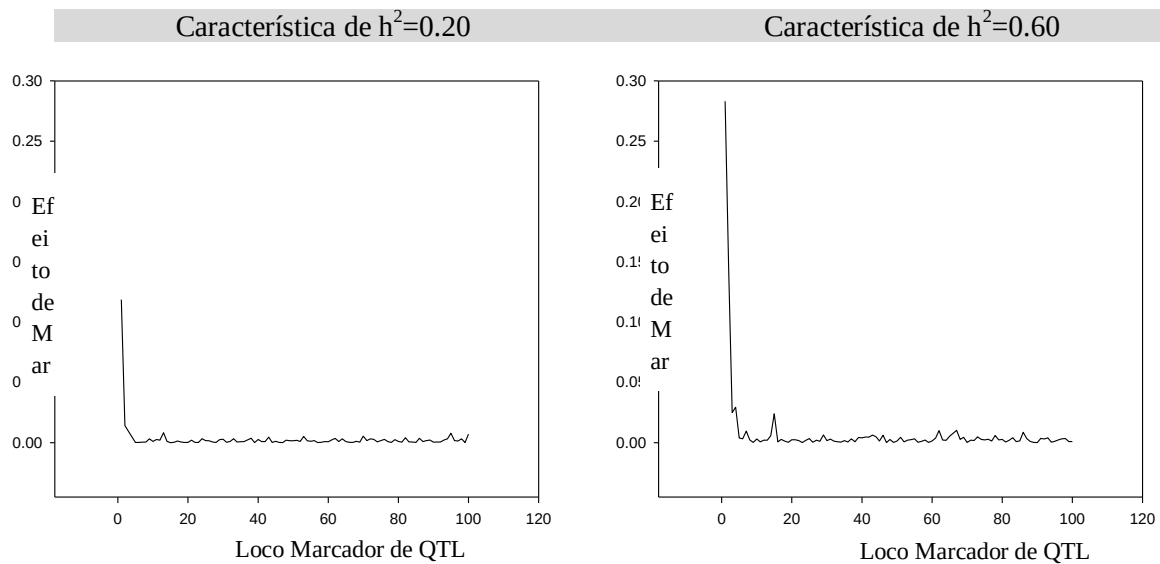


Figura 3- Efeito de marcadores para o fenótipo corrigido com distribuição exponencial. Método de estimação BLASSO.

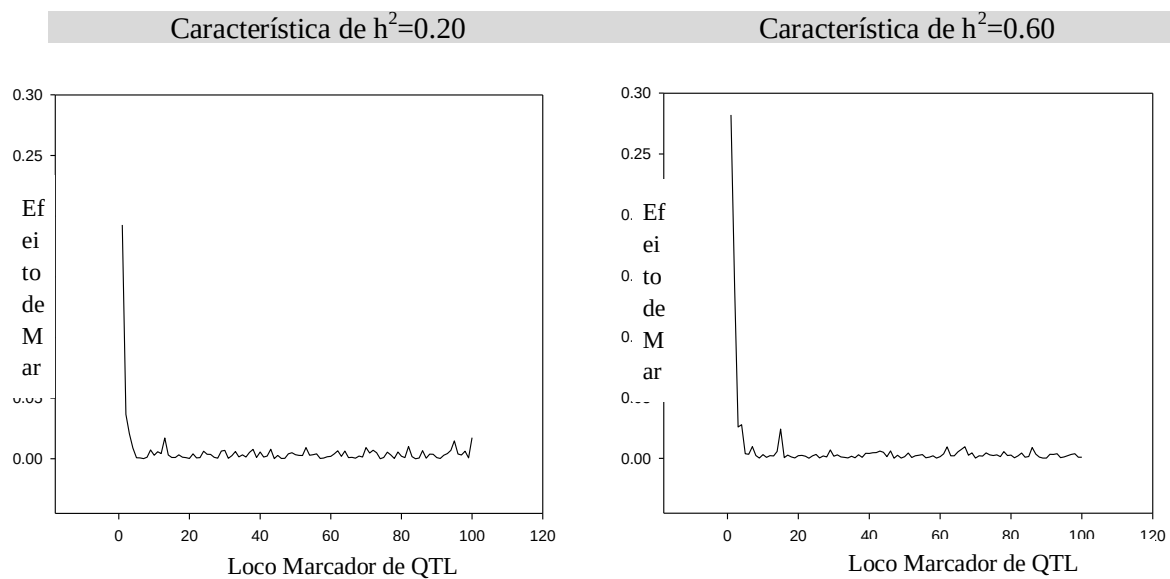


Figura 4- Efeito de marcadores para o fenótipo corrigido com distribuição uniforme. Método de estimação BLASSO.

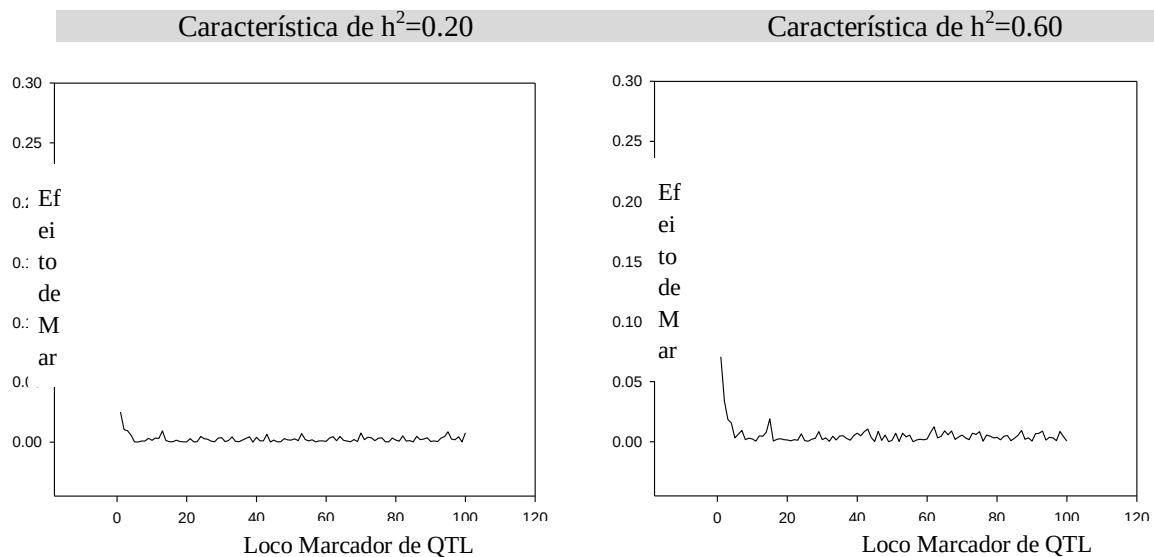


Figura 5- Efeito de marcadores para o fenótipo corrigido com distribuição exponencial. Método de estimação RR-BLUP.

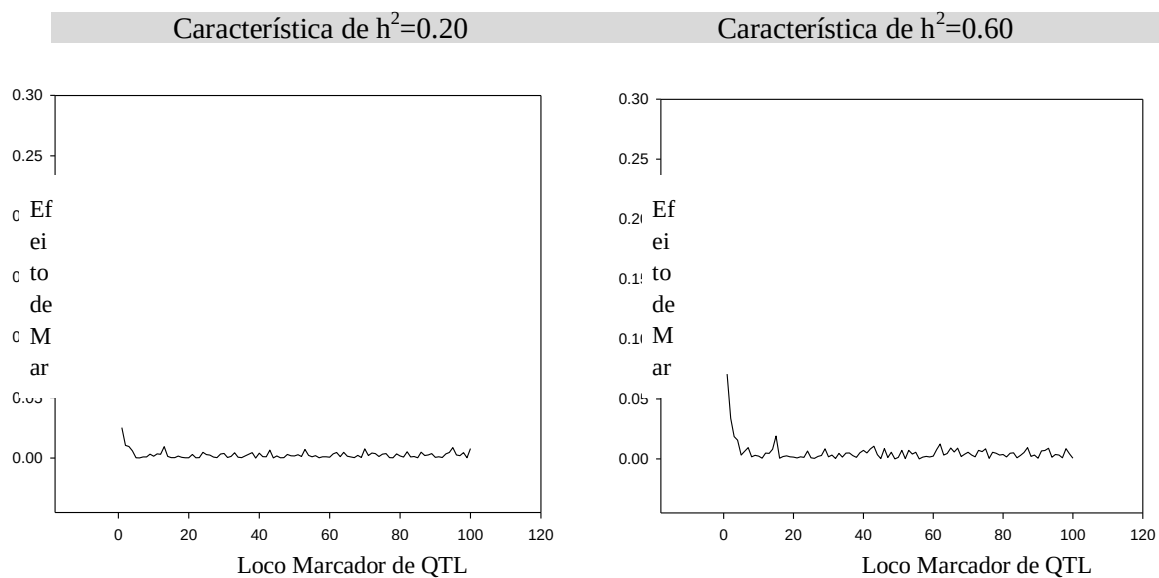


Figura 6- Efeito de marcadores para o fenótipo corrigido com distribuição uniforme. Método de estimação RR-BLUP.

4.3 Valores genéticos, genômicos e fenotípicos

Com o intuito de avaliar a eficiência da seleção genômica em detrimento da seleção fenotípica, além do viés e da precisão da estimação genômica, gráficos contemplando a relação entre o valor genético real (V_{gen}), fenótipo (V_{fen}) e valor genômico (V_{GG}) foram confeccionados.

A seleção fenotípica é realizada por meio de técnicas clássicas de melhoramento e apresenta comprovada eficiência, refletida pelos contínuos ganhos genéticos em produtividade, obtidos ao longo de vários anos, em várias culturas. Porém, como o progresso genético é contínuo, o nível de melhoramento das espécies vem aumentando e as diferenças a serem detectadas são cada vez menores. Com isso, os marcadores moleculares podem ser empregados para aumentar a eficiência dos programas de melhoramento.

Para que as informações de marcadores moleculares sejam úteis no processo seletivo, a contribuição destes para o melhoramento deve ser mais efetiva em relação à seleção fenotípica. No processo de seleção genômica ampla a utilização de marcadores se dá, resumidamente, em três passos: (i) uso das marcas para deduzir o genótipo de cada indivíduo (genotipagem); (ii) estimar cada efeito de QTL na característica e (iii) soma de todos os efeitos de QTL na obtenção do valor genômico (V_{GG}) dos candidatos à seleção. No processo de seleção quanto a caracteres quantitativos em que a influência do ambiente é maior, espera-se maior contribuição dos marcadores moleculares, em razão das dificuldades envolvidas no processo de seleção fenotípica.

Constatou-se que para herdabilidades de 0,60 houve tendência de maior adequação dos valores genômicos (V_{GG}) em relação aos valores genéticos reais (V_{gen}) do que na menor herdabilidade (0,20). Este comportamento foi mais evidente para a metodologia BLASSO, de forma que os valores genômicos acompanharam melhor os extremos da curva dos valores genéticos (Figuras 7, 8 e 9). No entanto, quando se avalia o comportamento da curva dos valores genéticos genômicos na validação tipo V3, herdabilidade 0,20, verifica-se uma perda de superioridade em relação à seleção fenotípica, ou seja, os valores fenotípicos (V_{fen}) estavam mais próximos dos valores genéticos reais (V_{gen}) (Figura 9).

Vários estudos de simulação (BERNARDO; YU 2007; WONG e BERNARDO 2008; MAYOR; BERNARDO 2009; ZHONG et al. 2009) reportam importantes ganhos

genéticos associados ao uso da seleção genômica. A superioridade em relação a seleção fenotípica retrata a eficiência das metodologias de análise em estudo.

Ao se analisar o viés (diferença entre VGG e valores genotípicos) e a precisão (variação entre valores estimados) através dos gráficos, observou-se que para a herdabilidade de 0,20 as menores precisões foram apresentadas na validação 2 e 3 da metodologia BLASSO (Figuras 7, 8, 9, 10, 11, 12). Tanto na metodologia RR-BLUP quanto na BLASSO não se obteve viés na validação do tipo 2 (Tabela 6). Por outro lado, para a herdabilidade de 0,60 apenas a validação 2 da metodologia BLASSO não foi precisa. Altas variâncias também foram obtidas nas validações 2 e 3 do RR-BLUP. Com relação ao viés, este foi observado em todas as validações utilizando a análise RR-BLUP. Na análise bayesina não se constatou viés em nenhuma forma de validação para a herdabilidade de 0,60.

De forma simplificada pode-se afirmar que a predição genômica foi satisfatória para a herdabilidade de 0,60 pelo método BLASSO, pelas validações V1 e V3, conduzindo a predições precisas e não viesadas (Tabela 4).

Tabela 6- Viés e precisão associados às metodologias de análise RR-BLUP e BLASSO e formas de validação V1, V2 e V3.

Método	Validação	Herdabilidade	Viés	Precisão
BLASSO	V1	0,20	Sim	Sim
		0,60	Não	Sim
	V2	0,20	Pequeno	Não
		0,60	Pequeno	Não
	V3	0,20	Sim	Não
		0,60	Não	Sim
RR-BLUP	V1	0,20	Sim	Sim
		0,60	Sim	Sim
	V2	0,20	Sim	Sim
		0,60	Pequeno	Não
	V3	0,20	Sim	Sim
		0,60	Sim	Não

Para a herdabilidade de 0,20, o melhor método é o RR-BLUP, que é preciso em todas as validações, porém é viciado também em todas elas; nessa situação, o BLASSO é preciso apenas na validação V1 e as predições são viesadas em todas as situações. De maneira geral, as melhores validações são a V1 e a V3; o método RR-BLUP é mais preciso e o BLASSO é menos viciado.

A metodologia BLASSO procede à seleção de modelos, de maneira que as marcas com efeitos nulos são retiradas do modelo, com isso têm-se um menor número de parâmetros. Este fato acarreta um menor N efetivo ao se analisar a variância da amostra, gerando maiores valores de variância e, conseqüentemente, uma menor precisão em relação ao RR-BLUP. Por outro lado, o viés é dado pela capacidade do método em aprender a arquitetura ou a forma de distribuição dos efeitos genéticos. Diante disso, verifica-se que o BLASSO foi mais eficiente em retratar a estrutura genética, que no caso foi a de distribuição exponencial dos efeitos genéticos.

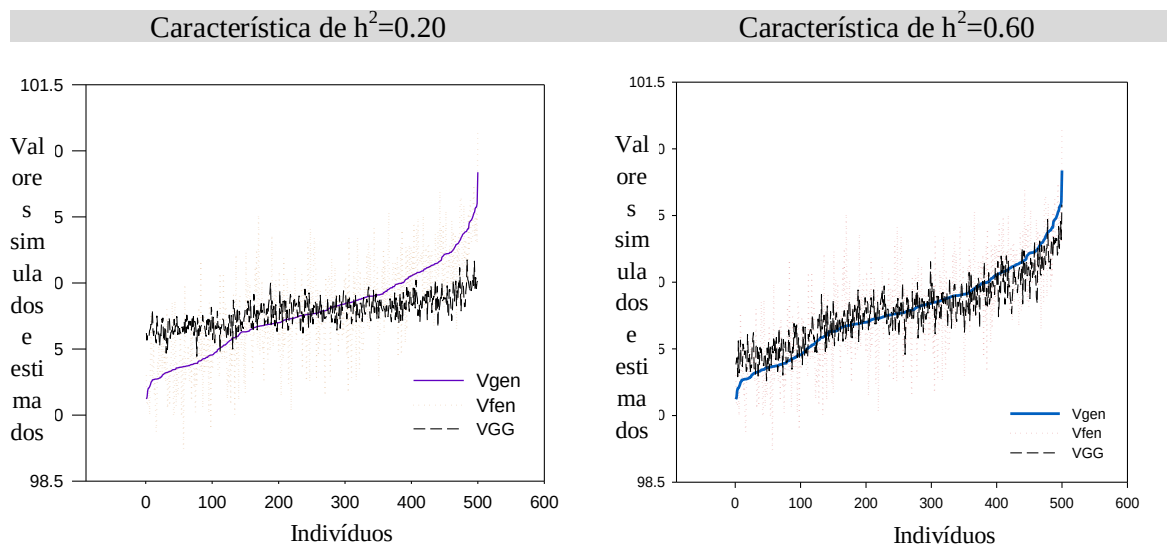


Figura 7- Valores genéticos (Vgen), valores fenotípicos (Vfen) e valores genéticos genômicos (VGG) para fenótipos corrigidos com distribuição exponencial. Método de estimação BLASSO com validação V1.

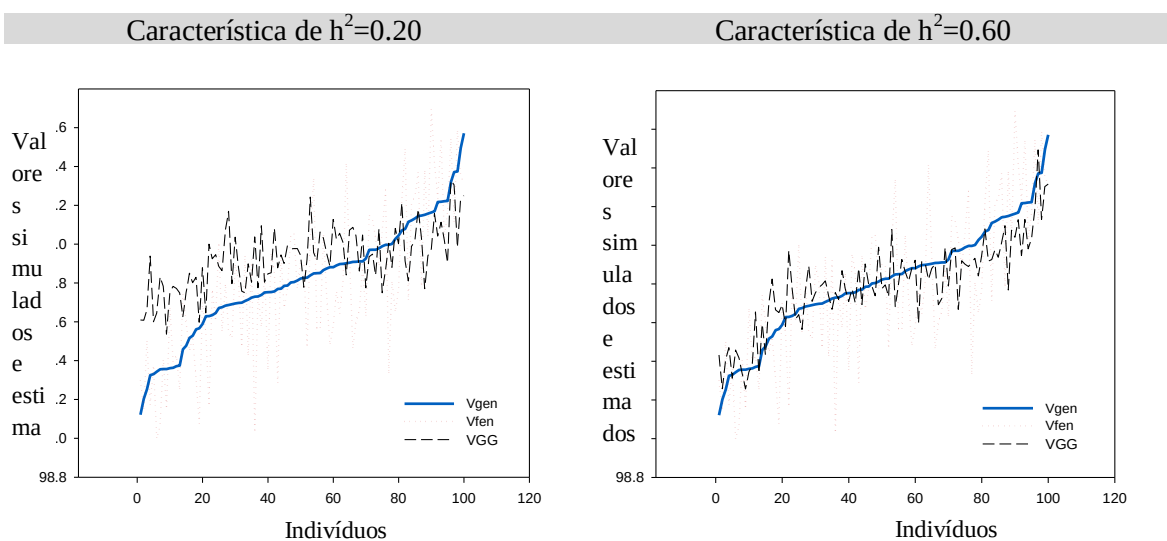


Figura 8- Valores genéticos (Vgen), valores fenotípicos (Vfen) e valores genéticos genômicos (VGG) para fenótipos corrigidos com distribuição exponencial. Método de estimação BLASSO com validação V2.

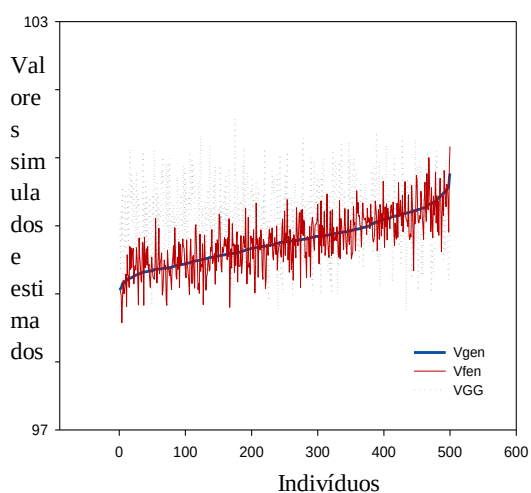
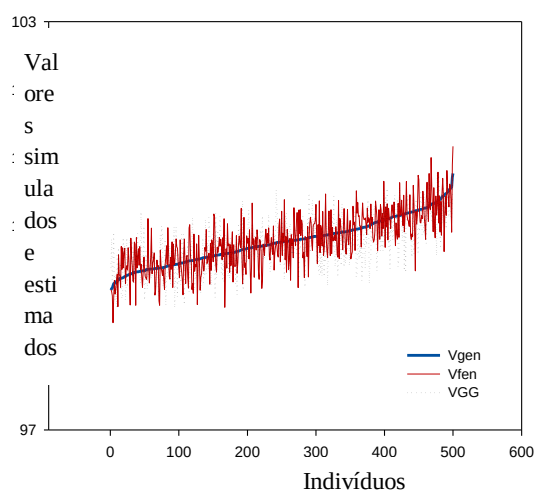
Característica de $h^2=0.20$ Característica de $h^2=0.60$ 

Figura 9- Valores genéticos (Vgen), valores fenotípicos (Vfen) e valores genéticos genômicos (VGG) para fenótipos corrigidos com distribuição exponencial. Método de estimação BLASSO com validação V3.

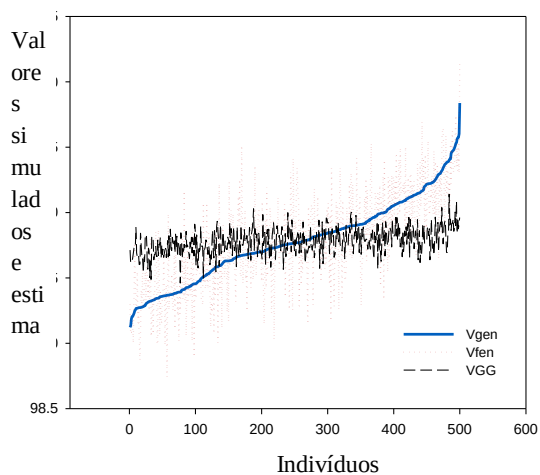
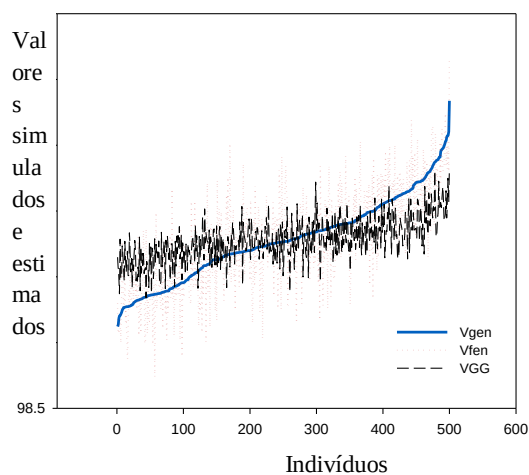
Característica de $h^2=0.20$ Característica de $h^2=0.60$ 

Figura 10- Valores genéticos (Vgen), valores fenotípicos (Vfen) e valores genéticos genômicos (VGG) para fenótipos corrigidos com distribuição exponencial. Método de estimação RR-BLUP com validação.

Característica de $h^2=0.20$

Característica de $h^2=0.60$

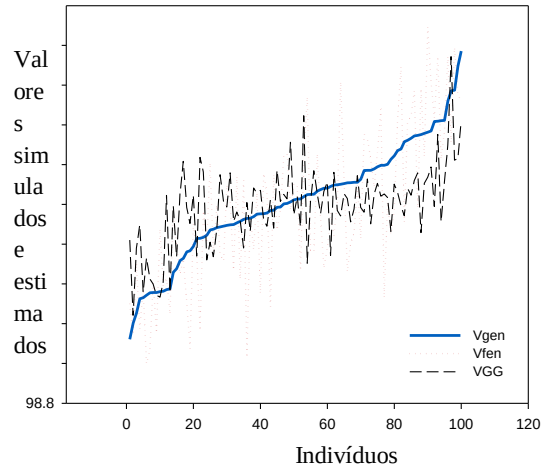
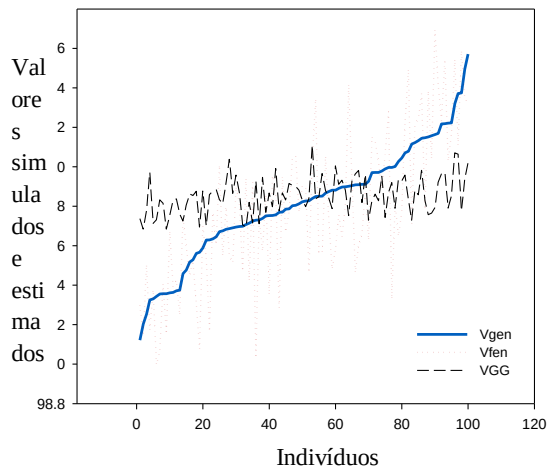


Figura 11- Valores genéticos (Vgen), valores fenotípicos (Vfen) e valores genéticos genômicos (VGG) para fenótipos corrigidos com distribuição exponencial. Método de estimação RR-BLUP com validação V2.

Característica de $h^2=0.20$

Característica de $h^2=0.60$

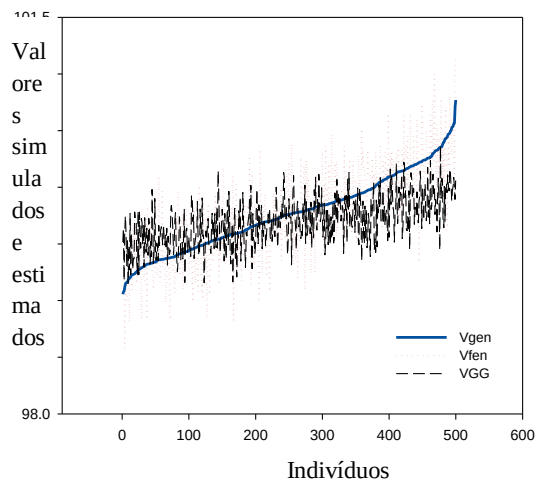
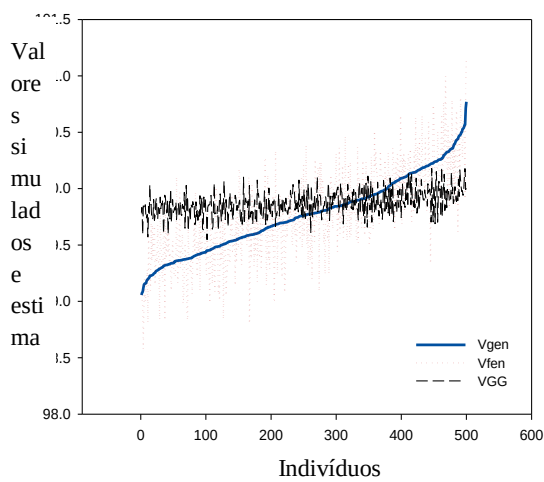


Figura 12- Valores genéticos (Vgen), valores fenotípicos (Vfen) e valores genéticos genômicos (VGG). Método de estimação RR-BLUP com validação V3.

4.4 Acurácia da seleção genômica ampla sem correção de fenótipos

Os fatores que afetam a acurácia na estimação dos valores genômicos incluem a herdabilidade da característica, o tamanho da população de referência, o método utilizado na estimação dos efeitos alélicos das marcas, desequilíbrio de ligação entre marcas e as características quantitativas (QTL) e a distribuição dos efeitos dos QTL, como, por exemplo, a arquitetura genética da característica (BASTIAANSEN et al. 2012). Dentro deste contexto observa-se o comportamento diferenciado das acurácias de acordo com a metodologia de estudo aplicada, distribuição genética dos efeitos do caráter quantitativo e tamanho da população de treinamento.

Observou-se que para a característica quantitativa cujos genes possuem efeitos com distribuição uniforme a metodologia RR-BLUP, no geral, apresentou maiores valores de acurácia quando comparada com a análise bayesiana (Tabela 7). A maior magnitude de acurácia para os genes com distribuição uniforme foi de 0,63, na metodologia RR-BLUP, contra 0,61, na metodologia BLASSO. Alguns autores apontam para a superioridade de métodos bayesianos em relação ao método BLUP (MEUWISSEN et al. 2001; USAI et al. 2009; CROSSA et al. 2010). Entretanto, há relatos quanto a inversão deste comportamento esperado (ZHONG et al. 2009; HABIER et al. 2010). Zhong et al. obtiveram valores de 0,62 e 0,61 de acurácia para RR-BLUP e Bayes B, respectivamente. Segundo estes autores, uma distribuição de efeitos aleatórios mais complicada, como a utilizada em métodos bayesianos, só é útil quando as marcas estão fortemente associadas com o QTL. Tal associação ocorre quando o efeito do QTL é elevado e quando as marcas estão em alto desequilíbrio de ligação com o QTL de importância. Desta forma, o comportamento inverso também foi verificado, ou seja, quando os genes apresentaram efeitos com distribuição exponencial, o BLASSO foi superior em relação ao RR-BLUP.

Tabela 7- Tabela de acurácia da seleção genômica ampla nos diferentes cenários sem a correção de fenótipos.

		Herdabilidade							
		0,20		0,40		0,60			
		Distribuição dos efeitos de QTL		Distribuição dos efeitos de QTL		Distribuição dos efeitos de QTL			
	Metodologia de Análise	Uniforme	Exponencial	Uniforme	Exponencial	Uniforme	Exponencial	Média	Desvio Padrão
Validação 1 (V1)	RR-BLUP	0,40	0,40	0,52	0,54	0,60	0,60	0,51	0,039
	BLASSO	0,34	0,48	0,45	0,82	0,45	0,90	0,57	0,037
Validação 2 (V2)	RR-BLUP	0,37	0,43	0,46	0,47	0,48	0,51	0,45	0,040
	BLASSO	0,34	0,67	0,38	0,53	0,44	0,53	0,48	0,039
Validação 3 (V3)	RR-BLUP	0,42	0,39	0,56	0,52	0,63	0,58	0,52	0,038
	BLASSO	0,34	0,36	0,50	0,60	0,61	0,73	0,52	0,038
Média		0,37	0,46	0,48	0,58	0,54	0,64		
Desvio Padrão		0,042	0,040	0,039	0,037	0,038	0,034		

No geral, a arquitetura genética de poucos genes de maior efeito (distribuição exponencial) foi o cenário que apresentou maiores valores de acurácia. Neste caso, a aplicação da metodologia bayesiana (BLASSO) foi a que apresentou maiores acurácias (0,73; 0,82 e 0,90) em relação à metodologia RR-BLUP (0,54; 0,56; 0,60). De maneira simplificada, pode-se afirmar que análise bayesiana e QTL com efeitos com distribuição exponencial, são responsáveis por valores mais elevados na magnitude média das acurácias. Zhong et al. (2009) afirmam que a acurácia relativa dos métodos depende da magnitude dos efeitos das marcas. Quando as marcas estão em alto desequilíbrio com um pequeno número de QTL que capturam a maior parte da variância genética, ou seja, marcadores de grande efeito, os métodos bayesianos são mais acurados, enquanto o método RR-BLUP é mais acurado quando muitas marcas possuem efeitos pequenos.

Verificou-se que os valores médios de acurácia decresceram com a herdabilidade, de forma que, na menor herdabilidade ($h^2=0,20$) obteve-se a acurácia máxima de 0,46 e, na maior ($h^2=0,60$), este valor chegou a 0,64 (Tabela 7). Goodard (2009) define a acurácia dos valores genômicos estimados como função do número de dados fenotípicos e da herdabilidade. Um decréscimo na herdabilidade leva a menores acurácias na estimação dos valores genômicos. Da mesma forma, os valores de acurácia serão aprimorados quando a herdabilidade da característica é elevada, no entanto, neste caso, a acurácia da seleção fenotípica também será alta. A superioridade na seleção genômica se dá, principalmente, para caracteres de baixa herdabilidade.

Bernardo e Yu (2007) ao compararem a seleção genômica em diferentes cenários de herdabilidade também verificaram que a acurácia descreceu com menores herdabilidades. Claramente, caracteres com baixa herdabilidade são um problema tanto na seleção genômica, quanto na seleção fenotípica, pois em ambos os casos leva a um decréscimo na acurácia. Entretanto, um maior investimento inicial na população de treinamento pode gerar ganhos na acurácia dos valores genômicos. Em contraste, para selecionar caracteres de baixa herdabilidade na seleção fenotípica, várias repetições ou laboriosos testes de progênie devem ser conduzidos a cada ciclo de seleção (RUTKOSKI et al. 2011).

Verificou-se que os maiores valores médios de acurácia foram obtidos na validação V1 (0,57) seguido da validação V3 (0,52). Na validação V1 foram incorporados 400 indivíduos na estimação dos valores genômicos, deixando apenas 100 indivíduos para a população de validação. De maneira semelhante, a validação V2, realizada no software SELEGEN GENÔMICA, através da validação cruzada de

Jackknife, direcionou 400 indivíduos para a população de estimação e 100 para a população de validação. E, finalmente, na validação V3 uma nova população foi gerada (G2-geração dois) com o intuito de constituir uma nova população de validação; com isso, tanto população de treinamento, quanto de validação possuíram o quantitativo de 500 indivíduos.

Maiores valores de acurácia podem ser obtidos ao se elevar o tamanho da população de estimação. Meuwissen et al. (2001) observaram que a acurácia dos valores genômicos estimados foi maior quando 2000 dados, ao invés de 1000, foram usados para estimar o efeito do QTL de uma característica com herdabilidade de 0,50. Sob condições de baixo desequilíbrio (como, por exemplo, elevado tamanho efetivo), mais dados ainda seriam necessários. Um estudo realizado por Zhong et al. (2009) examinou os efeitos de diferentes parâmetros na acurácia da seleção genômica e verificou que dobrar o tamanho da população de estimação aumentou a acurácia.

4.5 Acurácia da seleção genômica ampla com correção de fenótipos

O BLUP tradicional aumenta a acurácia ao incorporar os ancestrais e o parentesco na sua formulação. Entretanto, incluir a informação de famílias eleva a correlação entre os valores genéticos estimados dos membros das famílias, fazendo com que a seleção de aparentados seja maior (WRAY E THOMPSON, 1990). Daetwyler et al. (2007) revisaram este assunto e determinaram que a seleção genômica difere da seleção fenotípica simples e do BLUP tradicional por usar as marcas de maneira mais acurada na estimação da segregação mendeliana, que é, o desvio dos valores dos irmãos dentro das famílias. A correção dos fenótipos permite que somente a parcela devido à segregação mendeliana seja incluída no modelo. A segregação mendeliana, gerada pela segregação aleatória, é criada a cada novo ciclo. Selecionar com base somente nesta variação sustenta o progresso genético pela redução na seleção de aparentados e consequentemente, diminui a endogamia (WOOLLIAMS et al., 1999).

No geral, a metodologia BLASSO apresentou-se superior em relação à metodologia RR-BLUP no que se refere aos valores de acurácia (Tabela 8). Este comportamento está em oposição ao que foi observado na análise sem a correção dos fenótipos, onde as metodologias BLASSO e RR-BLUP divergiram nos valores de acurácia ao se considerar o modo de distribuição dos efeitos dos genes (genes de efeitos

Tabela 8- Tabela de acurácia da seleção genômica ampla nos diferentes cenários com a correção de fenótipos.

		Herdabilidade							
		0,20		0,40		0,60			
		Distribuição dos efeitos de QTL		Distribuição dos efeitos de QTL		Distribuição dos efeitos de QTL			
	Metodologia de Análise	Uniforme	Exponencial	Uniforme	Exponencial	Uniforme	Exponencial	Média	Desvio Padrão
Validação 1 (V1)	RR-BLUP	0,34	0,43	0,55	0,54	0,58	0,58	0,50	0,039
	BLASSO	0,38	0,58	0,82	0,83	0,89	0,90	0,73	0,030
Validação 2 (V2)	RR-BLUP	0,36	0,39	0,51	0,51	0,57	0,57	0,49	0,039
	BLASSO	0,39	0,58	0,83	0,83	0,90	0,90	0,74	0,030
Validação 3 (V3)	RR-BLUP	0,27	0,40	0,52	0,50	0,55	0,55	0,47	0,040
	BLASSO	0,26	0,37	0,57	0,57	0,70	0,70	0,53	0,038
Média		0,33	0,46	0,63	0,63	0,70	0,70		
Desvio Padrão		0,042	0,040	0,035	0,035	0,032	0,032		

pequenos e iguais e genes de grandes efeitos). Desta forma, o método BLASSO foi superior mesmo quando a distribuição dos efeitos genéticos eram de genes de efeitos uniformes. A maior amplitude entre acurácias, em relação aos métodos, foi observada na herdabilidade 0,60, validação do tipo 2, onde genes de distribuição uniforme apresentou acurácia de 0,57 para o RR-BLUP e 0,90 para o BLASSO.

A correção proporciona o uso da segregação mendeliana, que captura a associação de alelos de marcas e de QTL, ou seja, capta os efeitos genéticos explicados pelo desequilíbrio de ligação e não pelo parentesco (RESENDE, 2008). Maiores valores de desequilíbrio contribuem diretamente para um maior valor na acurácia. Com isso, a correção de fenótipos para estrutura de população apresentou maiores valores de acurácia em relação à análise onde essa correção não foi efetuada. Neste cenário, onde a correção foi aplicada, a metodologia BLASSO destaca-se como a metodologia mais robusta em relação à distribuição dos efeitos genéticos, apresentando-se mais acurada tanto para genes com distribuição uniforme quanto para genes com distribuição exponencial, evento que não foi observado quando não se procedeu à correção dos fenótipos.

Uma vez que a correção de fenótipos diminui a contribuição do parentesco na análise, pressupõe-se que a metodologia RR-BLUP foi mais penalizada por possuir maior contribuição do parentesco na estimação dos dados. Habier et al. (2007) desenvolveram uma regressão aproximada afim de quantificar a importância relativa das fontes parentesco genético e desequilíbrio de ligação encontrando que, sob o modelo de simulação aplicado, 39% e 21% da acurácia da seleção genômica foi devida a captura do parentesco genético para o RR-BLUP e para Bayes B, respectivamente. Por outro lado, o método bayesiano foi mais eficiente em capturar o efeito do desequilíbrio de ligação. Consequentemente, sua acurácia persiste por mais tempo que aquela da regressão aleatória.

A arquitetura genética, distribuição uniforme ou exponencial, não influenciou de modo significativo na diferenciação de acurácias, portanto, os valores para cada distribuição de efeito genético foram semelhantes dentro de cada herdabilidade em estudo. Para a herdabilidade de 0,40 a acurácia média para ambas formas de distribuição dos efeitos genéticos foi de 0,63 e para a herdabilidade de 0,60 foi de 0,70. A maior variação foi obtida na herdabilidade 0,20 onde os valores foram de 0,33 para o gene de distribuição uniforme e de 0,46 para o gene de distribuição exponencial. Essa divergência de comportamento pode estar relacionada com a própria herdabilidade

aplicada em cada cenário, pois como relatado anteriormente, o aumento da herdabilidade eleva, como consequência, os valores de acurácias.

Goodard e Hayes (2007), afirmam que na situação em que a maioria dos QTL possuem efeito zero, como na característica exponencial, mínimos quadrados e BLUP resultam em pequenos valores de estimação para esses efeitos “zero”, no entanto, o acúmulo dos efeitos não zerados ocasionam ruídos nas estimativas, o que pode ter levado a uma menor acurácia do RR-BLUP em relação ao BLASSO. Se todos os QTL são de genes com efeitos de uma distribuição exponencial, o estimador LASSO (TIBSHIRANI1996) seria mais apropriado. Melhores estimativas são alcançadas quando os vários QTL possuem efeito zero, o que equivale a retirá-los do modelo. O mesmo comportamento foi retratado por Bastiaansen et al.(2012) ao comparar a metodologia BLUP com a bayesiana. Os referidos autores verificaram que a distribuição dos efeitos genéticos influenciou na estimação dos valores genômicos, de forma que a distribuição desigual, poucos QTL de maior efeito associados a característica, apresentou maior acurácia quando se utilizou o método bayesiano.

Os diferentes cenários de validação também não influenciaram de forma pronunciada nos valores médios de acurácia dentro de cada metodologia, sendo que estes foram de 0,50; 0,49 e 0,47 para o método RR-BLUP e de 0,73; 0,74; 0,53 para o BLASSO nas formas de validação V1, V2 e V3, respectivamente. Como tanto na validação V1 quanto na validação V2 uma amostra de 400 indivíduos foram usados na população de estimação e 100 foram direcionados para a população de validação, valores aproximados entre essas duas formas de validação eram esperados. Por outro lado, com a correção de fenótipos, a validação V3 não se destacou como no cenário onde não houve a correção. Este evento está relacionado ao tipo de população de validação da validação V3. Esta foi obtida através do acasalamento ao acaso entre irmãos completos (G2), gerando maiores valores de parentesco e, consequentemente maiores quando se procedeu à correção de fenótipos, a contribuição da genealogia perdeu seu peso no cômputo da acurácia. Associado a este fato, tem-se também que o acasalamento ao acaso diminuiu o desequilíbrio de ligação nesta população de validação, o que também contribuiu para uma menor acurácia.

Com a correção de fenótipos, depreende-se que o maior peso nas acurácias é a proporção da segregação mendeliana capturada. Com a redução da contribuição do parentesco a metodologia RR-BLUP apresentou decréscimos em seus valores e a metodologia BLASSO, por outro lado, apresentou-se mais eficiente.

De acordo com os resultados, verifica-se que tanto a herdabilidade, quanto o tamanho da população de treinamento e distribuição dos efeitos genéticos influenciaram nos valores de acurácia, corroborando com Resende (2008) que afirma que a acurácia da seleção genômica ampla depende de cinco fatores: (i) da herdabilidade do caráter; (ii) do número de locos controlando o caráter e da distribuição de seus efeitos; (iii) do número de indivíduos na população de treinamento; (iv) do tamanho efetivo populacional; (v) do espaçamento entre marcadores, o qual depende do seu número e do tamanho do genoma. Além destes fatores, pôde-se observar que o tamanho da população de validação também influenciou nos valores de acurácia.

CONCLUSÕES

A simulação das populações foi eficiente ao retratar a estrutura genética das populações. Tanto os dados genotípicos quanto os dados fenotípicos foram preservados de acordo com o cenário simulado.

Sem correção de fenótipos para estrutura de população a distribuição de efeitos de QTL exponencial conduziu a maiores acurácias que a distribuição de efeitos uniforme. Adicionalmente, o método BLASSO mostrou-se mais acurado que o RR-BLUP para distribuição de efeitos genéticosexponencial e menos com distribuição uniforme; neste caso, as validações V1 (400 indivíduos na população de estimação e 100 na população de validação) e V3 (500 indivíduos da geração um, que constituiu a população de estimação, e 500 indivíduos da geração dois que representou a população de validação) mostraram-se mais acuradas.

Com correção de fenótipos para estrutura de populaçãoos QTL com efeitos genéticos com distribuições exponencial e uniforme conduziram a acurácias mais similares e o método BLASSO mostrou-se mais acurado que o RR-BLUP para ambas as formas de distribuição dos efeitos genéticos, de maneira que foi o método de estimação mais robusto. Nesta situação, de correção de fenótipos,as validações V1 e V2 (validação de Jackknife constituída por 400 indivíduos na população de estimação e 100 na de validação) apresentaram-se mais acuradas. Adicionalmente, a correção de fenótipos

apresentou maiores valores de acurácia devido a uma maior eficiência em capturar o desequilíbrio de ligação.

De maneira, o método RR-BLUP foi mais preciso e o BLASSO menos viciado. Um método combinando essas duas vantagens necessita ser desenvolvido.

REFERÊNCIAS

ARDLIE, K.G., KRUGLYAK, L.; SEIELSTAD, M. Patterns of linkage disequilibrium in the human genome. **Nat. Rev.Genet.**, v. 3, p. 299–309, 2002.

BASTIAANSEN, J.W.M.; COSTER,A.; CALUS,M.P.L; ARENDONK, J.A.; BOVENHUIS,H. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. **Genetics Selection Evolution**, v. 44, p.3:, 2012.

BERNARDO, R. Genomewide Selection for Rapid Introgression of Exotic Germplasm in Maize. **Crop Science**, v.49, n.2, p.419-425. Mar-Apr. 2009.

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

BERNARDO, R.;YU, J. M. Prospects for genomewide selection for quantitative traits in maize. **Crop Science**, v.47, n.3, p.1082-1090. May-Jun. 2007.

BROMAN, K. W.;SPEED, T. R. A model selection approach for the identification of quantitative trait loci in experimental crosses. **Journal of the Royal Statistical Society Series B-Statistical Methodology**, v.64, p.641-656. 2002.

CROSSA, J.;DE LOS CAMPOS, G.;PEREZ, P.;GIANOLA, D.;BURGUENO, J.;ARAUS, J. L.;MAKUMBI, D.;SINGH, R. P.;DREISIGACKER, S.;YAN, J.

B.;ARIEF, V.;BANZIGER, M.;BRAUN, H. J. Prediction of Genetic Values of Quantitative Traits in Plant Breeding Using Pedigree and Molecular Markers. **Genetics**, v.186, n.2, p.713-U406. Oct. 2010.

CRUZ, C.D. **Princípios de genética quantitativa**. Viçosa: UFV, 394p. 2005.

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

DAETWYLER HD, VILLANUEVA B, BIJMA P, J. A. WOOLLIAMS. Inbreeding in genome-wide selection. **J Anim Breed Genet**; v.124, p. 369–76, 2007.

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, n.1, p.375-385. May. 2009.

DEKKERS, J. C. M. Commercial application of marker-and gene-assisted selection in livestock: strategies and lessons. **Journal of Animal Science**, v. 82, n. 13 suppl, p. E313-E328, 2004.

ENDELMAN, J.B. Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome** 4:250-255. doi: 10.3835/plantgenome2011.08.0024. 2011.

FALCONER, D. S. **Introduction to quantitative genetics**. 3. ed. Harlow: Longman, 1989. 438 p.

FLINT-GARCIA, S. A.;THORNSBERRY, J. M.;BUCKLER, E. S. T. Structure of linkage disequilibrium in plants. **Annu Rev Plant Biol**, v.54, p.357-374. 2003.

FRARY, A.; NESBITT, T. C.; GRANDILLO, S.; VAN DER KNAAP, E.; CONG, B.; LIU, J.; MELLER, J.; ELBER, R.; ALPERT, K. B. fw2. 2: a quantitative trait locus key to the evolution of tomato fruit size. **Science**, v. 289, n. 5476, p. 85-88, 2000.

GAUT, B. S.; LONG, A. D. The lowdown on linkage disequilibrium. **The Plant Cell Online**, v. 15, n. 7, p. 1502-1506, 2003.

GIANOLA, D.;FERNANDO, R. L.;STELLA, A. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, v.173, n.3, p.1761-1776. Jul. 2006.

GIANOLA, D.;VAN KAAM, J. B. Reproducing kernel hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics**, v.178, n.4, p.2289-2303. Apr. 2008.

GODDARD, M. E. Genomic selection: prediction of accuracy and maximization of long term response. **Genetica**, Dordrecht, v. 136, n. 2, p. 245-257, 2009.

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

GUPTA, P. K.; RUSTGI, S.; KULWAL, P. L. Linkage disequilibrium and association studies in higher plants: present status and future prospects. **Plant molecular biology**, v. 57, n. 4, p. 461-485, 2005.

HABIER, D.;FERNANDO, R. L.;DEKKERS, J. C. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, v.177, n.4, p.2389-2397. Dec. 2007.

HABIER, D.;FERNANDO, R. L.;KIZILKAYA, K.;GARRICK, D. J. Extension of the bayesian alphabet for genomic selection. **BMC Bioinformatics**, v.12, p.186. 2011.

HAYES, B. J.;VISSCHER, P. M.;MCPARTLAN, H. C.;GODDARD, M. E. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. **Genome Research**, v.13, n.4, p.635-643. Apr. 2003.

HEFFNER, E. L.;SORRELLS, M. E.;JANNINK, J. L. Genomic Selection for Crop Improvement. **Crop Science**, v.49, n.1, p.1-12. Jan-Feb. 2009.

HELTSHE, J. F.;FORRESTER, N. E. Estimating Species Richness Using the Jackknife Procedure. **Biometrics**, v.39, n.1, p.1-11. 1983.

HOERL, A. E.;KENNARD, R. W. Ridge Regression - Biased Estimation for Nonorthogonal Problems. **Technometrics**, v.12, n.1, p.55-&. 1970.

JANNINK, J.-L.; WALSH, B. Association mapping in plant populations. In: M.S. Kang (Ed.), **Quantitative Genetics, Genomics and Plant Breeding**, CAB International, pp. 59–68, 2002.

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

KOLBEHDARI, D.;SCHAEFFER, L. R.;ROBINSON, J. A. Estimation of genome-wide haplotype effects in half-sib designs. **Journal of Animal Breeding and Genetics**, v.124, n.6, p.356-361. Dec. 2007.

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

LEGARRA, A.;MISZTAL, I. Technical note: Computing strategies in genome-wide selection. **Journal of Dairy Science**, v.91, n.1, p.360-366. Jan. 2008.

LIU Y.;JANSEN G.B.; LIN C.Y. The covariance between relatives conditional on genetic markers. **Genet. Sel. Evol.**, v.34, p.657–678, 2002.

LIU, X.;FU, J.;GU, D.;LIU, W.;LIU, T.;PENG, Y.;WANG, J.;WANG, G. Genome-wide analysis of gene expression profiles during the kernel development of maize (*Zea mays* L.). **Genomics**, v.91, n.4, p.378-387. Apr. 2008.

LONG, N.;GIANOLA, D.;ROSA, G. J. M.;WEIGEL, K. A.;AVENDANO, S. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of Animal Breeding and Genetics**, v.124, n.6, p.377-389. Dec. 2007.

MAYOR, P.J., BERNARDO, R. Genome-wide selection and marker-assisted recurrent selection in double haploid versus F2 population. **Crop Science**, v. 49, p.1719-1725, 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, n.4, p.1819-1829. Apr. 2001.

MOSER, G.;TIER, B.;CRUMP, R. E.;KHATKAR, M. S.;RAADSMA, H. W. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. **Genetics Selection Evolution**, v.41. Dec 31. 2009.

MUIR, W. M. Comparasion of genomic and traditional BLUP-estimated breeding values accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v.124, p.342-355. 2007.

PARK, T.;CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, v.103, n.482, p.681-686. Jun. 2008.

PIYASATIAN, N.;FERNANDO, R. L.;DEKKERS, J. C. M. Genomic selection for marker-assisted improvement in line crosses. **Theoretical and Applied Genetics**, v.115, n.5, p.665-674. Sep. 2007.

RAFALSKI, A.; MORGANTE, M. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. **TRENDS in Genetics**, v. 20, n. 2, p. 103-111, 2004.

RAMALHO, M. A. P.;SANTOS, J. B.;PINTO, C. A. B. P. **Genética na agropecuária**. Lavras:UFLA, 2000. 472 p.

RESENDE, M. D. **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**.Colombo:Embrapa florestas, 2008. 330 p.

RESENDE, M. D. V. de. **Selegen–Reml/Blup: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos**. Colombo: Embrapa Florestas, 2007. 360 p.

RESENDE, M. D. V. Seleção genômica ampla (GWS) e modelos lineares mistos. In:**Matemática e Estatística na Análise de Experimentos e no Melhoramento Genético**. Colombo: Embrapa Florestas, 2007. p. 517-534.

RESENDE, M. D.;LOPES, P. S.;SILVA, R. L.;PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, v.56, p.63-78. 2008.

RUTKOSKI, J.E.; HEFFNER,E.L.; SORRELLS, M.E. Genomic selection for durable stem rust resistance in wheat. **Euphytica**, v.179, p.161–173, 2011.

SCHAEFFER, L. R. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genetics**, v.123, n.4, p.218-223. Aug. 2006.

SOLBERG, T. R.;SONESSON, A. K.;WOOLLIAMS, J. A.;MEUWISSEN, T. H. Reducing dimensionality for prediction of genome-wide breeding values. **Genet Sel Evol**, v.41, p.29. 2009.

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

USAI, M. G.;GODDARD, M. E.;HAYES, B. J. LASSO with cross-validation for genomic selection. **Genetics Research**, v.91, n.6, p.427-436. Dec. 2009.

WEISS, K.M.; CLARK, A.G. Linkage disequilibrium and the mapping of complex human traits. **Trends Genet**, v.18, p.19–24, 2002.

WHITTAKER, J. C.; THOMPSON, R.; DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetical Research**, v.75, n.2, p.249-252. Apr. 2000.

WONG, C.; BERNARDO, R. Genomewide selection in oil palm: increasing selection gain per unit time and cost with small populations. **TAG Theoretical and Applied Genetics**, v. 116, n. 6, p. 815-824, 2008.

WOOLLIAMS, J.A., BIJMA P.; VILLANUEVA, B. Expected genetic contributions and their impact on gene flow and genetic gain. **Genetics**, v.153, p.1009–1020, 1999.

WRAY, N.R.; THOMPSON, R.. Prediction of rates of inbreeding in selected populations. **Genet. Res.**, v.55, p.41–54, 1990.

XU, S. Z. Estimating polygenic effects using markers of the entire genome. **Genetics**, v.163, n.2, p.789-801. Feb. 2003.

YANO, M.; KATAYOSE, Y.; ASHIKARI, M.; YAMANOUCHI, U.; MONNA, L.; FUSE, T.; BABA, T.; YAMAMOTO, K.; UMEHARA, Y.; NAGAMURA, Y. Hd1, a major photoperiod sensitivity quantitative trait locus in rice, is closely related to the Arabidopsis flowering time gene CONSTANS. **The Plant Cell Online**, v. 12, n. 12, p. 2473-2483, 2000.

YOUSEF, G.G. ; JUVIK, J.A. Comparison of phenotypic and marker- assisted selection for quantitative traits in sweet corn. **Crop Sci.** , v.41, p.645-655, 2001.

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.