

Validação e correção de fenótipos na seleção genômica ampla

Ísis Fernanda de Almeida⁽¹⁾, Cosme Damião Cruz⁽²⁾ e Marcos Deon Vilela de Resende⁽³⁾

⁽¹⁾Instituto Federal do Triângulo Mineiro, Campus Uberlândia, Fazenda Sobradinho s/nº, Zona Rural, Caixa Postal 1020, CEP 38400-970 Uberlândia, MG, Brasil. E-mail: isis@iftm.edu.br ⁽²⁾Universidade Federal de Viçosa, Avenida Peter Henry Rolfs, s/nº, CEP 36570-900 Viçosa, MG, Brasil. E-mail: cdcruz@ufv.br ⁽³⁾Embrapa Florestas, Estrada da Ribeira, km 111, Guaraituba, Caixa Postal 319, CEP 83411-000 Colombo, PR, Brasil. E-mail: marcos.deon@gmail.com

Resumo – O objetivo deste trabalho foi avaliar a influência da distribuição dos efeitos de QTL, do tipo de população de validação e da correção dos fenótipos sobre a acurácia da seleção genômica ampla. Duas populações de irmãos completos, com 500 indivíduos, foram simuladas, tendo-se considerado, genotipicamente, 1.000 locos marcadores – 100 ligados a QTL. Os efeitos de QTL apresentaram distribuição uniforme ou exponencial. Na validação 1, uma amostra com 100 indivíduos constituiu a população de validação; na validação 2, aplicou-se a validação cruzada, com amostra de 100 indivíduos em cinco repetições; e na 3, uma segunda geração constituiu a população de validação. As metodologias de análise utilizadas foram RR-Blup e Blasso, com modelos mistos para correção dos fenótipos. Sem correção fenotípica, a distribuição exponencial proporcionou maiores acurácias, e o método Blasso foi mais acurado com essa distribuição; enquanto o RR-Blup foi mais acurado com a distribuição uniforme. Nesse cenário sem correção, as validações 1 e 3 foram mais acuradas. Com correção, as distribuições exponencial e uniforme produziram acurácias similares, e o método Blasso mostrou-se mais acurado para ambas. Nesse cenário, as validações 1 e 2 foram mais acuradas. No geral, o método RR-Blup foi mais acurado, e o Blasso menos viciado.

Termos para indexação: acurácia, Blasso, genotipagem em larga escala, marcadores moleculares, RR-Blup.

Validation and phenotypic correction in genome-wide selection

Abstract – The objective of this work was to evaluate the effect of the distribution of QTL effects, of the type of validation population, and of phenotype adjustment in the accuracy of genome-wide selection. Two populations of full siblings with 500 individuals were simulated, with 1,000 loci markers being genotypically considered – 100 linked to QTL. The QTL effects had uniform or exponential distribution. For validation 1, a 100-individual sample constituted the validation population; in validation 2, cross validation was applied, with a 100-individual sample in five replicates; and in validation 3, a second generation formed the validation population. The analysis methodologies used were RR-Blup and Blasso, with mixed models for phenotype correction. Without phenotypic correction, the exponential distribution led to higher accuracies, and the Blasso method showed greater accuracy with this distribution; while RR-Blup was more accurate with uniform distribution. In this scenario without correction, validations 1 and 3 were more accurate. With phenotypic correction, exponential and uniform distributions led to similar accuracies, and the Blasso method proved more accurate for both of them. In this scenario, validations 1 and 2 were more accurate. Generally, the RR-Blup method was more accurate, and the Blasso method was less biased.

Index terms: accuracy, Blasso, large-scale genotyping, molecular markers, RR-Blup.

Introdução

Com os avanços de tecnologias de genotipagem em larga escala, tornou-se possível uma cobertura completa do genoma, o que levou à criação de uma nova forma de utilização da informação genética gerada com essa metodologia. A seleção genômica ampla (“genome wide selection”, GWS), proposta por Meuwissen et al. (2001), consiste na utilização simultânea de centenas – ou milhares – de marcadores,

os quais cobrem o genoma de maneira densa, de forma que todos os genes de um caráter quantitativo estejam em desequilíbrio de ligação com pelo menos uma parte dos marcadores.

Na GWS, obtém-se primeiramente o polimorfismo dos marcadores para, em seguida, estimarem-se seus efeitos com base em dados fenotípicos de uma população conhecida como população de estimação. Uma vez que os efeitos são modelados e estimados, eles precisam ser

testados em uma população de validação, para que se possam selecionar os marcadores que explicam parte da variância genética do caráter em estudo, e então incorporar efetivamente essa informação à etapa de seleção do programa de melhoramento. Na população de validação, utiliza-se um conjunto de dados menor do que na população de estimação, e apenas indivíduos genotipados e fenotipados para a característica de interesse são contemplados. Essa população consiste em uma amostra independente, utilizada para testar e verificar as acurácias das equações de predição de valores genômicos.

Durante as análises, os fenótipos devem ser corrigidos para os efeitos genéticos dos genitores, ocasião em que se trabalha basicamente com o efeito da segregação mendeliana desregressada, que capta a associação de alelos de marcas e dos locos de caracteres quantitativos (“quantitative trait loci”, QTL); ou seja, busca-se capturar efeitos genéticos explicados pelo desequilíbrio de ligação, e não pelo parentesco ou genealogia (Resende, 2008).

A implementação da seleção genômica, no entanto, impõe vários desafios estatísticos e computacionais, como a dimensionalidade do modelo, a colinearidade entre marcas e a complexidade das características quantitativas. Para que esses aspectos sejam devidamente contemplados, vários métodos têm sido propostos, os quais diferem entre si pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo. Entre eles, o método RR-Blup (“ridge regression-best linear unbiased prediction”) estima simultaneamente os efeitos de todas as marcas (Whittaker et al., 2000; Meuwissen et al., 2001), as quais são consideradas como efeitos aleatórios com variância comum; ou seja, assume-se que todos os marcadores contribuem igualmente para a variação genética. No entanto, essa hipótese pode não ser apropriada para casos em que algumas marcas estão localizadas em regiões não associadas à variância genética, enquanto outras estão localizadas em regiões associadas ao QTL (Goddard & Hayes, 2007). Para contornar este problema, muitos autores propuseram metodologias que utilizam efeito “shrinkage” específico para cada marcador. No contexto bayesiano, isto pode ser implementado com uso de distribuições a priori para os efeitos dos marcadores, a exemplo do método Lasso (“least absolute shrinkage and selection operator”) Bayesiano (Blasso), de Park & Casella (2008).

Após estimação e validação de um conjunto de marcadores, eles são utilizados em uma população que dispõe apenas de genotipagem (população de seleção). Zhong et al. (2009) relatam que a população de estimação pode afetar a acurácia da seleção genômica; sendo assim, pressupõe-se que o tipo de população de validação também pode influenciar seus valores genômicos.

O objetivo deste trabalho foi avaliar a influência da distribuição dos efeitos de QTL, do tipo de população de validação e da correção dos fenótipos sobre a acurácia da seleção genômica ampla.

Material e Métodos

Foram realizadas dez simulações de estruturas populacionais. A estrutura de cada simulação foi criada para apresentar os conjuntos de dados fenotípicos e genotípicos de cada indivíduo dentro de uma população. A primeira população simulada (G1), constituída por famílias de irmãos completos, foi derivada do cruzamento, planta a planta, entre duas populações genitoras P1 e P2, em equilíbrio de Hardy-Weinberg, e gerou uma população com 25 famílias de irmãos completos, com 20 indivíduos em cada família. A segunda, a geração dois (G2), foi obtida por meio do acasalamento ao acaso dos irmãos completos, dentro de cada família.

A seleção genômica requer o uso de uma população de estimação para estimar os efeitos dos marcadores, e uma população de validação, para analisar a eficiência da estimação destes efeitos na recuperação do valor genômico. Com isso, a geração G1 foi obtida com o intuito de constituir tanto a população de estimação como a de validação de duas formas de validação – V1 e V2, descritas adiante –, bem como a população de estimação de uma terceira forma de validação (V3). A geração G2, por sua vez, compôs a população de validação de V3.

Em cada simulação de dados, foram considerados 1.000 locos marcadores – espaçados igualmente em cinco grupos de ligação – que expressavam dois alelos codominantes. Entre os 1.000 locos genômicos simulados, 100 foram caracterizados como associados à característica quantitativa (QTL). Foi estabelecido que os 100 primeiros locos genotipados foram os controladores dessa característica. Os dados genotípicos e fenotípicos foram simulados, tendo-

se considerado: ausência de dominância e diferentes herdabilidades ($h^2 = 0,20$; $0,40$; e $0,60$) e formas de distribuição do efeito de QTL. Neste último caso, simularam-se dois cenários de efeitos de QTL: 1, genes com efeitos de distribuição uniforme, para reproduzir características controladas por vários genes de pequenos efeitos; e 2, genes com efeitos de distribuição exponencial, para retratar caracteres controlados por poucos genes de grandes efeitos, bem como de vários genes de pequenos efeitos.

Nas análises genômicas, foram utilizadas as metodologias RR-Blup e Blasso. Na metodologia RR-Blup foi usado o seguinte modelo para estimar os efeitos dos marcadores, conforme Resende et al. (2008): $y = Xb + Za + \varepsilon$, em que y é o vetor de observações fenotípicas, b é o vetor de efeitos fixos, a é o vetor dos efeitos aleatórios dos marcadores e ε refere-se ao vetor de resíduos aleatórios. X e Z são as matrizes de incidência para b e a . A estrutura de médias e variâncias no modelo em questão é definida como: $a \sim N(0, G)$; e $\varepsilon \sim N(0, R = I\sigma_\varepsilon^2)$; $E(y) = Xb$; $Var(y) = V = ZGZ' + R$ e $G = I\sigma_g^2/n$, em que σ_g^2 refere-se à variância genética total do caráter, e σ_ε^2 é a variância residual.

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

O método Lasso combina “shrinkage” (regularização) com seleção de variáveis, e envolve o seguinte problema de otimização, via minimização de:

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

é a soma dos valores absolutos dos coeficientes de regressão. O parâmetro de suavização λ controla a intensidade da regularização. Quando este último parâmetro é igual a zero, não há regularização. No Lasso Bayesiano (Blasso), esse parâmetro controla a precisão da distribuição a priori atribuída aos coeficientes de regressão. A implementação desse tipo de regularização envolve encurtamento mais forte, para que alguns coeficientes de regressão tenham valores iguais a zero. Isso pode ser implementado via análise Bayesiana do Lasso (De los Campos et al., 2009). Esta implementação impõe, como

distribuição a priori dos p coeficientes de regressão, um produto de densidades exponenciais duplas:

$$p(\beta|\lambda) = \prod_{j=1}^p \frac{\lambda}{2} \exp(-\lambda |\beta_j|).$$

No primeiro cenário de validação (V1), 100 indivíduos, constituídos pelas últimas cinco famílias simuladas, foram retirados da amostra para servir como população de validação. Assim, a população de estimação foi composta por 400 indivíduos, com uma população de validação de 100. No segundo cenário (V2), foi praticada a validação cruzada, pela reamostragem de um grupo de indivíduos com o procedimento Jackknife (Heltsh & Forrester, 1983), que se baseia na divisão do conjunto de C dados amostrais, em g grupos de tamanho igual a k , de forma que $C = gk$. Em cada um dos g grupos, k indivíduos são retirados para a formação da população de validação. No presente estudo, tomou-se $k=100$, a fim de se manter o padrão de comparação com a V1 (população de validação composta por 100 indivíduos). No terceiro e último cenário (V3), avaliou-se o uso de uma população de validação independente (G2), gerada pelo acasalamento entre os indivíduos da população de estimação (G1); neste caso, foram utilizados 500 indivíduos para a população de estimação e 500 para a população de validação.

As validações V1 e V3 foram efetuadas com o uso do programa R, e a V2 com o Selegen Genômica (Resende, 2007). A eficiência da GWS foi avaliada por meio da acurácia seletiva, dada pelos coeficientes de correlação que envolviam valores genéticos genômicos verdadeiros (V_{gen} fixados na simulação) e valores genômicos preditos (VGG) em cada metodologia de análise.

Para a correção dos dados de fenótipo, utilizou-se o software Selegen-REML/Blup, que usa análise de modelos mistos para identificar o efeito estimado de cada indivíduo dentro da família. Após a correção desses dados para o efeito de genitores, eles foram utilizados nos estudos de cada cenário analisado. Em caso de testes de progênie de irmãos germanos, o modelo ajustado é dado por $y = Xb + Zg + Wc + \varepsilon$. O valor individual corrigido (\hat{f}_{ijk}^*), para o valor genético médio de seus genitores (\hat{g}_j e \hat{g}_k) com matriz de incidência Z , é dado pela expressão: $\hat{f}_{ijk}^* = (y_i - 0,5 \hat{g}_j - 0,5 \hat{g}_k - \hat{b}_i - \hat{c}_i)$, em que \hat{b}_i e \hat{c}_i são os efeitos estimados de blocos (se houver) e de capacidade específica (se o modelo for com dominância), com as respectivas

matrizes de incidência X e W . Neste estudo, não foram simulados efeitos de bloco e de capacidade específica.

As simulações foram implementadas com o programa Genes (Cruz, 2013), e as análises estatísticas foram feitas com: Selegen Genômica e Selegen-REML-Blup (Resende, 2007); e R, via pacote RR-Blup (Endelman et al., 2011) e via pacote BLR (De los Campos et al., 2009).

Resultados e Discussão

As estimativas dos efeitos de marcadores variaram de acordo com a pressuposição do modelo

genético associado a cada metodologia de análise, e apresentaram comportamento semelhante nas diferentes herdabilidades simuladas. A título de ilustração, o gráfico referente à herdabilidade de 0,60 é apresentado na Figura 1. Por um lado, a análise realizada com o RR-Blup foi mais precisa em retratar o comportamento da característica uniforme, em que se observa menor variação entre os efeitos de marcadores estimados. Por outro lado, o Blasso foi mais consistente em descrever as variações de genes com distribuição exponencial, em que a variação entre os genes de maior e menor efeito é mais acentuada.

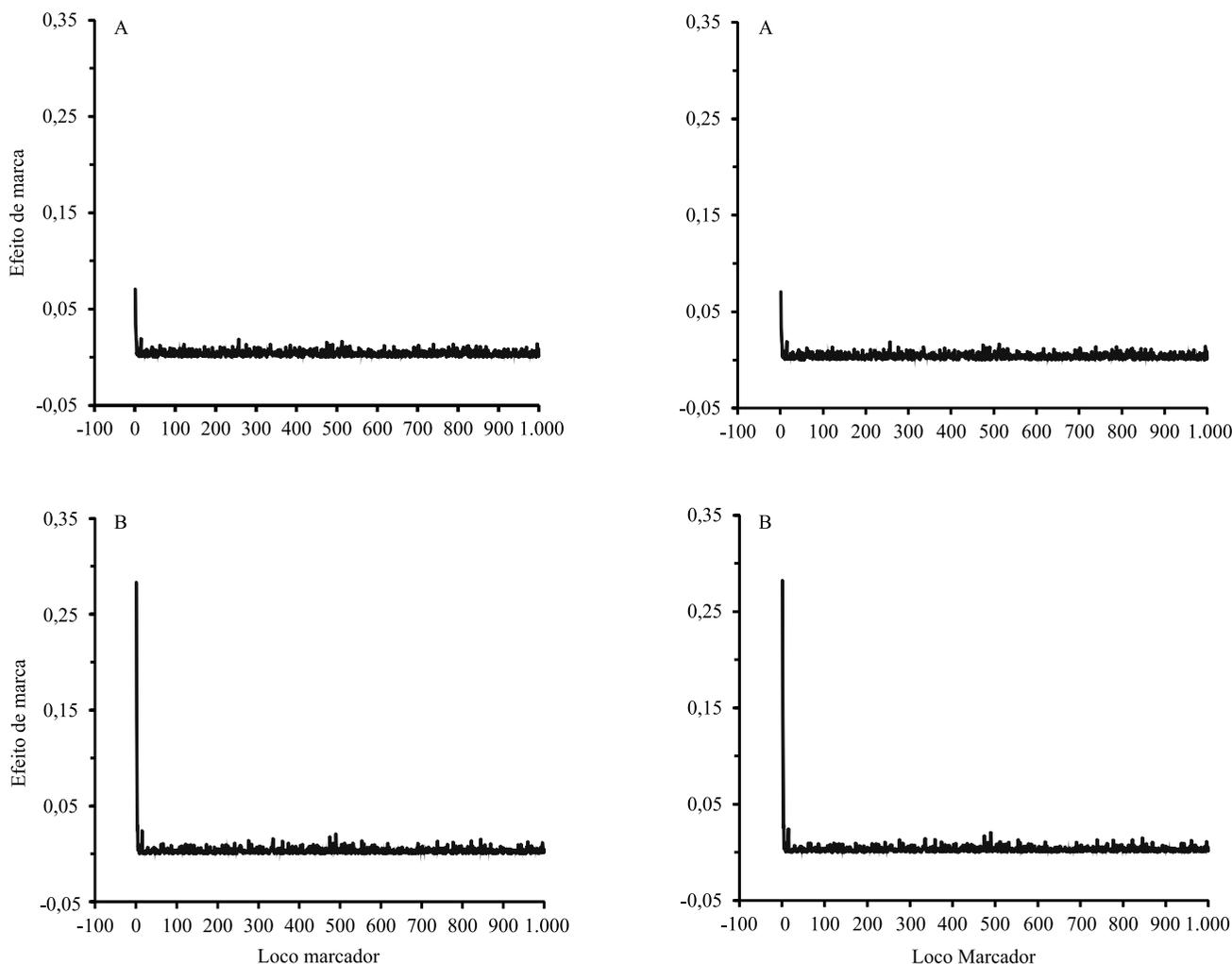


Figura 1. Efeito de marcadores com a correção de fenótipos, na herdabilidade 0,60, com distribuição uniforme nos gráficos à esquerda, e exponencial nos gráficos à direita: A, metodologia RR-Blup; e B, metodologia Blasso.

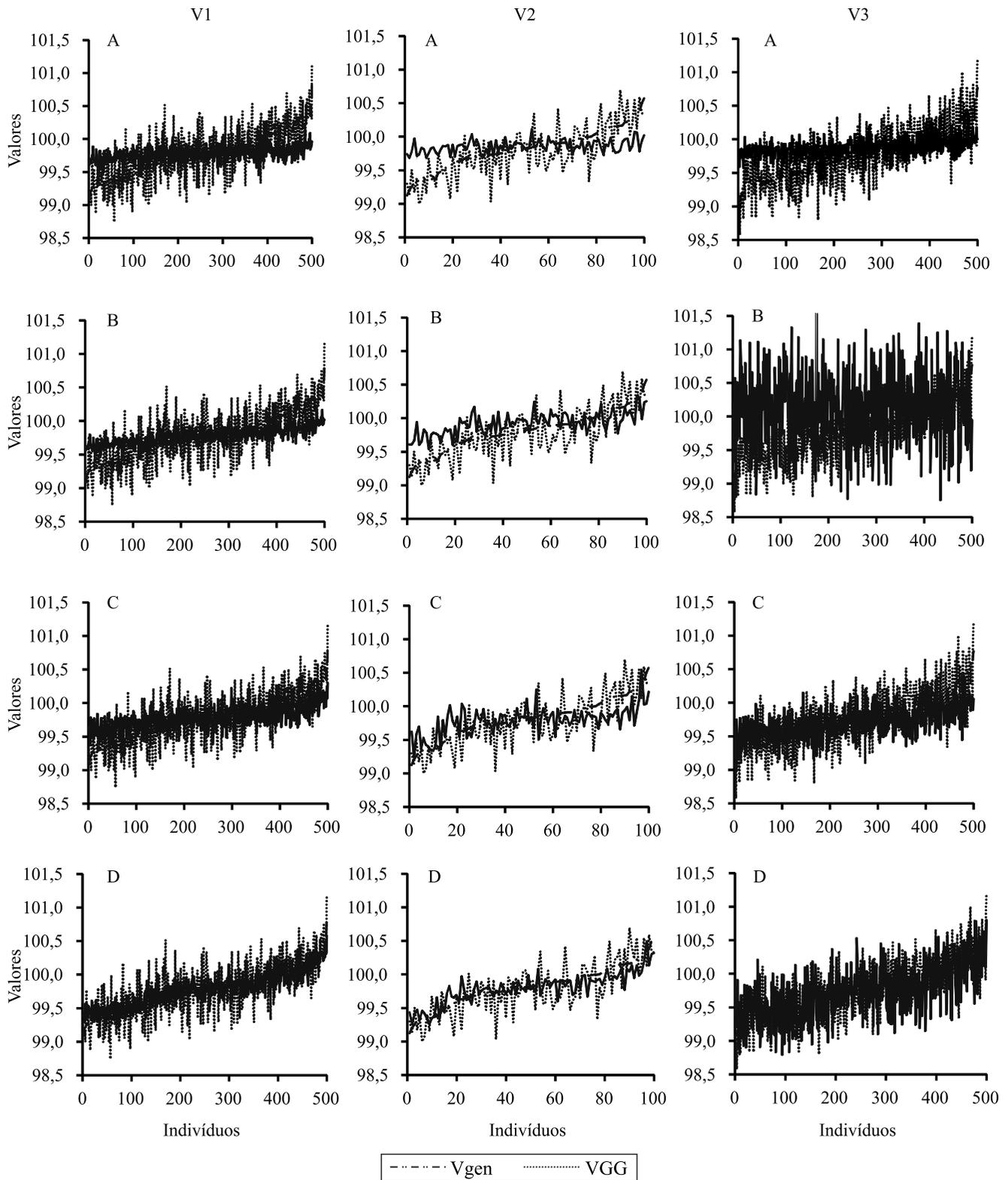


Figura 2. Valores genéticos (Vgen) e valores genéticos genômicos (VGG), para fenótipos corrigidos, e distribuição exponencial dos efeitos de QTL. Da esquerda para a direita, validação V1, V2 e V3: A, $h^2=0,20$, RR-Blup; B, $h^2=0,20$, Blasso; C, $h^2=0,60$, RR-Blup; e D, $h^2=0,60$, Blasso.

O viés, ou vício da análise, é dado pela diferença entre os valores genéticos genômicos preditos (VGG) e valores genotípicos simulados (Vgen); com isso, quanto maior a diferença entre esses valores, mais viciada é a metodologia. A acurácia, por outro lado, é a variação entre os valores genéticos genômicos preditos (VGG); ou seja, quanto maior a variação na predição, menor a acurácia. Os gráficos apresentados adiante retratam esses dois fenômenos (Figura 2). Para herdabilidade (h^2) de 0,20, as menores acurácias foram obtidas nas validações V1 e V3 da metodologia Blasso (Tabela 1). Nessa herdabilidade, o melhor método foi o RR-Blup, que foi acurado em todas as validações; porém, ele também se mostrou enviesado em todas elas. O Blasso, com esse valor de herdabilidade, foi acurado apenas na V2, e as predições foram enviesadas nos demais cenários.

Tabela 1. Viés e precisão associados às metodologias de análise RR-Blup e Blasso nas validações V1, V2 e V3 com correção de fenótipos e efeito de QTL exponencial.

Método	Validação 1		Validação 2		Validação 3	
	Viés	Precisão	Viés	Precisão	Viés	Precisão
Herdabilidade 0,20						
RR-Blup	Sim	Sim	Sim	Sim	Sim	Sim
Blasso	Pequeno	Não	Sim	Sim	Sim	Não
Herdabilidade 0,60						
RR-Blup	Pequeno	Não	Pequeno	Sim	Pequeno	Não
Blasso	Pequeno	Não	Não	Sim	Não	Não

A predição genômica foi satisfatória com h^2 de 0,60, pelo método Blasso e validação V2, com predições acuradas e não enviesadas. De maneira geral, o método RR-Blup foi mais acurado, e o Blasso menos viciado. Essa última metodologia precede a seleção de modelos, de maneira que as marcas com efeitos nulos são retiradas e, com isso, trabalha-se com menor número de parâmetros. Este fato resulta em menor número efetivo, com reflexo na variância da amostra, a qual aumenta e, conseqüentemente, diminui a acurácia do método em comparação ao RR-Blup. Por outro lado, o viés é dado pela capacidade do método em aprender a arquitetura ou a forma de distribuição dos efeitos genéticos. Uma vez que o Blasso apresentou menor viés, depreende-se que ele foi mais eficiente em retratar a distribuição exponencial dos efeitos genéticos.

Na análise sem correção de fenótipos, observou-se que, para genes com efeitos de distribuição uniforme, a metodologia RR-Blup apresentou, em geral, maiores valores de acurácia que o Blasso (Tabela 2). Alguns autores apontam para a superioridade de métodos bayesianos, em relação ao método Blup (Meuwissen et al., 2001; Usai et al., 2009; Crossa et al., 2010). Entretanto, há relatos de inversão nesse comportamento (Habier et al., 2007; Zhong et al., 2009). Zhong et al. (2009) obtiveram valores de acurácia de 0,62 e 0,61, com RR-Blup e Bayes B, respectivamente. Segundo os autores, uma distribuição de efeitos aleatórios

Tabela 2. Acurácia da validação na seleção genômica ampla sem a correção de fenótipos, nas diferentes herdabilidades avaliadas, com as distribuições dos efeitos de QTL uniforme e exponencial, sob diferentes cenários de validação⁽¹⁾.

Metodologia	Herdabilidade						Média
	0,20		0,40		0,60		
	Uniforme	Exponencial	Uniforme	Exponencial	Uniforme	Exponencial	
Validação V1							
RR-Blup	0,40	0,40	0,52	0,54	0,60	0,60	0,51
Blasso	0,34	0,48	0,45	0,82	0,45	0,90	0,57
Validação V2							
RR-Blup	0,37	0,43	0,46	0,47	0,48	0,51	0,45
Blasso	0,34	0,67	0,38	0,53	0,44	0,53	0,48
Validação V3							
RR-Blup	0,42	0,39	0,56	0,52	0,63	0,58	0,52
Blasso	0,34	0,36	0,50	0,60	0,61	0,73	0,52
Média	0,37	0,46	0,48	0,58	0,54	0,64	-

⁽¹⁾V1, amostra com 100 indivíduos, constituiu a população de validação; V2, validação cruzada, com amostra de 100 indivíduos, em cinco repetições; e V3, uma segunda geração, constituiu a população de validação.

mais complicada, como a utilizada em métodos bayesianos, somente é útil quando as marcas estão fortemente associadas com o QTL. Essa associação ocorre quando o efeito do QTL é elevado e quando as marcas estão em alto desequilíbrio de ligação com o QTL. O comportamento inverso também é verificado; ou seja, quando os genes apresentaram efeitos com distribuição exponencial, o Blasso é superior, em comparação ao RR-Blup, o que está de acordo com o pressuposto de cada metodologia.

Em geral, a arquitetura genética de poucos genes de maior efeito (distribuição exponencial) foi o cenário que apresentou maiores valores de acurácia, o que corrobora os autores citados nas discussões do parágrafo anterior, os quais afirmam que, quando as marcas estão em alto desequilíbrio, com pequeno número de QTL que capturam a maior parte da variância genética – ou seja, marcadores de grande efeito –, os métodos bayesianos são mais acurados; enquanto o RR-Blup é mais acurado quando muitas marcas possuem efeitos pequenos.

Os valores médios de acurácia decresceram com a herdabilidade. Goodard (2009) define a acurácia dos valores genômicos estimados como função do número de dados fenotípicos e da herdabilidade. Assim, um decréscimo nos valores de herdabilidade levaria a menores acurácias na estimação dos valores genômicos. Bernardo & Yu (2007), ao compararem a seleção genômica com diferentes cenários de herdabilidade, também verificaram que a acurácia decresce com a herdabilidade.

Os maiores valores médios de acurácia foram obtidos na validação V1, seguidos dos obtidos na validação V3. Maiores valores de acurácia podem ser obtidos ao se elevar o tamanho da população de estimação. Meuwissen et al. (2001) observaram que a acurácia dos valores genômicos estimados foi maior quando 2.000 dados, em vez de 1.000, foram usados para estimar o efeito do QTL de uma característica com herdabilidade de 0,50. Sob condições de baixo desequilíbrio (como no caso de elevado tamanho efetivo), ainda mais dados seriam necessários. Da mesma forma, Zhong et al. (2009) verificaram que dobrar o tamanho da população de estimação aumenta a acurácia.

Em geral, a acurácia da seleção genômica ampla com correção de fenótipos foi maior com a metodologia Blasso do que com a RR-Blup (Tabela 3).

Este comportamento é contrário ao que foi observado na análise sem a correção dos fenótipos, em que as metodologias Blasso e RR-Blup divergiram nos valores de acurácia, de acordo com o modo de distribuição dos efeitos dos QTLs.

A correção proporciona o uso da segregação mendeliana, que captura a associação de alelos dos marcadores codominantes e de QTL; ou seja, ela capta os efeitos genéticos explicados pelo desequilíbrio de ligação, e não pelo parentesco (Resende, 2008). Maiores valores de desequilíbrio contribuem diretamente para maior valor na acurácia. Com isso, a correção de fenótipos para estrutura de população apresentou maiores valores de acurácia, em comparação à análise em que essa correção não foi efetuada. Adicionalmente, nesse cenário, a metodologia Blasso destaca-se como a mais robusta quanto à distribuição dos efeitos genéticos, e apresenta-se mais acurada tanto para genes com distribuição uniforme quanto para genes com distribuição exponencial, o que não foi observado quando houve correção dos fenótipos.

Uma vez que a correção de fenótipos diminui a contribuição do parentesco para a análise, pressupõe-se que a metodologia RR-Blup foi a mais penalizada, já que ela possui contribuição do parentesco na estimação dos dados. O Blup tradicional aumenta a acurácia ao incorporar os ancestrais e o parentesco na sua formulação. Entretanto, a inclusão de informação de famílias eleva a correlação entre os valores genéticos estimados entre os membros das famílias, o que faz com que a seleção de aparentados seja maior (Wray & Thompson, 1990). Habier et al. (2007) desenvolveram uma regressão aproximada para quantificar a importância relativa das fontes parentesco genético e do desequilíbrio de ligação, e evidenciaram que, sob o modelo de simulação aplicado, 39 e 21% da acurácia da seleção genômica foi devida à captura do parentesco genético, para o RR-Blup e para Bayes B, respectivamente. Entretanto, os autores relataram que o método bayesiano foi mais eficiente em capturar o efeito do desequilíbrio de ligação e, conseqüentemente, que sua acurácia persiste por mais tempo que a observada com a regressão aleatória.

Goodard & Hayes (2007) afirmam que, quando a maioria dos QTLs possuem efeito zero, situação característica na distribuição exponencial, mínimos quadrados e Blup resultam em pequenos valores de estimação para esses efeitos “zero”; no entanto, o

acúmulo dos efeitos não zerados ocasiona ruídos nas estimativas, o que pode ter levado à menor acurácia do RR-Blup em relação ao Blasso. Se todos os QTLs são de genes com efeitos distribuídos exponencialmente, o estimador Lasso seria mais apropriado (Tibshirani, 1996). Melhores estimativas são alcançadas quando os vários QTLs possuem efeito zero, o que equivale a retirá-los do modelo. O mesmo comportamento foi retratado por Bastiaansen et al. (2012) ao compararem a metodologia Blup com outra bayesiana. Os autores verificaram que a distribuição dos efeitos genéticos influencia a estimação dos valores genômicos, de forma que a distribuição de poucos QTLs de maior efeito apresenta maior acurácia quando o método bayesiano é aplicado.

A arquitetura genética – distribuições uniforme ou exponencial –, entretanto, não exerceu influência pronunciada na diferenciação de acurácias; portanto, os valores para cada distribuição foram semelhantes, dentro das herdabilidades avaliadas. De forma semelhante, os diferentes cenários de validação também não influenciaram de forma pronunciada os valores médios de acurácia, em cada metodologia. Uma vez que tanto na validação V1 quanto na validação V2, uma amostra de 400 indivíduos foi usada na população de estimação, e 100 indivíduos foram direcionados para a população de validação, valores aproximados entre essas duas formas de validação eram, de fato, esperados.

No entanto, quando houve correção de fenótipos, a validação V3 não apresentou valores superiores de acurácia como os observados quando não houve correção. Esse resultado está relacionado ao tipo de população de validação em V3, que foi obtida pelo acasalamento ao acaso entre irmãos completos (G2), o que gerou maiores valores de parentesco quando se procedeu à correção de fenótipos. A correção da genealogia, portanto, contribuiu para redução da acurácia. Associado a este fato, o acasalamento ao acaso diminuiu o desequilíbrio de ligação nesta população de validação, o que também contribuiu para uma menor acurácia. Com a correção de fenótipos, depreende-se que o maior peso nas acurácias reflete a proporção da segregação mendeliana capturada. Com a redução da contribuição do parentesco, a metodologia RR-Blup diminuiu seus valores, enquanto a metodologia Blasso foi mais eficiente.

No presente trabalho, tanto a herdabilidade quanto a distribuição dos efeitos genéticos influenciaram os valores de acurácia, assim como o tipo da população de validação. Quanto a este último fator, observou-se que, na ausência de correção de fenótipos, os valores médios mais elevados estiveram associados às validações V3, V1 e V2, nesta ordem; enquanto, com a correção, estiveram associados às validações V2, V1 e V3.

Em termos práticos, a raiz quadrada da herdabilidade iguala-se ao coeficiente de correlação entre o valor

Tabela 3. Acurácia da validação na seleção genômica ampla com a correção de fenótipos, nas diferentes herdabilidades avaliadas, com as distribuições dos efeitos de QTL uniforme e exponencial, sob diferentes cenários de validação⁽¹⁾.

Metodologia	Herdabilidade						Média
	0,20		0,40		0,60		
	Uniforme	Exponencial	Uniforme	Exponencial	Uniforme	Exponencial	
Validação V1							
RR-Blup	0,34	0,43	0,55	0,54	0,58	0,58	0,50
Blasso	0,38	0,58	0,82	0,83	0,89	0,90	0,73
Validação V2							
RR-Blup	0,36	0,39	0,51	0,51	0,57	0,57	0,49
Blasso	0,39	0,58	0,83	0,83	0,90	0,90	0,74
Validação V3							
RR-Blup	0,27	0,40	0,52	0,50	0,55	0,55	0,47
Blasso	0,26	0,37	0,57	0,57	0,70	0,70	0,53
Média	0,33	0,46	0,63	0,63	0,70	0,70	-

⁽¹⁾V1, amostra com 100 indivíduos, constituiu a população de validação; V2, validação cruzada, com amostra de 100 indivíduos, em cinco repetições; e V3, uma segunda geração, constituiu a população de validação.

genético e o valor fenotípico (acurácia). Neste contexto, somente na validação V1 com distribuição de efeito de QTL exponencial, na análise Blasso, obteve-se acurácia da seleção genômica superior à raiz quadrada da herdabilidade, na ausência de correção de fenótipos, independentemente da herdabilidade. Isso equivale a dizer que, nas demais situações, a seleção fenotípica seria mais acurada. Na presença de correção, essa superioridade ocorreu nas validações V1 e V2, com a metodologia Blasso, para características uniformes e exponenciais, nas herdabilidades 0,40 e 0,60; e somente para características exponenciais, na herdabilidade 0,20. Dessa forma, espera-se que a seleção genômica seja vantajosa ao programa de melhoramento quando se realiza a correção de fenótipos, e isso se aplica às validações V1 e V2, utilizando-se a análise Blasso.

Conclusões

1. Em geral, a correção de fenótipos eleva os valores de acurácia.

2. Sem correção de fenótipos, a análise Blasso é mais acurada para QTL de distribuição exponencial, e o RR-Blup para distribuição uniforme; em contrapartida, quando se procede à correção, essa análise é mais acurada tanto para distribuição de efeitos exponencial quanto para uniforme.

3. Na ausência de correção, as validações mais acuradas são V3, V1 e V2, nesta ordem; com a correção, a ordem muda para V2, V1 e V3.

4. De maneira geral, o método RR-Blup é mais preciso, enquanto o Blasso é menos viciado.

Agradecimentos

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes), pelo apoio financeiro.

Referências

BASTIAANSEN, J.W.M.; COSTER, A.; CALUS, M.P.L.; ARENDONK, J.A.M. van; BOVENHUIS, H. Long-term response to genomic selection: effects of estimation method and reference population structure for different genetic architectures. *Genetics Selection Evolution*, v.44, p.1-13, 2012. DOI: 10.1186/1297-9686-44-3.

BERNARDO, R.; YU, J.M. Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, v.47, p.1082-1090, 2007. DOI: 10.2135/cropsci2006.11.0690.

CROSSA, J.; DE LOS CAMPOS, G.; PÉREZ, P.; GIANOLA, D.; BURGUEÑO, J.; ARAUS, J.L.; MAKUMBI, D.; SINGH, R.P.; DREISIGACKER, S.; YAN, J.; ARIEF, V.; BANZIGER, M.; BRAUN, H.J. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, v.186, p.713-724, 2010. DOI: 10.1534/genetics.110.118521.

CRUZ, C.D. GENES - a software package for analysis in experimental statistics and quantitative genetics. *Acta Scientiarum. Agronomy*, v.35, p.271-276, 2013. DOI: 10.4025/actasciagron.v35i3.21251.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D.; CROSSA, J.; LEGARRA, A.; MANFREDI, E.; WEIGEL, K.; COTES, J.M. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, v.182, p.375-385, 2009. DOI: 10.1534/genetics.109.101501.

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

GODDARD, M. Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica*, v.136, p.245-257, 2009. DOI: 10.1007/s10709-008-9308-0.

GODDARD, M.E.; HAYES, B.J. Genomic selection. *Journal of Animal Breeding and Genetics*, v.124, p.323-330, 2007. DOI: 10.1111/j.1439-0388.2007.00702.x.

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

HELTSHE, J.F.; FORRESTER, N.E. Estimating Species Richness Using the Jackknife Procedure. *Biometrics*, v.39, p.1-11, 1983. DOI: 10.2307/2530802.

KOLBEHDARI, D.; SCHAEFFER, L.R.; ROBINSON, J.A. Estimation of genome-wide haplotype effects in half-sib designs. *Journal of Animal Breeding and Genetics*, v.124, p.356-361, 2007. DOI: 10.1111/j.1439-0388.2007.00698.x.

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

MUIR, W.M. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *Journal of Animal Breeding and Genetics*, v.124, p.342-355, 2007. DOI: 10.1111/j.1439-0388.2007.00700.x.

PARK, T.; CASELLA, G. The Bayesian Lasso. *Journal of the American Statistical Association*, v.103, p.681-686, 2008. DOI: 10.1198/016214508000000337.

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

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

RESENDE, M.D.V. de; LOPES, P.S.; SILVA, R.L. da; PIRES, I.E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, n.56, p.63-77, 2008.

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

USAI, M.G.; GODDARD, M.E.; HAYES, B. J. LASSO with cross-validation for genomic selection. **Genetics Research**, v.91, p.427-436, 2009. DOI: 10.1017/S0016672309990334.

WHITTAKER, J.C.; THOMPSON, R.; DENHAM, M.C. Marker-assisted selection using ridge regression. **Genetical Research**, v.75, p.249-252, 2000. DOI: 10.1017/S0016672399004462.

WRAY, N.R.; THOMPSON, R. Prediction of rates of inbreeding in selected populations. **Genetical Research**, v.55, p.41-54, 1990. DOI: 10.1017/S0016672300025180.

ZHONG, S.; DEKKERS, J.C.M.; FERNANDO, R.L.; JANNINK, J.-L. Factors affecting accuracy from genomic selection in populations derived from multiple inbred lines: a barley case study. **Genetics**, v.182, p.355-364, 2009. DOI: 10.1534/genetics.108.098277.

Recebido em 20 de julho de 2015 e aprovado em 13 de julho de 2016