CAILLET DORNELLES MARINHO

TRIAGEM DE MARCADORES NA SELEÇÃO GENÔMICA AMPLA

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

VIÇOSA
MINAS GERAIS - BRASIL
2014

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

T

M337t
2014

Marinho, Caillet Dornelles, 1987-
Triagem de marcadores na seleção genômica ampaillet Dornelles Marinho. - Viçosa, MG, 2014.
vii, 32f. ; 29 cm.

Orientador: Luiz Alexandre Peternelli.
Tese (doutorado) - Universidade Federal de Viçosa.
Referências bibliográficas: f.29-32.

1. Eucalipto - Melhoramento genético. 2. Seleção genômica. I. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 634.973766

CAILLET DORNELLES MARINHO

TRIAGEM DE MARCADORES NA SELEÇÃO GENÔMICA AMPLA

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

APROVADA: 09 de maio de 2014.

Edilson Romais Schmildt

Moysés Nascimento

Felipe Lopes da Silva

Leonardo Lopes Bhering

Luiz Alexandre Peternelli

(Orientador)

“Pouco conhecimento faz com que as pessoas se sintam orgulhosas. Muito conhecimento, que se sintam humildes. É assim que as espigas sem grãos erguem desdenhosamente a cabeça para o céu, enquanto que as cheias as baixam para a terra, sua mãe”.

- Leonardo da Vinci -

Aos meus amores Fabrícia e João Pedro

pelo carinho, companheirismo,

paciência e inspiração.

Dedico

AGRADECIMENTOS

À minha esposa Fabrícia e ao meu filho João Pedro pela caminhada sempre juntos, me ensinando sempre a ser melhor e aprendendo o significado de família. Sem vocês não estaria aqui.

À minha amada mãe, pela educação, carinho, apoio, amor, confiança e todas as outras coisas que jamais poderei retribuir.

Ao meu pai, Cláudio (*in memoriam*) e irmão Rômulo (*in memoriam*), por olharem por mim aí de cima.

À minha inestimável sogra, Cormaria, por todo apoio, carinho, solicitude e amizade.

Aos companheiros de curso Bruno, Paulo, Édimo e Janeo, pela amizade e solidariedade em compartilhar novos conhecimentos.

Ao meu orientador, Luiz Alexandre Peternelli, pelos ensinamentos, amizade, confiança e apoio de sempre.

Aos professores Marcos Deon Vilela de Resende e Fabyano Fonseca e Silva, por me iniciarem nos estudos de seleção genômica.

Ao meu co-orientador, Márcio Henrique Pereira Barbosa e a equipe do CECA, pela receptividade e aprendizado.

Aos professores Felipe Lopes da Silva, Leonardo Lopes Bhering e Moysés Nascimento, pela disponibilidade e conselhos de sempre.

Ao professor Edilson Romais Schimildt pela presteza em participar da banca com tamanha boa vontade.

Aos pesquisadores Elizabete Keiko Takahashi e Dário Grattapaglia e suas respectivas instituições, CENIBRA e EMBRAPA, pela gentileza em fornecer os dados e informações preciosas para o trabalho.

À Universidade Federal de Viçosa – MG, pela oportunidade de formação sem igual.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES, pela concessão da bolsa de doutorado.

E, finalmente, a Deus, por me fazer levantar todos os dias da minha vida com grandes objetivos e grandes pessoas ao meu lado.

ÍNDICE

RESUMO

MARINHO, Caillet Dornelles, D.Sc., Universidade Federal de Viçosa, maio de 2014. **Triagem de marcadores na seleção genômica ampla.** Orientador: Luiz Alexandre Peternelli. Co-orientadores: Marcos Deon Vilela de Resende e Márcio Henrique Pereira Barbosa.

A seleção genômica ampla (GWS) foi proposta visando aumentar a eficiência dos programas de melhoramento genético e otimizar o processo de seleção utilizando informações pré-estimadas de centenas ou milhares de marcadores moleculares, sendo os marcadores DArTs e SNPs os mais apropriados para tal propósito. No entanto, apesar do avanço da biologia molecular e a relativa redução do preço de obtenção de marcadores a genotipagem de muitos indivíduos com alta densidade de marcadores pode não ser economicamente viável em alguns programas de melhoramento, dependendo da espécie. Os objetivos desse trabalho foram: i)triagem de marcadores para seleção genômica em eucalipto via seleção pelo efeito dos marcadores e via desequilíbrio de ligação (LD) com a verificação da capacidade preditiva após redução dimensional; ii) verificaçãodo ganho genético esperado baseado em seleção fenotípica e com uso da GWS, com e sem seleção de marcadores; iii)comparar a utilização de marcadores DArTs e SNPs na GWS em uma mesma população de eucalipto; iv) proposição de uma modificação para o método RR-BLUP-B preconizado por Resende et al. (2010).Os resultados demonstram que a seleção de marcadores pode manter a capacidade preditiva dos métodos de GWS muito próximas as obtidas utilizando-se todos os marcadores. Os ganhos genéticos por unidade de tempo, independente do método utilizado, foram superiores aos da seleção tradicional e podem ser mantidos utilizando a seleção de marcadores. Ademais, neste estudo, a abordagem de redução via desequilíbrio de ligação (PTMOD) foi superior ao RR-BLUP-B (efeito de marcas), porém, com menor redução no número de SNPs selecionados.Para a população de eucalipto e variáveis empregadas neste trabalho, os marcadores DArTs com menor densidade apresentaram melhores predições na GWS em relação a utilização de marcadores SNPs com maior densidade. O método RR-BLUP-B modificadosuperestimou a capacidade preditiva em relação ao método RR-BLUP-B original. Portanto, recomenda-se a validação independente como melhor forma de validação.

ABSTRACT

MARINHO, Caillet Dornelles, D.Sc., Universidade Federal de Viçosa, May, 2014. **Screening markers in genome wide selection.** Adviser: Luiz Alexandre Peternelli. Co-advisers: Marcos Deon Vilela de Resende and Márcio Henrique Pereira Barbosa.

A genome wide selection (GWS) has been proposed to increase the efficiency of breeding programs and to optimize the selection process using pre-estimated information from hundreds or thousands molecular markers. The most suitable and usedfor this purpose are DArTs and SNPs markers. However, despite the advances in molecular biology and related reduction in the price of obtaining markers for genotyping many individuals with high density of markers, it may not be economically feasible in some breeding programs yet, depending on the species. Given the above, the objectives of this work were:i) screening markers for genomic selection in eucalypt selection by the effect of markers and by linkage disequilibrium (LD) with verification of predictive capacity after number of markers reduction; ii) verification of the expected genetic gain based on phenotypic selection and use of GWS, with and without markers selection; iii) to compare the use of DArTand SNP markers in GWS in the same population of Eucalyptus; iv) proposing new validation for RR-BLUP-B method proposed by Resende et al . (2010). Results demonstrate that the selection of markers can maintain the predictive ability of the methods GWS very close to those obtained using all markers. The genetic gain per unit of time, regardless of the method used, were superior to the traditional selection and can be maintained using selection markers. Moreover, in this study the reduction approach via linkage disequilibrium (PTMOD) was higher than RR-BLUP-B (effect of markers), but with a smaller reduction in the number of selected SNPs. For the population of eucalyptus and variables used in this study, DArTsmarkers with lower density showed better predictions in GWS regarding the use of SNP markers with higher density. The RR-BLUP-B modified overestimated the predictive ability compared to the original RR-BLUP-B method. Soitis recommended an independent validation as the best form of validation.

1. INTRODUÇÃO

A seleção genômica ampla (*genome wide selection* – GWS) ou seleção genômica (GS) foi proposta por Meuwissen et al. (2001) com o intuito de aumentar a eficiência dos programas de melhoramento genético e otimizar o processo de seleção. A GWS visa predizer o fenótipo futuro de indivíduos provenientes de populações em melhoramento, utilizando informações pré-estimadas de marcadores moleculares. Para tanto, a GWS utiliza centenas ou milhares de marcadores, os quais cobrem o genoma de forma ampla, garantindo que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores, permitindo que estes expliquem quase a totalidade da variação genética do caráter. Dessa forma, utilizando-se regressões aleatórias (RR) com preditores do tipo BLUP (*Best Linear Unbiased Prediction*) (RR-BLUP), todas as marcas são colocadas no modelo estatístico e através dos efeitos genéticos aditivos dos marcadores, o valor genético genômico (VGG) ou fenótipo futuro do indivíduo é predito.

Atualmente, marcadores do tipo SNP (*Single Nucleotide polimorfism*) são utilizados na GWS devido a sua abundância no genoma, baixa taxa de mutação, facilidade de genotipagem e baixo custo. Os marcadores DArTs (*Diversity Array Technology*) são outra alternativa para utilização na GWS, pois também são abundantes, obtidos com alta velocidade e sem conhecimento prévio de sequenciamento do genoma alvo. No entanto, os marcadores DArTs são dominantes, o que não permite distinguir os indivíduos heterozigotos dos homozigotos, o que é possível com os marcadores SNPs. De toda forma, milhares de marcadores podem ser utilizados para cobrir o genoma de um organismo de tal modo que a probabilidade de se encontrar um marcador em desequilíbrio de ligação (LD) com um QTL (*Quantitative Trait Loci*) é muito alta.

Assim, aestimação dos efeitos genéticos dos marcadores sobre o fenótipo é realizada com base em dados genotípicos e fenotípicos provenientes de várias famílias de indivíduos pertencentes a uma grande amostra da população de seleção (RESENDE et al., 2008). Esses efeitos genéticos dos marcadores sobre o fenótipo são utilizados na predição dos fenótipos futuros de indivíduos genotipados, sendo que a predição e a seleção podem ser realizadas em fases

muito precoces, sem a necessidade de fenotipagem, acelerando o processo de melhoramento (RESENDE et al., 2011).

Para culturas perenes como o eucalipto, essa ferramenta genômica pode ser muito útil, pois para obtenção dos fenótipos em cada ciclo, espera-se, no mínimo três anos para coleta. Além disso, os métodos de melhoramento mais utilizados para tal cultura são os métodos de seleção recorrente, que são trabalhosos e apresentam baixos ganhos com a seleção por unidade de tempo. Com isso, uma alternativa para acelerar os ganhos genéticos por unidade de tempo é a utilização da GWS, que se enquadra perfeitamente no melhoramento do eucalipto, onde o ciclo pode ser reduzido com utilização de indutores hormonais de floração. Assim, utilizando a GWS para obtenção dos fenótipos futuros, pode-se intercruzar os melhores indivíduos precocemente e avançar gerações em menor tempo.

Alguns trabalhos práticos com GWS no melhoramento vegetal foram realizados por Resende Júnior (2010) e Resende et al. (2012) com eucalipto, Cavalcanti et al. (2012) com caju e Resende Júnior et al. (2012a; 2012b) com pinus, Oliveira et al. (2012) com mandioca, Fritsche-Neto et al (2012) com milho, Kumar et al. (2012) com maçã e Gouy et al. (2013) com cana-de-açúcar. Todos comprovaram as vantagens propostas pelo método, pois mesmo nos casos onde a acurácia seletiva permaneceu abaixo daquela obtida pela seleção baseada unicamente em fenótipos, o ganho foi altamente alavancado devido à redução no tempo (seleção precoce). Além disso, em situações em que a coleta do fenótipo exige a destruição dos indivíduos, como no caso do melhoramento de milho para eficiência no uso do nitrogênio (Fritsche-Neto et al., 2012), em que as raízes são arrancadas para avaliação, a GWS permite seleção apenas coletando o DNA, sem destruição da planta.

Apesar do avanço da biologia molecular e da relativa redução do preço de obtenção de marcadores, a genotipagem de muitos indivíduos com alta densidade de marcadores, pode não ser economicamente viável em alguns programas de melhoramento durante vários ciclos. Sendo assim, a seleção de marcadores informativos pode ser uma alternativa para diminuir o custo de novas genotipagem para aplicação da GWS em novas etapas subsequentes.

Ademais, é extremamente desejável na GWS, investigar se um número menor de marcadores informativos pode ser utilizado e ainda proporcionar uma predição acurada, pois assim, pensando em um custo fixo, reduzindo-se a

densidade de marcas, um número maior de indivíduos poderá ser genotipado e mais haplotipos serão amostrados. Além disso, haverá redução nos problemas advindos da multicolinearidade dos dados e redução no tempo de análise computacional.

Nesse aspecto, pesquisas têm sido desenvolvidas visando a triagemde marcadores via seleção pelo efeito das marcas (RESENDE et al., 2010; CAVALCANTI et al., 2012; OLIVEIRA et al., 2012; RESENDE et al., 2012; RESENDE JÚNIOR et al., 2012a), o que foi denominado por Resende Júnior et al. (2012a) como RR-BLUP-B, ou seja, usa-se o método RR-BLUP com prévia seleção de marcadores pelos seus efeitos, ou via teste de associação pela abordagem via *genome wide association studies*(GWAS) para selecionar os marcadores significativos e depois ajustar o modelo somente com estes marcadores (SUBEDI et al., 2012). No entanto, novas abordagens ainda foram pouco investigadas na seleção genômica, principalmente, no campo vegetal.

Diante do exposto, os objetivos desse trabalho foram: i)triagem de marcadores para seleção genômica em eucalipto via seleção pelo efeito dos marcadores e via desequilíbrio de ligação (LD) com a verificação da capacidade preditiva após redução dimensional; ii) verificar o ganho genético esperado baseado em seleção fenotípica e com uso da GWS, com e sem seleção de marcadores; iii)comparar a utilização de marcadores DArTs e SNPs na GWS em uma mesma população de eucalipto; iv) proposição de uma modificação para o método RR-BLUP-B preconizado por Resende et al. (2010).

2. MATERIAL E MÉTODOS

2.1. Material vegetal

Foi utilizada neste trabalho uma população de eucalipto proveniente do programa de melhoramento da empresa CENIBRA. Esta população foi composta por 4900indivíduos (F1) provenientes de 43 famílias obtidas de um dialelo incompleto entre 11 progenitores de *E. grandis* e *E. urophylla*. O delineamento experimental utilizado foi o de blocos incompletos com 36 repetições, parcela com um único indivíduo e espaçamento de 3×2 m.

Os fenótipos foram coletados aos três anos, sendo que a altura da planta (ALT) foi obtida por um clinômetro Suunto PM-5 e o diâmetro altura do peito (DAP), obtida a 130 cm do solo, foi obtida com fita métrica.

Para a genotipagem utilizou-se 766 indivíduos e foram obtidos3129 DArTs e 49042 SNPs por meio da parceria CENIBRA/EMBRAPA.Os DArTs foram previamente filtrados e somente os de alta qualidade (3129) foram utilizados.

Para retirar os SNPs pouco polimórficos e com muitos dados perdidos (*missing data*) foi adotado um controle de qualidade com *Call Rate*>0,90 e MAF>0,01. O *Call Rate* (taxa de atendimento) é utilizado para eliminar marcadores com grande quantidade de valores perdidos, já a MAF (frequência do alelo menor) está relacionada com o polimorfismo dos locos marcadores na população.

Essas medidas de controle de qualidade visam assegurar a retirada de marcadores sem relevância para as análises.Entretanto, não existe um valor ideal estabelecido na literatura para ser utilizado. Nos estudos de Faria et al. (2013), os autores citam que, geralmente utiliza-se *Call Rate* igual a 0,95 e MAF de 0,01 ou 0,05. No entanto, a combinação MAF igual a 0,01 e *Call Rate* igual a 0,90 demonstrou estimativas semelhantes para todas as variáveis estudadas em comparação com a utilização de todos os marcadores.

2.2. Parâmetros genéticos

A partir dos dados experimentais foram calculados os parâmetros genéticos e consequente cômputo dos ganhos genéticos esperados por meio da acurácia seletiva via seleção fenotípica embasada em melhor predição linear não viesada (BLUP), para posterior comparação com os métodos de seleção genômica.

As análises foram realizadas empregando-se a metodologia de modelos mistos (Lynch e Walsh, 1998) por meio do software SELEGEN-REML/BLUP (Resende, 2007), seguindo o seguinte modelo linear: $y = Xf + Za + Wb + e$, em que, y é o vetor de observações; f é o vetor dos efeitos assumidos como fixos (média geral e efeitos experimentais), a é o vetor de efeitos aleatórios genéticos aditivos dos indivíduos; b é o vetor de efeitos aleatórios de blocos; e é o vetor de efeitos residuais aleatórios; e X, Z e W são as matrizes de incidência para os respectivos efeitos. As estruturas de variâncias para o modelo em questão são dadas por:

$$y \mid f, V \sim N(Xf, V);$$

$$a \mid A, \sigma_a^2 \sim N(0, A\sigma_a^2);$$

$$b \mid \sigma_b^2 \sim N(0, I\sigma_b^2);$$

$$e \mid \sigma_e^2 \sim N(0, I\sigma_e^2);$$

$$Cov(a, b') = 0;\ Cov(a, e') = 0;\ Cov(b, e') = 0;$$

$$V = ZA\sigma_a^2 Z' + WI\sigma_b^2 W' + I\sigma_e^2 = ZGZ' + WBW + R;$$

$$G = A\sigma_a^2;$$

$$B = I\sigma_b^2;$$

$$R = I\sigma_e^2;$$

em que, A representa a matriz das relações genéticas aditivas entre os indivíduos; I, a matriz identidade.

As equações de modelos mistos para obtenção dos BLUPs foram dadas por:

$$\begin{bmatrix} X'X & X'Z & X'W \\ Z'X & Z'Z + A^{-1}\lambda_1 & Z'W \\ W'X & W'Z & W'W + I\lambda_2 \end{bmatrix} \begin{bmatrix} \hat{f} \\ \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \\ W'y \end{bmatrix},$$

em que:

$$\lambda_1 = \frac{\sigma_e^2}{\sigma_a^2} = \frac{1 - h^2 - b^2}{h^2};$$

$$\lambda_2 = \frac{\sigma_e^2}{\sigma_b^2} = \frac{1 - h^2 - b^2}{b^2}$$

$h_2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_b^2 + \sigma_e^2}$: herdabilidade individual no sentido restrito através de blocos incompletos.

$b_2 = \frac{\sigma_b^2}{\sigma_a^2 + \sigma_b^2 + \sigma_e^2}$: coeficiente de determinação dos efeitos de blocos incompletos.

σ_{a}^2, variância genética aditiva; σ_{b}^2, variância entre blocos; σ_{e}^2, variância residual.

Os componentes de variância foram estimados por máxima verossimilhança restrita - REML (LYNCH E WASH, 1998).

2.3. Modelo estatístico da GWS

No presente trabalho os efeitos dos marcadores foram estimados pelos seguintes métodos: RR-BLUP (MEUWISSEN et al., 2001) e LASSO Bayesiano - BLASSO (PARK E CASELLA, 2008). Em ambos os métodos, a informação genotípica foi ajustada pelo seguinte modelo linear misto:

$$\mathrm{y} = \mathrm{Wb} + \mathrm{Xm} + \varepsilon \ ,$$

em que y é o vetor de fenótipos desregressados conforme Resende et al. (2010), com dimensão $\mathrm{I} \times 1$, na qual I é o número de indivíduos genotipados e fenotipados; b é o vetor de efeitos fixos com matriz de incidência W; m é o vetor de efeitos de marcas com matriz de incidência X com dimensão $\mathrm{J} \times 1$, em que J é o número de marcadores, e ε é o vetor de efeitos de erros aleatórios.

Assim, utilizando os métodos citados acima, a predição do valor genético genômico (VGG) do indivíduo ifoi obtido por:

$$\hat{y} = \text{?} + \sum_{j=1}^{J} X_{ij}\hat{m},$$

em que, $\sum_{i=1}^{J} X_{ij}\hat{m}$ é a linha da matriz X, correspondente ao indivíduo i.

Para todos os métodos, o esquema de validação 10-*fold*(KOHAVI, 1995) foi utilizado de forma que o conjunto de dados foi particionado em dois. O primeiro (subconjunto de treinamento), composto pela maioria dos indivíduos (~90%), foi utilizado para estimar os efeitos de marcadores. O segundo subconjunto (~10%), compôs os dados de validação, onde os indivíduos tiveram seus fenótipos preditos com base nos efeitos pré-estimados dos marcadores do conjunto de treinamento. Esse procedimento foirepetido 10 vezes, aleatoriamente, cada vez com diferentes indivíduos pertencendo ao conjunto de validação, até que todos os indivíduos tivessem seus fenótipos preditos (LEGARRA et al., 2008; USAI et al., 2009; VERBYLA et al., 2010; RESENDE JÚNIOR et al., 2012).

2.3.1. RR-BLUP

O método *Ridge Regressionbest linear unbiased predictor* (RR-BLUP) utiliza preditores do tipo BLUP dos efeitos de marcadores como covariáveis aleatórias, ou seja, os fenótipos são regressados com base nos efeitos de marcadores (MEUWISSEN et al., 2001). As estruturas de médias e variâncias são definidas a seguir como preconizado por Resende et al. (2008):

$$m \sim N(0, G = I\sigma_m^2) E(y) = \mathrm{Wb}$$

$$\varepsilon \sim N(0, R = I\sigma_\varepsilon^2) \mathrm{Var}(y) = \mathrm{V} = \mathrm{XGR}' + R,$$

em que, σ_m^2 é a variância comum explicada por cada marcador e σ_ε^2 é a variância residual.As equações de modelo misto genômicas para predição de m são dadas por:

$$\begin{bmatrix} W'W & W'X \\ X'W & X'X + I\frac{\sigma_\varepsilon^2}{\sigma_a^2/\eta} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{m} \end{bmatrix} = \begin{bmatrix} W'y \\ X'y \end{bmatrix},$$

em que σ_a^2 refere-se a variância aditiva total da característica e η corresponde ao número total de marcadores utilizados na análise. A matriz X foi padronizada conforme Resende et al. (2010). As análises do método RR-BLUP foram feitas por meio do *software* R pacote rrBLUP (ENDELMAN, 2011).

2.3.2. LASSO Bayesiano (BLASSO)

Este método propicia estimativas específicas para cada marcador (heterogeneidade de variância) com seleção de covariáveis, pois promove maior *shrinkage* das estimativas dos efeitos de marcadores próximos de zero e menor *shrinkage* das estimativas dos marcadores com maiores efeitos.

O método BLASSO foi realizado utilizando a mesma equação do modelo citado anteriormente em **2.3**, porém, com as seguintes estruturas:

$$m|\lambda \sim \prod_i^n \frac{\lambda}{2}\exp(-\lambda|a_i|); \qquad \varepsilon|\sigma_\varepsilon^2 \sim \mathrm{MVN}(0, I\sigma_\varepsilon^2)$$

$$\mathrm{Var}(m) = \frac{2}{\lambda^2},$$

em que, MNV indica uma distribuição normal multivariada e λ é o parâmetro de penalização. Como nessa formulação existe um componente de variância (τ_i^2) extra, associado a cada marcador, tem-se:

$$P(m|\tau) \sim N(0, D); \qquad \mathrm{diag}(D) = (\tau_1^2 \ldots \tau_n^2)$$

$$P(\tau|\lambda) = \prod_i \left(\frac{\lambda^2}{2}\right) \exp\left(\frac{-\lambda^2 \tau_i^2}{2}\right),$$

em que, $\mathrm{Var}(m_i) = \sigma_{mi}^2 = \tau_i^2$.

O método BLASSO assume para os efeitos de marcadores uma distribuição *apriori* dupla exponencial. Ademais, os hiperparâmetros foram escolhidos conforme proposto por Pérez et al. (2010), em que, para cálculo das*prioris* de lambda e da variância do erro,foram utilizadas, respectivamente, as seguintes expressões:

$$\hat{\lambda} = \sqrt{2V_e V^{-1} \sum_j^p \bar{x}_j^2},$$

$$S_e = V_e(df_e + 2),$$

na qual $\bar{x}_j$ denota a média dos valores da j-ésima coluna da matriz X. E V_e e V correspondem a variância residual e a variância genética, respectivamente, obtida pelo método RR-BLUP. O grau de liberdade(df_e) foi igual atrês para garantir variâncias finitas (PÉREZ et al., 2010).

O algoritmo *Gibbs Sampler* com 100000 iterações, *burn-in* igual a 20000 (descarte das 20000 iterações) e *thin* igual a 100 foi utilizado. O termo *thin* refere-se à amostragem de 100 em 100 iterações para evitar amostras correlacionadas. As análises do método BLASSO foram realizadas por meio do *software* R pacote BLR (DE LOS CAMPOS e RODRIGUEZ, 2012).

2.4. Seleção de marcadores

2.4.1. RR-BLUP-B

A estratégia proposta por Resende et al. (2010) foi empregada como segue: obtenção do BLUP dos efeitos de marcadores usando todos os SNPs no conjunto de treinamento via RR-BLUP; ordenação dos marcadores por maiores módulos de seus efeitos estimados e seleção dos 1000 marcadores de maiores efeitos; re-estimação dos efeitos desses 1000 marcadores separados dos demais e obtenção do valor genético genômico (VGG) dos indivíduos do grupo de validação. Esse procedimento, como proposto inicialmente,foi realizado dentro de cada *fold*, portanto, em cada ciclo, 1000 marcadores de maiores efeitos foram selecionados.

2.4.2. RR-BLUP-B modificado

Esse procedimento é apenas uma modificação do método anterior e consiste em calcular uma média dos efeitos de marcadores (em módulo) após os 10-*fold*s, e selecionar os 1000 marcadores de maior efeito para a re-estimação destes e posterior validação. Sendo assim, os 1000 marcadores serão os mesmos na validação, o que difere do RR-BLUP-B original.

Neste trabalho, para seleção dos 1000 marcadores de maior efeito foi utilizada duas estratégias: i) a média do módulo dos efeitos de marcadores após os 10-*fold*s (RR-BLUP-B1); e, ii)a média dos postos dos efeitos de marcadores após os 10-*fold*s (RR-BLUP-B2). Na Figura 1, o esquema de validação proposto é apresentado.

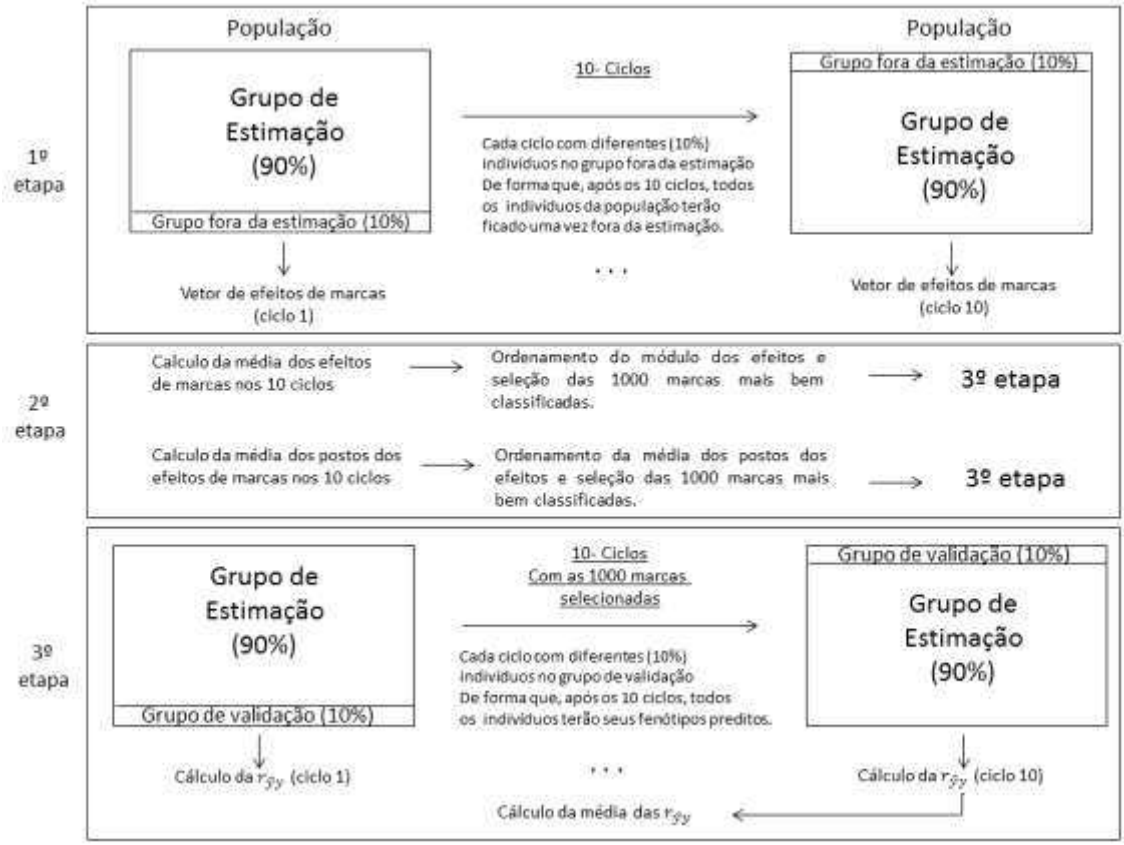

Figura 1 – Esquema RR-BLUP-B modificado.

2.4.3. Seleção de marcadores via desequilíbrio de ligação (PTMOD)

Esse método foi proposto por Peternelli e Rosa (2013) em sua forma preliminar. A abordagem proposta pode ser dividida em duas etapas: seleção e ajuste. A descrição é apresentada a seguir.

A etapa de seleção consiste no agrupamento de p SNPs, os quais apresentam maior desequilíbrio de ligação (LD) dentro de grupos (clusters) e menor LD entre grupos. Os cálculos da medida relativa de LD entre pares de SNPs proposta por Hill e Robertson (1968) e representada por R^2 foram realizados utilizando o pacote “genetics” do *software* R (WARNES et al., 2013) e os valores de R^2 foram salvos em uma matriz **S**, denominada matriz de similaridade.

O agrupamento foi realizado com base na matriz de dissimilaridade **D**, obtida pela transformação da matriz **S** por meio da função monotônica $\mathrm{m(D)} = \exp(-\lambda \times S)$ (PETERNELLI e ROSA, 2013). Nesta função, que apresenta propriedades interessantes para a transformação proposta pelo método em

questão, λ é uma constante definida pelo pesquisador. Nesse trabalho foi utilizado λ= 3. De acordo com o valor de lambda, os altos valores de R² podem ser transformados em valores próximos, alocando os SNPs no mesmo grupo, ou seja, a transformação propicia um pré-agrupamento, forçando marcadores com alto LD a se agruparem. A Figura 2 abaixo descreve a transformação proposta pela função.

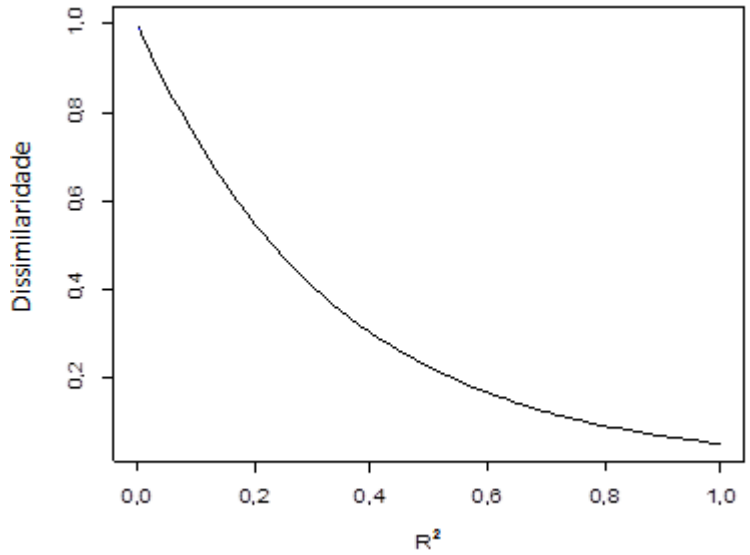

Figura 2. Descreve a função monotônica utilizada para transformar LD (medida de similaridade) em dissimilaridade. Função monotônica $m(D) = \exp(-\lambda \times S)$ para λ = 3.

Posteriormente o método de agrupamento de Ward (Ward, 1963) foi utilizado para agrupar os SNPs com base na matriz **D**. Em seguida, dentro de cada cluster formado pelo método de Ward, o procedimento *backward elimination*, baseando em regressão de mínimos quadrados, foi utilizado para exclusão dos SNPs não significativos. Esse método mantém no modelo somente as variáveis (SNPs) com *p*-valor parcial significativo (KLEINBAUM et al., 1997). Para tanto, a variável resposta (y) foi o fenótipo observado e o nível de significância foi arbitrário, neste estudo, visando comparações foi igual a $\alpha = 0,05;0,10;0,20;0,30;0,40;0,50;0,60;0,70$ Assim, cada cluster contribuicomq_iSNPs significantes (i = 1 até K - número total de clusters).

Por fim, todos os SNPs selecionados de cada cluster foram combinados, formando um único grupo final com $q < p$ SNPs, em que $q = \sum_{i=1}^{k} q_i$. Finalmente, o método BLASSO, descrito anteriormente em **2.3.2**, foi utilizado com esses SNPs selecionados para o ajuste final do modelo (etapa de ajuste).

2.5. Avaliação dos métodos

A comparação entre os métodos foi feita com base na capacidade preditiva $(r_{y\hat{y}})$ que é a correlação entre os valores genéticos genômicos preditos e os valores fenotípicos desregressados do grupo de validação. Também foi avaliado como medida de viés, o coeficiente de regressão do fenótipo regredido no genótipo predito (β) para os dados centrados na média e a acurácia $(r_{g\hat{g}})$ calculada pela razão entre capacidade preditiva$(r_{y\hat{y}})$e raiz da herdabilidade (DEKKERS, 2007).

Outra comparação entre os métodos foi realizada pelo coeficiente de coincidência, sendo utilizada a intensidade de seleção de 20 e 30 indivíduos dentro de cada grupo de validação, ou seja, 26% e 39%. Portanto, foi feita a comparação da coincidência dos indivíduos selecionados via fenótipo desregressados com os indivíduos selecionados pelo uso da GWS para todos os diferentes métodos.

Além disso, foram comparadas as acurácias da GWS em relação à seleção tradicional via BLUP para averiguar os ganhos genéticos esperados por unidade de tempo.

3. RESULTADOS E DISCUSSÃO

3.1. Controle de qualidade

No presente estudo, antes do controle de qualidade existiam 49042 SNPs e esse número foi reduzido para 24577 SNPs. Pela Figura 3 nota-se a redução dos SNPs após controle de qualidade dentro de cada cromossomo, sendo que a média de redução foi de 50% (Figura 3).

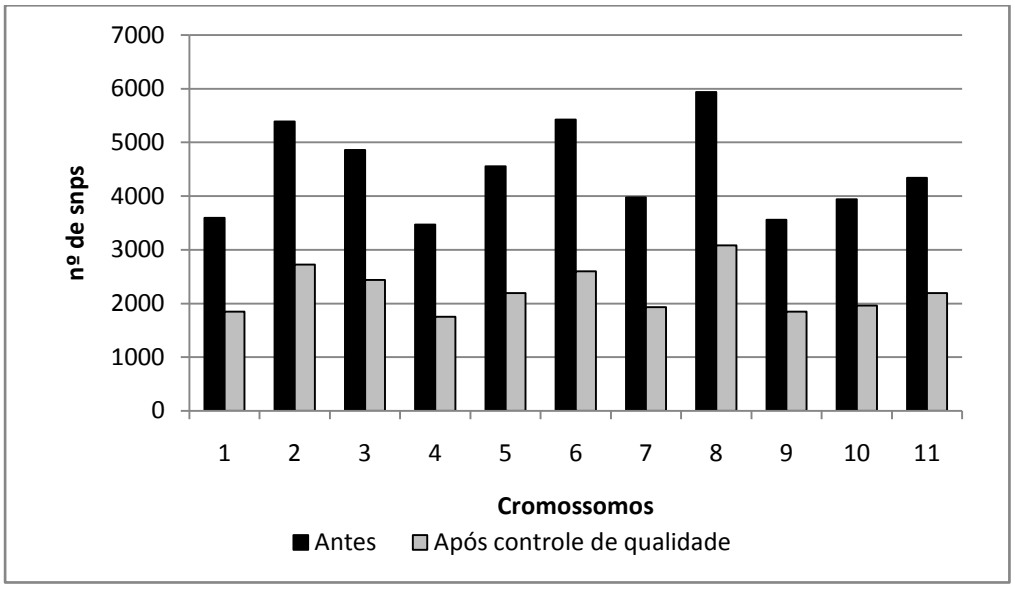

Figura 3 – Número de SNPs por cromossomo antes e após controle de qualidade.

3.2. Número de SNPs selecionados

O número de marcadores SNPs utilizados na seleção pelo método RR-BLUP-B foi determinado pelo ponto em que, o menor número de marcadores fosse selecionado mantendo capacidade preditiva muito próxima da obtida pelo método aplicado com todos os marcadores. Para tanto, foram utilizados subsets de marcadores (1000, 2500, 5000, 7500, 10000, 12500, 15000, 17500, 20000 e 22500) com maiores efeitos e validados em uma fração aleatória (10%) da população. Na Figura 4, verifica-se que tanto para ALT quanto para DAP esse ponto ficou em torno de 1000 SNPs.

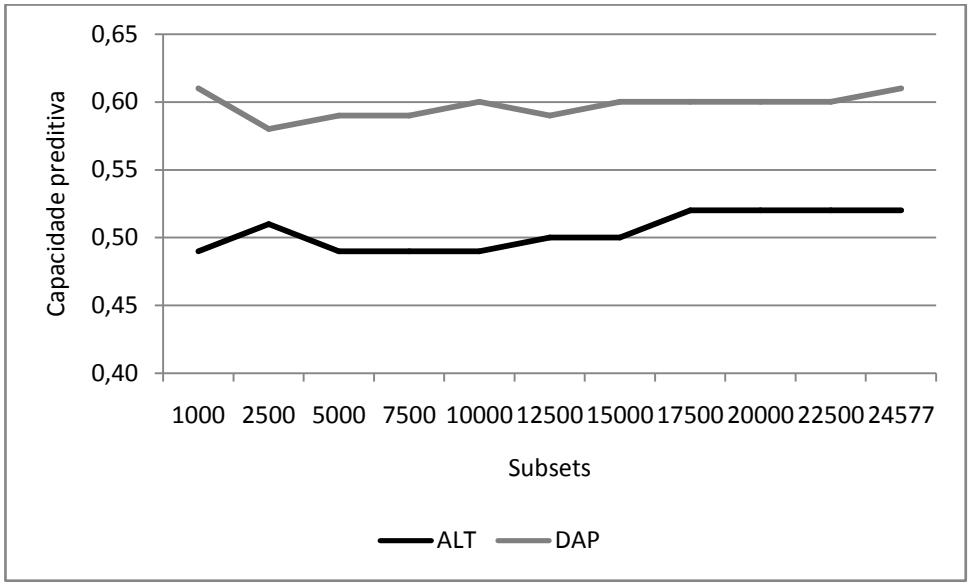

Figura 4–Resposta na capacidade preditiva ($r_{y\hat{y}}$)pelo método RR-BLUP-Bperante diferentes subsets de seleção de marcadores para altura (ALT) e diâmetro à altura do peito (DAP) em eucalipto.

No método PTMOD, não há uma predefinição do número de marcadores a ser escolhido, ficando a cargo dos *p*-valores utilizados na regressão *backward*. Neste estudo, pode-se averiguar pela Figura 5, que o aumento do p-valor causou um incremento médio de 908 SNPs selecionados para ALT e 1099 SNPs para DAP. Se for considerado apenas o menor p-valor (0,05), a redução média no número de SNPs para ambas as características foi de 67%. Já com o maior *p*-valor (0,70) a redução foi de 30%.

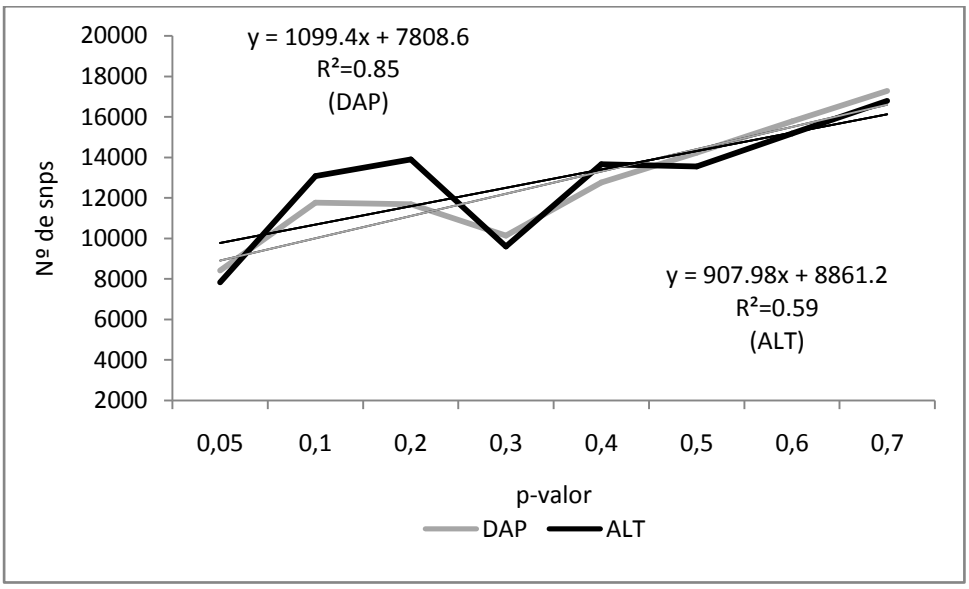

Figura 5 - Número de SNPs selecionados (média dos *folds*) pelo PTMOD para as variáveis altura (ALT) e diâmetro à altura do peito (DAP) para eucalipto.

Considerando a média dos p-valores por *fold* o método PTMOD obteve uma redução média no número de marcadores de 48,10% e 47,32% para DAP e ALT, respectivamente (Figura 6). Praticamente em todos os *fold*s (com exceção dos *folds* 1 e 2) foram selecionados mais SNPs para altura (ALT) do que para diâmetro à altura do peito (DAP) (Figura 6).

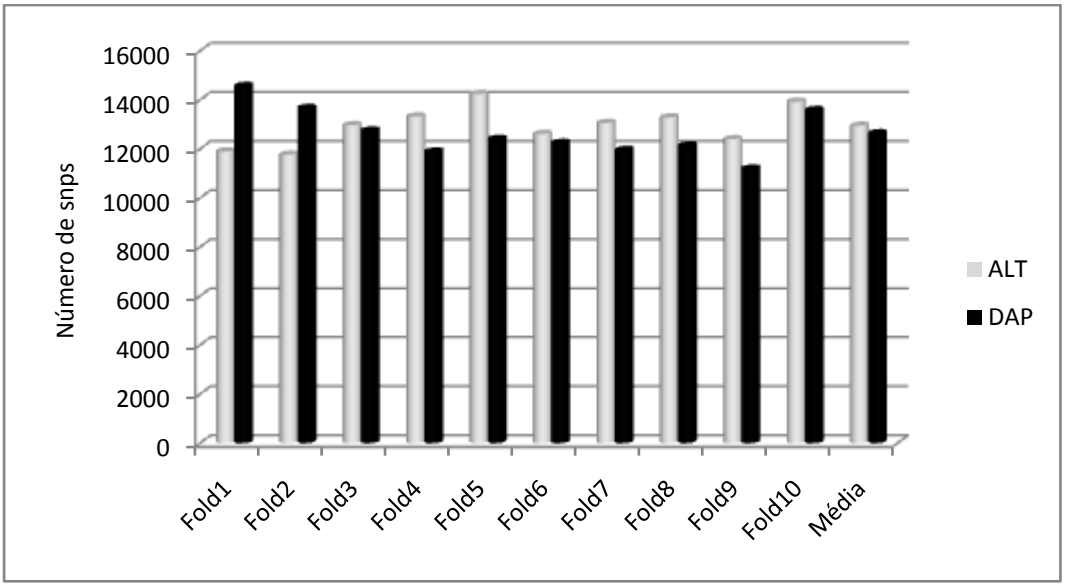

Figura 6 – Número de SNPs selecionados (média dos p-valores) por *fold*pelo PTMOD para as variáveis altura (ALT) e diâmetro à altura do peito (DAP) para eucalipto.

3.3. Análise comparativa entre os diferentes métodos de seleção genômica

As Figuras 7 e8mostram o comportamento das estimativas de capacidade preditiva dos diferentes métodos perante os 10-*fold*s realizados. Para o método PTMOD o valor da acurácia foi calculado pela média dos 10-*folds* utilizando o *p*-valor que proporcionou melhor capacidade preditiva e menor número de marcadores selecionados. Nota-se para ALT que os métodos apresentaram comportamento muito semelhante, com exceção do 4-*fold*, onde o método BLASSO foi bastante superior aos demais. Observa-se também, que o método RR-BLUP-B esteve quase sempre com capacidade preditiva inferior aos demais métodos em todos os *folds* (Figura 7).

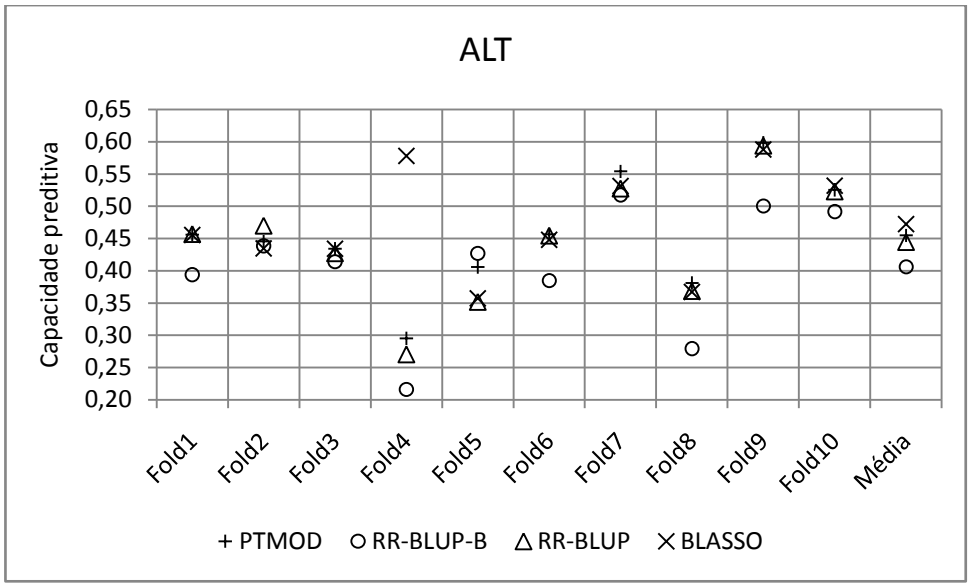

Figura 7 – Capacidade preditiva ($r_{y\hat{y}}$) para os métodos PTMOD, RR-BLUP-B, RR-BLUP e BLASSO para altura (ALT) em eucalipto.

Considerando a variável DAP, os métodos apresentaram novamente comportamento semelhante e mais uma vez o método BLASSO divergiu no 4-*fold* dos demais métodos, ficando também com capacidade preditiva bem superior nesse *fold* (Figura 8).

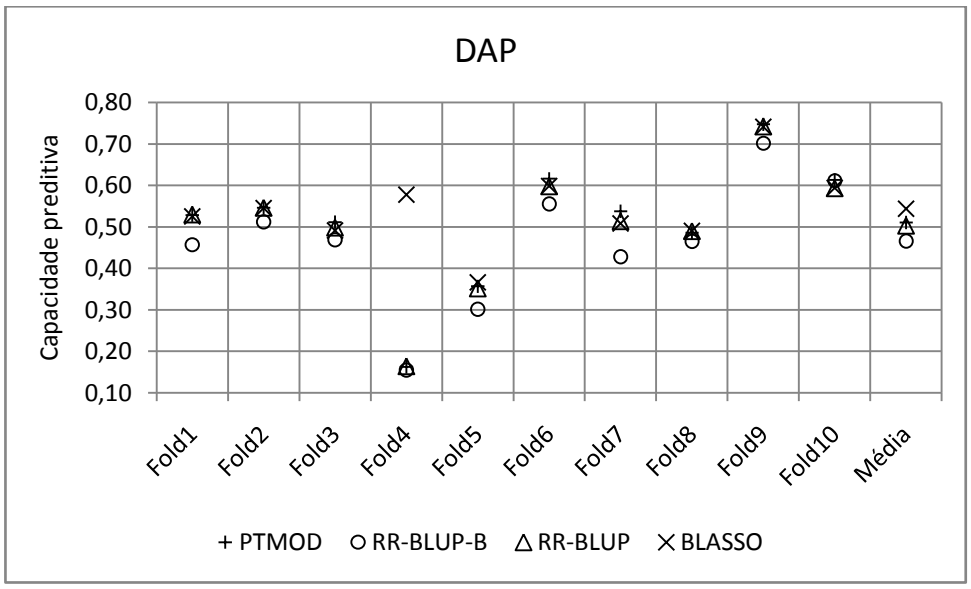

Figura 8 – Capacidade preditiva para os métodos PTMOD, RR-BLUP-B, RR-BLUP e BLASSO para diâmetro à altura do peito (DAP) em eucalipto.

As estimativas da capacidade preditiva, coeficiente de regressão e acurácia para média dos 10-*fold*s dos diferentes métodos podem ser observadas na Tabela 1. Para ambas as variáveis (ALT e DAP), verificou-se um ligeiro incremento na capacidade preditiva (em valor absoluto) pelo método PTMOD e ligeira redução pelo método RR-BLUP-B em relação ao método RR-BLUP sem seleção de marcadores. O método BLASSO foi o de maior capacidade preditiva entre os métodos (Tabela 1).

O coeficiente de regressão do fenótipo regredido no genótipo predito foi calculado como medida de viés de cada método. Os valores de β iguais a um representam, teoricamente, que não há viés na predição. Os métodos RR-BLUP, BLASSO e PTMOD não apresentaram coeficiente de regressão significativamente diferente de um, indicando assim, uma predição não viesada para ambas as características. Entretanto, para o método RR-BLUP-B, houve viés significativo (Tabela 1).

Tabela 1. Estimativas das médias para os 10-*fold*s da capacidade preditiva via correlação de Pearson ($r_{y\hat{y}}$), coeficiente de regressão do fenótipo regredido no genótipo predito (β) utilizando diferentes métodos para a variável altura (ALT) e diâmetro à altura do peito (DAP) em eucalipto

Métodos[1]	ALT			DAP		
	nº de SNPs	$r_{y\hat{y}}$	β	nº de SNPs	$r_{y\hat{y}}$	β
RR-BLUP	24577	0,44	1,06	24577	0,50	0,98
BLASSO	24577	0,47	0,99	24577	0,54	0,95
RR-BLUP-B	1000²	0,41	0,71*	1000²	0,48	0,70*
PTMOD	9094³	0,46	1,05	11075³	0,52	1,00

[1] RR-BLUP (MEUWISSEN et al., 2001); BLASSO (PARK E CASELLA, 2008); RR-BLUP-B (RESENDE et al., 2010); PTMOD (PETERNELLI E ROSA, 2013); ² Média por *fold*; ³ Média dos *fold*s utilizando o *p*-valor que proporcionou melhor capacidade preditiva e menor número de marcadores selecionados; * Significativo para o teste t a 5% de probabilidade para H0:β=1.

Nos resultados obtidos por Resende Júnior et al (2012a) em trabalho com *Pinus* utilizando 4853 SNPs, os autores empregaram o mesmo esquema de validação 10-*fold*s e obtiveram resultados de capacidade preditiva iguais a 0,39, 0,38 e 0,38 para ALT utilizando os métodos RR-BLUP, RR-BLUP-B (4630 SNPs selecionados) e BLASSO, respectivamente. Para DAP os resultados obtidos foram iguais 0,46 para todos os métodos e o número de SNPs selecionados para o RR-BLUP-B foi de 3800.Portanto, aqueles autores obtiveram resultados muito semelhantes entre os métodos RR-BLUP e RR-BLUP-B. Já no presente estudo, o método RR-BLUP-B reduziu um pouco mais a capacidade preditiva em relação ao RR-BLUP.

Percebe-se ainda, neste trabalho, que os métodos para seleção de SNPs obtiveram redução mais drástica no número de marcadores, e ainda assim, foi possível manter a capacidade preditiva das análises de seleção genômica (Tabela 1).

Ressalta-se também, que o método PTMOD foi superior ao RR-BLUP-B em capacidade preditiva e não apresentou viés na predição. No entanto, o número de SNPs selecionados foi maior (Tabela 1). É necessário considerar que pode não ser interessante reduzir fortemente o número de marcadores, conforme feito no RR-BLUP-B, pois como a GWS é baseada em alto desequilíbrio de ligação (LD) para capturar a maioria dos genes relacionados às características quantitativas e manter predições acuradas nas gerações

posteriores, uma intensa redução pode provocar perda considerável em capacidades preditivas para as próximas gerações.

As estimativas de capacidade preditiva do método PTMOD para todos os *fold*s e todos os *p*-valores utilizados estão apresentadas na Tabela 2. Nota-se que a utilização de diferentes *p*-valores não causou diferenças significativas na capacidade preditiva. No entanto, o número de marcadores selecionados variou consideravelmente dependendo do *p*-valor utilizado. Para ALT, ao utilizar o *p*-valor de 0,05 foram selecionados, na média dos 10-*fold*s, 7827 SNPs. Já com *p*-valor igual a 0,70, sobraram 16794 SNPs. Para DAP foram selecionados 8417 e 16982 SNPs, respectivamente. Portanto, considerando *p*-valor igual a 0,7, houve uma redução no número de SNPs iniciais de aproximadamente 68 e 66% para ALT e DAP, respectivamente.

Tabela 2. Estimativas das capacidades preditivas ($r_{y\hat{y}}$)do método PTMOD para altura (ALT) e diâmetro à altura do peito (DAP) em eucalipto, utilizando diferentes *p*-valores em cada *fold*

ALT	*p*-valor							
	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7
*Fold*1	0,46	**0,47**	**0,47**	0,46	0,46	0,46	0,46	0,46
*Fold*2	0,42	**0,45**	0,44	0,42	0,44	0,42	0,43	0,43
*Fold*3	0,42	**0,43**	**0,43**	**0,43**	**0,43**	**0,43**	**0,43**	**0,43**
*Fold*4	**0,30**	**0,30**	0,27	0,29	0,27	0,27	0,28	0,27
*Fold*5	**0,41**	0,38	0,37	0,38	0,36	0,36	0,36	0,35
*Fold*6	**0,46**	0,44	**0,46**	**0,46**	0,44	0,45	0,45	0,45
*Fold*7	**0,55**	0,53	0,52	0,53	0,53	0,53	0,53	0,53
*Fold*8	**0,38**	0,37	0,35	0,37	0,36	0,36	0,36	0,36
*Fold*9	**0,60**	**0,60**	0,59	**0,60**	0,59	0,59	**0,60**	0,58
*Fold*10	**0,53**	0,52	**0,53**	**0,53**	0,52	0,52	0,52	0,52
DAP	*p*-valor							
	0,05	0,1	0,2	0,3	0,4	0,5	0,6	0,7
*Fold*1	0,51	0,52	0,52	0,52	0,52	0,52	**0,53**	**0,53**
*Fold*2	0,54	0,54	0,55	**0,55**	0,54	0,54	0,53	0,54
*Fold*3	0,50	**0,51**	0,50	0,50	0,50	0,50	0,50	0,50
*Fold*4	0,16	**0,17**	0,16	0,16	0,14	0,16	0,16	0,16
*Fold*5	0,35	**0,38**	0,35	0,32	0,34	0,35	0,35	0,35
*Fold*6	**0,62**	**0,62**	0,61	**0,62**	0,61	0,61	0,60	0,60
*Fold*7	**0,54**	0,52	0,52	0,53	0,52	0,53	0,52	0,52
*Fold*8	0,47	0,48	0,48	0,48	0,47	0,48	**0,49**	**0,49**
*Fold*9	0,73	0,74	0,74	0,73	**0,75**	**0,75**	0,74	0,74
*Fold*10	**0,61**	0,60	0,60	0,60	0,60	0,60	0,60	0,60

Em negrito estão os valores máximos de capacidades preditivas.

A comparação entre a seleção fenotípica e seleção genômica de 20 indivíduos superiores em cada *fold* foi realizada para verificar o índice de coincidência da seleção embasada na predição dos fenótipos futuros (VGGs) dos indivíduos com a seleção pelo fenótipo observado. Ao selecionar os melhores 20 indivíduos por meio de seus VGGs para DAP, nota-se uma coincidência média de 11 indivíduos para todos os métodos (RR-BLUP, RR-BLUP-B, BLASSO e PTMOD) (Tabela 3). Ao aumentar o número de indivíduos selecionados (30 indivíduos), percebe-se uma coincidência média de 19 acertos para o método RR-BLUP e BLASSO, 18 para o RR-BLUP-B e 20 para PTMOD.

No mesmo sentido para variável ALT, ao selecionar os melhores 20 indivíduos por meio de seus VGGs, tem-se uma coincidência média de 10 acertos para o método RR-BLUP e RR-BLUP-B, nove para o BLASSO e para o PTMOD (Tabela 3). Já com aumento do número de indivíduos selecionados (30 indivíduos), há uma coincidência média de 17 acertos para o método RR-BLUP e PTMOD, 18 para o BLASSO e para o RR-BLUP-B.

Tabela 3.Índice de coincidência (IC) nos 10-*fold*s para diferentes métodos de seleção genômica para altura (ALT) e altura diâmetro do peito (DAP) em eucalipto

Métodos	DAP		ALT	
	IC1[1]	IC2[2]	IC1	IC2
RR-BLUP	0,56	0,63	0,50	0,58
BLASSO	0,55	0,64	0,47	0,60
RR-BLUP-B	0,54	0,61	0,50	0,60
PTMOD	0,57	0,65	0,47	0,58

[1] IC1 = índice de coincidência (média dos 10-*fold*s) entre a seleção dos 20 melhores indivíduos via fenótipos observados e preditos via seleção genômica; [2] IC2 = índice de coincidência (média dos 10-*fold*s) entre a seleção dos 30 melhores indivíduos via fenótipos observados e preditos via seleção genômica.

Percebe-se por esses resultados, que os métodos de seleção genômica demonstram eficiência em predizer os fenótipos futuros, uma vez que, apresentaram boa coincidência na seleção dos indivíduos superiores por seus fenótipos preditos com a seleção pelos fenótipos observados. É necessário ressaltar que, esses fenótipos apresentaram herdabilidade no sentido restrito,

calculada por meio do procedimento REML/BLUP, com valores iguais a 0,53 (DAP) e 0,42 (ALT). Dessa forma, considerando que as características utilizadas nesse estudo apresentaram herdabilidades moderadas, a predição dos fenótipos futuros nessas condições ou com herdabilidades superiores, representa boa coincidência com a observação dos fenótipos no campo, garantindo assim, um bom índice de acerto ao selecionar indivíduos pelos seus fenótipos preditos sem ter que levá-los ao campo.

Para avaliar o desempenho da seleção genômica em relação ao melhoramento tradicional foi estimado a acurácia da seleção baseada em BLUP (RESENDE et al., 2008) para comparar com a acurácia da seleção genômica. Considerando uma redução de 1/2 no ciclo reprodutivo, o incremento em eficiência por unidade de tempo para os diferentes métodos variou de 66-92% para ALT e de 65-85% para DAP. Com a possibilidade de redução de 1/3 no ciclo, o incremento obtido foi de 149-188% para ALT e de 148-178% para DAP (Tabela 4). Desse modo, independente do método utilizado, os ganhos genéticos por unidade de tempo foram superiores aos da seleção tradicional e podem ser mantidos utilizando a seleção de marcadores. Ademais, a redução de 1/2 ou 1/3 no ciclo reprodutivo na cultura do eucalipto é possível de ser feito devido à utilização de indutores hormonais.

Tabela 4. Eficiência da seleção genômica quando comparada com seleção fenotípica tradicional para altura (ALT) e diâmetro à altura do peito (DAP) em eucalipto

Características	Métodos	$r_{g\hat{g}}$ (BLUP)[1]	$r_{g\hat{g}}$ (GS)[2]	Eficiência[*]	Incremento em relação à seleção fenotípica (%)	Eficiência[**]	Incremento em relação à seleção fenotípica (%)
ALT	RR-BLUP	0,76	0,68	1,79	79	2,68	168
	RR-BLUP-B	0,76	0,63	1,66	66	2,49	149
	BLASSO	0,76	0,73	1,92	92	2,88	188
	PTMOD	0,76	0,71	1,87	87	2,80	180
DAP	RR-BLUP	0,80	0,69	1,73	73	2,59	159
	RR-BLUP-B	0,80	0,66	1,65	65	2,48	148
	BLASSO	0,80	0,74	1,85	85	2,78	178
	PTMOD	0,80	0,71	1,78	78	2,66	166

[1] Acurácia ($r_{g\hat{g}}$) com base na melhor predição linear não viesada (BLUP); [2] Acurácia ($r_{g\hat{g}}$) da seleção genômica (GS) (DEKKERS, 2007).[*]Eficiência calculada assumindo redução de 1/2 no ciclo reprodutivo;[**]Eficiência calculada assumindo redução de 1/3 no ciclo reprodutivo.

Para visualizar o comportamento dos efeitos dos SNPs após seleção e re-estimação de seus efeitos, foram plotados na Figura 9 as estimativas dos efeitos dos SNPs pelo RR-BLUP (com todos os SNPs) e pelo RR-BLUP-B, com 1000 SNPs selecionados,utilizando-se toda a população na estimação.

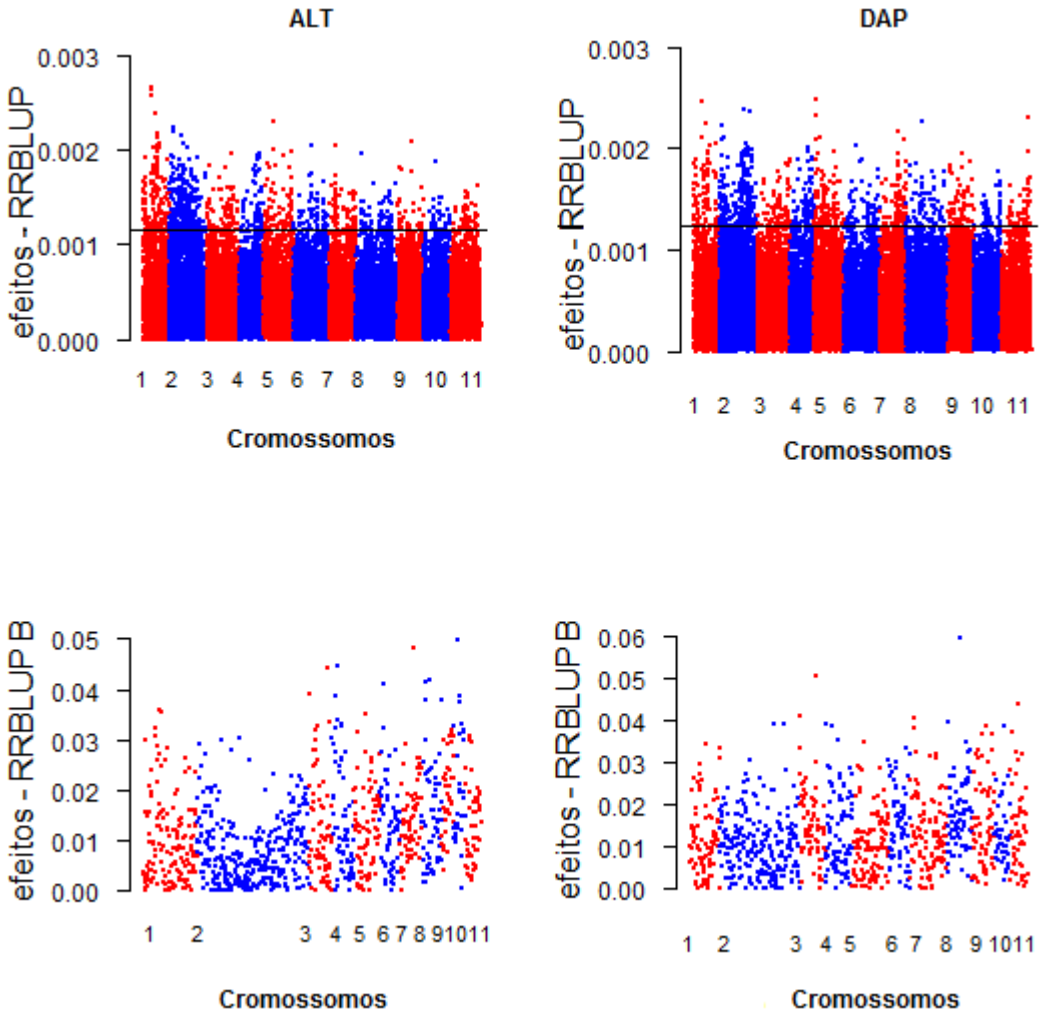

Figura 9– Efeitos dos SNPs para o RR-BLUP e RR-BLUP-Bpara variáveisaltura (ALT) e diâmetro à altura do peito (DAP)e ponto de corte (linha preta) para seleção dos 1000 SNPs.

Nota-se que após a seleção dos SNPs e re-estimação de seus efeitos, a importância relativa de cada SNPs se torna diferente. Por exemplo, para ALT, antes da seleção (RR-BLUP) alguns SNPs do cromossomo 1 possuíam os maiores efeitos, no entanto após a seleção (RR-BLUP-B) alguns SNPs do cromossomo 10 passaram a ter os maiores efeitos (Figura 9). Isto demonstra a grande influência que os SNPs causam uns aos outros nas análises, o que

pode influenciar e mascarar efeitos de possíveis SNPs ligados a QTLs e dificultar, por exemplo, a abordagem via *genome wide association studies* (GWAS) para descoberta de QTLs. Desse modo, a seleção de SNPs pode também contribuir para uma melhor abordagem associativa.

Ao identificar os SNPs selecionados para ALT e DAP,observa-se que há regiões importantes diferenciadas ao longo dos cromossomos para as duas características (Figura 9). No entanto, dentre os 1000 marcadores SNPs selecionados (no método RR-BLUP-B) para as duas variáveis em questão, houve 263 SNPs coincidentes. Portanto, pode-se concluir que existem genes comuns responsáveis pelo controle dessas características. Ademais, os SNPs selecionados estão distribuídos ao longo de todos os 11 cromossomos, porém, com maior concentração no cromossomo 2, para ambas as variáveis (Figura 10).

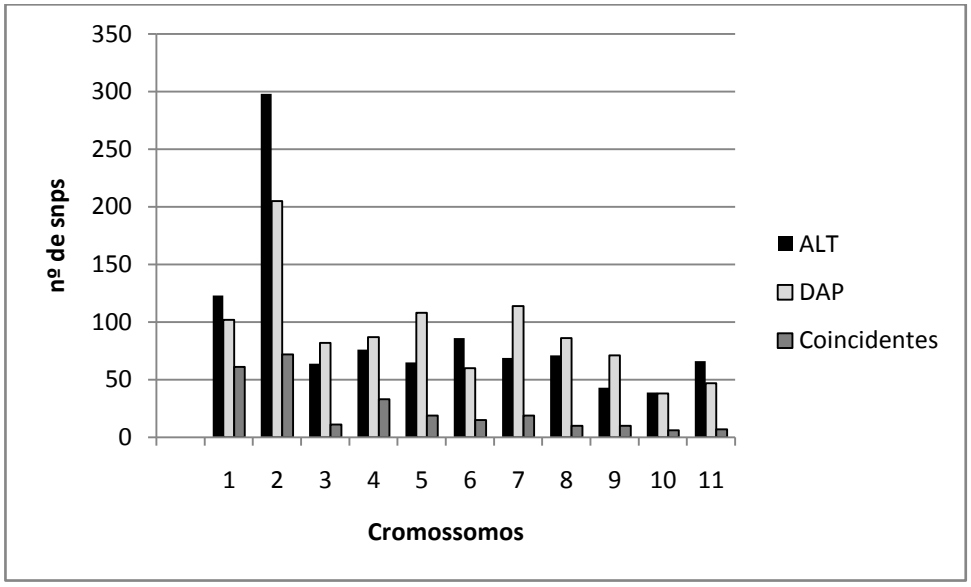

Figura 10 – Númerode SNPs selecionados por cromossomo após seleção por meio do método RR-BLUP-B para as variáveis altura (ALT) e diâmetro à altura do peito (DAP) em eucalipto utilizando toda a população na estimação.

Resende et al. (2012) identificaram marcadores DArTs com efeitos significativos para ALT e DAP ao longo dos 11 cromossomos. Neste estudo, notou-se, para a variável DAP, maior número de SNPs selecionados nos cromossomos 2, 5 e 7 (Figura 9). Grattapaglia et al. (1996) relataram a detecção de QTLs exatamente nestes cromossomos para DAP. Gion et al. (2011) também relataram QTLs nos cromossomos 2, 5 e 6. Já Freeman et al.

(2013) e Cappa et al. (2013) evidenciaram QTLs nos cromossomos 2, 3, 4, 5, 6, 7, 8, 10 e 11 e 3, 5, 7 e 10, respectivamente.

Outra opção de seleção de SNPs foi investigada no presente estudo utilizando o método RR-BLUP-B da seguinte maneira:i)análise com osSNPs coincidentes presentes entre os 1000 SNPs selecionados via RR-BLUP-B tanto para ALT quanto para DAP; ii) análise com todos os SNPs selecionados tanto em ALT quanto em DAP. Isso foi feito em cada *fold* e as estimativas da capacidade preditiva e viés foi calculado pela média de todos os 10-*fold*s (Tabela 5).

Tabela 5. Estimativas da capacidade preditiva ($r_{y\hat{y}}$),coeficiente de regressão do fenótipo regredido no genótipo predito (β) pelo método RR-BLUP-B para as variáveis altura (ALT) e altura diâmetro do peito (DAP) em eucalipto em duas situações

Situações[1]	ALT		DAP	
	$r_{y\hat{y}}$	β	$r_{y\hat{y}}$	β
RR-BLUP-B	0,41	0,71*	0,47	0,70*
I	0,41	0,78*	0,40	0,79*
II	0,47	0,77*	0,47	0,73*

[1]I = Análise com osSNPs coincidentes que estavam presentes dentre os 1000 SNPs selecionados via RR-BLUP-B-modificado tanto para ALT quanto para DAP no *fold*1; II = Análise com todos os SNPs selecionados tanto em ALT quanto em DAP.* Significativo para o teste t a 5% de probabilidade para H0: β=1.

Ao utilizar apenas os SNPs coincidentes para predizer os fenótipos para ALT, a capacidade preditiva obtida foi igual à obtida com os 1000 SNPs selecionados, e recuperou 85% da obtida em DAP. Além disso, as predições apresentaram menor viés. Ademais, ao utilizar todos os SNPs selecionados (exclusivos para ALT + exclusivos para DAP + coincidentes), as análises retomaram toda a capacidade preditiva para DAP e suplantou a obtida com os 1000 SNPs para ALT (Tabela 5).

Esse resultado demonstra que apesar da seleção de marcadores ser característica específica, a união desses marcadores selecionados pode proporcionar predições eficientes para ambas as características. Dessa forma, pode-se selecionarSNPs para mais de uma variável e fazer uma nova genotipagem mantendo a capacidade preditiva para ambas as variáveis.

Dessa forma, conclui-se pelo presente estudo, que a seleção de marcadores pode manter a capacidade preditiva dos métodos de GWS muito próximas as obtidas utilizando-se todos os marcadores. Assim, torna-se possível a redução nos custos de futuras genotipagens para posteriores gerações. Ademais, neste estudo o método PTMOD foi superior ao RR-BLUP-B, porém, com menor redução no número de SNPs selecionados, o que garante maior LD nas futuras genotipagens.

3.4. SNPs *versus* DArTs

A utilização de marcadores SNPs e DArTs foi comparada no presente estudo para averiguar as implicações na seleção genômica. O genoma do eucalipto possui tamanho aproximado de 1300 cM (BRONDANI et al., 2006), logo, a genotipagem com SNPs possibilitou a obtenção de marcadores com espaçamento médio de 0,05 cM. Já a genotipagem com DArT o espaçamento entre marcadores foi igual a 0,4 cM. Dessa forma, a genotipagem com SNPs foi bem mais densa do que a com DArTs.

Nesse contexto esperava-se uma melhor capacidade preditiva na análise com os marcadores SNPs, porém, como demonstrado na Tabela 6, as capacidade preditivas foram um pouco superiores utilizando-se DArTs. Sendo assim, supõe-se que a genotipagem com SNPs se aproximou de muitas regiões não codificadoras (íntrons) enquanto a genotipagem com DArTs, foi mais precisa na captura de genes relacionados com as características ALT e DAP.

Tabela 6. Estimativas da capacidade preditiva ($r_{y\hat{y}}$), coeficiente de regressão do fenótipo regredido no genótipo predito (β) pelo método RR-BLUP para as variáveis altura (ALT) e diâmetro à altura do peito (DAP) em uma mesma população de eucalipto utilizando marcadores SNPs e DArTs

	SNP		DArT	
	ALT	DAP	ALT	DAP
Nº de marcadores	24577	24577	3129	3129
$r_{y\hat{y}}$	0,44	0,50	0,48	0,53
β	1,05	0,98	1,00	1,00

Portanto, conclui-se que, para a população de eucalipto e variáveis empregadas neste trabalho, os marcadores DArT com menor densidade apresentaram melhores predições na GWS em relação a utilização de marcadores SNPs com maior densidade.

3.5. Modificação no método RR-BLUP-B

Com a modificação proposta neste trabalho o método RR-BLUP-B obteve resultados de capacidade preditiva igual a 0,67 e 0,76 para ALT e DAP, respectivamente. O coeficiente de regressão médio foi igual a 1,04 para ambas as características. Dessa forma, o incremento médio na capacidade preditiva com a modificação ficou em cerca de 60% em relação ao método RR-BLUP-B original (Tabela 7). Além disso, não houve diferença entre selecionar os SNPs pela média dos efeitos (RR-BLUP-B1) ou pela média dos postos dos efeitos (RR-BLUP-B2).

Tabela 7. Estimativas da capacidade preditiva ($r_{y\hat{y}}$), coeficiente de regressão do fenótipo regredido no genótipo predito (β) pelo método RR-BLUP-B e RR-BLUP-B modificado para as variáveis altura (ALT) e diâmetro à altura do peito (DAP) em eucalipto

Métodos	ALT		DAP	
	$r_{y\hat{y}}$	β	$r_{y\hat{y}}$	β
RR-BLUP-B	0,41	0,71*	0,48	0,70*
RR-BLUP-B1[1]	0,67	1,04	0,76	1,04
RR-BLUP-B2[1]	0,67	1,04	0,76	1,04

[1] RR-BLUP-B1 = modificação do RR-BLUP-B proposta no presente trabalho, utilizando o módulo da média dos efeitos dos snps para seleção; [1] RR-BLUP-B2 = modificação do RR-BLUP-B proposta no presente trabalho, utilizando o módulo da média dos postos dos efeitos dos snps para seleção. * Significativo para o teste t a 5% de probabilidade para H0: β=1.

Esse resultado pode ser explicado pelo fato de que, no RR-BLUP-B, os marcadores são ranqueados e selecionados em cada *fold*, ou seja, para cada ciclo diferentes marcadores são selecionados. Já na modificação proposta no

presente trabalho (RR-BLUP-B modificado), os marcadores são ranqueados pela média dos efeitos em todos os *fold*s. Assim, espera-se que, em média, os marcadores com efeitos maiores em todos os *fold*s sejam selecionados, o que garante melhor capacidade preditiva, já que os marcadores foram testados com todos os indivíduos da população.

Entretanto, como essa validação utilizou reamostragem, ou seja, os indivíduos do grupo de validação também participaram, em certo momento, do grupo de estimação, foi feito um novo procedimento para testar o método RR-BLUP-B e a modificação proposta. Nesse sentido, retirou-se uma amostra aleatória (~10%) que ficou totalmente fora da estimação, sendo utilizada somente para validação após aplicação dos dois métodos (RR-BLUP-B e modificação) na porção de estimação.

Com essa porção totalmente fora da estimação os dois métodos apresentaram mesma capacidade preditiva, sendo de 0,48 para ALT e 0,61 para DAP. Portanto, indica-se a validação independente como melhor forma de validação, já que esta não proporciona uma capacidade preditiva superestimada.

4. REFERÊNCIAS BIBLIOGRÁFICAS

BRONDANI R.P. et al. (2006). A microsatellite-based consensus linkage map for species of Eucalyptus and a novel set of 230 microsatellite markers for the genus. BMC Plant Biology 6: 20.

CAPPA, E.P. (2013). Impacts of Population Structure and Analytical Models in Genome-Wide Association Studies of Complex Traits in Forest Trees: A Case Study in *Eucalyptus globulus*. Plos One, 8(11): e81267.

CAVALCANTI, J.J.V. et al. (2012). Predição simultânea dos efeitos de marcadores moleculares e seleção genômica ampla em cajueiro. Revista Brasileira de Fruticultura, 34: 840-846.

DE LOS CAMPOS, G.; RODRIGUEZ, P.P. (2012). BLR: Bayesian Linear Regression. R package version 1.3.Disponível em: http://CRAN.R-project.org/package=BLR.

DEKKERS, J.C.M. (2007). Prediction of response to marker-assisted and genomic selection using selection index theory.Journal of Animal Breeding and Genetics, 124: 331-341.

ENDELMAN, J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4: 250-255.

FARIA, G.M.P. et al. (2013). Controle de qualidade de dados genotípicos para estudos genômicos em clones de eucalipto. In: 7º Congresso Brasileiro de Melhoramento de Plantas, Uberlândia - MG, p.1068-1071.

FREEMAN, J.S. (2013). Stability of quantitative trait loci for growth and wood propertiesacross multiple pedigrees and environments in Eucalyptus globulus. New Phytologist, 198: 1121-1134.

FRITSCHE-NETO, R. et al. (2012). Genome wide selection for root traits in topical maize under stress conditions of nitrogen and phosphorus. ActaScientiarum Agronomy. 34: 389-396.

GION, J.M. et al. (2011). Comprehensive genetic dissection of wood properties in a widely-grown tropical tree: *Eucalyptus*. BMC Genomics, 12: 301.

GOUY, M. et al. (2013).Experimental assessment of the accuracy of genomic selection in sugarcane. Theoretical Applied Genetics, 126:2575-2586.

GRATTAPAGLIA, D. (1996). Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. Genetics 144: 1205-1214.

HILL, W.G.; ROBERTSON, A. (1968).A linkage disequilibrium in finitive populations. Theoretical Applied Genetics, 33: 226-231.

KLEINBAUM, D.G. et al. (1997). Applied regression analysis and other multivariable methods. Pacific Grove: Duxbury Press, 816p.

KOHAVI, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, San Francisco, p. 1137-1143.

KUMAR, S. et al. (2012).Genomic selection for fruit quality traits in apple (*Malus x domestica*Borkh.).Plos One, 7(5):e36674.

LEGARRA, A. et al. (2008). Performance of genomic selection in mice. Genetics, 180:611-618.

LYNCH, M.; WALSH, B. (1998).Genetics and analysis of quantitative traits. Sunderland: Sinauer Associates, 980p.

MEUWISSEN, T.H.E. et al. (2001). Prediction of total genetic value using genome-wide dense marker maps.Genetics, 157: 1819-1829.

OLIVEIRA, E.J. et al. (2012). Genome-wide selection in cassava.Euphtyca, 187:263-276.

PARK, T.; CASELLA, G. (2008).The Bayesian LASSO. Journal of the American Statistical Association, 103: 681-686.

PÉREZ, P. et al. (2010). Genomic-Enabled prediction basedon molecular markers and pedrigree using the bayesian linear regression package in R. The Plant Genome, 3:106-116.

PETERNELLI, L.A.; ROSA, G. (2013). A novel approach for subset selection of SNP markers for cost-effective implementation of genomic selection. In: 55th Annual Maize Genetics Conference, 2013, Saint Charles, IL. Abstracts of the 55th Annual Maize Genetics Conference, 2013.v.1. p. 1. Disponível em: http://www.maizegdb.org/maize_meeting/abstracts/2013Program.pdf.

R DEVELOPMENT CORE TEAM (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Áustria, 2013. Disponível em http://www.R-project.org.

RESENDE JÚNIOR, M.F.R. (2010). Seleção genômica ampla no melhoramento vegetal. Dissertação (Mestrado em Genética e Melhoramento), Universidade Federal de Viçosa. 67f.

RESENDE JÚNIOR, M.F.R. et al. (2012a). Accuracy of genomic selection method in a standard dataset of loblolly pine.Genetics, 190: 1503-1510.

RESENDE JÚNIOR, M.F.R. et al. (2012b). Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments.New Phytologist, 193:617-624.

RESENDE, M.D.V. (2007). Software SELEGEN-REML/BLUP: Sistema Estatístico e Seleção Genética Computadorizada via Modelos Lineares Mistos. Embrapa Florestas.

RESENDE, M.D.V. et al. (2005). Seleção recorrente e o melhoramento genético do eucalipto no Brasil. In: Simpósio sobre Atualização em Genética e Melhoramento de Plantas, 2005, Lavras. Anais, v.1, p.59-84.

RESENDE, M.D.V. et al. (2008). Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. Pesquisa Florestal Brasileira, 56:63-77.

RESENDE, M.D.V. et al. (2010). Computação da seleção genômica ampla (GWS). Colombo: Embrapa Florestas. 79p.

RESENDE, M.D.V. et al. (2011). Métodos estatísticos na seleção genômica ampla. Colombo: Embrapa Florestas. 106p.

RESENDE, M.D.V. et al. (2012). Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência Bayesiana (MCMC), regressão aleatória multivariada e estatística espacial. Viçosa: Departamento de Estatística. Disponível em: http://www.det.ufv.br/ppestbio/corpo_docente.php.

SUBEDI, S. et al. (2012).SNP selection for predicting a quantitative trait. Journal of Applied Statistics, 4:600-613.

USAI, M.G. et al. (2009). LASSO with cross-validation for genomic selecion. Genetic Research, 91:427-436.

VERBYLA, K.L. et al. (2010). Predicting energy balance for dairy cowsusing high-density single nucleotide polymorphism information. Journal of Dairy Science, 93:2757-2764.

WARD, J.H. (1963). Hierarquical grouping to optimize an objective function. Journal of the American Statistical Association, 58:236-244.

WARNES et al. (2013).Population Genetics R package.Disponível em: http://cran.r-project.org/web/packages/genetics/genetics.pdf.