LEÍSA PIRES LIMA

SELEÇÃO GENÔMICA NÃO PARAMÉTRICA VIA DISTÂNCIA GENÉTICA ENTRE SUBPOPULAÇÕES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

VIÇOSA
MINAS GERAIS – BRASIL
2017

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

T
L732s
2017

Lima, Leísa Pires, 1993-
Seleção genômica não paramétrica via distância genética entre subpopulações / Leísa Pires Lima. – Viçosa, MG, 2017.
xi, 88f. : il. ; 29 cm.

Inclui apêndices.
Orientador: Camila Ferreira Azevedo.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Estatística não-paramétrica. 2. Genômica. 3. Genética de populações. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Estatística Aplicada e Biometria. II. Título.

CDD 22 ed. 519.54

LEÍSA PIRES LIMA

SELEÇÃO GENÔMICA NÃO PARAMÉTRICA VIA DISTÂNCIA GENÉTICA ENTRE SUBPOPULAÇÕES

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

APROVADA: 15 de fevereiro de 2017.

Fabyano Fonseca e Silva

Antônio Policarpo Souza Carneiro
(Coorientador)

Camila Ferreira Azevedo
(Orientadora)

Aos meus pais, Rossir e Yêda.

AGRADECIMENTOS

A Deus por sempre iluminar meu caminho e dar força e amparo para passar por todos os obstáculos, desânimo, cansaço e desespero. Sem Ele eu não chegaria até aqui.

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

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

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

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

A minha família, pela torcida e apoio.

Aos meus amigos do PPESTBIO pelos momentos de descontração, pelas trocas de experiência, pelas palavras de conforto e constantes incentivos.

Aos meus amigos, Marcos, Shirleny, Gaby, Lucas, Bruna, Renata e Luana pela torcida e amizade.

A Doutora e orientadora Camila Ferreira Azevedo, por todo conhecimento transmitido, pela paciência, confiança, pelos ensinamentos, conselhos, críticas e sugestões que foram primordiais para o meu crescimento tanto profissional quanto pessoal. Obrigada por tudo!

Aos Doutores e coorientadores Marcos Deon Vilela de Resende e Antônio Policarpo Souza Carneiro, pela disponibilidade, confiança, incentivo e pelos saberes transmitidos.

Aos membros da banca examinadora, Professor Antônio Policarpo Souza Carneiro e ao Professor Fabyano Fonseca e Silva, pela disponibilidade e pelas sugestões para o enriquecimento deste trabalho.

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

A CAPES e a FUNARBE, pelo apoio financeiro.

Enfim, deixo aqui meus sinceros agradecimentos a todos aqueles que de certa forma contribuíram direta ou indiretamente para a execução e concretização deste trabalho.

BIOGRAFIA

LEÍSA PIRES LIMA, filha de Yêda Pires da Luz Lima e de Rossir Pires Lima, nasceu em Ubá, Minas Gerais, em 02 de maio de 1993.

Em março de 2010, ingressou no curso de Licenciatura em Matemática na Universidade Federal de Viçosa, Viçosa-MG, graduando-se em agosto de 2014.

Em março de 2015, iniciou o curso de Mestrado no Programa de Pós-Graduação em Estatística Aplicada e Biometria na Universidade Federal de Viçosa, submetendo-se à defesa de dissertação em 15 de fevereiro de 2017.

SUMÁRIO

RESUMO

LIMA, Leísa Pires, M.Sc., Universidade Federal de Viçosa, fevereiro de 2017. **Seleção genômica não paramétrica via distância genética entre subpopulações.** Orientadora: Camila Ferreira Azevedo. Coorientadores: Marcos Deon Vilela de Resende e Antônio Policarpo Souza Carneiro.

A seleção genômica ampla (*Genome Wide Selection* – GWS) consiste na análise de um grande número de marcadores SNPs (*Single Nucleotide Polymorphisms*) amplamente distribuídos no genoma. As principais metodologias propostas e utilizadas na GWS se dividem em metodologias paramétricas, semi-paramétricas ou metodologias de redução de dimensionalidade. Dessa forma, um dos objetivos desse trabalho foi avaliar metodologias não paramétricas, denominadas Delta-p e Regressão Categórica Tripla (TCR), além de compará-las com métodos tradicionalmente aplicados a GWS, tais como G-BLUP (*Genomic Best Linear Unbiased Predictor*) e BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*). O primeiro capítulo deste trabalho consiste em uma revisão de literatura sobre a GWS apresentando sua definição e importância no melhoramento genético, abordando sobre o desenvolvimento dos métodos propostos e avaliados e também retratando sobre o processo de validação utilizado para a comparação das metodologias. No segundo capítulo, foi proposto e analisado o método Delta-p e um índice de seleção, denominado índice Delta-p/G-BLUP que combina os valores genômicos provenientes do método G-BLUP com os valores genômicos estimados via Delta-p. Sob o contexto Bayesiano, foi incorporado ao LASSO Bayesiano, por meio de uma distribuição a *priori* altamente informativa, os valores genômicos estimados via G-BLUP, essa abordagem foi denominada método Bayes Híbrido. Para avaliar a eficiência dos métodos estatísticos, no que se refere à estimação dos valores genômicos aditivos e devidos à dominância, foram utilizados dados simulados, sendo estabelecidos oito cenários (dois níveis de herdabilidade × duas arquiteturas genéticas × ausência de dominância e dominância completa) sendo cada cenário simulado dez vezes. Os resultados do segundo capítulo indicaram que o índice Delta-p/G-BLUP e o Bayes Híbrido se mostraram eficientes para predição dos valores genômicos podendo ser usados vantajosamente na GWS. Ademais, no terceiro capítulo, foi avaliada a eficiência do método TCR em comparação com os métodos G-BLUP e BLASSO utilizando quatro cenários (dois níveis de herdabilidade × modelo infinitesimal ×

ausência de dominância e dominância completa) sendo cada cenário simulado dez vezes. Os resultados indicaram que o método TCR mostrou-se adequado para a estimação dos componentes de variação genômica e da herdabilidade. Em vista disso, uma metodologia baseada em uma modificação do método G-BLUP, denominada TCR/G-BLUP, foi proposta e consiste em estimar a herdabilidade via TCR e fixá-la nas equações de modelos mistos do método G-BLUP. A eficiência dos métodos G-BLUP e TCR/G-BLUP foram comparadas utilizando dados reais, seis caracteristicas avaliadas em mandioca (*Manihot esculenta*). O experimento foi instalado segundo um delineamento em blocos casualizados com três repetições e 10 plantas por parcela. Os resultados indicaram que o método TCR/G-BLUP foi capaz de aumentar a acurácia e fornecer valores genômicos não viesados se comparados ao método G-BLUP, sendo, portanto recomendado para a aplicação na GWS.

ABSTRACT

LIMA, Leísa Pires, M.Sc., Universidade Federal de Viçosa, February, 2017. **Non-parametric genomic selection via genetic distance between subpopulations.** Adviser: Camila Ferreira Azevedo. Co-advisers: Marcos Deon Vilela de Resende and Antônio Policarpo Souza Carneiro.

The genomic wide selection (GWS) consists in analyzing of a large number of single nucleotide polymorphisms (SNPs) markers widely distributed in the genome. The main methodologies proposed and used in GWS are divided into parametric methodologies, semi-parametric methodologies or dimensionality reduction methodologies. Thus, one of the objectives of this work was to evaluate non-parametric methodologies, called Delta-p and Triple Categorical Regression (TCR), and to compare them with methods traditionally applied to GWS, such as G-BLUP (*Genomic Best Linear Unbiased Predictor*) and Bayesian LASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*). The first chapter of this work consists of a literature review about GWS presenting its definition and importance in genetic improvement, discussing the development of the proposed and evaluated methods and also describing the validation process used to compare the methodologies. In the second chapter, were proposed and analyzed the Delta-p method and a selection index, called the Delta-p / G-BLUP index, combining the genomic values derived from the G-BLUP method with the estimated genomic values via Delta-p. Under the Bayesian context, it was incorporated into the Bayesian LASSO, by means of a highly informative a priori distribution, the genomic values estimated by G-BLUP, this approach was called the Hybrid Bayes method. In order to evaluate the efficiency of the statistical methods, in the estimation of the additive and dominance genomic values, simulated data were used, being established eight scenarios (two levels of heritability × two genetic architectures × absence of dominance and complete dominance) each scenario being simulated ten times. The results of the second chapter indicated that the Delta-p/G-BLUP index and the Hybrid Bayes proved to be efficient for predicting the genomic values and could be advantageously used in GWS. In addition, in the third chapter, the efficiency of the TCR method was evaluated in comparison to the G-BLUP and BLASSO methods using four scenarios (two levels of heritability × infinitesimal model × absence of dominance and complete dominance), each scenario being simulated ten times. The results indicated that the TCR method

proved adequate for the estimation of the components of genotype variation and heritability. Therefore, a methodology based on a modification of the G-BLUP method, called TCR/G-BLUP, was proposed and consists of estimating the heritability by means of TCR and fixing it in the mixed model equations of the G-BLUP method. The efficiency of the G-BLUP and TCR/G-BLUP methods were compared using real data, six characteristics evaluated in cassava (*Manihot esculenta*). The experiment was installed according to a randomized block design with three replicates and 10 plants per plot. The results indicated that the TCR / G-BLUP method was able to increase accuracy and provide non-biased genomic values when compared to the G-BLUP method and is therefore recommended for GWS application.

INTRODUÇÃO GERAL

A crescente evolução das tecnologias de sequenciamento e genotipagem promoveu um grande avanço da genética molecular, o qual tem beneficiado o melhoramento animal e vegetal. Visando o aumento do ganho genético por unidade de tempo, baixo custo e alta eficiência na seleção de indivíduos geneticamente superiores, é que Meuwissen et al. (2001) idealizaram a Seleção Genômica Ampla (GWS). A GWS faz uso de informações diretas do DNA por meio dos marcadores moleculares. Os marcadores que mais se destacam são os SNPs (*Single Nucleotide Polymorphisms*), pois são amplamente distribuídos no genoma, possuem baixa taxa de mutação e são codominantes. Desta forma, a GWS consiste na análise de um grande número desses marcadores, capturando genes que afetam um caráter quantitativo.

Conforme Meuwissen et al. (2001) é possível supor que alguns marcadores moleculares estão em desequilíbrio de ligação (*Linkage Disequilibrium* - LD) com locos de características quantitativas (*Quantitative Trait Loci* - QTL) permitindo a sua utilização direta na predição dos valores genéticos genômicos (GEBVs) dos indivíduos sujeitos a seleção, incluindo os indivíduos que ainda não tenham seus fenótipos avaliados. No entanto, o número de marcadores é geralmente muito maior do que o número de indivíduos genotipados e fenotipados e tais marcadores são altamente correlacionados, necessitando da utilização de métodos estatísticos apropriados para análise dos dados no âmbito da genômica (Gianola et al.,2003).

As metodologias estatísticas propostas e utilizadas para aplicação da GWS se dividem em três principais linhas, as metodologias paramétricas o qual pressupõem distribuições de probabilidade para os efeitos aleatórios, como por exemplo, os métodos G-BLUP/RR-BLUP e métodos bayesianos (BLASSO, BayesA, Bayes B,

entre outros), as metodologias semi-paramétricas como, por exemplo, Regressão Kernel e RKHS (*Reproducing kernel Hilbert space*) e as metodologias de redução de dimensionalidade, como por exemplo, a Regressão via componentes principais e independentes. No entanto, novas metodologias não paramétricas podem ser desenvolvidas e avaliadas. O sucesso da GWS se dá pela permanente busca e escolha de metodologias apropriadas que estimem valores genômicos com elevada acurácia e com ausência de viés e que posteriormente serão utilizados no melhoramento genético na seleção de indivíduos geneticamente superiores.

O capítulo 1 consiste em uma revisão de literatura sobre a GWS apresentando sua definição e importância no melhoramento genético, abordando sobre o desenvolvimento dos métodos propostos: Delta-p, Índice combinando os valores genômicos estimados via Delta-p (Índice Delta-p/G-BLUP) e via BLUP Genômico (G-BLUP) e a Regressão Categórica Tripla (TCR). Além dos métodos propostos também foi descrito de forma detalhada as metodologias tradicionalmente aplicadas a GWS, G-BLUP e LASSO Bayesiano (BLASSO). Neste capítulo também foi retratado sobre o processo de validação independente que é utilizado para a comparação entre os métodos.

O capítulo 2 teve por objetivo propor e avaliar as metodologias Delta-p e Índice Delta-p/G-BLUP em comparação com o método G-BLUP. No método, denominado Delta-p, a população de estimação é dividida em duas subpopulações em que são obtidas suas respectivas frequências alélicas. Os efeitos dos marcadores são estimados de forma não paramétrica utilizando a diferença entre as frequências alélicas e ganho genético associados as duas subpopulações. Uma grande vantagem é que essa abordagem não demanda método computacional iterativo. Já o índice de seleção, denominado índice Delta-p/G-BLUP, é baseado na seleção combinada proposta por

Lush (1947) e Falconer (1989) foi estabelecido. O índice combina informações dos valores genômicos preditos provenientes do método G-BLUP e do método Delta-p.

Utilizando a abordagem Bayesiana e seguindo essa mesma linha do índice Delta-p/G-BLUP, ao combinar informações provenientes de duas metodologias estatísticas, também foi estabelecido no capítulo 2 um método que combina os valores genômicos preditos via método G-BLUP e via método BLASSO. Este método foi denominado Bayes Híbrido e nele, a distribuição dos valores genômicos aditivos preditos pelo G-BLUP seria assumida como distribuição *a priori* por meio dos hiperparâmetros das distribuições *a priori* dos componentes de variância. Enquanto, que os valores fenotípicos (y) como os dados para a verossimilhança.

Ademais, no terceiro capítulo o método introduzido por Resende et al. (2014), chamado Regressão Categórica Tripla (TCR), foi combinado com o método Delta-p proposto e assim avaliado. Essa abordagem regressa os fenótipos nas três categorias de genótipos marcadores, visando capturar os efeitos genéticos em um loco b com categorias genotípicas BB, Bb e bb, em que B é o alelo favorável. O TCR é computacionalmente vantajoso e pode ser um melhor estimador dos componentes da variação genotípica e da herdabilidade. Essa metodologia foi comparada com os métodos G-BLUP e BLASSO. Além disso, também foi proposto um método que combina os resultados do TCR com o G-BLUP. Neste método, denominado TCR/G-BLUP, a herdabilidade estimada via TCR é utilizada nas equações de modelos mistos genômicas (G-BLUP) com intuito de tornar os valores genômicos mais acurados. A eficiência dos métodos G-BLUP e TCR/G-BLUP foram comparados utilizando dados reais de seis caracteristícas avaliadas em mandioca.

CAPÍTULO 1
REVISÃO DE LITERATURA

1. Seleção Genômica Ampla

1.1.Definição e Importância

A contribuição principal da genética molecular em benefício do melhoramento genético é a utilização direta das informações de DNA no processo de identificação de indivíduos geneticamente superiores. O uso de marcadores moleculares permite alta eficiência seletiva, rapidez na obtenção de ganhos genéticos e baixo custo. Os marcadores que mais se destacam são os SNPs (*Single Nucleotide Polymorphisms*), pois são amplamente distribuídos no genoma, possuem baixa taxa de mutação e codominância. Com este propósito, Meuwissen et al. (2001) idealizaram a seleção genômica ampla (*Genome Wide Selection* – GWS) que consiste na análise de um grande número de marcadores moleculares distribuídos no genoma, capturando os genes que afetam um caráter quantitativo.

Segundo Meuwissen et al. (2001), os indivíduos candidatos a seleção podem ser identificados por meio dos marcadores moleculares que estão em desequilíbrio de ligação com locos de características quantitativas (*Quantitative Trait Loci* - QTL). Dessa forma, a predição dos valores genéticos genômicos (GEBVs) dos indivíduos é feita com base na estimação dos efeitos desses marcadores no fenótipo, o que consequentemente aumentaria a acurácia da avaliação genética.

No entanto, geralmente, o número de marcadores é muito maior que o número de indivíduos genotipados e fenotipados (alta dimensionalidade), o que impossibilita a utilização adequada de métodos tradicionais baseados em quadrados mínimos (*Least Squares* – LS) a fim de estimar o efeito de cada SNP no fenótipo. Ademais, segundo Gianola et al. (2003), a alta colinearidade entre os marcadores (multicolinearidade),

requer a utilização de metodologias estatísticas diferenciadas para a análise de dados na área de seleção genômica, sendo a escolha dessas metodologias essenciais para o sucesso da GWS.

O método G-BLUP (*Genomic Best Linear Unbiased Predictor)* foi inicialmente aplicado por Nejati-Javaremi et al. (1997) e Fernando (1998) e, no contexto da seleção genômica por Habier et al. (2007), Van Raden (2008), Goddard (2009), Goddard et al. (2009), Hayes et al. (2009) e Stranden e Garrick (2009). Nos dias atuais, constitui-se de um método tradicionalmente aplicado à seleção genômica e que possui um apelo prático entre os programas de melhoramento. A versão Bayesiana da regressão LASSO (*Least Absolute Shrinkage and Selection Operator* – Tibshirani, 1996) aplicada à seleção genômica foi proposta por de los Campos et al. (2009). O método LASSO Bayesiano é vantajoso se comparado aos métodos BayesA e BayesB, uma vez que o BLASSO é assintoticamente livre de informação *a priori*, ou seja, propicia melhor aprendizado com os dados ou melhor aprendizado bayesiano.

Recentemente na literatura, os métodos G-BLUP e BLASSO têm sido amplamente aplicados a GWS e recomendados para a predição genômica (Azevedo et al., 2015; de los Campos et al. 2012; Gianola, 2013; Gianola et al., 2009). No entanto, novas metodologias para Seleção Genômica, denominadas índice Delta-p/G-BLUP e Regressão Categórica Tripla (TCR) foram propostas por Resende (2015) e Resende et al. (2014), mas até o momento não foram aplicadas e avaliadas na seleção genômica. Essas metodologias são puramente conceituais e engenhosas. Dessa forma, serão apresentadas a seguir descrições metodológicas básicas sobre os métodos G-BLUP, Delta-p, Índice Delta-p/G-BLUB, TCR e BLASSO.

1.2.Método G-BLUP

O método G-BLUP (*Genomic Best Linear Unbiased Predictor)* ou BLUP Genômico foi inicialmente aplicado no contexto da Seleção Genômica por Habier et al. (2007), Van Raden (2008), Goddard (2009), Goddard et al. (2009), Hayes et al. (2009) e Stranten e Garrick (2009), com intuito de tornar as predições dos valores genômicos mais acuradas do que os estimados via BLUP tradicional. A superioridade do método se deve ao fato de que no G-BLUP, incluindo efeitos aditivos e devido à dominância, as matrizes de parentesco baseadas em pedigree (A e D) são substituídas pelas matrizes de parentesco genômico, G_a e G_d, recuperando as informações genéticas realizadas entre os indivíduos (Van Raden, 2008).

A predição dos efeitos genômicos, aditivo e devido à dominância, via G-BLUP, usando as informações fenotípicas e genotípicas para cada indivíduo, pode ser feita por meio do seguinte modelo linear misto (Resende, 2008; Resende et al., 2010):

$$y = Xb + Za + Zd + e \quad (1)$$

em que:

y é o vetor de fenótipos ($N\times1$, em que N é o número de indivíduos genotipados e fenotipados);

b é o vetor de efeitos fixos ($p\times1$, em que p é o número de efeitos fixos considerados) com matriz de incidência X ($N\times p$);

a é o vetor de efeitos genômicos aditivos dos indivíduos ($N\times 1$) com matriz de incidência Z ($N\times N$), sendo a estrutura de variância dada por $a\sim N(0, G_a\sigma_a^2)$ em que σ_a^2 é a variância aditiva e $G_a(N\times N)$ é a matriz de parentesco genômica para os efeitos aditivos construída a partir de informações de marcadores moleculares;

d é o vetor de efeitos genômicos devido à dominância dos indivíduos com matriz de incidência Z ($N \times N$), sendo a estrutura de variância dada por $d \sim N(0, G_d\sigma_d^2)$ em que

σ_d^2 é a variância devido à dominância e G_d ($N \times N$) é a matriz de parentesco genômica para os efeitos devido à dominância;

e é o vetor de efeitos residuais aleatórios com $e \sim N(0, I\sigma_e^2)$ sendo σ_e^2 a variância residual.

As equações de modelos mistos para predição de a e d via o método G-BLUP equivalem a:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'X & Z'Z & Z'Z + G_d^{-1}\dfrac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix},$$

sendo os componentes de variância, σ_d^2, σ_a^2 e σ_e^2, estimados via REML (*Restricted maximum likelihood*). Assim, o valor genômico predito total de cada indivíduo *j* (*j*= 1, … , *N*) é dado por:

$$\widehat{GEBV}_j = \hat{g} = \hat{a}_j + \hat{d}_j.$$

Conforme Vitezica et al. (2013), as matrizes de parentesco genômicas para efeitos aditivos e para efeitos devido à dominância, G_a e G_d, são dadas respectivamente por:

$$G_a = \frac{WW'}{\sum_{i=1}^{n}(2p_iq_i)} \quad e \quad G_d = \frac{SS'}{\sum_{i=1}^{n}(2p_iq_i)^2} \ .$$

em que W é a matriz de incidência para os vetores de efeitos aditivos de marcadores (α), S é a matriz de incidência para os vetores de efeitos devido à dominância dos marcadores (δ), p_i e q_i são as frequências alélicas do locos *i*.

Diversas parametrizações para as matrizes de incidência W e S estão disponíveis, no entanto, a única que está de acordo com a teoria de genética quantitativa clássica (Falconer e Mackay, 1996) é a apresentada a seguir (Van Raden, 2008; Vitezica et al., 2013; Wang e Da, 2014; Da et al., 2014; Resende et al., 2014):

$$W = \begin{cases} Se\ MM, então\ 2 - 2p \rightarrow 2q \\ Se\ Mm, então\ 1 - 2p \rightarrow q - p \\ Se\ mm, então\ 0 - 2p \rightarrow -2p \end{cases} \qquad S = \begin{cases} Se\ MM, então\ 0 \rightarrow 2q^2 \\ Se\ Mm, então\ 1 \rightarrow 2pq \\ Se\ mm, então\ 0 \rightarrow -2p^2 \end{cases}.$$

Por meio desta parametrização, o efeito aditivo do marcador m_{ai} em um loco i é igual ao efeito de substituição alélica (α), ou seja:

$$m_{ai} = \alpha_i = a_i + (q_i - p_i)d_i,$$

em que a_i e d_i são valores genotípicos do homozigoto e do heterozigoto, respectivamente, para o loco i. Enquanto, a quantidade m_{di} pode ser diretamente definida como:

$$m_{di} = d_i.$$

1.3.Método BLASSO

Meuwissen et al. (2001) apresentam diversos métodos como possíveis abordagens para predição de valores genéticos genômicos com base em informações genômicas. O modelo (1) apresentado anteriormente é um modelo em nível de indivíduos, no entanto, um modelo em nível de marcas também pode ser definido em GWS. Para isso, o seguinte modelo linear básico foi proposto:

$$y = Xb + Wm_a + Sm_d + e, \qquad (2)$$

em que:

y é o vetor de fenótipos ($N \times 1$, em que N é o número de indivíduos genotipados e fenotipados);

b é o vetor de efeitos sistemáticos ($p \times 1$, em que p é o número de efeitos fixos considerados) com matriz de incidência X $(N \times p)$;

m_a é o vetor de efeitos genéticos aditivos dos marcadores ($m \times 1$, em que m é o

número de marcadores moleculares) com matriz de incidência W;

m_d é o vetor de efeitos genéticos devido à dominância dos marcadores com matriz de incidência S;

e é o vetor de resíduos.

A inferência Bayesiana trata o vetor de parâmetros desconhecidos como quantidades aleatórias e qualquer informação inicial sobre elas pode ser representada por modelos probabilísticos. Assim, é assumido distribuições de probabilidade a todas as quantidades desconhecidas, tais como:

$$y|m_{a1}, m_{a2}, \dots, m_{am}, m_{d1}, m_{d2}, \dots, m_{dm} \sim Normal(Xb + Wm_a + Sm_d, I\sigma^2);$$

$$p(b) \propto k \text{ (distribuição } \textit{a priori flat} \text{ ou não informativa);}$$

$$e|\sigma^2 \sim MVN(0, I\sigma^2);$$

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

em que k é uma constante, v_e e S_e^2 são os hiperparâmetros, MVN representa a distribuição normal multivariada e σ^2 tem como distribuição a *priori* de uma qui-quadrado invertida escalada.

A versão bayesiana da regressão via LASSO (BLASSO - Park e Casella, 2008) para seleção genômica ampla foi proposta por de los Campos et al. (2009). O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*) inclui um termo de variância comum para ambos os termos do modelo (2), ou seja, para os efeitos genéticos de marcadores e de efeitos residuais.

A distribuição a *priori* usada no LASSO Bayesiano apresenta maior massa de densidade no valor 0 e caudas robustas, impondo um maior *shrinkage* nos coeficientes de regressão próximos de 0 e um menor *shrinkage* nos coeficientes de regressão distantes de 0. Assim, são estimadas médias a *posteriori*, produzindo valores muito

pequenos, mas não zero como o LASSO original. Usando uma formulação em termos de um modelo hierárquico aumentado, tem-se:

$$m_{ai}|\tau_a \sim N(0, D_a\sigma^2),\ m_{di}|\tau_d \sim N(0, D_d\sigma^2)$$

$$p(\tau_a^2\,|\lambda_a^2) = \prod_i \left(\frac{\lambda_a^2}{2}\right) e^{\left[\frac{-\lambda_a^2\tau_{ai}^2}{2}\right]} \text{ e } p(\tau_d^2\,|\lambda_d^2) = \prod_i \left(\frac{\lambda_d^2}{2}\right) e^{\left[\frac{-\lambda_d^2\tau_{di}^2}{2}\right]}$$

em que $D_a = diag\left(\tau_{1a}^2, \tau_{2a}^2, \ldots, \tau_{ma}^2\right)$ e $D_d = diag\left(\tau_{1d}^2, \tau_{2d}^2, \ldots, \tau_{md}^2\right)$, λ_a e λ_d são os parâmetros de "suavização" e podem ser estimados por meio do conjunto de dados via método MCMC (*Markov Chain Monte Carlo* - usando uma *priori* não informativa) ou via método MCEM (*Markov Chain Expectation Maximization* - não requer uma informação a *priori*).

Isso conduz a uma distribuição exponencial dupla para os efeitos de marcadores (Park e Casella, 2008), como a seguir:

$$p(m_{ai}|\lambda_a^2) = \int_R N(0, \sigma^2\tau_{ai}^2)\, Exp\left(\frac{\lambda_a^2}{2}\right) d\tau_{ai}^2 \propto \frac{1}{2\left(\frac{\sigma}{\lambda_a}\right)} e^{\left(-\frac{m_{ai}}{\sigma/\lambda_a}\right)},$$

$$p(m_{di}|\lambda_d^2) = \int_R N(0, \sigma^2\tau_{di}^2)\, Exp\left(\frac{\lambda_d^2}{2}\right) d\tau_{di}^2 \propto \frac{1}{2\left(\frac{\sigma}{\lambda_d}\right)} e^{\left(-\frac{m_{di}}{\sigma/\lambda_d}\right)}$$

$$m_{ai}|\lambda_a^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda_a}\right) \text{ e } m_{di}|\lambda_d^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda_d}\right).$$

O LASSO Bayesiano é vantajoso se comparado com os métodos bayesianos propostos por Meuwissen et al. (2001) por ser assintoticamente livre de informação a *priori*, isto é, proporciona uma melhor aprendizagem por meio do conjunto de dados (Gianola, 2013; Gianola et al., 2009). Isto ocorre porque nos modelos hierárquicos, como BLASSO, a informação a *priori* é atribuída aos hiperparâmetros de modo que a influência desta informação desaparece assintoticamente (Resende et al., 2012).

A variância genética aditiva de cada loco marcador é dada por $\sigma_{mai}^2 = \tau_{ai}^2\sigma^2$ e $\sigma_{mdi}^2 = \tau_{di}^2\sigma^2$ com $i = 1, 2, \ldots, m$. Dessa forma, a variância genética aditiva e de

dominância podem ser estimadas utilizando as relações $\sigma_a^2 = \sum_{i=1}^{m} 2p_i q_i \, \sigma_{mai}^2$ e $\sigma_d^2 = \sum_{i=1}^{m} [2p_i q_i]^2 \, \sigma_{mdi}^2$, sendo assim, $\sigma_a^2 = \sum_{i=1}^{m} 2p_i q_i \tau_{ai}^2 \sigma^2$ e $\sigma_d^2 = \sum_{i=1}^{m} [2p_i q_i]^2 \, \tau_{di}^2 \sigma^2$. Os valores genômicos aditivos e devido à dominância são estimados via as seguintes expressões $\hat{a} = W\hat{m}_a$ e $\hat{d} = S\hat{m}_d$, respectivamente.

As distribuições condicionais completas *a posteriori* para os parâmetros do BLASSO são apresentados em detalhes por de los Campos et al. (2009).

1.4.Método Delta-p

O método Delta-p utiliza o conceito de mudança na frequência alélica devido à seleção e o conceito teórico de ganho genético (contraste entre médias de duas subpopulações). Dessa forma, a população de estimação inicialmente é dividida em duas subpopulações, uma com os indivíduos acima da média geral (subpopulação 1, com valor genético médio u_1 superior) e outra com os indivíduos abaixo da média geral (subpopulação 2, com valor genético médio u_2 inferior). A diferença ($u_1 - u_2$) entre os valores genéticos médios das duas subpopulações é devida à maior frequência alélica (p) dos alelos favoráveis (e menor frequência dos alelos desfavoráveis) na subpopulação 1 em relação à subpopulação 2. Assim, ($u_1 - u_2$) é explicada por ($\Delta p = p_1 - p_2$), sendo delta-p (Δp) a diferença de frequências alélicas entre as duas subpopulações. Os valores de Δp são calculados para cada loco marcador e aqueles com sinais positivos são alocados como favoráveis, ou seja, os efeitos de substituição alélica (α_i) desses marcadores são tomados como positivos. Da mesma forma, aqueles com sinais negativos de Δp tem seus $\alpha_i's$ atribuídos como negativos.

Dessa forma, redefine-se a codificação da matriz de incidência W para os vetores de efeitos aditivos de marcadores, compatibilizando-se o arquivo de marcas formado por 0 (mm), 1 (Mm) e 2 (MM) de forma a se ter um arquivo de “genes” dado

por 0 (bb), 1 (Bb) e 2 (BB), sendo o alelo B o alelo favorável e a alocação em BB ou bb é ditada pelo sinal de α_i. Logicamente, o acerto na alocação em BB ou bb é probabilístico. Em média (esperança matemática) haverá acerto na maioria dos locos, sendo que o maior número de erros será naqueles locos marcadores de efeitos muito pequenos, tendendo a zero.

O método Delta-p necessita da estimação dos efeitos aditivos de marcadores $\alpha_i = m_{ai}$ e posteriormente da predição dos valores genômicos dos indivíduos (a). Tomando Δp_i como indicador da magnitude relativa de α_i (quanto maior o Δp_i, maior a chance de o valor de α_i ser maior), tem-se que a proporção $\Delta p_i/\Delta p_m$ entre os Δp_i (com seus respectivos sinais positivos ou negativos) e o delta-p médio (Δp_m, computado usando os módulos de Δp_i), multiplicado pelo α médio (α_m) fornece os efeitos α_i. Assim, $\alpha_i = (\Delta p_i/\Delta p_m)\,\alpha_m$.

A quantidade (α_m) pode ser obtida usando a expressão teórica de ganho genético. Para um loco, tem-se $u_j = (p_j - q_j)\,\alpha_i = (2p_j - 1)\,\alpha_i$, sendo $j = 1,2,\ldots,N$ (N é o número de indivíduos). Assim, $(u_1 - u_2) = (2p_1 - 2p_2)\alpha_i = 2(p_1 - p_2)\alpha_i = 2\Delta p_i\alpha_i$. Na soma dos locos tem-se $(u_1 - u_2) = 2\sum(\Delta p_i\alpha_i)$. A quantidade Δp_m é definida como: $\Delta p_m = sq(1-q)$, em que $s = \alpha k/\sigma_F$ é o coeficiente seletivo do loco, assumindo um modelo aditivo-dominante. Assim, $\Delta p_m = \alpha kpq/\sqrt{\sigma_F^2}$, sendo k o índice de seleção e σ_F^2 a variância fenotípica. O ganho genético é definido como $G_s = k\sigma_a^2/\sqrt{\sigma_F^2}$. Dessa forma, $k/\sqrt{\sigma_F^2} = G_s/\sigma_a^2$. Substituindo na expressão para Δp_m, tem-se: $\Delta p_m = \alpha pq G_s/\sigma_a^2$.

O ganho da subpopulação 1 em relação à subpopulação 2 é duplicado pois é composto de duas partes: seleção à direita truncada no ponto zero de uma curva normal padrão, com diferencial de seleção ($u_1 - u_0$) e seleção à esquerda truncada no ponto zero, com diferencial de seleção ($u_0 - u_2$). O ganho genético associado ao primeiro

diferencial de seleção é dado por $G_{s1} = h_a^2(u_1 - u_0)$ e aquele associado ao segundo diferencial de seleção é dado por $G_{s2} = h_a^2(u_0 - u_2)$.

Esses ganhos são simétricos e devem ser somados para a obtenção do ganho total $2G_s$, dado por: $2G_s = G_{s1} + G_{s2} = h_a^2(u_1 - u_2)$. Assim, $G_s = \left(\frac{1}{2}\right) h_a^2(u_1 - u_2)$. Substituindo G_s em Δp_m, $\Delta p_m = \alpha pq G_s/\sigma_a^2$, tem-se $\Delta p_m = \alpha pq(1/2)h_a^2(u_1 - u_2)/\sigma_a^2$ sendo $\sigma_a^2 = 2pq\alpha^2$, substituindo e dividindo por Δp_m, tem-se $1 = [\alpha pq(1/2)h_a^2(u_1 - u_2)]/(2pq\alpha^2\Delta p_m)$. Simplificando tem-se $1 = (1/2)h_a^2(u_1 - u_2)/(2\alpha\Delta p_m)$. Isolando α tem-se $\alpha = (1/2)h_a^2(u_1 - u_2)/(2\Delta p_m)$ para n locos e o α médio por loco é dado por $\alpha_m = 0{,}5h_a^2\ (u_1 - u_2)/\ [n_{marcas}2\Delta p_m]$.

Assim, os efeitos aditivos α dos marcadores são estimados na população de treinamento (indivíduos com fenótipos e genótipos conhecidos) de forma não paramétrica. Com base nesses efeitos são estimados os valores genéticos aditivos (a) dos indivíduos da população de validação (indivíduos com fenótipos e genótipos conhecidos) ou indivíduos da população de seleção (indivíduos com genótipos conhecidos) por meio da seguinte expressão $\hat{a} = W\hat{\alpha}$, sendo W a matriz de incidência dos efeitos genéticos aditivos dos marcadores. A abordagem não demanda método computacional iterativo e usa apenas os conceitos de distância genética (magnitude de Δp_i) e ganho genético associados as duas subpopulações. Não usa também o *shrinkage* diferenciado com base nas frequências alélicas, fato que beneficia locos com maiores MAFs (*Minor allele frequency*).

O algoritmo completo é:

(i) subdivisão da população de treinamento em duas de acordo com o fenótipo corrigido para efeitos ambientais;

(ii) cálculo de Δp_i e Δp_m;

(iii) cálculo de α_m; (iv) cálculo de $\alpha_i = (\Delta p_i/\Delta p_m)\alpha_m$.

Pode-se realizar também a seleção de marcas com base em Δp_i ou α_i e a estimação da variância aditiva via uma função quadrática dos α_i. O método será tanto melhor quanto maior for o tamanho da população de estimação.

Para os efeitos de dominância deve-se usar $\Delta(2pq)_i$ (diferença entre as frequências de genótipos heterozigotos das subpopulações 1 e 2) em lugar de Δp_i. O fundamento advém de o fato relatado a seguir. Quanto melhor o fenótipo e simultaneamente maior a frequência de genótipos heterozigotos na população, mais favorável é o valor do genótipo corrigido para os efeitos genéticos aditivos, ou seja, maiores são os desvios de dominância. Desta forma, para inferir sobre os efeitos de dominância, o interesse é encontrar esta associação (frequência de heterozigotos e valores dos fenótipos) e, para isso, o fenótipo deve ser regressado na frequência de heterozigotos, corrigida para os efeitos aditivos. Assim, a matriz de incidência S (distribuição Bernoulli) é inerente à heterozigose ou a frequência de heterozigotos dentre as três classes genotípicas. Define-se então g_m como o valor genotípico do heterozigoto, o qual corrigido para o efeito aditivo (α_i) do loco fornece o desvio de dominância do loco i ($\delta_i = m_{di}$) e dos indivíduos (vetor d), dado por $\hat{d} = S\hat{\delta}$, em que S é a matriz de incidência dos efeitos devido à dominância dos marcadores.

Em resumo, tem-se:

	Aditivo		Dominância	
Marcas	$\hat{\alpha}_m = \frac{0{,}5h_a^2(u_1 - u_2)}{n_{marcas} 2\Delta p_m}$	$\hat{\alpha}_i = \frac{\Delta p_i}{\Delta p_m}\hat{\alpha}_m$	$\hat{g}_m = \frac{0{,}5h_a^2(u_1 - u_2)}{n_{marcas}\Delta(2pq)_m}$	$\hat{\delta}_i = \frac{\Delta(2pq)_i}{\Delta(2pq)_m}\hat{g}_m$
Indivíduo	-	$\hat{a} = W\hat{\alpha}$	-	$\hat{d} = S\hat{\delta}$
Incidência	-	$W = \begin{bmatrix} 2q \\ (q-p) \\ -2p \end{bmatrix}$	-	$S = \begin{bmatrix} -2q^2 \\ 2pq \\ -2p^2 \end{bmatrix}$

* nos cálculos de Δp_m e $\Delta(2pq)_m$, usar módulo. Modelo:$(1/2)h_a^2(u_1 - u_2) = n\left(2\Delta p_m \hat{\alpha}_m + \Delta(2pq)_m \hat{\delta}_m\right) + erro$.

Para os efeitos aditivos, uma interpretação resumida do quadro é apresentada a seguir:

$\hat{\varsigma}_m = \frac{h_a^2(u_1 - u_2)}{n_{marcas}}$: ganho genético médio por loco, da subpopulação 1 em relação a 2.

$\hat{\alpha}_m = \frac{\hat{\varsigma}_m}{2\Delta p_m} = \frac{0{,}5h_a^2(u_1 - u_2)}{n_{marcas}2\Delta p_m}$: efeito de substituição alélica médio por loco.

$\hat{\alpha}_i = \frac{\Delta p_i}{\Delta p_m}\hat{\alpha}_m$: efeito de substituição alélica específico para o loco i.

1.5.Método do Índice Delta-p/G-BLUP

Com base no princípio da seleção combinada (Lush, 1947; Falconer, 1989), o método G-BLUP pode ser usado conjuntamente ao procedimento Delta-p. Um índice de seleção da forma $I = b_1\hat{a}_1 + b_2\hat{a}_2$ pode ser estabelecido usando as informações dos valores genômicos aditivos preditos provenientes do G-BLUP ($\hat{a}_1$) e Delta-p ($\hat{a}_2$), ponderados pelos pesos b_1 e b_2, respectivamente.

A seleção combinada é da forma $I = b_1 Indivíduo + b_2 Média\_família$, em que as diferentes fontes de informação, $Indivíduo$ e $Média\_família$, são correlacionadas em algum grau, mas essas correlações são consideradas na construção do índice. O índice $I = b_1\hat{a}_1 + b_2\hat{a}_2$ é similar, podendo ser simbolizado também por $I = b_1 f(Indivíduo) + b_2 f(Média\_subpopulação; \Delta p)$. Nota-se a similaridade entre esse índice e o índice da seleção combinada. O índice pode ser representado também por $I = b_1 f(y; X) + b_2 f((u_1 - u_2); \Delta p)$ em que y é o vetor de dados fenotípicos individuais, X é a matriz de dosagem alélica dentro de indivíduo centralizada, $(u_1 - u_2)$ é o contraste de médias de subpopulações e Δp é o diferencial de frequência alélica entre as subpopulações. Verifica-se que os dois componentes do índice usam diferentes informações e por isso podem ser combinados. O índice

$I = b_1\hat{a}_1 + b_2\hat{a}_2$ combina também componentes obtidos via diferentes conceitos e estimadores.

Os pesos do índice são dados por $b = P^{-1}C$, em que:

$$C = \begin{bmatrix} \sigma^2_{a_1} r^2_{\hat{a}_1 a} \\ \sigma^2_{a_2} r^2_{\hat{a}_2 a} \end{bmatrix}; P = \begin{bmatrix} \sigma^2_{a_1} r^2_{\hat{a}_1 a} & cov(a_1, a_2) r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} \\ cov(a_1, a_2) r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} & \sigma^2_{a_2} r^2_{\hat{a}_2 a} \end{bmatrix} =$$

$$\begin{bmatrix} \sigma^2_{a_1} r^2_{\hat{a}_1 a} & \sigma^2_{\Delta} r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} \\ \sigma^2_{a_2} r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} & \sigma^2_{a_2} r^2_{\hat{a}_2 a} \end{bmatrix},$$

em que C é uma matriz de variâncias e confiabilidades dos valores genômicos estimados via Delta-p e G-BLUP, P é uma matriz de covariâncias entre os valores genômicos estimados via Delta-p e G-BLUP e a é o valor genético aditivo verdadeiro dos indivíduos.

Dividindo P e C por $\sigma^2_{a_1}$ e sendo $\Delta^2 = \sigma^2_{a_2}/\sigma^2_{a_1}$, tem-se:

$$C = \begin{bmatrix} r^2_{\hat{a}_1 a} \\ \Delta^2 r^2_{\hat{a}_2 a} \end{bmatrix}; P = \begin{bmatrix} r^2_{\hat{a}_1 a} & \Delta^2 r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} \\ \Delta^2 r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} & \Delta^2 r^2_{\hat{a}_2 a} \end{bmatrix}.$$

Os pesos e acurácia ou correlação r_{Ia} entre o índice e o valor genético aditivo verdadeiro a, são:

$$b_1 = \frac{1 - r^2_{\hat{a}_2 a}\Delta^2}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}; b_2 = \frac{1 - r^2_{\hat{a}_1 a}}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2};$$

$$r_{Ia} = \left[\frac{r^2_{\hat{a}_1 a} + r^2_{\hat{a}_2 a}\Delta^2 - 2 r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}\right]^{1/2} = \left[1 - \frac{(1 - r^2_{\hat{a}_1 a})(1 - r^2_{\hat{a}_2 a}\Delta^2)}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}\right]^{1/2}.$$

Assim, são necessárias as confiabilidades (quadrado da acurácia) de $\hat{a}_1$ ($r^2_{\hat{a}_1 a}$, obtida via G-BLUP) e de $\hat{a}_2$($r^2_{\hat{a}_2 a}$, obtida via Delta-p) e a proporção entre as variâncias de $\hat{a}_2$ e $\hat{a}_1$. Se $\Delta^2 = \sigma^2_{a_2}/\sigma^2_{a_1}$ produzir resultado maior que 1, deve-se fazer $\Delta^2 = 1$. Outra opção é calcular $cov(a_1, a_2)$ diretamente via $cov(\hat{a}_1, \hat{a}_2)$.

Considerando os efeitos devido à dominância o procedimento é análogo ao descrito acima, deve-se apenas utilizar o valor genômico devido à dominância ($\hat{d}$) ao

invés do valor genômico aditivo ($\hat{a}$).

1.6.Método de Regressão Categórica Tripla (TCR)

Resende et al. (2014) introduziram um novo e simples método, chamado Regressão Categórica Tripla (TCR), para estimação da herdabilidade genômica em uma população totalmente não aparentada e não estruturada (Por exemplo, populações base para melhoramento ou composta; populações F2 de espécies autógamas).

Esse método, comparativamente a outros, permite avaliar a presença de estrutura de população, pois, permite capturar a herdabilidade genômica devido exclusivamente ao desequilíbrio de ligação excluindo as informações de co-segregação e relações de parentesco IBD (*Identical by Descent* - efeitos de família). Assim, esse método propicia um limite inferior na magnitude da herdabilidade genômica.

Em vez de regressar os fenótipos em milhares de locos marcadores, o método TCR regressa fenótipos nas três categorias de genótipos marcadores, MM, Mm e mm, visando capturar os efeitos genéticos em um loco b com categorias genotípicas BB, Bb e bb, em que B é o alelo favorável. Os métodos tradicionais regressam, através de todos os locos, os fenótipos no número de M em cada loco marcador. O TCR regressa no número total de indivíduos em cada categoria genotípica. Isto é coerente com a filosofia do modelo genético infinitesimal (caracteres governados por muitos genes de pequenos efeitos e sem locos de grandes efeitos) e então, com a filosofia do G-BLUP e RR-BLUP. Entretanto, o TCR é computacionalmente vantajoso e pode ser um melhor estimador dos componentes da variação genotípica e da herdabilidade.

O método TCR pode ser combinado com a metodologia do método Delta-p e com isso dar origem a um procedimento mais eficiente. Pelo procedimento do TCR

aliado ao método Delta-p, a população de estimação inicialmente é dividida em duas subpopulações, uma com os indivíduos ou famílias acima da média geral (subpopulação 1, com valor fenotípico médio u_1 superior) e outra com os indivíduos ou famílias abaixo da média geral (subpopulação 2, com valor fenotípico médio u_2 inferior). A diferença $(u_1 - u_2)$ entre os valores fenotípicos médios das duas subpopulações é devida à maior frequência alélica (p) dos alelos favoráveis (e menor frequência dos alelos desfavoráveis) na subpopulação 1 em relação à subpopulação 2. Assim, $(u_1 - u_2)$ é explicada por $(\Delta p = p_1 - p_2)$, sendo delta-p (Δp) a diferença de frequências alélicas p_1 e p_2 entre as duas subpopulações. Os valores de Δp são calculados para cada loco marcador e aqueles com sinais positivos são alocados como favoráveis, ou seja, os efeitos de substituição alélica (α_i) desses marcadores são tomados como positivos. Da mesma forma, aqueles com sinais negativos de Δp tem seus $\alpha_i's$ atribuídos como negativos. Assim, a codificação utilizada pelo método Delta-p também pode ser usada no método TCR (*Triple Categorical Regression*) proposto por Resende et al. (2015), ou seja, redefine-se a codificação da matriz de incidência dos marcadores W, compatibilizando-se o arquivo de marcas formado por 0 (mm), 1 (Mm) e 2 (MM) de forma a se ter um arquivo de "genes" dado por 0 (bb), 1 (Bb) e 2 (BB), sendo que a alocação em BB ou bb é ditada pelo sinal de α_i. Logicamente, o acerto na alocação em BB ou bb é probabilístico. Em média (esperança matemática) haverá acerto na maioria dos locos, sendo que o maior número de erros será naqueles locos marcadores de efeitos muito pequenos, tendendo a zero. A abordagem não demanda método computacional iterativo e usa apenas o conceito de distância genética (sinal de Δp_i) associada as duas subpopulações.

O algoritmo completo é:

(i) subdivisão da população de treinamento em duas, de acordo com o fenótipo corrigido para efeito ambiental;

(ii) cálculo de Δp_i;

(iii) se o sinal de Δp_i for negativo, trocar 0 por 2 e 2 por 0 em cada coluna de marcador com Δp_i negativo;

(iv) determinar a quantidade (n_{BB}) do código 2 na linha correspondente a cada indivíduo j do arquivo de marcas e fazer o mesmo para os códigos 1 e 0, obtendo n_{Bb} e n_{bb};

(v) o modelo de regressão categórica tripla é definido como:

$$y = 1\mu + \beta_{BB} n_{BB} I_{(BB)} + \beta_{Bb} n_{Bb} I_{(Bb)} + \beta_{bb} n_{bb} I_{(bb)} + e,$$

em que $I_{(BB)}$, $I_{(Bb)}$ e $I_{(bb)}$ são variáveis indicadoras. Se a categoria analisada é BB então $I_{(BB)} = 1$ e $I_{(Bb)} = I_{(bb)} = 0$. Analogamente, o mesmo pode ser definido para as demais categorias genotípicas.

Dessa forma, a estimação dos coeficientes de regressão ($\hat{\beta}$) via método dos mínimos quadrados ordinários (MQO), que se refere ao valor genético global de cada categoria genotípica, é dada por:

$$\hat{\beta}_{BB} = Cov(y_j, n_{BB-j})/Var(n_{BB-j})$$

$$\hat{\beta}_{Bb} = Cov(y_j, n_{Bb-j})/Var(n_{Bb-j})$$

$$\hat{\beta}_{bb} = Cov(y_j, n_{bb-j})/Var(n_{bb-j});$$

(vi) obtenção dos valores genotípicos ($\hat{u}_{BB-kj}, \hat{u}_{Bb-kj}$ e $\hat{u}_{bb-kj}$) por categoria genotípica dos marcadores, na soma de todos os k locos em cada indivíduo j, conforme segue, pela regressão (por meio de $\hat{\beta}$) dos fenótipos no número n de cada categoria: $\hat{u}_{BB-kj} = \hat{\beta}_{BB} n_{BB-j} = 2\alpha_{B-kj} + \delta_{BB-kj}$: valor genotípico total da categoria BB nos n_{BB} locos no indivíduo j;

$\hat{u}_{Bb-kj} = \hat{\beta}_{Bb} n_{Bb-j} = \alpha_{B-kj} + \alpha_{b-kj} + \delta_{Bb-kj}$: valor genotípico total da categoria Bb nos n_{Bb} locos no indivíduo j;

$\hat{u}_{bb-kj} = \hat{\beta}_{bb} n_{bb-j} = 2\alpha_{b-kj} + \delta_{bb-kj}$: valor genotípico total da categoria bb nos n_{bb} locos no indivíduo j, sendo $\delta_{BB} = -2q^2 d$, $\delta_{bb} = -2p^2 d$ e $\alpha_k = \alpha_{Bk} - \alpha_{bk}$ (Falconer, 1989);

(vii) disposição dos valores genotípicos totais de cada indivíduo em um vetor:

(viii) cômputo das variâncias genéticas conforme detalhado mais adiante;

(ix) estimação das herdabilidades dadas por: $\hat{h}_a^2 = \hat{\sigma}_{u_{ajs}}^2 / \sigma_y^2$ e $\hat{h}_d^2 = \hat{\sigma}_{u_{djs}}^2 / \sigma_y^2$, em que σ_y^2 é a variância entre os valores fenotípicos individuais.

A composição dos genótipos em termos de suas frequências, efeitos aditivos e de dominância e variâncias encontra-se no Quadro 1 a seguir. Essas informações foram usadas para compor os estimadores das variâncias genéticas pelo método TCR.

Quadro 1. Efeitos e variâncias genéticas paramétricas (teóricas).

Genótipo	**Frequências**	**Valor Genotípico**	**Efeito Aditivo**	**Efeito de dominância**
BB	p^2	a	$2\alpha_B = 2q\alpha$	$\delta_{BB} = -2q^2 d$
Bb	$2pq$	d	$\alpha_B + \alpha_b = (q-p)\alpha$	$\delta_{Bb} = 2pqd$
bb	q^2	$-a$	$2\alpha_b = -2p\alpha$	$\delta_{bb} = -2p^2 d$
Genótipo	**Frequências**	**Variância Aditiva**		**Variância de Dominância**
BB	p^2	$p^2(2\alpha_B)^2 = p^2(2q\alpha)^2$		$p^2(-2q^2 d)^2$
Bb	$2pq$	$2pq(\alpha_B + \alpha_b)^2 = 2pq[(q-p)\alpha]^2$		$2pq(2pqd)^2$
bb	q^2	$q^2(2\alpha_b)^2 = q^2(-2p\alpha)^2$		$q^2(-2p^2 d)^2$
Soma		$\sigma_{u_a}^2 = 2pq\alpha^2$		$\sigma_{u_d}^2 = (2pqd)^2$

Estimadores dos efeitos genéticos

Sendo a e $-a$ os valores genotípicos (Quadro 1) de BB e bb relacionados aos efeitos aditivos, tem-se que a soma $\hat{\mu}_{aj} = f(\hat{\alpha}_k) = \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ fornece uma estimativa dos efeitos aditivos dos indivíduos. Esses podem ser usados para o cômputo da acurácia seletiva e do viés da predição.

Sendo d o valor genotípico do heterozigoto (Quadro 1) Bb relacionado ao efeito de dominância, tem-se que $\hat{\mu}_{dj} = \hat{u}_{Bb-kj}$ fornece uma estimativa dos efeitos de dominância dos indivíduos. Esses podem ser usados para o cômputo da acurácia seletiva e do viés da predição dos efeitos de dominância. Com p tendendo a q, ou seja, $p \approx q \approx 0{,}5$, a quantidade $\hat{\mu}_{dj} = \hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ também é um estimador desses efeitos.

Estimadores das variâncias genéticas

Variância aditiva

Conforme o Quadro 1, $\sigma^2_{u_a} = 2pq\alpha^2$ e sendo $\hat{\mu}_{aj} = f(\hat{\alpha}_k) = \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$, tem-se que $\sigma^2_{u_a} = 2pqf(\hat{\alpha}_k) = 2pqVar(\hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ é um estimador para a variância genética aditiva, em que $\hat{\alpha}_k$ é um estimador intrínseco para o efeito de substituição alélica nos k locos.

Variância de dominância

Também conforme o Quadro 1, $\sigma^2_{u_d} = (2pqd)^2$. O contraste $2\hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ fornece uma estimativa de d e, portanto, $\sigma^2_{u_d} = (2pq)^2Var(2\hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ é um estimador para a variância genética de dominância. Com $p \approx q \approx 0{,}5$, a quantidade $\sigma^2_{u_d} = (2pq)^2 4Var(\hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ também é um estimador para $\sigma^2_{u_d}$.

Variância genotípica total

A variância do somatório $\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ fornece informação sobre a variância genotípica total como função $f(p, d, \alpha)$ de p, d e α. Assim, $Var(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj}) = f(p, d, \alpha)$ e as variâncias genéticas aditiva e de dominância podem ser extraídas de $f(p, d, \alpha)$ via $\sigma^2_{u_a} = 2pqVar(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ e $\sigma^2_{u_d} = (2pq)^2 Var(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$, respectivamente**.** Sendo assim, a variância genotípica total é dada por $\sigma^2_{u_g} = [2pq + (2pq)^2]Var(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$.

1.7.Validação

Para a implementação da metodologia de GWS, três populações, conhecidas por população de estimação, de validação e seleção podem ser definidas:

i) População de estimação é composta por indivíduos genotipados e fenotipados. É nesta população em que os efeitos dos marcadores no fenótipo são estimados.

ii) População de validação também é composta por indivíduos genotipados e fenotipados. Os valores genéticos genômicos dos indivíduos desta população são preditos usando os efeitos de marcadores estimados na população de estimação. Após isso, as correlações entre os valores preditos e os valores fenotípicos são obtidas para que seja possível calcular medidas de eficiência afim de comparar as metodologias abordadas.

iii) Na população de seleção há apenas indivíduos genotipados candidatos à seleção e não é necessário, ter nessa população, os fenótipos avaliados. As equações de predição provenientes da população de estimação são então

usadas na predição dos valores genéticos genômicos ou fenótipos futuros dos candidatos da seleção.

Conforme Resende et al. (2010) as três populações podem ser distintas uma das outras (validação independente), desempenhar duas funções ao mesmo tempo, isto é, apenas uma população usada para estimação e validação (procedimentos *Jacknife*) ou exercer três funções ao mesmo tempo sendo uma só população usada para estimação, validação e seleção (sem validação). A estratégia mais indicada é a validação independente e será esta a estratégia adotada no estudo apresentado no próximo capítulo.

2. Referências Bibliográficas

AZEVEDO, C.F.; RESENDE, M.D.V.; SILVA, F.F.; VIANA, J.M.S.; VALENTE, M.S.F., RESENDE JUNIOR, M.F.R.; MUÑOZ, P. Ridge, LASSO and Bayesian Additive-Dominance Genomic Models. **BMC Genetics**, v.16, p.105, 2015.

DA, Y.; WANG, C.; WANG, S.; HU, G. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. **PLOS ONE**, v. 9, n. 1, p. e87666, 2014.

DE LOS CAMPOS, G. et al. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. **Genetics**, v.182, n.1, p.375-385, 2009.

DE LOS CAMPOS, G.; HICKEY, J.M.; PONG-WONG, R.; DAETWYLER, H.D.; CALLUS, M.P.L. Whole genome regression and prediction methods applied to plant and animal breeding. **Genetics,** v.193, p. 327–45, 2012.

FALCONER, D. S.; MACKAY, T. F. C. **Introduction to Quantitative Genetics**, Ed 4. Longmans Green, Harlow, Essex, UK, 1996.

FALCONER, D. S. **Introduction to Quantitative Genetics**. Longmans, New York, p.438, 1989.

FERNANDO, R.L. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. **Proceedings of the 6th World Congress on Genetics Applied to Livestock Production**, Armidale, NSW, Australia. p. 329-336, 1998.

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

GIANOLA, D.; PEREZ‐ENCISO, M.; TORO, M.A. On marker‐assisted prediction of genetic value: beyond the ridge. **Genetics**, v.163, p.347‐365, 2003.

GIANOLA, D. Priors in whole-genome regression: the bayesian alphabet returns. **Genetics**, v.194, n. 3, p.573-96, 2013.

GODDARD, M. E. Genomic selection: prediction of accuracy and maximization of long term response. **Genetics**, Dordrecht, v.136, n.2, p.345-357, 2009.

GODDARD, M. E.; WRAY, N. R.; VERBYLA, k.; VISSCHER, P. M. Estimating effects and making predictions from genome-wide marker data. **Statistical Science**, Hayward, v. 24, p. 517-529, 2009.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of Genetic Relationship on Genome-Assisted Breeding Values. **Genetics**, v.117, p.2389-2397, 2007.

HAYES B. J.; VISSCHER P. M.; GODDARD M. E. Increased accuracy of artificial selection by using the realized relationship matrix. **Genetics Research**, v.91, p. 47-60, 2009.

LUSH, J.L. Family merit and individual merit as basis for selection. **America Naturalist**, v.81, p. 241-261, 1947.

MEUWISSEN, T.H.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819–29, 2001.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J.P. Effect of total allelic relationship on accuracy of evaluation and response to selection. **Journal of Animal Science**, v. 75, n. 7, p. 1738 – 1745, 1997.

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

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

RESENDE, M.D.V.; RESENDE JUNIOR, M.F.R.; AGUIAR, A.M.; ABAD, J.I.M.; MISSIAGIA, A.A.; SANSALONI, C.P.; PETROLI, C.D.; GRATTAPALIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, p.79, 2010.

RESENDE, M.D.V. (ORG.); SILVA, F.F (ORG.); AZEVEDO, C.F. (ORG.). **Estatística Matemática, Biométrica e Computacional**: **Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência.**1. ed. Visconde do Rio Branco: Suprema, v.1, p.881, 2014.

RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial.** Viçosa: Universidade Federal de Viçosa/Departamento de Estatística 291 p, 2012. Disponível em: <http://www.det.ufv.br/ppestbio/corpo_docente.php>.

RESENDE, M.D.V.; RAMALHO, M. A. P.; GUILHERME, S. R.; ABREU, A. F. B. Multi generation index in the within progenies bulk method for breeding of self-pollinated plants. **Crop Science**, v. 55, p. 1202-1211, 2015.

RESENDE, M. D. V. **Genética Quantitativa e de Populações**. 1. ed. Visconde do Rio Branco: Suprema, v.1, p.422, 2015.

STRANDEN, I.; GARRICK, D.J. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. **Journal of Dairy Science**, v.92, p. 2971–2975, 2009.

TIBSHIRANI, R. Regression shrinkage and selection via the LASSO. **Journal of the Royal Statistics Society Series B,** p. 267-288, 1996.

VAN RADEN, P. M. Efficient Methods to compute genomic predictions. **Journal of Dairy Science,** Champaign, v.91, n.11, p. 4414-4423, 2008.

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

WANG, C., DA, Y. Quantitative Genetics Model as the Unifying Model for Defining Genomic Relationship and Inbreeding Coefficient. **PLoS ONE**, v. 9, n. 12, e114484, 2014.

CAPÍTULO 2

REGRESSÃO NÃO-PARAMÉTRICA VIA MUDANÇA NA FREQUÊNCIA ALÉLICA ENTRE SUBPOPULAÇÕES

Resumo

A utilização de informações do DNA foi a principal contribuição da genética molecular no que se refere ao melhoramento genético. Neste sentido, idealizou-se a Seleção Genômica Ampla (GWS), a qual permite analisar um grande número de marcadores SNPs (*Single Nucleotide Polymorphisms*) que se encontram amplamente distribuídos no genoma. O grande diferencial da GWS está associado a estimar o efeito de cada marcador molecular no fenótipo e assim capturar genes que afetam um caráter quantitativo. Os principais métodos estatísticos aplicados na GWS pressupõem distribuições para os efeitos aleatórios do modelo ou são baseados em regressão implícita ou em redução de dimensionalidade. Assim, o objetivo desse trabalho foi avaliar uma metodologia não paramétrica, denominada Método Delta-p, além de compará-la com o método G-BLUP (*Genomic Best Linear Unbiased Predictor),* método tradicionalmente aplicado a GWS. Além disso, foi proposto um índice de seleção, denominado índice Delta-p/G-BLUP que combina os valores genômicos (GEBVs) provenientes do método G-BLUP com os valores genômicos do procedimento Delta-p. Sob o contexto Bayesiano, foi proposto um método que utiliza a distribuição de valores genômicos estimados via G-BLUP como distribuição *a priori* no método bayesiano BLASSO, esse método foi denominado método Bayes Híbrido. A eficiência dos métodos propostos, no que se refere à estimação dos GEBVs, foi avaliada via simulação de dados. Foram estabelecidos oito cenários, compostos pela combinação de dois níveis de herdabilidade × duas arquiteturas genéticas × ausência de dominância e dominância completa, sendo cada cenário simulado dez vezes. Os

métodos foram comparados por meio de validação independente e considerando medidas de eficiência, tais como acurácia da predição, viés e herdabilidade genômica. Os resultados indicaram que o Índice Delta-p/G-BLUP pode ser utilizado vantajosamente no contexto da GWS visto que conduziu a valores genômicos, aditivos e devido à dominância, mais acurados do que o método G-BLUP. O método Bayes Hibrido mostrou-se igualmente acurado ao LASSO Bayesiano, no entanto, conduz a estimativas de valores genômicos com viés mais próximo de um.

Palavras-chaves: Predição genômica, Índice de seleção, Estatística não paramétrica, Frequência alélica e Ganho genético.

1. INTRODUÇÃO

A evolução das técnicas de sequenciamento e genotipagem promoveu um grande avanço na genética molecular, o que consequentemente, beneficiou o melhoramento genético, uma vez que se tornou possível a utilização direta das informações do DNA na seleção de indivíduos geneticamente superiores. Meuwissen et al. (2001) propuseram a Seleção Genômica Ampla (GWS) visando viabilizar a seleção precoce direta dos indivíduos, aumentar o ganho genético por unidade de tempo e a eficiência na predição dos valores genéticos genômicos (*Genomic estimated breeding values* - GEBVs).

A GWS permite analisar centenas ou milhares de marcadores SNPs (*Single Nucleotide Polymorphisms*) que se encontram amplamente distribuídos no genoma, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação (*Linkage Disequilibrium* - LD) com pelo menos uma parte dos marcadores. O grande diferencial da GWS está associado a estimar o efeito de cada marcador molecular no fenótipo e assim permitir a predição dos valores genéticos genômicos

dos indivíduos sujeitos a seleção, incluindo os indivíduos que ainda não tenham seus fenótipos avaliados.

No entanto, a aplicação prática destas informações genômicas é um desafio, pois geralmente não é possível a utilização adequada de métodos tradicionais baseados em quadrados mínimos (*Least Squares*– LS) para estimar o efeito de cada SNP no fenótipo. Isso ocorre, devido ao número de marcadores ser geralmente muito maior do que o número de indivíduos genotipados e fenotipados e tais marcadores serem altamente correlacionados (Gianola et al., 2003).

Segundo Resende et al. (2014), os principais métodos estatísticos aplicados na GWS pressupõem distribuições para os efeitos aleatórios do modelo ou são baseados em regressão implícita ou em redução de dimensionalidade. Os métodos G-BLUP - *Genomic Best Linear Unbiased Predictor* (Habier et al., 2007; Van Raden, 2008; Goddard, 2009) e LASSO Bayesiano (de los Campos et al., 2009) têm sido amplamente aplicados a GWS e recomendados para a predição de valores genômicos (Azevedo et al., 2015; de los Campos et al. 2012; Gianola, 2013; Gianola et al., 2009).

No entanto, uma nova metodologia não paramétrica e puramente conceitual, denominada Método Delta-p, foi proposta por Resende (2015) para aplicação na Seleção Genômica, porém até o momento não foi avaliada. O método é baseado na distância genética entre duas subpopulações, utilizando o conceito de mudança na frequência alélica devido à seleção e o conceito teórico de ganho genético. É possível combinar os valores genômicos provenientes do método G-BLUP com os valores genômicos do procedimento Delta-p e estabelecer um índice de seleção, denominado índice Delta-p/G-BLUP. Enquanto que sob o contexto Bayesiano, pode-se combinar os GEBVs estimados via G-BLUP e os GEBVs via método bayesiano BLASSO, esse método foi denominado Bayes Híbrido. Neste método a distribuição dos valores

genômicos provenientes do método G-BLUP é assumida como distribuição *a priori* altamente informativa no método BLASSO.

Diante do exposto, o presente trabalho tem como principal objetivo propor e avaliar o método Delta-p e o índice Delta-p/G-BLUP comparado ao método G-BLUP quanto à eficiência na estimação dos valores genômicos aditivos e devido à dominância, utilizando dados simulados para oito cenários diferentes (dois níveis de herdabilidade × duas arquiteturas genéticas × ausência de dominância e dominância completa). Ademais, o trabalho também tem por finalidade avaliar o comportamento da distribuição *a priori* do método Bayes Híbrido em relação à eficiência na estimação dos valores genéticos aditivos, utilizando dados simulados para dois cenários com herdabilidade baixa e modelo infinitesimal com ausência de dominância e com dominância completa.

2. MATERIAIS E MÉTODOS

2.1.Dados Simulados

O conjunto de dados foi simulado usando o *software Real Breeding* (Viana, 2011) e sua geração foi descrita por Azevedo et al. (2015). Um total de 2000 marcadores SNPs equidistantes separados por 0,1 centiMorgan (cM) entre os dez cromossomos foram simulados. Os QTLs foram distribuídos nas regiões abrangidas pelo SNPs. Um montante de 1000 indivíduos de 20 famílias de irmãos completos foram genotipados e fenotipados.

Características com duas arquiteturas genéticas foram simuladas, uma seguindo um modelo infinitesimal (locos não ligados, com efeitos iguais) e outra com cinco genes de efeitos maiores, responsável por 50% da variabilidade genética. No primeiro caso, para cada um dos 100 QTLs foi atribuído um efeito aditivo de pequena

magnitude no fenótipo (sob a definição de Distribuição Normal). Para o segundo caso, pequenos efeitos aditivos foram designados para os restantes 95 locos. Os efeitos foram normalmente distribuídos com média zero e variância genética permitindo o nível de herdabilidade desejado. O valor fenotípico foi obtido adicionando ao valor genotípico um efeito ambiental proveniente de uma distribuição normal $N(0, \sigma_e^2)$, em que a variação σ_e^2 foi definida de acordo com dois níveis de herdabilidade no sentido restrito em torno de 0,20 e 0,35, respectivamente. Os níveis de herdabilidade foram escolhidos para representar uma característica com baixa herdabilidade e outra com herdabilidade moderada, casos em que se espera que a seleção genômica seja superior à seleção fenotípica (Azevedo et al., 2015).

As magnitudes das herdabilidades no sentido restrito e no sentido amplo estão associadas com um grau médio de nível de dominância *(d/a)* de aproximadamente 1 (domínio completo) e 0 (ausência de dominância) em uma população com frequências alélicas intermediárias. As simulações assumiram independência entre efeitos aditivos e devido à dominância, quando este era considerado, com efeitos de dominância tendo a mesma distribuição que os efeitos aditivos (ambos foram normalmente distribuídos com média zero). Marcadores com MAF (*Minor Allele Frequency* – Frequência do Menor Alelo) menor do que 5% foram excluídos das análises.

2.2.Cenários

Oito cenários diferentes foram utilizados nas análises: dois níveis de herdabilidades (cerca de 0,30 e 0,50, associados à herdabilidades em sentido restrito de 0,20 e 0,35, respectivamente) × duas arquiteturas genéticas × ausência de dominância e dominância completa. A descrição dos cenários é apresentada na Tabela 1.

Tabela 1 – Cenários com as respectivas médias das herdabilidades aditivas (h_a^2), devido a dominância (h_d^2) e total (h_g^2), arquiteturas genéticas (características controladas por genes de pequeno efeito – herança poligênica e características controladas por genes de pequeno e maior efeito - herança mista) e níveis de dominância (ausência de dominância e dominância completa).

Cenário	Arquitetura Genética	Nível de dominância	h_a^2	h_d^2	h_g^2
Cenário 1	Herança poligênica	Ausência	0,22	-	0,22
Cenário 2	Herança poligênica	Ausência	0,33	-	0,33
Cenário 3	Herança mista	Ausência	0,20	-	0,20
Cenário 4	Herança mista	Ausência	0,35	-	0,35
Cenário 5	Herança poligênica	Completa	0,21	0,10	0,31
Cenário 6	Herança poligênica	Completa	0,35	0,17	0,52
Cenário 7	Herança mista	Completa	0,20	0,13	0,33
Cenário 8	Herança mista	Completa	0,33	0,21	0,54

Estes oito cenários foram analisados considerando-se os dois métodos, Delta-p e G-BLUP. Cada tipo de população (ou cenário) foi simulado 10 vezes. Nove repetições foram utilizadas como populações de treinamento (estimação dos efeitos aditivos e devido à dominância dos marcadores), e uma repetição foi utilizada como população de validação (estimação dos valores genômicos aditivos e devido à dominância dos indivíduos desta população). As estimativas baseadas em cada uma das nove repetições foram validadas para obtenção das estimativas de acurácia, de viés e de herdabilidades genômicas. Assim, essas medidas foram calculadas em cada repetição da simulação e depois foi obtida a média desses valores.

2.3.Método G-BLUP

A predição dos valores genômicos por meio do método G-BLUP (*Genomic Best Linear Unbiased Predictor)* é baseada no modelo linear misto dado por:

$$y = 1\mu + Za + Zd + e,$$

em que:

y é o vetor de fenótipos ($N\times1$, em que N é o número de indivíduos genotipados e fenotipados);

μ é a média geral e 1 é o vetor com dimensão $N\times1$ cujos seus elementos são iguais a 1;

a é o vetor de efeitos genômicos aditivos dos indivíduos ($N\times 1$) com matriz de incidência Z ($N\times N$), sendo a estrutura de variância dada por $a\sim N(0, G_a\sigma_a^2)$ em que σ_a^2 é a variância aditiva e G_a ($N\times N$) é a matriz de parentesco genômica para os efeitos aditivos;

d é o vetor de efeitos genômicos devido à dominância dos indivíduos com matriz de incidência Z ($N \times N$), sendo a estrutura de variância dada por $d \sim N(0, G_d\sigma_d^2)$ em que σ_d^2 é a variância devido à dominância e G_d ($N \times N$) é a matriz de parentesco genômica para os efeitos devido à dominância;

e é o vetor de efeitos residuais aleatórios com $e \sim N(0, I\sigma_e^2)$ e σ_e^2 é a variância residual.

As equações de modelos mistos para predição de a e d via o método G-BLUP equivalem a:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\dfrac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'X & Z'Z & Z'Z + G_d^{-1}\dfrac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix},$$

sendo os componentes de variância, σ_d^2, σ_a^2 e σ_e^2, estimados via REML (*Restricted maximum likelihood*).

Conforme Vitezica et al. (2013), as matrizes de parentesco genômicas para efeitos aditivos e para efeitos devido à dominância, G_a e G_d, são dadas respectivamente por:

$$G_a = \frac{WW'}{\sum_{i=1}^{n}(2p_iq_i)} \quad e \quad G_d = \frac{SS'}{\sum_{i=1}^{n}(2p_iq_i)^2},$$

em que p_i e q_i as frequências alélicas do locos i, W é a matriz de incidência para os vetores de efeitos aditivos de marcadores (α) e S é a matriz de incidência para os vetores de efeitos devido à dominância de marcadores (δ) parametrizadas conforme Van Raden (2008), Vitezica et al. (2013), Wang e Da (2014), Da et al. (2014) e Resende et al. (2014) e que é apresentada a seguir:

$$W = \begin{cases} Se\ MM, então\ 2 - 2p \rightarrow 2q \\ Se\ Mm, então\ 1 - 2p \rightarrow q - p \\ Se\ mm, então\ 0 - 2p \rightarrow -2p \end{cases} \qquad S = \begin{cases} Se\ MM, então\ 0 \rightarrow 2q^2 \\ Se\ Mm, então\ 1 \rightarrow 2pq \\ Se\ mm, então\ 0 \rightarrow -2p^2 \end{cases}.$$

2.4.Método Delta-p

O método Delta-p utiliza o conceito de mudança na frequência alélica devido à seleção e o conceito teórico de ganho genético (contraste entre médias de duas subpopulações). Dessa forma, a população de estimação inicialmente é dividida em duas subpopulações, uma com os indivíduos acima da média geral (subpopulação 1, com valor genético médio u_1 superior) e outra com os indivíduos abaixo da média geral (subpopulação 2, com valor genético médio u_2 inferior). A diferença ($u_1 - u_2$) entre os valores genéticos médios das duas subpopulações é devida à maior frequência alélica (p) dos alelos favoráveis (e menor frequência dos alelos desfavoráveis) na subpopulação 1 em relação à subpopulação 2. Assim, ($u_1 - u_2$) é explicada por ($\Delta p = p_1 - p_2$), sendo delta-p (Δp) a diferença de frequências alélicas entre as duas subpopulações. Os valores de Δp são calculados para cada loco marcador e aqueles com sinais positivos são alocados como favoráveis, ou seja, os efeitos de substituição alélica (α_i) desses marcadores são tomados como positivos. Da mesma forma, aqueles com sinais negativos de Δp tem seus $\alpha_i's$ atribuídos como negativos.

Dessa forma, redefine-se a codificação da matriz de incidência W para os vetores de efeitos aditivos de marcadores, compatibilizando-se o arquivo de marcas

formado por 0 (mm), 1 (Mm) e 2 (MM) de forma a se ter um arquivo de "genes" dado por 0 (bb), 1 (Bb) e 2 (BB), sendo o alelo B o alelo favorável e a alocação em BB ou bb é ditada pelo sinal de α_i. Logicamente, o acerto na alocação em BB ou bb é probabilístico. Em média (esperança matemática) haverá acerto na maioria dos locos, sendo que o maior número de erros será naqueles locos marcadores de efeitos muito pequenos, tendendo a zero.

O método Delta-p necessita da estimação dos efeitos α_i e posteriormente da predição dos valores genômicos dos indivíduos (a). Tomando Δp_i como indicador da magnitude relativa de α_i (quanto maior o Δp_i, maior a chance de o valor de α_i ser maior), tem-se que a proporção $\Delta p_i/\Delta p_m$ entre os Δp_i (com seus respectivos sinais positivos ou negativos) e o delta-p médio (Δp_m, computado usando os módulos de Δp_i), multiplicado pelo α médio (α_m) fornece os efeitos α_i. Assim, $\alpha_i = (\Delta p_i/\Delta p_m)\,\alpha_m$.

A quantidade (α_m) pode ser obtida usando a expressão teórica de ganho genético. Para um loco, tem-se $u_j = (p_j - q_j)\,\alpha_i = (2p_j - 1)\,\alpha_i$, sendo $j = 1,2,\ldots,N$. Assim, $(u_1 - u_2) = (2p_1 - 2p_2)\alpha_i = 2(p_1 - p_2)\alpha_i = 2\Delta p_i\alpha_i$. Na soma dos locos tem-se $(u_1 - u_2) = 2\sum(\Delta p_i\alpha_i)$. A quantidade Δp_m é definida como: $\Delta p_m = sq(1-q)$, em que $s = \alpha k/\sigma_F$ é o coeficiente seletivo do loco, assumindo um modelo aditivo-dominante. Assim, $\Delta p_m = \alpha kpq/\sqrt{\sigma_F^2}$, sendo k o índice de seleção e σ_F^2 a variância fenotípica. O ganho genético é definido como $G_s = k\sigma_a^2/\sqrt{\sigma_F^2}$. Dessa forma, $k/\sqrt{\sigma_F^2} = G_s/\sigma_a^2$. Substituindo na expressão para Δp_m, tem-se: $\Delta p_m = \alpha pqG_s/\sigma_a^2$.

O ganho da subpopulação 1 em relação à subpopulação 2 é duplicado pois é composto de duas partes: seleção à direita truncada no ponto zero de uma curva normal padrão, com diferencial de seleção ($u_1 - u_0$) e seleção à esquerda truncada no ponto

zero, com diferencial de seleção ($u_0 - u_2$). O ganho genético associado ao primeiro diferencial de seleção é dado por $G_{s1} = h_a^2(u_1 - u_0)$ e aquele associado ao segundo diferencial de seleção é dado por $G_{s2} = h_a^2(u_0 - u_2)$.

Esses ganhos são simétricos e devem ser somados para a obtenção do ganho total $2G_s$, dado por: $2G_s = G_{s1} + G_{s2} = h_a^2(u_1 - u_2)$. Assim, $G_s = \left(\frac{1}{2}\right) h_a^2(u_1 - u_2)$. Substituindo G_s em Δp_m, $\Delta p_m = \alpha pq G_s/\sigma_a^2$, tem-se:

$$\Delta p_m = \alpha pq(1/2)h_a^2(u_1 - u_2)/\sigma_a^2$$

sendo $\sigma_a^2 = 2pq\alpha^2$, substituindo e dividindo por Δp_m, tem-se:

$$1 = [\alpha pq(1/2)h_a^2(u_1 - u_2)]/(2pq\alpha^2\Delta p_m)$$

Simplificando tem-se $1 = (1/2)h_a^2(u_1 - u_2)/(2\alpha\Delta p_m)$. Isolando α da expressão anterior tem-se $\alpha = (1/2)h_a^2(u_1 - u_2)/(2\Delta p_m)$ para n locos e o α médio por loco é dado por:

$$\alpha_m = 0{,}5h_a^2\,(u_1 - u_2)/\,[n_{marcas}2\Delta p_m].$$

Assim, os efeitos aditivos dos marcadores α são estimados na população de treinamento (indivíduos com fenótipos e genótipos conhecidos) de forma não paramétrica. Com base nesses efeitos os valores genéticos (a) dos indivíduos da população de validação (indivíduos com fenótipos e genótipos conhecidos) ou indivíduos da população de seleção (indivíduos com genótipos conhecidos) são estimados por meio da expressão via $\hat{a} = W\hat{\alpha}$, sendo W a matriz de incidência dos efeitos genéticos aditivos dos marcadores. A abordagem não demanda método computacional iterativo e usa apenas os conceitos de distância genética (magnitude de Δp_i) e ganho genético associados as duas subpopulações. Não usa também o *shrinkage* diferenciado com base nas frequências alélicas, fato que beneficia locos com maiores MAFs (*Minor allele frequency*).

O algoritmo completo é:

(iv) subdivisão da população de treinamento em duas, de acordo com o fenótipo corrigido para efeitos ambientais;

(v) cálculo de Δp_i e Δp_m;

(vi) cálculo de α_m; (iv) cálculo de $\alpha_i = (\Delta p_i/\Delta p_m)\alpha_m$.

Pode-se realizar também a seleção de marcas com base em Δp_i ou α_i e a estimação da variância aditiva via uma função quadrática dos α_i. O método será tanto melhor quanto maior for o tamanho da população de estimação.

Para os efeitos de dominância deve-se usar $\Delta(2pq)_i$ (diferença entre as frequências de genótipos heterozigotos das subpopulações 1 e 2) em lugar de Δp_i. O fundamento advém de o fato relatado a seguir. Quanto melhor o fenótipo e simultaneamente maior a frequência de genótipos heterozigotos na população, mais favorável é o valor do genótipo corrigido para os efeitos genéticos aditivos, ou seja, maiores são os desvios de dominância. Desta forma, para inferir sobre os efeitos de dominância, o interesse é encontrar esta associação (frequência de heterozigotos e valores dos fenótipos) e, para isso, o fenótipo deve ser regressado na frequência de heterozigotos, corrigida para os efeitos aditivos. Assim, a matriz de incidência S (distribuição Bernoulli) é inerente à heterozigose ou a frequência de heterozigotos dentre as três classes genotípicas. Define-se então g_m como o valor genotípico do heterozigoto, o qual corrigido para o efeito aditivo (α_i) do loco fornece o desvio de dominância do loco i (δ_i) e dos indivíduos (vetor d), dado por $\hat{d} = S\hat{\delta}$, em que S é a matriz de incidência dos efeitos devido à dominância dos marcadores.

2.5.Método do Índice Delta-p/G-BLUP

Com base no princípio da seleção combinada (Lush, 1947; Falconer, 1989), o método G-BLUP pode ser usado conjuntamente ao procedimento Delta-p. Um índice

de seleção da forma $I = b_1\hat{a}_1 + b_2\hat{a}_2$ pode ser estabelecido usando as informações dos valores genômicos aditivos preditos provenientes do G-BLUP ($\hat{a}_1$) e Delta-p ($\hat{a}_2$), ponderados pelos pesos b_1 e b_2, respectivamente.

O índice pode ser representado também por $I = b_1 f(y; X) + b_2 f((u_1 - u_2); \Delta p)$ em que y é o vetor de dados fenotípicos individuais, X é a matriz de dosagem alélica dentro de indivíduo centralizada, $(u_1 - u_2)$ é o contraste de médias de subpopulações e Δp é o diferencial de frequência entre as subpopulações. Verifica-se que os dois componentes do índice usam diferentes informações e por isso podem ser combinados. O índice $I = b_1\hat{a}_1 + b_2\hat{a}_2$ combina também componentes obtidos via diferentes conceitos e estimadores.

Os pesos do índice são dados por $b = P^{-1}C$, em que:

$$C = \begin{bmatrix} \sigma^2_{a_1} r^2_{\hat{a}_1 a} \\ \sigma^2_{a_2} r^2_{\hat{a}_2 a_2} \end{bmatrix}; P = \begin{bmatrix} \sigma^2_{a_1} r^2_{\hat{a}_1 a} & cov(a_1, a_2) r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} \\ cov(a_1, a_2) r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} & \sigma^2_{a_2} r^2_{\hat{a}_2 a} \end{bmatrix} =$$

$$\begin{bmatrix} \sigma^2_{a_1} r^2_{\hat{a}_1 a} & \sigma^2_{\Delta} r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} \\ \sigma^2_{a_2} r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} & \sigma^2_{a_2} r^2_{\hat{a}_2 a} \end{bmatrix},$$

em que C é uma matriz de variâncias e confiabilidades dos valores genômicos estimados via Delta-p e G-BLUP, P é uma matriz de covariâncias entre os valores genômicos estimados via Delta-p e G-BLUP e a é o valor genético aditivo verdadeiro dos indivíduos.

Dividindo P e C por $\sigma^2_{a_1}$ e sendo $\Delta^2 = \sigma^2_{a_2}/\sigma^2_{a_1}$, tem-se:

$$C = \begin{bmatrix} r^2_{\hat{a}_1 a} \\ \Delta^2 r^2_{\hat{a}_2 a_2} \end{bmatrix}; P = \begin{bmatrix} r^2_{\hat{a}_1 a} & \Delta^2 r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} \\ \Delta^2 r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a} & \Delta^2 r^2_{\hat{a}_2 a} \end{bmatrix}.$$

Os pesos e acurácia ou correlação r_{Ia} entre o índice e o valor genético aditivo verdadeiro a, são:

$$b_1 = \frac{1 - r^2_{\hat{a}_2 a}\Delta^2}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}; b_2 = \frac{1 - r^2_{\hat{a}_1 a}}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2};$$

$$r_{Ia} = \left[\frac{r^2_{\hat{a}_1 a} + r^2_{\hat{a}_2 a}\Delta^2 - 2r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}\right]^{1/2} = \left[1 - \frac{(1 - r^2_{\hat{a}_1 a})(1 - r^2_{\hat{a}_2 a}\Delta^2)}{1 - r^2_{\hat{a}_1 a} r^2_{\hat{a}_2 a}\Delta^2}\right]^{1/2}.$$

Assim, são necessárias as confiabilidades de $\hat{a}_1$ ($r^2_{\hat{a}_1 a}$, obtida via G-BLUP) e de $\hat{a}_2$ ($r^2_{\hat{a}_2 a}$, obtida via Delta-p) e a proporção entre as variâncias de $\hat{a}_2$ e $\hat{a}_1$. Se $\Delta^2 = \sigma^2_{a_2}/\sigma^2_{a_1}$ produzir resultado maior que 1, deve-se fazer $\Delta^2 = 1$. Outra opção é calcular $cov(a_1, a_2)$ diretamente via $cov(\hat{a}_1, \hat{a}_2)$.

Considerando os efeitos devido à dominância o procedimento é análogo ao descrito acima, deve-se apenas utilizar o valor genômico devido à dominância ($\hat{d}$) ao invés do valor genômico aditivo ($\hat{a}$).

2.6.Método BLASSO

A versão bayesiana da regressão via LASSO (BLASSO - Park e Casella, 2008) para seleção genômica ampla foi proposta por de los Campos et al. (2009). O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*) inclui um termo de variância comum para os efeitos genéticos de marcadores e de efeitos residuais. Portanto, usando o modelo linear básico para predição dos efeitos dos marcadores, $y = Xb + Wm_a + Sm_d + e$, sendo, y é o vetor de fenótipos, b é o vetor de efeitos sistemáticos com matriz de incidência X, m_a é o vetor de efeitos genéticos aditivos dos marcadores com matriz de incidência W, m_d é o vetor de efeitos genéticos devido à dominância dos marcadores com matriz de incidência S e e é o vetor de resíduos. As distribuições *a priori* dos parâmetros do modelo são apresentadas a seguir:

$$e|\sigma^2 \sim MVN(0, I\sigma^2)$$

$$m_{ai}|\lambda_a, \sigma^2 \sim \prod_i \left(\frac{\lambda_a}{2\sigma}\right) e^{\left[\frac{-\lambda_a|m_{ai}|}{\sigma}\right]}$$

$$m_{di}|\lambda_d, \sigma^2 \sim \prod_i \left(\frac{\lambda_d}{2\sigma}\right) e^{\left[\frac{-\lambda_d|m_{di}|}{\sigma}\right]}$$

em que MNV representa a distribuição normal multivariada, λ_a e λ_d são parâmetros de "suavização" que são estimados por meio do conjunto de dados via método MCMC usando uma *priori* não informativa, σ^2 tem como distribuição a *priori* uma qui-quadrado invertida escalada.

Usando uma formulação em termos de um modelo hierárquico aumentado, tem-se:

$$m_{ai}|\tau_a \sim N(0, D_a\sigma^2),\ m_{di}|\tau_d \sim N(0, D_d\sigma^2)$$

$$p(\tau_a^2\ |\lambda_a^2) = \prod_i \left(\frac{\lambda_a^2}{2}\right) e^{\left[\frac{-\lambda_a^2\tau_{ai}^2}{2}\right]} \text{ e } p(\tau_d^2\ |\lambda_d^2) = \prod_i \left(\frac{\lambda_d^2}{2}\right) e^{\left[\frac{-\lambda_d^2\tau_{di}^2}{2}\right]}$$

em que $D_a = diag\left(\tau_{1a}^2, \tau_{2a}^2, \dots, \tau_{ma}^2\right)$ e $D_d = diag\left(\tau_{1d}^2, \tau_{2d}^2, \dots, \tau_{md}^2\right)$. Isso leva a uma distribuição exponencial dupla para os efeitos marcadores (Park e Casella, 2008), como a seguir:

$$m_{ai}|\lambda_a^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda_a}\right) \text{ e } m_{di}|\lambda_d^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda_d}\right).$$

A variância genética aditiva e devido à dominância de cada locos marcador é dada respectivamente por $\sigma_{mai}^2 = \tau_{ai}^2\sigma^2$ e $\sigma_{mdi}^2 = \tau_{di}^2\sigma^2$ com $i = 1, 2, \dots, m$. Dessa forma, a variância genética aditiva e de dominância podem ser estimadas utilizando as relações $\sigma_a^2 = \sum_{i=1}^m 2p_iq_i\,\sigma_{mai}^2$ e $\sigma_d^2 = \sum_{i=1}^m [2p_iq_i]^2\,\sigma_{mdi}^2$, sendo assim, $\sigma_a^2 = \sum_{i=1}^m 2p_iq_i\tau_{ai}^2\sigma^2$ e $\sigma_d^2 = \sum_{i=1}^m [2p_iq_i]^2\,\tau_{di}^2\sigma^2$. Os valores genômicos aditivos e devido à dominância são estimados via as seguintes expressões $\hat{a} = W\hat{m}_a$ e $\hat{d} = S\hat{m}_d$, respectivamente.

As distribuições condicionais completas *a posteriori* para os parâmetros do BLASSO são apresentados em detalhes por de los Campos et al. (2009).

Conforme Gianola (2013), quaisquer diferenças nas inferências *a posteriores* entre os métodos bayesianos são devidas a influência e a diferença entre as distribuições *a priori*. Com base nesses resultados, pode-se afirmar que diferentes métodos podem ser equipados com a mesma filosofia apenas alterando drasticamente a distribuição *a priori* por meio dos hiperparâmetros (Azevedo et al., 2015). A capacidade preditiva dos métodos bayesianos está associada a escolha adequada dos hiperparâmetros (Azevedo et al., 2015). Ao assumir oito (8) graus de liberdade na distribuição *a priori* de qui-quadrado invertida escalada para a variância genética conduz a distribuições com caudas suficientemente espessas associadas a uma distribuição semelhante a t de *student* para efeitos de marcadores (Gianola, 2013). Enquanto, que ao assumir menos dois (-2) graus de liberdade transforma a distribuição de qui-quadrado invertida escalada em uma distribuição uniforme.

Neste trabalho foram utilizadas 100.000 iterações para os algoritmos MCMC (*Markov* chain Monte Carlo), das quais 20.000 foram descartadas (*burn-in*) para garantir o aquecimento da cadeia e com selecção de uma em cada 10 iterações (*thin*). A análise de covergência foi realizada via o critério proposto por Geweke (1992).

2.7.Método Bayes Híbrido ou BLASSO/G-BLUP

O método Bayes Híbrido consiste em combinar os valores genômicos estimados via G-BLUP e os valores genômicos estimados via método bayesiano BLASSO. A distribuição *a posteriori* é dita condicional do parâmetro dadas as observações (y) e é proporcional ao produto da função de verossimilhança pela distribuição *a priori* dos parâmetros. Esta distribuição é a base de toda inferência a respeito dos parâmetros. Assim, no Bayes Hibrido, o G-BLUP seria considerado na *posteriori* por meio da distribuição *a priori* altamente informativa/exata, enquanto, que

a função de verossimilhança, em nível de marcas individuais (ou grupos ou regiões) poderia explicar alguma informação adicional que o G-BLUP não detectou. Dessa forma, a distribuição dos valores genômicos aditivos preditos ($\hat{a}$) via método G-BLUP seria assumida como distribuição *a priori* por meio dos hiperparâmetros das distribuições *a priori* dos componentes de variância e os valores fenotípicos (y) como os dados para a verossimilhança. Ou seja, usar a predito via G-BLUP e a predito via método Bayesiano, um deles como *priori* e outro como verossimilhança. Sendo esta uma "atualização da distribuição *a priori*".

Para avaliar o método Bayes Híbrido foram utilizados dados simulados para dois cenários com herdabilidade baixa e modelo infinitesimal com ausência de dominância e com dominância completa (cenário 1 e cenário 5).

2.8.Recursos Computacionais

Toda a implementação dos métodos utilizados foi realizada no software R (R Development Core Team, 2010), uma vez que este é de fácil acesso, por se tratar de um software livre (http://cran.r-project.org), e também por suportar a manipulação de grandes conjuntos de dados, como é caso dos arquivos de marcadores SNPs.

O G-BLUP foi realizado no pacote *rrBLUP* com a função *mixed.solve*, o LASSO Bayesiano e o Bayes Hibrído no pacote *BGLR* com a função *BLR*. O algoritmo utilizado para o desenvolvimento do método Delta-p está no Apêndice I.

2.9.Comparação das metodologias de seleção genômica ampla

Os métodos foram comparados por meio de uma validação independente em que as nove primeiras replicatas foram assumidas como populações de estimação e utilizadas para estimar os efeitos dos marcadores SNPs no fenótipo. A décima

replicata foi assumida como população de validação e utilizada para predizer os valores genéticos genômicos via estimativas dos efeitos dos marcadores obtidos na população de estimação. Assim, foi viável calcular medidas de eficiência das metodologias para a predição genômica, tais como acurácia ($r_{\hat{a}a}\ e\ r_{\hat{d}d}$), viés de predição ($b_{y\hat{a}}\ e\ b_{y\hat{d}}$), herdabilidade genômica aditiva e devido à dominância ($h^2_{aM}\ e\ h^2_{dM}$) e eficiência relativa (ER) das estimativas baseadas em cada um dos nove cenários simulados.

As medidas utilizadas são descritas a seguir: (i) a acurácia que é dada pela correlação entre os GEBVs e os valores genéticos paramétricos; (ii) o viés de predição o qual é definido como sendo o coeficiente da regressão entre o fenótipo e o GEBV, sendo que para coeficientes de regressão abaixo de 1 (< 1) entende-se que os GEBVs foram superestimados, para coeficientes de regressão acima de 1 (> 1) conclui-se que os GEBVs foram subestimados e para os coeficientes iguais a 1 ($= 1$) conclui-se que os GEBVs são não viesados; (iii) a herdabilidade molecular aditiva é dada por $h^2_{aM} = \frac{\sigma^2_{a_M}}{\sigma^2_{a_M}+\sigma^2_{d_M}+\sigma^2_e}$, em que $\sigma^2_{a_M} = \sum_{i=1}^{n} 2p_i q_i m_i^2$ é a variância genômica aditiva, m_i^2 é o quadrado do efeito do i-ésimo marcador, p_i e q_i são as frequências alélicas do i-ésimo marcador. A herdabilidade molecular devido à dominância é dada por $h^2_{dM} = \frac{\sigma^2_{d_M}}{\sigma^2_{a_M}+\sigma^2_{d_M}+\sigma^2_e}$, em que $\sigma^2_{d_M} = \sum_{i=1}^{n}(2p_i q_i d_i)^2$ e d_i é o valor genotípico do heterozigoto; (iv) a eficiência relativa é dada pelo quociente entre as acurácias provenientes dos métodos Índice Delta-p/G-BLUP e G-BLUP.

As medidas de acurácia, viés, herdabilidade e eficiência relativa foram obtidas para cada replicata em cada cenário e os resultados gerais relatados foram a média desses valores.

3. RESULTADOS E DISCUSSÃO

Os resultados médios e respectivos desvios-padrão estimados de acurácia, viés de predição e herdabilidades obtidos por meio dos dois métodos, Delta-p e G-BLUP, associados aos valores genômicos aditivos preditos considerando os modelos aditivo e aditivo-dominante e também os pesos, acurácia e eficiência obtidos via índice Delta-p/G-BLUP estão apresentados na Tabela 2.

Os resultados revelaram um aumento da acurácia com o uso das marcas selecionadas. Assim, as diferenças entre frequências alélicas foram efetivas em indicar as magnitudes dos efeitos das marcas. Isto corrobora o princípio e fundamento do método Delta-p. Weller et al. (2014a, b) e Weller (2016) reportaram sobre a seleção de marcas baseada nas diferenças entre frequências alélicas entre grupos de touros jovens e mais velhos, ou seja, entre duas subpopulações contrastantes.

A correlação entre os valores genômicos preditos via método Delta-p e G-BLUP foi elevada ($r_{\hat{a}_p \hat{a}_g}$), variando de 0,80 a 0,88, porém essa correlação não foi perfeita. Dessa forma, combinar esses valores em um índice dando pesos distintos a cada um deles pode ser vista como uma melhoria nas predições dos GEBVs. De acordo com a Tabela 1, para modelo aditivo, o aumento da acurácia na predição utilizando o índice Delta-p/G-BLUP (0,80 a 0,87) foi em média de 0,03 unidades em relação à acurácia do método G-BLUP (0,77 a 0,85). Enquanto que para o modelo aditivo-dominante esse acréscimo foi mais evidente, sendo de 0,08 unidades. O índice Delta-p/G-BLUP, considerando o modelo aditivo, produziu eficiência média relativa de 103% a 105%, enquanto que no modelo aditivo-dominante essa eficiência aumentou para 109% a 116%. Nota-se que o índice Delta-p/G-BLUP ocasionou uma melhoria na predição dos GEBVs via G-BLUP, visto que o índice proporcionou acurácias

superiores. É importante relatar que uma grande vantagem é que esses pontos percentuais a mais não têm custo computacional adicional.

Conforme já mencionado, o método do índice Delta-p/G-BLUP é fundamentado no índice de seleção combinada, esta metodologia pondera os valores individuais e os valores das suas respectivas famílias (Lush, 1945). Segundo (Lush,

Tabela 2: Médias e desvios-padrão de acurácia ($r_{\hat{a}a}$), viés ($b_{y\hat{a}}$), herdabilidade aditiva (h^2_{aM}), correlação entre os valores genômicos aditivos estimados via método Delta-p e via G-BLUP ($r_{\hat{a}_1\hat{a}_2}$), proporção entre as variâncias aditivas (Δ^2) estimadas, pesos (b_1 e b_2) e acurácia (r_{Ia}) do método do índice Delta-p/G-BLUP, eficiência relativa (ER) entre o índice e o G-BLUP.

Modelo	Cenários	Método	$r_{\hat{a}a}$	$b_{y\hat{a}}$	h^2_{aM}	$r_{\hat{a}_1\hat{a}_2}$	Δ^2	b_1	b_2	r_{Ia}	ER
Aditivo	1	Delta-p	0,67±0,03	1,13±0,10	0,11±0,01	0,86±0,03	0,62±0,09	0,87±0,02	0,49±0,02	0,80±0,01	1,05±0,01
		GBLUP	0,77±0,01	1,02±0,06	0,17±0,02						
	2	Delta-p	0,71±0,01	1,00±0,08	0,25±0,03	0,82±0,03	0,71±0,12	0,85±0,03	0,42±0,02	0,85±0,01	1,03±0,01
		GBLUP	0,83±0,01	1,02±0,09	0,35±0,04						
	3	Delta-p	0,69±0,02	1,15±0,07	0,11±0,01	0,88±0,01	0,67±0,17	0,85±0,04	0,47±0,03	0,83±0,01	1,04±0,01
		GBLUP	0,79±0,01	1,05±0,11	0,18±0,03						
	4	Delta-p	0,72±0,01	1,05±0,06	0,24±0,02	0,86±0,02	0,69±0,09	0,86±0,03	0,38±0,03	0,87±0,01	1,03±0,01
		GBLUP	0,85±0,01	1,00±0,05	0,36±0,03						
Aditivo-Dominante	5	Delta-p	0,55±0,06	0,73±0,10	0,16±0,03	0,84±0,04	1,09±0,32	0,78±0,05	0,69±0,04	0,73±0,02	1,15±0,04
		GBLUP	0,63±0,02	1,00±0,08	0,13±0,02						
	6	Delta-p	0,59±0,03	0,70±0,08	0,33±0,07	0,80±0,03	1,32±0,35	0,68±0,12	0,67±0,06	0,80±0,04	1,16±0,06
		GBLUP	0,69±0,02	1,02±0,06	0,26±0,03						
	7	Delta-p	0,55±0,07	1,10±0,17	0,07±0,01	0,85±0,04	0,60±0,17	0,88±0,03	0,68±0,05	0,66±0,04	1,10±0,03
		GBLUP	0,60±0,04	1,02±0,25	0,12±0,01						
	8	Delta-p	0,57±0,05	0,89±0,11	0,15±0,04	0,81±0,02	0,68±0,27	0,85±0,08	0,64±0,04	0,71±0,05	1,09±0,04
		GBLUP	0,65±0,03	0,89±0,07	0,23±0,04						

Cenários com características controladas por genes de pequenos efeitos - Cenário 1 ($h^2_{aM} = 0{,}22$), Cenário 2 ($h^2_{aM} = 0{,}33$), Cenário 5 ($h^2_{aM} = 0{,}21$ e $h^2_{dM} = 0{,}10$), Cenário 6 ($h^2_{aM} = 0{,}35$ e $h^2_{dM} = 0{,}17$). Cenários com características controladas por genes de pequenos e grandes efeitos - Cenário 3 ($h^2_{aM} = 0{,}20$), Cenário 4 ($h^2_{aM} = 0{,}35$), Cenário 7 ($h^2_{aM} = 0{,}20$ e $h^2_{dM} = 0{,}13$), Cenário 8 ($h^2_{aM} = 0{,}33$ e $h^2_{dM} = 0{,}21$)

1964; Falconer, 1989) esse processo apresenta resultados superiores a vários outros métodos de seleção, como por exemplo, seleção individual ou seleção entre ou dentro de famílias. Da mesma forma, no presente trabalho, o índice do método Delta-p/G-BLUP que é feito com base no índice de seleção combinada obteve resultados superiores se comparado ao método G-BLUP e ao Delta-p. Ganhos de 5% em acurácia já são significativos no melhoramento, equivalendo muitas vezes ao ganho que se obtém em um ciclo completo de melhoramento (Resende et al., 2015). Sob seleção genômica realizada anualmente ou a cada curto espaço de tempo, esses ganhos obtidos são acumulativos e crescem rapidamente.

Considerando o modelo aditivo observou-se que para todos os cenários, o método G-BLUP foi o que apresentou estimativa de herdabilidade aditiva mais próxima da herdabilidade paramétrica (Tabela 2). Por sua vez, considerando o modelo aditivo-dominante, observou-se que para os cenários cujas características são controladas por genes de pequenos efeitos (Cenários 5 e 6), o método Delta-p foi o que apresentou estimativa de herdabilidade aditiva mais próxima da herdabilidade paramétrica, enquanto que para os cenários cujas as características são controladas por genes de pequenos e grandes efeitos (Cenários 7 e 8), o método G-BLUP obteve melhor desempenho.

Considerando o modelo aditivo, os métodos G-BLUP e Delta-p tiveram seus valores genômicos aditivos preditos não viesados (Tabela 2), exceto para o método Delta-p nos cenários 1 e 3, em que o método subestimou estes valores ($b_{y\hat{a}} > 1$). Considerando o modelo aditivo-dominante, o método G-BLUP teve seus valores genômicos aditivos preditos não viesados, exceto para o cenário 8, em que o método superestimou estes valores ($b_{y\hat{a}} < 1$). Enquanto, que o método Delta-p teve seus

valores genômicos aditivos preditos superestimados para os cenários 5, 6 e 8 e subestimados no cenário 7.

Na Tabela 3 são apresentados os resultados médios obtidos por meio dos métodos para os valores genômicos devido à dominância considerando o modelo aditivo-dominante e também os pesos, acurácia e eficiência obtida via índice Delta-p/G-BLUP.

Neste contexto, a melhoria na predição dos valores genômicos com a utilização do índice (Tabela 3) foi menos evidente, de apenas 0,01 unidades na acurácia de predição em relação à acurácia do G-BLUP. Hill et al. (2008), Bennewitz e Meuwissen (2010), Wellmann e Bennewitz (2012) discutem sobre a relevância da inclusão da dominância na Genômica Quantitativa. No entanto, conforme os resultados reportados por Azevedo et al. (2015) os valores genômicos devido à dominância são difíceis de serem preditos e consequentemente estão associados a baixos valores de acurácia e viés distantes de um. Este fato corrobora os baixos valores de acurácia encontrados para o índice Delta-p/G-BLUP e para o GBLUP. Devido a esta baixa magnitude, o índice produziu alta eficiência média (104% - 362%), mas que são valores pouco informativos, além disso, estão associadas a estimativas de desvios-padrão de alta magnitudes.

Apesar dos valores genômicos devido à dominância considerando o modelo aditivo-dominante terem sido estimados de forma inacurada, observa-se que as acúracias entre os métodos foram similares. Este resultado está de acordo com os resultados da literatura (Azevedo et al., 2015; Gianola, 2013; de los Campos et al., 2012), que apontam a semelhança de vários métodos em termos de acurácia no que se refere a predição de valores genômicos. Dessa forma, nota-se que os principais

Tabela 3: Médias e desvios-padrão de acurácia ($r_{\hat{a}d}$), viés ($b_{y\hat{a}}$), herdabilidade aditiva (h^2_{dM}), correlação entre os valores genômicos aditivos estimados via método Delta-p e via G-BLUP ($r_{d_p\hat{a}_g}$), proporção entre as variâncias aditivas (Δ^2) estimadas, pesos (b_1 e b_2) e acurácia (r_{Id}) do método índice Delta-p/G-BLUP, eficiência relativa (ER) entre o índice e o G-BLUP.

Modelo	Cenários	Método	$r_{\hat{a}d}$	$b_{y\hat{a}}$	h^2_{dM}	$r_{d_p\hat{a}_g}$	Δ^2	b_1	b_2	r_{Id}	ER
Aditivo-Dominante	5	Delta-p	0,13±0,04	1,28±0,31	0,13±0,04	0,67±0,11	0,20±0,08	1,00±0,00	0,96±0,01	0,21±0,02	1,04±0,04
		GBLUP	0,20±0,02	0,84±0,13	0,20±0,02						
	6	Delta-p	0,16±0,02	1,18±0,24	0,03±0,01	0,66±0,06	0,18±0,09	1,00±0,00	0,94±0,01	0,25±0,01	1,04±0,03
		GBLUP	0,24±0,01	0,62±0,11	0,17±0,03						
	7	Delta-p	0,02±0,08	0,95±0,18	0,03±0,01	0,60±0,12	0,39±0,12	1,00±0,01	1,00±0,00	0,07±0,06	3,62±8,84
		GBLUP	0,05±0,07	0,71±0,42	0,15±0,21						
	8	Delta-p	0,01±0,05	0,89±0,16	0,06±0,02	0,48±0,12	0,35±0,14	1,00±0,00	1,00±0,00	0,07±0,02	1,48±1,27
		GBLUP	0,06±0,03	0,50±0,11	0,19±0,04						

Cenários com características controladas por genes de pequenos efeitos - Cenário 5 ($h^2_{aM} = 0{,}21$ e $h^2_{dM} = 0{,}10$), Cenário 6 ($h^2_{aM} = 0{,}35$ e $h^2_{dM} = 0{,}17$). Cenários com características controladas por genes de pequenos e grandes efeitos Cenário 7 ($h^2_{aM} = 0{,}20$ e $h^2_{dM} = 0{,}13$), Cenário 8 ($h^2_{aM} = 0{,}33$ e $h^2_{dM} = 0{,}21$).

critérios para comparação dos métodos nestes cenários serão as herdabilidades devido à dominância e o viés de predição.

Analisando as herdabilidades devido à dominância observa-se que para o cenário 5, o método Delta-p foi o que apresentou estimativa de herdabilidade mais próxima da herdabilidade paramétrica. Enquanto que para os cenários 6, 7 e 8, as estimativas mais próximas foram as do método G-BLUP. No entanto, nos cenários 5, 7 e 8 as herdabilidades devido à dominância foram ligeiramente superestimadas. Azevedo et al. (2015) também verificou resultados semelhantes para os métodos bayesianos considerados.

O método Delta-p subestimou os valores genômicos devido à dominância nos cenários cujas características eram controladas por genes de pequenos efeitos (Cenários 5 e 6) e superestimou nos cenários cujas características eram controladas por genes de pequenos e grandes efeitos. O GBLUP superestimou os valores em todos os cenários, sendo os vieses consideravelmente menores aos provenientes do método Delta-p.

Na Tabela 4 estão apresentados os resultados obtidos de herdabilidades, acurácias e viés para os métodos Bayes Híbrido e LASSO Bayesiano considerando os cenários 1 e 5. Nota-se que nos dois cenários avaliados, os métodos apresentaram valores de herdabilidades e acurácias similares, sendo as herdabilidades muito próximas aos valores paramétricos. Isso ocorre, pois, o LASSO Bayesiano é assintoticamente livre de informação a *priori*, ou seja, proporciona uma melhor aprendizagem por meio do conjunto de dados (Gianola, 2013; Gianola et al., 2009). Entretanto, os resultados encontrados para os vieses mostraram uma ligeira diferença, sendo que o método Bayes Híbrido foi o que obteve viés mais próximo de um, exceto para o cenário 1 com 8 graus

de liberdade. Assim, ao se utilizar uma *priori* altamente informativa ocasiona uma melhoria na predição dos valores genômicos aditivos em relação ao LASSO Bayesiano.

Tabela 4: Graus de Liberdade da distribuição *a priori* qui-quadrado para variância genética (*gl*), Herdabilidade aditiva (h^2_{aM}), Acurácia ($r_{\hat{a}a}$) e Viés ($b_{y\hat{a}}$) com respectivos desvios-padrão dos valores genômicos aditivos estimados via método LASSO Bayesiano (BLASSO) e Bayes Híbrido considerando os cenários 1 e 5.

Cenários	*Gl*	Bayes Híbrido			LASSO Bayesiano		
		h^2_{aM}	$r_{\hat{a}a}$	$b_{y\hat{a}}$	h^2_{aM}	$r_{\hat{a}a}$	$b_{y\hat{a}}$
Cenário 1	8	0,23±0,03	0,59±0,02	0,76±0,09	0,22±0,02	0,60±0,02	0,77±0,07
	-2	0,22±0,02	0,61±0,02	0,79±0,03	0,24±0,03	0,60±0,02	0,75±0,06
Cenário 5	8	0,17±0,04	0,48±0,06	0,68±0,14	0,18±0,04	0,47±0,05	0,65±0,11
	-2	0,17±0,03	0,48±0,04	0,67±0,07	0,19±0,02	0,47±0,04	0,63±0,04

Cenário 1: características controladas por genes de pequenos efeitos - $h^2_{aM} = 0{,}22$; Cenário 5: características controladas por genes de pequenos efeitos - $h^2_{aM} = 0{,}21$ e $h^2_{dM} = 0{,}10$).

4. CONCLUSÕES

O método proposto, índice Delta-p/G-BLUP, é de fácil implementação no contexto da Seleção Genômica, demanda custo computacional reduzido além de conduzir a valores genômicos, aditivos e devido à dominância, mais acurados do que o método G-BLUP. O método Bayes Hibrido mostrou-se igualmente acurado ao LASSO Bayesiano, além de conduzir a estimativas de valores genômicos com viés mais próximo de um. Já o método Delta-p proporcionou menores valores de acurácias e viés mais distantes de um do que o método G-BLUP.

5. REFERÊNCIAS

AZEVEDO, C.F.; RESENDE, M.D.V.; SILVA, F.F.; VIANA, J.M.S.; VALENTE, M.S.F.; RESENDE JUNIOR, M.F.R.; MUÑOZ, P. Ridge, LASSO And Bayesian Additive-Dominance Genomic Models. **BMC Genetics**, v.16, p.105, 2015.

BENNEWITZ, J; MEUWISSEN, T.H.E. The distribution of QTL additive and dominance effects in porcine F2 crosses. **Journal of Animal Breeding and Genetics**, v. 127, n. 3, p. 171-179, 2010.

DA, Y.; WANG, C.; WANG, S.; HU, G. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. **PLoS ONE**, v. 9, n. 1, p. e87666, 2014.

DE LOS CAMPOS, G. et al. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. **Genetics**, v.182, n 1, p.375-385, 2009.

DE LOS CAMPOS, G.; HICKEY, J.M.; PONG-WONG, R.; DAETWYLER, H.D.; CALLUS, M.P.L. Whole genome regression and prediction methods applied to plant and animal breeding. **Genetics,** v.193, p. 327–45, 2012.

FALCONER, D. S. **Introduction to Quantitative Genetics**. Longmans, New York, p.438, 1989.

GEWEKE, J. **Evaluating the Accuracy of Sampling-Based Approaches to the Calculation of Posterior Moments.** In Bayesian Statistics 4 (eds. J.M. Bernardo, J. Berger, A.P. Dawid and A.F.M. Smith), Oxford: Oxford University Press, 169-193, 1992.

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

GIANOLA, D.; PEREZ‐ENCISO, M.; TORO, M.A. On marker‐assisted prediction of genetic value: beyond the ridge. **Genetics**, v.163, p.347‐365, 2003.

GIANOLA, D. Priors in whole-genome regression: the bayesian alphabet returns. **Genetics**, v. 194, n.3, p.573–96, 2013.

GODDARD, M. E.; Genomic selection: prediction of accuracy and maximization of long term response. **Genetics**, Dordrecht, v. 136, n. 2, p. 345-357, 2009.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of Genetic Relationship on Genome-Assisted Breeding Values. **Genetics**, v.117, p.2389-2397, 2007.

HILL, W.G.; GODDARD, M.E.; VISSCHER, P.M. Data and theory point to mainly additive genetic variance for complex traits. **PLOS Genetics**, v.4, n.2, p. e1000008, 2008.

LUSH, J.L. **Animal breeding plans.** 3.ed. Ames: State College Press, p.443, 1945.

LUSH, J.L. Family merit and individual merit as basis for selection. **America Naturalist**, v.81, p. 241-261, 1947.

LUSH, J. L. Melhoramento dos animais domésticos. **Rio de Janeiro: CEDEGRA**, p. 566, 1964.

MEUWISSEN, TH.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819–29, 2001.

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

R Development Core Team. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria, Available: http://www.R-project.org, 2010.

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

RESENDE, M. D. V.; SILVA, F. F.; LOPES, P. S.; AZEVEDO, C. F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência**

Bayesiana (MCMC), Regressão Aleatória Multivariada (RRM) e Estatística Espacial. Viçosa: Universidade Federal de Viçosa/Departamento de Estatística 291 p, 2012. Disponível em: <http://www.det.ufv.br/ppestbio/corpo_docente.php>.

RESENDE, M.D.V.; RESENDE JUNIOR, M.F.R.; AGUIAR, A.M.; ABAD, J.I.M.; MISSIAGIA, A.A.; SANSALONI, C.P.; PETROLI, C.D.; GRATTAPALIA, D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, p.79, 2010.

RESENDE, M.D.V. (ORG.); SILVA, F.F (ORG.); AZEVEDO, C.F. (ORG.). **Estatística Matemática, Biométrica e Computacional**: **Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência.** 1. ed. Visconde do Rio Branco: Suprema, v.1, p.881, 2014.

RESENDE, M.D.V.; RAMALHO, M. A. P.; GUILHERME, S. R.; ABREU, A. F. B. Multi generation index in the within progenies bulk method for breeding of self-pollinated plants. **Crop Science**, v. 55, p. 1202-1211, 2015.

RESENDE, M. D. V. **Genética Quantitativa e de Populações**. 1. ed. Visconde do Rio Branco: Suprema, v. 1, p.422, 2015.

VAN RADEN, P. M. Efficient Methods to compute genomic predictions. **Journal of Dairy Science,** Champaign, v.91, n.11, p. 4414-4423, 2008.

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

VIANA, J. M. S. **Programa para análises de dados moleculares e quantitativos Real Breeding**. Viçosa: UFV, 2011.

WANG, C.; DA, Y. Quantitative Genetics Model as the Unifying Model for Defining Genomic Relationship and Inbreeding Coefficient. **PLoS ONE**, v. 9, n. 12, p. e114484, 2014.

WELLER, J. I.; GLICK. G.; EZRA, E.; SEROUSSI, E.; SHEMESH, M.; ZERON, Y.; RON, M. Predictive ability of selected subsets of single nucleotide polymorphisms (SNPs) in a moderately sized dairy cattle population. **Animal**, v.8:2, p.208–216, 2014a.

WELLER, J. I.; EZRA, E.; SEROUSSI, E.; SHEMESH, M.; RON, M. Improving predictive ability of selected subsets of single nucleotide polymorphisms in a moderately sized dairy cattle population. **Proceedings, 10th World Congress of Genetics Applied to Livestock Production**, 2014b.

WELLER, J.I. **Genomic selection in animals**. Wiley, p.175, 2016.

WELLMANN, R; BENNEWITZ, J. Bayesian models with dominance effects for genomic evaluation of quantitative traits. **Genetics Research**, v. 94, p. 21-37, 2012.

CAPÍTULO 3
MÉTODO DA REGRESSÃO CATEGÓRICA TRIPLA (TCR) APLICADA A SELEÇÃO GENÔMICA

Resumo

A seleção genômica (GWS) é uma técnica de grande importância no melhoramento animal e vegetal, permitindo a eficiência na avaliação genética e na predição de ganhos genéticos. Esta metodologia baseia-se em valores genômicos estimados via valores fenotípicos e um grande número (n) de marcadores amplamente distribuídos no genoma. Os valores genéticos genômicos (GEBVs) de N indivíduos são estimados via modelos funcionais apropriados, que estimam o efeito de cada marcador em fenótipos, permitindo a identificação precoce dos indivíduos geneticamente superiores. No entanto, a predição genômica conduz a desafios estatísticos, como a estimabilidade, devido ao problema da alta dimensionalidade ($N<<n$), e multicolinearidade entre as covariáveis, uma vez que os marcadores moleculares são altamente correlacionados. Estes desafios exigem a utilização de métodos estatísticos para considerar a regularização no processo de estimação. Assim, o objetivo desse trabalho foi avaliar uma metodologia, denominada Regressão Categórica Tripla (TCR), além de compará-la com os métodos G-BLUP (*Genomic Best Linear Unbiased Predictor*) e BLASSO (*Genomic Best Linear Unbiased Predictor Bayesian Least Absolute Shrinkage and Selection Operator*), métodos que vêm sendo aplicados a GWS. A eficiência dos métodos propostos, no que se refere à estimação dos GEBVs, foi avaliada em populações simuladas considerando quatro cenários diferentes (modelo infinitesimal × dois níveis de herdabilidade × ausência de dominância e dominância completa). Também neste trabalho, com intuito de tornar os valores genômicos via G-BLUP mais acurados, foi proposta melhoria ao método G-BLUP por meio da herdabilidade estimada via TCR (TCR/G-BLUP). O método foi avaliado em dados reais (características de

mandioca) visando elucidar a importância do método nos programas de melhoramento. Os métodos foram comparados por meio de validação independente e considerando medidas de eficiência, tais como acurácia da predição, viés e herdabilidade genômica. Os resultados indicaram que o método TCR mostrou-se adequado para a estimação dos componentes da variação genotípica e da herdabilidade e o método TCR/G-BLUP mostrou-se eficiente para predição dos valores genômicos, podendo ser usados vantajosamente.

Palavras-chaves: Predição genômica, G-BLUP, BLASSO, melhoramento genético de mandioca, herdabilidade genômica.

1. INTRODUÇÃO

A Seleção Genômica Ampla (*Genome Wide Selection* – GWS), proposta por Meuwissen et al. (2001), consiste na 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 (*Linkage Disequilibrium* - LD) com pelo menos uma parte dos marcadores. Esses marcadores em desequilíbrio de ligação com os QTLs (*Quantitative Trait Loci*), tanto de grandes quanto de pequenos efeitos, explicarão quase a totalidade da variação genética de um caráter quantitativo.

Conforme Meuwissen et al. (2001), é possível supor que alguns marcadores moleculares estão em desequilíbrio de ligação com locos de características quantitativas, permitindo a sua utilização direta na predição dos valores genéticos genômicos (*Genomic Estimated Breeding Values* - GEBVs) dos indivíduos sujeitos a seleção, incluindo os indivíduos que ainda não tenham seus fenótipos avaliados. No entanto, a aplicação prática destas informações genômicas é um desafio, pois geralmente não é possível a utilização

adequada de métodos tradicionais baseados em quadrados mínimos (*Least Squares* – LS) para estimar o efeito de cada SNP no fenótipo.

No contexto da predição genômica, o problema principal é a estimação de um grande número de efeitos de marcadores (*n*) a partir de um limitado número de indivíduos fenotipados e genotipados (*N*) e também as colinearidades advindas do LD entre os marcadores (Gianola et al., 2003). Estes problemas estatísticos conduzem a uma escassez de graus de liberdade para estimar os efeitos de todos os marcadores e à uma instabilidade das estimativas. Uma solução para contornar essa questão é o ajuste destes efeitos como aleatórios ou sob o enfoque bayesiano, como realizado pelo G-BLUP (Habier et al., 2007; Van Raden, 2008; Goddard, 2009), RR-BLUP (Whittaker et al., 2000; Meuwissen et al,. 2001) e LASSO Bayesiano (de los Campos et al., 2009a), que não consomem graus de liberdade e contornam a multicolinearidade. Esses métodos estimam simultaneamente *n* efeitos baseados em *N* observações, sendo $n >>$ N, porém é importante ressaltar que quanto menor a razão *n*/*N*, mais favorável é a condição estatística além de fornecer estimativas mais acuradas.

Resende et al. (2014) introduziram um novo e simples método, chamado Regressão Categórica Tripla (TCR), para estimação da herdabilidade genômica em uma população totalmente não aparentada e não estruturada. No entanto, o método TCR pode ser combinado com a metodologia do método Delta-p e com isso dar origem a um procedimento mais eficiente. Pelo procedimento do TCR aliado ao método Delta-p, ao invés de regressar os *N* fenótipos em *n* locos marcadores (*n*/*N*), o método TCR regressa fenótipos nas três categorias de genótipos marcadores ($3/N << n/N$), MM, Mm e mm, visando capturar os efeitos genéticos em um loco b com categorias genotípicas BB, Bb e bb, em que B é o alelo favorável. Os métodos tradicionais regressam, através de todos os locos, os fenótipos no número de M em cada loco marcador. O TCR regressa no número

total de indivíduos em cada categoria genotípica. Isto é coerente com a filosofia do modelo genético infinitesimal (caracteres governados por muitos genes de pequenos efeitos e sem locos de grandes efeitos) e então, com a filosofia do G-BLUP e RR-BLUP. Além disso, o TCR é computacionalmente vantajoso e pode ser um melhor estimador dos componentes da variação genotípica e da herdabilidade.

Dessa forma, o presente trabalho tem como um dos objetivos avaliar a eficiência do método TCR e compará-lo aos métodos G-BLUP e BLASSO em populações simuladas considerando quatro cenários diferentes (modelo infinitesimal × dois níveis de herdabilidade × ausência de dominância e dominância completa). Ademais, com intuito de tornar os valores genômicos via G-BLUP mais acurados, foi proposta, neste trabalho, uma melhoria ao método G-BLUP por meio da herdabilidade estimada via TCR. Esta metodologia, denominada TCR/G-BLUP, consiste em estimar a herdabilidade via TCR e fixá-la no método G-BLUP. A eficiência dos métodos G-BLUP e TCR/G-BLUP foram comparados utilizando seis caracteristicas avaliadas em mandioca (*Manihot esculenta*).

2. MATERIAIS E MÉTODOS

2.1.Dados Simulados

O conjunto de dados foi simulado usando o *software Real Breeding* (Viana, 2011) e sua geração foi descrita por Azevedo et al. (2015). Um total de 2000 marcadores SNPs equidistantes separados por 0,1 centiMorgan (cM) entre os dez cromossomos foram simulados. Os QTLs foram distribuídos nas regiões abrangidas pelo SNPs. Um montante de 1000 indivíduos de 20 famílias de irmãos completos foram genotipados e fenotipados.

As características foram simuladas considerando a arquitetura genética seguindo um modelo infinitesimal, ou seja, locos não ligados com efeitos iguais. Cada um dos 100 QTLs foi atribuído um efeito aditivo de pequena magnitude no fenótipo (sob a definição

de Distribuição Normal). Os efeitos foram normalmente distribuídos com média zero e variância genética permitindo o nível de herdabilidade desejado. O valor fenotípico foi obtido adicionando ao valor genotípico um efeito ambiental proveniente de uma distribuição normal $N(0, \sigma_e^2)$, em que a variação σ_e^2 foi definida de acordo com dois níveis de herdabilidade no sentido restrito em torno de 0,20 e 0,35, respectivamente. Os níveis de herdabilidade foram escolhidos para representar uma característica com baixa herdabilidade e outra com herdabilidade moderada, casos em que se espera que a seleção genômica seja superior à seleção fenotípica (Azevedo et al., 2015).

As magnitudes das herdabilidades no sentido restrito e no sentido amplo estão associadas com um grau médio de nível de dominância *(d/a)* de aproximadamente 1 (domínio completo) e 0 (ausência de dominância) em uma população com frequências alélicas intermediárias. As simulações assumiram independência entre efeitos aditivos e devido à dominância, quando este era considerado, com efeitos de dominância tendo a mesma distribuição que os efeitos aditivos (ambos foram normalmente distribuídos com média zero). Marcadores com MAF (*Minor Allele Frequency* – Frequência do Menor Alelo) menor do que 5% foram excluídos das análises.

2.2.Cenários

Quatro cenários diferentes foram simulados e utilizados nas análises: dois níveis de herdabilidade (cerca de 0,30 e 0,50, associados à herdabilidades em sentido restrito de 0,20 e 0,35, respectivamente) × ausência de dominância e dominância completa.

A descrição dos cenários é apresentada na Tabela 1.

Tabela 1 – Cenários com as respectivas médias das herdabilidades aditivas (h_a^2), devido a dominância (h_d^2) e total (h_g^2), arquiteturas genéticas (características controladas por genes de pequeno efeito – herança poligênica) e níveis de dominância (ausência de dominância e dominância completa).

Cenário	Arquitetura Genética	Nível de dominância	h_a^2	h_d^2	h_g^2
Cenário 1	Herança poligênica	Ausência	0,22	-	0,22
Cenário 2	Herança poligênica	Ausência	0,33	-	0,33
Cenário 3	Herança poligênica	Completa	0,21	0,10	0,31
Cenário 4	Herança poligênica	Completa	0,35	0,17	0,52

Estes quatro cenários foram analisados considerando-se os métodos, TCR, G-BLUP e BLASSO. Cada tipo de população (ou cenário) foi simulado 10 vezes. Nove repetições foram utilizadas como populações de treinamento, e uma repetição foi utilizada como população de validação. As estimativas baseadas em cada uma das nove repetições foram validadas para obtenção das estimativas de acurácia, de viés e de herdabilidades genômicas. Assim, essas medidas foram calculadas em cada repetição da simulação e depois foi calculada a média desses valores.

2.3.Dados Reais

A seleção genômica foi realizada para seis caracteristicas avaliadas em mandioca (*Manihot esculenta*). O experimento foi instalado segundo um delineamento em blocos casualizados com três repetições e 10 plantas por parcela, incluindo 358 acessos de mandioca pertencentes à coleção de germoplasma da Embrapa, que foram genotipados para 390 marcadores moleculares SNPs. O experimento foi estabelecido em Cruz das Almas, Bahia, Brasil, sob as diretrizes da Embrapa. As características avaliadas foram: peso da parte área (PPA), produtividade total de raízes (PTR), teor de amilose (AML), teor de amido (AMD), teor de compostos cianogênicos (HCN) e produtividade

de amido (PROD-AMD). Maiores detalhes do experimento podem ser encontrado em Oliveira et al. (2012).

2.4.Regressão Categórica Tripla

Considerando o procedimento de Regressão Categórica Tripla (TCR), a população de estimação inicialmente é dividida em duas subpopulações, uma com os indivíduos ou famílias acima da média geral (subpopulação1, com valor fenotípico médio u_1 superior) e outra com os indivíduos ou famílias abaixo da média geral (subpopulação 2, com valor fenotípico médio u_2 inferior). A diferença $(u_1 - u_2)$ entre os valores fenotípicos médios das duas subpopulações é devida à maior frequência alélica (p) dos alelos favoráveis (e menor frequência dos alelos desfavoráveis) na subpopulação 1 em relação à subpopulação 2. Assim, $(u_1 - u_2)$ é explicada por $(\Delta p = p_1 - p_2)$, sendo delta-p (Δp) a diferença de frequências alélicas p_1 e p_2 entre as duas subpopulações. Os valores de Δp são calculados para cada loco marcador e aqueles com sinais positivos são alocados como favoráveis (do tipo B), ou seja, seus efeitos genéticos aditivos latentes ou efeitos de substituição alélica (α_i) são tomados como positivos. Da mesma forma, aqueles com sinais negativos de Δp tem seus (α_i) atribuídos como negativos. Assim, redefine-se a codificação da matriz de incidência dos marcadores W, compatibilizando-se o arquivo de marcas formado por 0 (mm), 1 (Mm) e 2 (MM) de forma a se ter um arquivo de "genes" dado por 0 (bb), 1 (Bb) e 2 (BB), sendo que a alocação em BB ou bb é ditada pelo sinal de α_i. Logicamente, o acerto na alocação em BB ou bb é probabilístico. Em média (esperança matemática) haverá acerto na maioria dos locos, sendo que o maior número de erros será naqueles locos marcadores de efeitos muito pequenos, tendendo a zero. A abordagem não demanda método computacional iterativo e faz uso apenas do conceito de distância genética (sinal de Δp_i) associada as duas subpopulações.

O algoritmo completo do método está descrito a seguir:

(i) subdivisão da população de treinamento em duas, de acordo com o fenótipo corrigido para efeitos ambientais;

(ii) cálculo de Δp_i;

(iii) se o sinal de Δp_i for negativo, trocar 0 por 2 e 2 por 0 em cada coluna de marcador com Δp_i negativo;

(iv) determinar a quantidade (n_{BB}) do código 2 na linha correspondente a cada indivíduo j do arquivo de marcas e fazer o mesmo para os códigos 1 e 0, obtendo n_{Bb} e n_{bb};

(v) o modelo de regressão categórica tripla é definido como:

$$y = 1\mu + \beta_{BB} n_{BB} I_{(BB)} + \beta_{Bb} n_{Bb} I_{(Bb)} + \beta_{bb} n_{bb} I_{(bb)} + e,$$

em que $I_{(BB)}$, $I_{(Bb)}$ e $I_{(bb)}$ são variáveis indicadoras. Se a categoria analisada é BB então $I_{(BB)} = 1$ e $I_{(Bb)} = I_{(bb)} = 0$. Analogamente, o mesmo pode ser definido para as demais categorias genotípicas.

Dessa forma, a estimação dos coeficientes de regressão ($\hat{\beta}$) estimados via método dos mínimos quadrados ordinários, que se refere ao valor genético global de cada categoria genotípica, é dada por:

$$\hat{\beta}_{BB} = Cov(y_j, n_{BB-j})/Var(n_{BB-j})$$

$$\hat{\beta}_{Bb} = Cov(y_j, n_{Bb-j})/Var(n_{Bb-j})$$

$$\hat{\beta}_{bb} = Cov(y_j, n_{bb-j})/Var(n_{bb-j});$$

(vi) obtenção dos valores genotípicos ($\hat{u}_{BB-kj}, \hat{u}_{Bb-kj}$ e $\hat{u}_{bb-kj}$) por categoria genotípica dos marcadores, na soma de todos os k locos em cada indivíduo j, conforme segue, pela regressão (através de $\hat{\beta}$) dos fenótipos no número n de cada categoria:

$\hat{u}_{BB-kj} = \hat{\beta}_{BB} n_{BB-j} = 2\alpha_{B-kj} + \delta_{BB-kj}$: valor genotípico total da categoria BB nos n_{BB} locos no indivíduo j;

$\hat{u}_{Bb-kj} = \hat{\beta}_{Bb} n_{Bb-j} = \alpha_{B-kj} + \alpha_{b-kj} + \delta_{Bb-kj}$: valor genotípico total da categoria Bb nos n_{Bb} locos no indivíduo j;

$\hat{u}_{bb-kj} = \hat{\beta}_{bb} n_{bb-j} = 2\alpha_{b-kj} + \delta_{bb-kj}$: valor genotípico total da categoria bb nos n_{bb} locos no indivíduo j, sendo $\delta_{BB} = -2q^2d$, $\delta_{bb} = -2p^2d$ e $\alpha_k = \alpha_{Bk} - \alpha_{bk}$ (Falconer, 1989);

(vii) disposição dos valores genotípicos totais de cada indivíduo em um vetor:

(viii) cômputo das variâncias genéticas conforme detalhado mais adiante;

(ix) estimação das herdabilidades dadas por: $\hat{h}_a^2 = \hat{\sigma}_{u_{ajs}}^2/\sigma_y^2$ e $\hat{h}_d^2 = \hat{\sigma}_{u_{djs}}^2/\sigma_y^2$, em que σ_y^2 é a variância entre os valores fenotípicos individuais.

A composição dos genótipos em termos de suas frequências, efeitos aditivos e de dominância e variâncias encontra-se no Quadro 1 a seguir. Essas informações foram usadas para compor os estimadores das variâncias genéticas pelo método TCR.

Quadro 1. Efeitos e variâncias genéticas paramétricas (teóricas).

Genótipo	**Frequências**	**Valor Genotípico**	**Efeito Aditivo**	**Efeito de dominância**
BB	p^2	a	$2\alpha_B = 2q\alpha$	$\delta_{BB} = -2q^2d$
Bb	$2pq$	d	$\alpha_B + \alpha_b = (q-p)\alpha$	$\delta_{Bb} = 2pqd$
bb	q^2	$-a$	$2\alpha_b = -2p\alpha$	$\delta_{bb} = -2p^2d$
Genótipo	**Frequências**	**Variância Aditiva**		**Variância de Dominância**
BB	p^2	$p^2(2\alpha_B)^2 = p^2(2q\alpha)^2$		$p^2(-2q^2d)^2$
Bb	$2pq$	$2pq(\alpha_B + \alpha_b)^2 = 2pq[(q-p)\alpha]^2$		$2pq(2pqd)^2$
bb	q^2	$q^2(2\alpha_b)^2 = q^2(-2p\alpha)^2$		$q^2(-2p^2d)^2$
Soma		$\sigma_{u_a}^2 = 2pq\alpha^2$		$\sigma_{u_d}^2 = (2pqd)^2$

Estimadores dos efeitos genéticos

Sendo a e $-a$ os valores genotípicos (Quadro 1) de BB e bb relacionados aos efeitos aditivos, tem-se que a soma $\hat{\mu}_{aj} = f(\hat{\alpha}_k) = \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ fornece uma estimativa dos efeitos aditivos dos indivíduos. Esses podem ser usados para o cômputo da acurácia seletiva e do viés da predição.

Sendo d o valor genotípico do heterozigoto (Quadro 1) Bb relacionado ao efeito de dominância, tem-se que $\hat{\mu}_{dj} = \hat{u}_{Bb-kj}$ fornece uma estimativa dos efeitos de dominância dos indivíduos. Esses podem ser usados para o cômputo da acurácia seletiva e do viés da predição dos efeitos de dominância. Com p tendendo a q, ou seja, $p \approx q \approx 0{,}50$, a quantidade $\hat{\mu}_{dj} = \hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ também é um estimador desses efeitos.

Estimadores das variâncias genéticas

Variância aditiva

Conforme o Quadro 1, $\sigma^2_{u_a} = 2pq\alpha^2$ e sendo $\hat{\mu}_{aj} = f(\hat{\alpha}_k) = \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$, tem-se que $\sigma^2_{u_a} = 2pqf(\hat{\alpha}_k) = 2pqVar(\hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ é um estimador para a variância genética aditiva, em que $\hat{\alpha}_k$ é um estimador intrínseco para o efeito de substituição alélica nos k locos.

Variância de dominância

Também conforme o Quadro 1, $\sigma^2_{u_d} = (2pqd)^2$. O contraste $2\hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ fornece uma estimativa de d e, portanto, $\sigma^2_{u_d} = (2pq)^2 Var(2\hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ é um estimador para a variância genética de dominância. Com $p \approx q \approx 0{,}5$, a quantidade $\sigma^2_{u_d} = (2pq)^2 4Var(\hat{u}_{Bb-kj} - \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$ também é um estimador para $\sigma^2_{u_d}$.

Variância genotípica total

A variância do somatório $\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj}$ fornece informação sobre a variância genotípica total como função $f(p, d, \alpha)$ de p, d e α. Assim, $Var\left(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj}\right) = f(p, d, \alpha)$ e as variâncias genéticas aditiva e de dominância podem ser extraídas de $f(p, d, \alpha)$ via $\sigma^2_{u_a} = 2pqVar\left(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj}\right)$ e $\sigma^2_{u_d} = (2pq)^2 Var(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$, respectivamente**.** Assim, a variância genotípica total é dada por $\sigma^2_{u_g} = [2pq + (2pq)^2] Var(\hat{u}_{Bb-kj} + \hat{u}_{BB-kj} + \hat{u}_{bb-kj})$.

2.5.Método G-BLUP

A predição dos efeitos genômicos aditivos via G-BLUP (*Genomic Best Linear Unbiased Predictor)*, usando as informações fenotípicas e genotípicas para cada indivíduo, é feita por meio do seguinte modelo linear misto (Resende, 2008; Resende et al., 2010):

$$y = 1\mu + Za + Zd + e,$$

em que, y é o vetor de fenótipos ($N \times 1$, em que N é o número de indivíduos genotipados e fenotipados); μ é a média geral e 1 é o vetor com dimensão $N \times 1$ cujos seus elementos são iguais a 1; a é o vetor de efeitos genômicos aditivos dos indivíduos ($N \times 1$) com matriz de incidência Z ($N \times N$), sendo a estrutura de variância dada por $a \sim N(0, G_a\sigma^2_a)$ em que σ^2_a é a variância aditiva e $G_a (N \times N)$ é a matriz de parentesco genômica para os efeitos aditivos; d é o vetor de efeitos genômicos devido à dominância dos indivíduos com matriz de incidência Z ($N \times N$), sendo a estrutura de variância dada por $d \sim N(0, G_d\sigma^2_d)$ em que σ^2_d é a variância devido à dominância e G_d ($N \times N$) é a matriz de parentesco genômica para os efeitos devido à dominância.

As equações de modelos mistos para predição de a e d via o método G-BLUP equivalem a:

$$\begin{bmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + G_a^{-1}\frac{\sigma_e^2}{\sigma_a^2} & Z'Z \\ Z'X & Z'Z & Z'Z + G_d^{-1}\frac{\sigma_e^2}{\sigma_d^2} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ Z'y \end{bmatrix},$$

sendo os componentes de variância, σ_d^2, σ_a^2 e σ_e^2, estimados via REML (*Restricted maximum likelihood*).

Conforme Vitezica et al. (2013), as matrizes de parentesco genômicas para efeitos aditivos e para efeitos devido à dominância, G_a e G_d, são dadas respectivamente por:

$$G_a = \frac{WW'}{\sum_{i=1}^{n}(2p_iq_i)} \quad e \quad G_d = \frac{SS'}{\sum_{i=1}^{n}(2p_iq_i)^2},$$

em que p_i e q_i as frequências alélicas do locos i, W é a matriz de incidência para os vetores de efeitos aditivos de marcadores (α) e S é a matriz de incidência para os vetores de efeitos devido à dominância de marcadores (δ) parametrizadas conforme Van Raden (2008), Vitezica et al. (2013), Wang e Da (2014), Da et al. (2014) e Resende et al. (2014) e que é apresentada a seguir:

$$W = \begin{cases} Se\ MM, então\ 2 - 2p \rightarrow 2q \\ Se\ Mm, então\ 1 - 2p \rightarrow q - p \\ Se\ mm, então\ 0 - 2p \rightarrow -2p \end{cases} \qquad S = \begin{cases} Se\ MM, então\ 0 \rightarrow 2q^2 \\ Se\ Mm, então\ 1 \rightarrow 2pq \\ Se\ mm, então\ 0 \rightarrow -2p^2 \end{cases}.$$

2.6.Método TCR/G-BLUP

Com intuito de tornar os valores genômicos via G-BLUP mais acurados, foi proposta uma melhoria ao método G-BLUP por meio da utilização da herdabilidade estimada via TCR, caracterizando-se no método TCR/G-BLUP. Neste método, adotou-se a estratégia de fixar a herdabilidade estimada pelo método TCR nas equações de modelos mistos genômicas do G-BLUP.

A eficiência dos métodos G-BLUP e TCR/G-BLUP foram comparadas utilizando seis caracteristicas avaliadas em mandioca (*Manihot esculenta*). O experimento foi instalado segundo um delineamento em blocos casualizados com três repetições e 10 plantas por parcela.

2.7.Método BLASSO

A versão bayesiana da regressão via LASSO (BLASSO - Park e Casella, 2008) para seleção genômica ampla foi proposta por de los Campos et al. (2009a). O BLASSO (*Bayesian Least Absolute Shrinkage and Selection Operator*) inclui um termo de variância comum para os efeitos genéticos de marcadores e de efeitos residuais. Portanto, usando o modelo linear básico para predição dos efeitos dos marcadores, $y = Xb + Wm_a + Sm_d + e$, sendo, y é o vetor de fenótipos, b é o vetor de efeitos sistemáticos com matriz de incidência X, m_a é o vetor de efeitos genéticos aditivos dos marcadores com matriz de incidência W, m_d é o vetor de efeitos genéticos devido à dominância dos marcadores com matriz de incidência S e e é o vetor de resíduos. As distribuições *a priori* dos parâmetros do modelo são apresentadas a seguir:

$$e|\sigma^2 \sim MVN(0, I\sigma^2)$$

$$m_{ai}|\lambda_a, \sigma^2 \sim \prod_i \left(\frac{\lambda_a}{2\sigma}\right) e^{\left[\frac{-\lambda_a|m_{ai}|}{\sigma}\right]}$$

$$m_{di}|\lambda_d, \sigma^2 \sim \prod_i \left(\frac{\lambda_d}{2\sigma}\right) e^{\left[\frac{-\lambda_d|m_{di}|}{\sigma}\right]}$$

em que MNV representa a distribuição normal multivariada, λ_a e λ_d são parâmetros de "suavização" e podem ser estimados por meio do conjunto de dados via método MCMC usando uma *priori* não informativa, σ^2 tem como distribuição a *priori* uma qui-quadrado invertida escalada.

Usando uma formulação em termos de um modelo hierárquico aumentado, tem-se:

$$m_{ai}|\tau_a \sim N(0, D_a\sigma^2),\ m_{di}|\tau_d \sim N(0, D_d\sigma^2)$$

$$p(\tau_a^2\,|\lambda_a^2) = \prod_i \left(\frac{\lambda_a^2}{2}\right) e^{\left[\frac{-\lambda_a^2\tau_{ai}^2}{2}\right]} \text{ e } p(\tau_d^2\,|\lambda_d^2) = \prod_i \left(\frac{\lambda_d^2}{2}\right) e^{\left[\frac{-\lambda_d^2\tau_{di}^2}{2}\right]}$$

em que $D_a = diag\left(\tau_{1a}^2, \tau_{2a}^2, \dots, \tau_{ma}^2\right)$ e $D_d = diag\left(\tau_{1d}^2, \tau_{2d}^2, \dots, \tau_{md}^2\right)$. Isso leva a uma distribuição exponencial dupla para os efeitos de marcadores (Park e Casella, 2008), como a seguir:

$$m_{ai}|\lambda_a^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda_a}\right) \text{ e } m_{di}|\lambda_d^2 \sim ExpDupla\left(0, \frac{\sigma}{\lambda_d}\right).$$

A variância genética aditiva de cada loco marcador é dada por $\sigma_{mai}^2 = \tau_{ai}^2\sigma^2$ e $\sigma_{mdi}^2 = \tau_{di}^2\sigma^2$ com $i = 1, 2, \dots, m$. Dessa forma, a variância genética aditiva e de dominância podem ser estimadas utilizando as relações $\sigma_a^2 = \sum_{i=1}^m 2p_iq_i\,\sigma_{mai}^2$ e $\sigma_d^2 = \sum_{i=1}^m [2p_iq_i]^2\,\sigma_{mdi}^2$, sendo assim, $\sigma_a^2 = \sum_{i=1}^m 2p_iq_i\tau_{ai}^2\sigma^2$ e $\sigma_d^2 = \sum_{i=1}^m [2p_iq_i]^2\,\tau_{di}^2\sigma^2$. Os valores genômicos aditivos e devido à dominância são estimados via as seguintes expressões $\hat{a} = W\hat{m}_a$ e $\hat{d} = S\hat{m}_d$, respectivamente.

As distribuições condicionais completas *a posteriori* para os parâmetros do BLASSO são apresentados em detalhes por de los Campos et al. (2009a).

Neste trabalho foram utilizadas 100.000 iterações para os algoritmos MCMC (*Markov* chain Monte Carlo), das quais 20.000 foram descartadas (*burn-in*) para garantir o aquecimento da cadeia e com selecção de uma em cada 10 iterações (*thin*). A análise de covergência foi realizada via o critério proposto por Geweke (1992).

2.8.Recursos Computacionais

Todas as rotinas dos métodos utilizados foram implementadas no *software* R (R Development Core Team, 2010), uma vez que este é de fácil acesso, por se tratar de um software livre (http://cran.r-project.org), e também por suportar a manipulação de grandes conjuntos de dados, como é caso dos arquivos de marcadores SNPs.

O G-BLUP foi realizado no pacote *rrBLUP* com a função *mixed.solve*, o LASSO Bayesiano no pacote *BGLR* com a função *BLR*. O algoritmo utilizado para o desenvolvimento do método TCR está no Apêndice II.

2.9.Comparação das metodologias de seleção genômica ampla

Os métodos foram comparados por meio de uma validação independente em que as nove primeiras replicatas foram assumidas como populações de estimação e utilizadas para estimar os efeitos dos marcadores SNPs no fenótipo. A décima replicata foi assumida como população de validação e utilizada para predizer os valores genéticos genômicos via estimativas dos efeitos dos marcadores obtidos na população de estimação. Assim, foi viável calcular medidas de eficiência das metodologias para a predição genômica, tais como acurácia ($r_{\hat{a}a}\ e\ r_{\hat{a}d}$), viés de predição ($b_{y\hat{a}}\ e\ b_{y\hat{a}}$), herdabilidade genômica aditiva e devido à dominância ($h^2_{aM}\ e\ h^2_{dM}$) e eficiência relativa (ER) das estimativas baseadas em cada um dos nove cenários simulados.

As medidas utilizadas são descritas a seguir: (i) a acurácia que é dada pela correlação entre os GEBVs e os valores genéticos paramétricos; (ii) o viés de predição o qual é definido como sendo o coeficiente da regressão entre o fenótipo e o GEBV, sendo que para coeficientes de regressão abaixo de 1 (< 1) entende-se que os GEBVs foram superestimados, para coeficientes de regressão acima de 1 (> 1) conclui-se que os GEBVs foram subestimados e para os coeficientes iguais a 1 ($= 1$) conclui-se que os

GEBVs são não viesados; (iii) a herdabilidade molecular aditiva é dada por $h^2_{aM} = \frac{\sigma^2_{a_M}}{\sigma^2_{a_M}+\sigma^2_{d_M}+\sigma^2_e}$, em que $\sigma^2_{a_M} = \sum_{i=1}^{n} 2p_i q_i m_i^2$ é a variância genômica aditiva, m_i^2 é o quadrado do efeito do i-ésimo marcador, p_i e q_i são as frequências alélicas do i-ésimo marcador. A herdabilidade molecular devido à dominância é dada por $h^2_{dM} = \frac{\sigma^2_{d_M}}{\sigma^2_{a_M}+\sigma^2_{d_M}+\sigma^2_e}$, em que $\sigma^2_{d_M} = \sum_{i=1}^{n}(2p_i q_i d_i)^2$ e d_i é o valor genotípico do heterozigoto.

As medidas de acurácia, viés, herdabilidade e eficiência relativa foram obtidas para cada replicata em cada cenário e os resultados gerais relatados foram a média desses valores.

Para a análise da eficiência do método TCR/G-BLUP, em que se utilizou dados reais associados a características fenotípicas de mandioca, foi considerado a capacidade preditiva ($r_{\hat{a}y}$), a qual consiste na correlação entre os valores genômicos estimados e os valores fenotípicos da população de validação.

3. RESULTADOS E DISCUSSÃO

Os resultados médios encontrados de acurácia, viés de predição e herdabilidade obtidos por meio dos três métodos, TCR, G-BLUP e BLASSO, associados aos valores genômicos aditivos preditos considerando ausência de dominância e dominância completa, estão apresentados na Tabela 2.

Para os efeitos aditivos, verifica-se que o método TCR mostrou-se muito superior aos métodos G-BLUP e BLASSO em termos de estimação da herdabilidade (sempre muito próximo da herdabilidade paramétrica pelo método TCR), exceto para o cenário 1. Além de propiciar estimativas de valores genômicos aditivos não viesadas, ou seja, valores ($b_{y\hat{a}}$) sempre muito próximos da unidade. A propriedade de não vício é importante quando a seleção envolve indivíduos de muitas gerações usando efeitos dos

marcadores estimados em uma só geração (Resende et al, 2012). Por outro lado, o método TCR propiciou menor acurácia do que os métodos G-BLUP e BLASSO, sendo o BLASSO o método que destacou-se produzindo maiores valores.

Tabela 2. Herdabilidade aditiva (h^2_{aM}), acurácia ($r_{\hat{a}a}$) e viés ($b_{y\hat{a}}$), com respectivos desvios-padrão, dos valores genômicos aditivos estimados via método TCR, G-BLUP e BLASSO considerando modelo aditivo e aditivo-dominante em dados simulados.

Modelo	**Cenário**	**Método**	$\boldsymbol{h^2_{aM}}$	$\boldsymbol{r_{\hat{a}a}}$	$\boldsymbol{b_{y\hat{a}}}$
Modelo aditivo	Cenário 1	TCR	0,31 ± 0,03	0,65 ± 0,02	1,09 ± 0,01
		G-BLUP	0,27 ± 0,04	0,64 ± 0,03	1,48 ± 0,04
		BLASSO	0,28 ± 0,03	0,76 ± 0,02	1,03 ± 0,06
Modelo aditivo	Cenário 2	TCR	0,47± 0,04	0,69± 0,02	0,77± 0,01
		G-BLUP	0,50 ± 0,04	0,79 ± 0,02	1,30 ± 0,02
		BLASSO	0,50 ± 0,05	0,82 ± 0,01	1,00 ± 0,08
Modelo aditivo - dominante	Cenário 3	TCR	0,23 ± 0,03	0,57 ± 0,05	1,09 ± 0,01
		G-BLUP	0,15 ± 0,05	0,63 ± 0,03	1,25 ± 0,35
		BLASSO	0,17 ± 0,09	0,63 ± 0,03	1,44 ± 0,65
Modelo aditivo - dominante	Cenário 4	TCR	0,35 ± 0,04	0,62 ± 0,02	1,09 ± 0,01
		G-BLUP	0,27 ± 0,03	0,70 ± 0,02	1,17 ± 0,13
		BLASSO	0,18 ± 0,05	0,69 ± 0,03	1,69 ± 0,45

Cenários cujas características são controladas por genes de pequenos efeitos - Cenário 1 (h^2_a= 0,22), Cenário 2 (h^2_a= 0,33), Cenário 3 ($h^2_a = 0{,}21$ e $h^2_d = 0{,}10$), Cenário 4: ($h^2_a = 0{,}35$ e $h^2_d = 0{,}17$).

Os resultados médios encontrados de acurácia, viés de predição e herdabilidades obtidos por meio dos três métodos, TCR, G-BLUP e BLASSO, associados aos valores genômicos devido à dominância preditos considerando dominância completa estão apresentados na Tabela 3.

Tabela 3. Herdabilidade devido à dominância (h^2_{dM}), acurácia ($r_{\hat{a}d}$) e viés ($b_{y\hat{a}}$), com respectivos desvios-padrão, dos valores genômicos devido à dominância estimados via método TCR, G-BLUP e BLASSO, razão entre as herdabilidades devido à dominância e aditiva (h^2_{dM}/h^2_{aM}), considerando o modelo aditivo-dominante em dados simulados.

Cenário	**Método**	$\boldsymbol{h^2_{dM}}$	$\boldsymbol{r_{\hat{a}d}}$	$\boldsymbol{b_{y\hat{a}}}$	$\boldsymbol{h^2_{dM}/h^2_{aM}}$
	TCR	**0,10±0,01**	0,40±0,02	0,90±0,14	**0,43**
Cenário 3	G-BLUP	0,13±0,06	0,31±0,04	0,70±0,30	0,87
	BLASSO	0,13±0,02	0,29±0,05	3,20±5,34	0,76
	TCR	**0,17±0,02**	0,40±0,02	0,96±0,12	**0,49**
Cenário 4	G-BLUP	0,20±0,02	0,40±0,04	0,74±0,22	0,74
	BLASSO	0,29±0,03	0,35±0,03	0,46±0,08	1,61

Cenários cujas características são controladas por genes de pequenos efeitos - Cenário 3 ($h^2_a = 0{,}21$ e $h^2_d = 0{,}10$), Cenário 4: ($h^2_a = 0{,}35$ e $h^2_d = 0{,}17$).

Para os efeitos de dominância, verifica-se que o método TCR apresentou, em média, estimativas da herdabilidade coincidentes com a herdabilidade paramétrica. Os métodos G-BLUP e BLASSO foram inferiores em relação à herdabilidade e apresentaram valores de viés distantes de 1. O método TCR propiciou também maior acurácia do que os métodos G-BLUP e BLASSO (cerca de 0,40 no método TCR, 0,31 a 0,40 no G-BLUP e 0,29 a 0,35 no BLASSO) e foi também capaz de extrair melhor a relação variância de dominância / variância aditiva. Ademais, o método TCR apresentou maiores valores de acurácia para os efeitos devido à dominância e a razão entre variâncias mais próximas dos valores paramétricos que os métodos bayesianos reportados por Azevedo et al. (2015) utilizando o mesmo conjunto de dados simulados. Assim, para os efeitos de dominância o método TCR mostrou superioridade para todos os quatro critérios.

Os resultados do estudo de simulação revelaram a adequabilidade dos estimadores propostos pelo método TCR para a estimação dos componentes da variação genotípica e

da herdabilidade. Segundo de los Campos et al. (2009b) a capacidade de estimar de forma acurada as herdabilidades pode ser um critério mais sensível para discriminar e avaliar metodologias estatísticas em GWS. Esta maior sensibilidade se deve ao fato de que as herdabilidades são parâmetros mais complexos do que os coeficientes de correlação simples. Além disso, de acordo com Makowsky et al. (2011) e Azevedo et al. (2015), a herdabilidade pode ser considerada como uma medida da qualidade do ajuste.

Os resultados médios encontrados de herdabilidade aditiva e herdabilidade devido à dominância associadas à estimação da herdabilidade via variação genotípica total estão apresentados na Tabela 3.

Tabela 4. Herdabilidade aditiva (h^2_{aM}), herdabilidade devido à dominância (h^2_{dM}), herdabilidade no sentido amplo (h^2_{gM}), com respectivos desvios-padrão, estimadas via método TCR considerando as variâncias genotípicas e o modelo aditivo-dominante em dados simulados.

Cenário	Estimador direto	$\boldsymbol{h^2_{aM}}$	$\boldsymbol{h^2_{dM}}$	$\boldsymbol{h^2_{gM}}$
Cenário 3	σ^2_{aM} e σ^2_{aM}	0,23±0,03	0,10± 0,01	0,33
	σ^2_{gM}	0,24±0,03	0,12± 0,01	0,36
Cenário 4	σ^2_{aM} e σ^2_{aM}	0,35±0,04	0,17± 0,02	0,52
	σ^2_{gM}	0,37±0,04	0,18± 0,02	0,55

Cenários cujas características são controladas por genes de pequenos efeitos - Cenário 3 ($h^2_a = 0{,}21$ e $h^2_d = 0{,}10$), Cenário 4: ($h^2_a = 0{,}35$ e $h^2_d = 0{,}17$).

A herdabilidade no sentido amplo, estimada diretamente pelo estimador da variância genotípica total, é um parâmetro genético importante para plantas de propagação vegetativa, como mandioca e cana, ou de autofecundação, nas quais o genótipo é herdado integralmente pelos descendentes. Enquanto, que a herdabilidade no sentido restrito, estimada diretamente pelo TCR, é um parâmetro genético importante

principalmente quando o interesse for a predição de ganho devido à seleção (Falconer e Mackay, 1996). No entanto, de acordo com os resultados apresentados na Tabela 4, verificam-se resultados aproximadamente iguais para ambos os métodos (TCR e TCR2). Dessa forma, utilizando o estimador da variância genotípica total é possível obter de forma adequada as herdabilidades aditiva (sentido restrito) e devido à dominância.

Os resultados médios encontrados de capacidade preditiva, viés de predição e herdabilidades obtidos por meio dos métodos, TCR e G-BLUP, associados aos valores genômicos aditivos preditos de seis caracteristicas avaliadas em mandioca (*Manihot esculenta*) estão apresentados na Tabela 5.

Verifica-se que o método TCR conduziu viés mais próximo de 1 e menores valores de capacidade preditiva do que o método G-BLUP. Diante da conclusão de que o TCR estima melhor a herdabilidade e o G-BLUP apresenta maior acurácia (dados simulados) ou maior capacidade preditiva (dados reais), adotou-se a estratégia de fixar a herdabilidade estimada pelo método TCR nas equações de modelos mistos genômicas do G-BLUP, gerando o método TCR/G-BLUP. Essa abordagem aumentou a capacidade preditiva e diminuiu o viés do método G-BLUP, sendo, portanto recomendável.

Segundo Oliveira et al. (2012), a cultura da mandioca tem uma importância fundamental para o país, pois é uma das mercadorias mais relevantes para a agricultura de subexistência e segurança alimentar. Dessa forma, são grandes as perspectivas da utilização da GWS em características de mandioca, visto que a estimação dos valores genômicos dos indivíduos permite a seleção das plantas geneticamente superiores na fase mudal e consequentemente aumenta o ganho da seleção por unidade de tempo. As estimativas de herdabilidades obtidas pelo método TCR/G-BLUP considerando as características de mandioca foram similares aos valores encontrados por Azevedo et al. (2016) e por de Oliveira et al. (2012).

Tabela 5. Herdabilidade aditiva (h^2_{aM}), capacidade preditiva ($r_{\hat{a}y}$) e viés ($b_{y\hat{a}}$) dos valores genômicos aditivos estimados via método TCR, G-BLUP e TCR/GBLUP em dados reais de mandioca.

Variável	**Método**	h^2_{aM}	$r_{\hat{a}y}$	$b_{y\hat{a}}$
PPA	TCR	0.36	0.44	1.25
	G-BLUP	0.17	0.60	1.68
	TCR/ G-BLUP	0.36	**0.65**	1.54
PTR	TCR	0.28	0.44	1.44
	G-BLUP	0.15	0.57	1.87
	TCR/ G-BLUP	0.28	**0.64**	1.64
AML	TCR	0.12	0.30	1.45
	G-BLUP	0.07	0.50	3.77
	TCR/ G-BLUP	0.12	**0.56**	3.02
AMD	TCR	0.20	0.40	1.52
	G-BLUP	0.23	0.66	2.10
	TCR/ G-BLUP	0.20	**0.65**	2.21
HCN	TCR	0.47	0.46	1.13
	G-BLUP	0.50	0.83	1.26
	TCR/ G-BLUP	0.47	**0.83**	1.28
PROD-AMD	TCR	0.30	0.45	1.39
	G-BLUP	0.15	0.57	1.80
	TCR/ G-BLUP	0.30	**0.64**	1.56

As características avaliadas foram: peso da parte área (PPA), produtividade total de raízes (PTR), teor de amilose (AML), teor de amido (AMD), teor de compostos cianogênicos (HCN) e produtividade de amido (PROD-AMD).

Os resultados das Tabelas 2 e 3 foram obtidos a partir de dados simulados cujo valor de $2pq$ médio é igual a 0,49, portanto, com p médio aproximadamente igual ao q médio de 0,5. Essa situação coaduna bem com as populações de base genética ampla tais quais compostos e populações F_2. Nessa situação o método TCR se adequa melhor e é uma alternativa recomendável. Esses resultados são válidos para o modelo genético infinitesimal de Fisher, o qual não admite genes de efeitos maiores, como nos cenários simulados.

O G-BLUP acomoda genes maiores ao passo que o TCR não. Isto é devido ao G-BLUP ajustar efeitos de marcadores via uma distribuição Normal, a qual admite, embora com baixa probabilidade, alguns valores extremos presentes nas caudas da distribuição. Enquanto que o TCR ajusta efeitos de marcadores via uma distribuição que tende para uma Uniforme.

4. CONCLUSÕES

Para os dados simulados conclui-se que o método TCR, em comparação com o G-BLUP e BLASSO, foi o que apresentou estimativa de herdabilidade mais próxima da paramétrica. Este método provou ser o melhor para estimar herdabilidade e componentes de variação genotípica em uma população. Para as características avaliadas em mandioca conclui-se que o método TCR/G-BLUP foi superior ao método G-BLUP aumentando a capacidade preditiva e produzindo viés mais próximo de um.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AZEVEDO, C. F. et al. New accuracy estimators for genomic selection with application in a cassava (Manihot esculenta) breeding program. **Genetics and molecular research: GMR**, v. 15, n. 4, 2016.

AZEVEDO, C.F.; RESENDE, M.D.V.; SILVA, F.F.; VIANA, J.M.S.; VALENTE, M.S.F., RESENDE JUNIOR, M.F.R.; MUÑOZ, P. Ridge, LASSO And Bayesian Additive-Dominance Genomic Models. **BMC Genetics**, v.16, p.105, 2015.

DA, Y.; WANG, C.; WANG, S.; HU, G. Mixed model methods for genomic prediction and variance component estimation of additive and dominance effects using SNP markers. **PLOS ONE**, v. 9, n. 1, p. e87666, 2014.

DE LOS CAMPOS, G.et al. Predicting Quantitative Traits with Regression Models for Dense Molecular Markers and Pedigree. **Genetics**, v. 182, n. 1, p. 375-385. 2009a.

DE LOS CAMPOS, G.; GIANOLA, D.; ROSA, G.J.M. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal Animal Science**, v. 87, n. 1883–1887, 2009b.

DE OLIVEIRA, E. J. et al. Genome-wide selection in cassava. **Euphytica**, v. 187, n. 2, p. 263-276, 2012.

FALCONER, D. S. **Introduction to Quantitative Genetics**. Longmans, New York, p.438, 1989.

GIANOLA, D.; PEREZ‐ENCISO, M.; TORO, M.A. On marker‐assisted prediction of genetic value: beyond the ridge. **Genetics**, v.163, p.347‐365, 2003.

GODDARD, M. E. Genomic selection: prediction of accuracy and maximization of long term response. **Genetics**, Dordrecht, v. 136, n. 2, p. 345-357, 2009.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. The impact of Genetic Relationship on Genome-Assisted Breeding Values. **Genetics**, v.117, p.2389-2397, 2007.

MAKOWSKY, R.; PAJEWSKI, N.M.; KLIMENTIDIS, Y.C.; VAZQUEZ, A.I.; DUARTE, C.W.; ALISSON, D.B. et al. Beyond missing heritability: prediction of complex traits. **Plos Genetics**, v. 7, n. 4, p. e1002051, 2011.

MEUWISSEN, TH.E.; HAYES, B.J.; GODDARD, M.E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, p.1819–29, 2001.

OLIVEIRA, E.J.; RESENDE, M.D.V.; SANTOS, V.S., FERREIRA, C.F.; OLIVEIRA , G.A.F.; SILVA, M.S.; OLIVEIRA, L.A.; VILDOSO, C.I.A. Genome-wide selection in cassava. **Euphytica**, jun. 2012.

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

R Development Core Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, Available: http://www.R-project.org, 2010.

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

RESENDE, M.D.V.; RESENDE JR, M.F.R.; SANSALONI, C.P.; PETROLI, C.D.; MISSIAGGIA, A.A.; AGUIAR, A.M.; ABAD, J.M.; TAKAHASHI, E.K.; ROSADO, A.M.; FARIA, D.A.; PAPPAS JR, G.J.; KILIAN, A.; GRATTAPAGLIA, D. Genomic selection for growth and wood quality in Eucalyptus: capturing the missing heritability and accelerating breeding for complex traits in forest trees. **New Phytologist**, v. 194, n. 1, p. 116–128, 2012.

RESENDE, M.D.V.; RESENDE JUNIOR, M.F.R.; AGUIAR, A.M.; ABAD, J.I.M.; MISSIAGIA, A.A.; SANSALONI, C.P.; PETROLI, C.D.; GRATTAPALIA,

D. **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, p.79, 2010.

RESENDE, M.D.V. (ORG.); SILVA, F.F (ORG.); AZEVEDO, C.F. (ORG.). **Estatística Matemática, Biométrica e Computacional**: **Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência.**1. ed. Visconde do Rio Branco: Suprema, v.1, p.881, 2014.

VAN RADEN, P. M. Efficient Methods to compute genomic predictions. **Journal of Dairy Science,** Champaign, v.91, n.11, p. 4414-4423, 2008.

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

VIANA, J. M. S. **Programa para análises de dados moleculares e quantitativos Real Breeding**. Viçosa: UFV, 2011.

WANG, C.; DA, Y. Quantitative Genetics Model as the Unifying Model for Defining Genomic Relationship and Inbreeding Coefficient. **PLoS ONE**. v. 9, n. 12, p. e114484, 2014.

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

APÊNDICE I

```
dados=read.table("M1D.txt",h=T)   #arquivo de dados da
população de treinamento

M1=as.matrix(dados[,-(1:4)]) #arquivo de marcas da população
de treinamento

# Separando as subpopulações da população de treinamento

fen = cbind(1:1000,dados[,4]) # vetor de fenótipos da
população de treinamento
colnames(fen)=c("id","fen")
y=fen[,2]

# vetor de fenótipos da população de treinamento em ordem
descrescente
order=fen[order(fen[,2],decreasing=TRUE),]

# subpopulação 1 com os maiores fenótipos
g1= order[1:(0.5*nrow(dados)),] # fenótipos
M1_novo = M1[g1[,1],] # genótipos

# subpopulação 2 com os menores fenótipos
g2=order[(0.5*nrow(dados)+1):nrow(dados),] # fenótipos
M2_novo = M1[g2[,1],] # genótipos


# frequências alélicas da população total, subpopulação 1 e
subpopulação 2
p10=matrix(0,ncol(M1_novo),1)
p20=matrix(0,ncol(M1_novo),1)
p0=matrix(0,ncol(M1_novo),1)
for(i in 1:ncol(M1_novo))
{
p10[i,]=(length(which(M1_novo[,i]==1))+2*length(which(M1_no
vo[,i]==2)))/(2*nrow(M1_novo))
p20[i,]=(length(which(M2_novo[,i]==1))+2*length(which(M2_no
vo[,i]==2)))/(2*nrow(M2_novo))
p0[i,]=(length(which(M1[,i]==1))+2*length(which(M1[,i]==2))
)/(2*nrow(M1))
}

# Matriz dominância
Wp0 = matrix(0,nrow(M1),ncol(M1))
Mp0 = matrix(0,nrow(M1),ncol(M1))

for(j in 1:ncol(M1))
{
Mp0[,j] = M1[,j]-2*p0[j,] # matriz aditiva
for(i in 1:nrow(M1))
```

```
{
if(M1[i,j]==2)(Wp0[i,j]=-2*(1-p0[j,])^2)
if(M1[i,j]==1)(Wp0[i,j]=2*p0[j,]*(1-p0[j,]))
if(M1[i,j]==0)(Wp0[i,j]=-2*(p0[j,])^2)
}}


# Exclusão dos marcadores pela MAF
MAF=0.05
maf=NULL
count=NULL
for(i in 1:length(p0))
{
maf[i]=min(p0[i,],1-p0[i,])
if(maf[i]>MAF){count[i]=i}
}

# Exclusão dos marcadores com deltap0=0

deltap0 = (p10 - p20)  # diferença entre a frequências

count1=NULL
for(i in 1:length(p0))
{
if(deltap0[i]!=0){count1[i]=i}
}

# Frequências dos marcadores que não foram excluídos pelo
controle
int_snp=intersect(as.vector(na.omit(count)),
as.vector(na.omit(count1)))
Mp=as.matrix(Mp0[,int_snp])
Wp=as.matrix(Wp0[,int_snp])
p1=as.matrix(p10[int_snp,])
p2=as.matrix(p20[int_snp,])
p=as.matrix(p0[int_snp,])


########### Efeitos Aditivos

deltap = (p1 - p2)  # diferença entre a frequências
deltap_med=mean(abs(deltap))# diferença média

# bm médio
ad=dados[,2]
h2a=var(ad)/var(y)
bm                =                (0.5*h2a)*(mean(g1[,2])-
mean(g2[,2]))/(length(p)*2*deltap_med)

# Cálculo de bi - efeitos aditivos de marcadores
bi = (deltap/deltap_med)*bm
```

```
########### Validação independente

#arquivo de dados da população de validação
dados10=read.table(paste("M",10,"D.txt",sep=""),h=T)
M10=as.matrix(dados10[,-c(1:4)]) # matriz de marcadores da
população de validação
M100=M10[,int_snp]

# frequências alélicas e matriz de incidência aditiva da
população de validação
p10 = matrix(0,ncol(M100),1)
M10p = matrix(0,nrow(M100),ncol(M100))
for(i in 1:ncol(M100))
{
p10[i,]=(length(which(M100[,i]==1))+2*length(which(M100[,i]
==2)))/(2*nrow(M100))
M10p[,i] = M100[,i]-2*p10[i,] # matriz aditiva
}

# valor genômico aditivo dos indivíduos da população de
validação
gbv_p = M10p%*%bi

# capacidade preditiva, acurácia e viés de predição
ad10=dados10[,2]
y10=dados10[,4]
(res_p=cbind(cor(gbv_p,y10),              cor(gbv_p,ad10),
cov(gbv_p,y10)/var(gbv_p)))

########### RR-BLUP / G-BLUP aditivo

library(rrBLUP)
rrblup=mixed.solve(y,Z=Mp)
gbv_g = M10p%*%rrblup$u
(res_g=cbind(cor(gbv_g,y10),cor(gbv_g,ad10),cov(gbv_g,y10)/
var(gbv_g)))

########### Indice aditivo

cor_pp = cor(gbv_p,ad10)
cor_gg = cor(gbv_g,ad10)

deltap2 = var(gbv_p)/var(gbv_g)

b1              =              (1-(cor_pp^2)*(deltap2))/(1-
(cor_pp^2)*(cor_gg^2)*(deltap2))
b2 = (1-(cor_gg^2))/(1-(cor_pp^2)*(cor_gg^2)*(deltap2))
I                                                          =
diag(b1[1],nrow(M10p))%*%as.matrix(gbv_g)+diag(b2[1],nrow(M
10p))%*%as.matrix(gbv_p)
```

```
num=(1-(cor_gg^2))*(1-(cor_pp^2)*deltap2)
den=1-(cor_gg^2)*(cor_pp^2)*deltap2
rIa=(1-num/den)^(1/2)
ef=rIa/cor_gg

########### Efeitos devido a dominância

delta2pq = 2*p1*(1-p1) - 2*p2*(1-p2)  # diferença entre 2pq

delta2pq_med=mean(abs(delta2pq))# diferença entre 2pq média

# bm médio
dom=dados[,3]
h2d=var(dom)/var(y)
bm_2pq                                =(0.5*h2d)*(mean(g1[,2])-
mean(g2[,2]))/(length(p)*delta2pq_med)

# Cálculo de bi_2pq - efeitos de marcadores devido a
dominância
bi_2pq = (delta2pq/delta2pq_med)*bm_2pq

########### Validação independente

# matriz de incidência devido a dominância da população de
validação
W10p = matrix(0,nrow(M100),ncol(M100))
for(j in 1:ncol(M100))
{
for(i in 1:nrow(M100))
{
if(M100[i,j]==2)(W10p[i,j]=-2*(1-p10[j,])^2)
if(M100[i,j]==1)(W10p[i,j]=2*p10[j,]*(1-p10[j,]))
if(M100[i,j]==0)(W10p[i,j]=-2*(p10[j,])^2)
}}

# valor genômico devido a dominância dos indivíduos da
população de validação
d_p = W10p%*%bi_2pq

# capacidade preditiva, acurácia e viés de predição
dom10=dados10[,3]
(res_d_p=cbind(cor(d_p,y10),cor(d_p,dom10),cov(d_p,y10)/var
(d_p)))


########### RR-BLUP / G-BLUP devido a dominância

rrblup_d=mixed.solve(fen[,2],Z=Wp)
d_g = W10p%*%rrblup_d$u
```

```
(res_d_g=cbind(cor(d_g,y10),cor(d_g,dom10),cov(d_g,dom10)/v
ar(d_g)))


########### Índice devido a dominância

cord_pp = cor(d_p,dom10)
cord_gg = cor(d_g,dom10)

deltap2d = var(d_p)/var(d_g)
b1d            =            (1-(cord_pp^2)*(deltap2d))/(1-
(cord_pp^2)*(cord_gg^2)*(deltap2d))
b2d                 =                  (1-(cord_gg^2))/(1-
(cord_pp^2)*(cord_gg^2)*(deltap2d))
Id                                                          =
diag(b1d[1],nrow(Wp))%*%as.matrix(d_g)+diag(b2d[1],nrow(Wp)
)%*%as.matrix(d_p)
numd=(1-(cord_gg^2))*(1-(cord_pp^2)*deltap2d)
dend=1-(cord_gg^2)*(cord_pp^2)*deltap2d
rId=(1-numd/dend)^(1/2)
efd=rId/cord_gg
```

APÊNDICE II

```
dados=read.table("M1D.txt",h=T)   #arquivo   de   dados   da
população de treinamento

M1=as.matrix(dados[,-(1:4)]) #arquivo de marcas da população
de treinamento

# Separando as subpopulações da população de treinamento
fen  =  cbind(1:1000,dados[,4])  #  vetor  de  fenótipos  da
população de treinamento
colnames(fen)=c("id","fen")
y=fen[,2]

# vetor de fenótipos da população de treinamento em ordem
descrescente
order=fen[order(fen[,2],decreasing=TRUE),]

# subpopulação 1 com os maiores fenótipos
g1= order[1:(0.5*nrow(dados)),] # fenótipos
M1_novo = M1[g1[,1],] # genótipos

# subpopulação 2 com os menores fenótipos
g2=order[(0.5*nrow(dados)+1):nrow(dados),] # fenótipos
M2_novo = M1[g2[,1],] # genótipos

# frequências alélicas da população total, subpopulação 1 e
subpopulação 2
p1=matrix(0,ncol(M1_novo),1)
p2=matrix(0,ncol(M1_novo),1)
p=matrix(0,ncol(M1_novo),1)
for(i in 1:ncol(M1_novo))
{
p1[i,]=(length(which(M1_novo[,i]==1))+2*length(which(M1_nov
o[,i]==2)))/(2*nrow(M1_novo))
p2[i,]=(length(which(M2_novo[,i]==1))+2*length(which(M2_nov
o[,i]==2)))/(2*nrow(M2_novo))
p[i,]=(length(which(M1[,i]==1))+2*length(which(M1[,i]==2)))
/(2*nrow(M1))
}

# Exclusão de marcas com MAF<0,05
MAF=0.05
maf=NULL
count=NULL
for(i in 1:ncol(M1))
{
maf[i]=min(p[i,],1-p[i,])
if(maf[i]>MAF){count[i]=i}
}
```

```
# diferença entre a frequências alélicas da subpopulação 1
e 2
deltap = (p1 - p2)

# Exclusão de marcas com deltap=0
count1=NULL
for(i in 1:ncol(M1))
{
if(deltap[i]!=0){count1[i]=i}
}

# marcadores após o controle de qualidade
int_snp=intersect(as.vector(na.omit(count)),
as.vector(na.omit(count1)))
M_novo=as.matrix(M1[,int_snp]) # genótipos da população
deltap_novo = deltap[int_snp] # diferença entre a
frequências


# Mudança no arquivo de marcas: se o sinal de deltap for
negativo, trocar 0 por 2 e
#2 por 0 em cada coluna de marcador
M=M_novo
for(i in 1:nrow(M))
{
for(j in 1:ncol(M))
{
if(deltap_novo[j]<0){
if(M_novo[i,j]==2){M[i,j]=0}
if(M_novo[i,j]==0){M[i,j]=2}
}
}
}

# Contagem de 0 (bb) , 1 (Bb) e 2 (BB) por indivíduos
pqn1=NULL
pqn2=NULL
pqn0=NULL
for(i in 1:nrow(M))
{
pqn1[i]=length(which(M[i,]==1))
pqn2[i]=length(which(M[i,]==2))
pqn0[i]=length(which(M[i,]==0))
}

# Coeficiente da regressão em cada classe
regBB=cov(y,pqn2)/var(pqn2)
regbb=cov(y,pqn0)/var(pqn0)
regBb=cov(y,pqn1)/var(pqn1)

# Valores genéticos por classe genotípica
```

```
vgBB=regBB*pqn2
vgbb=regbb*pqn0
vgBb=regBb*pqn1

# Valores genômico aditivo
a=(vgBB/2+vgbb/2)

# Acurácia - correlação entre o valor estimado e o valor
simulado
cora=cor(a,dados[,2])

# Viés - coeficiente da regressão entre o valor estimado e
o valor fenotípico
ba=cov(a,y)/var(a)

# Variância aditiva
va=var((1/sqrt(mean(2*p*(1-p))))*a)

# herdabilidade molecular aditiva
h2a=va/var(y)

# Valores genômico devido a dominância
d=0.5*(vgBb-vgBB+vgbb)

# Acurácia - correlação entre o valor estimado e o valor
simulado
cord=cor(d,dados[,3])

# Viés - coeficiente da regressão entre o valor estimado e
o valor fenotípico
bd=cov(d,y)/var(d)

# Variância aditiva
vd=var((1/mean(2*p*(1-p)))*d)

# Herdabilidade aditiva
h2d=vd/var(y)
```