Validação Cruzada e Independente na Seleção Genômica Ampla

Caillet Dornelles Marinho[1], Janeo Eustáquio de Almeida Filho[1], Camila Ferreira Azevedo[2], Marcos Deon Vilela de Resende[3], Fabyano Fonseca e Silva[4], Karina Carnielli Zamprogno Ferreira[5], Leonardo Novaes Rosse[5], Carolina Paola Sansaloni[6], César Daniel Petroli[7], Dario Grattapaglia[8]

Resumo

Este trabalho foi realizado com o objetivo de avaliar a capacidade preditiva e o viés da seleção genômica ampla (GWS) na validação independente e validação cruzada, quando números diferentes de indivíduos são utilizados nas equações de estimação, bem como a eficiência computacional das análises. Para tanto, foram utilizados dados de eucalipto, contendo 1000 clones fenotipados para diâmetro altura do peito (DAP), altura (ALT) e volume (VOL), e, 936 clones fenotipados para densidade básica pelo pilodyn® (PIL). Todos os clones foram genotipados com 2668 marcadores DArT. Os valores genéticos genômicos () foram preditos via *Ridge Regression* e a capacidade preditiva () foi calculada pela correlação entre os e os fenótipos observados (*f*). Para verificar o viés da predição, foi obtido pela regressão de *f* em . Na validação cruzada, quanto o tamanho (k) dos grupos foi igual a 1, levou-se de 150 (PIL) a 198 (DAP, ALT e VOL) minutos para conclusão da análise, para cada característica. Quando k foi igual a 10 e 12, o tempo gasto ficou abaixo dos 20 minutos e com k igual a 125 e 117, a análise levou de 0.69 a 1.12 minutos para ser realizada. A capacidade preditiva e variaram pouco para os diferentes k, sendo que, com k igual a 125 (DAP e ALT), 100 (VOL) e 117 (PIL), as predições foram praticamente idênticas as obtidas com k=1, com a vantagem de necessitarem de aproximadamente 1% do tempo de processamento computacional em relação a k=1. Na validação independente, verificou-se que aumenta com o incremento de indivíduos usados na população de estimação e que foram observados valores de maiores que na validação cruzada, porém, esses valores foram acompanhados de alto viés. Dessa forma, visando uma predição com menor viés possível e boa capacidade preditiva, as melhores porcentagens utilizadas na validação independente foram: 80% para DAP, ALT e PIL e 70% para VOL.

Introdução

A seleção genômica ampla (GWS) foi proposta por Meuwissen et al. (2001) com intuito de predizer o fenótipo futuro de indivíduos provenientes de populações em melhoramento, utilizando informações pré-estimadas de marcadores moleculares, que explicam os efeitos genéticos aditivos dos caracteres de interesse. 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 com preditores do tipo BLUP, todas as marcas são colocadas no modelo estatístico e através dos efeitos genéticos aditivos dos marcadores o valor genético genômico do indivíduo é predito.

Na GWS três populações podem ser definidas: (i) população de estimação, a qual deve ter seus fenótipos avaliados e marcadores obtidos, visando associar, por meio de regressão múltipla aleatória, cada marcador ao seu efeito predito no caráter de interesse; (ii) população de validação, também possui indivíduos fenotipados

[1] Doutorando do Programa de Pós-graduação em Genética e Melhoramento de Plantas– UFV/Viçosa. E-mail: caillet.marinho@yahoo.com.br; janeo.eustaquio@ymail.com

[2] Doutoranda do Programa de Pós-graduação em Estatística e Biometria – UFV/Viçosa. E-mail: camila.azevedo1504@gmail.com

[3] Pesquisador Pós-doutor da Embrapa Floresta e Professor Credenciado do Departamento de Estatística – UFV/Viçosa. E-mail: marcos.deon@gmail.com

[4] Professor Pós-doutor do Departamento de Estatística – UFV/Viçosa. E-mail: fabyanofonseca@ufv.br

[5] Pesquisador (a) Doutor (a) da Empresa Veracel Celulose S.A. Salvador, BA. E-mail:karina.zamprogno@veracel.com.br; leonardo.rosse@veracel.com.br

[6] Doutoranda do Programa de Pós-graduação em Biologia Molecular – UNB/Brasília. E-mail:carosansaloni@hotmail.com

[7] Doutorando do Programa de Pós-graduação em Biologia Molecular – UNB/Brasília. E-mail:petrolic@hotmail.com

[8] Pesquisador Pós-doutor da Embrapa Recursos Genéticos e Biotecnologia e Professor do curso de Pós-graduação em CiênciasGenômicas e Biotecnologia – UCB/Brasília. E-mail:dario.grattapaglia@embrapa.br

e genotipados, no entanto, utiliza-se os efeitos pré-estimados dos marcadores para predizer os fenótipos, assim, verifica-se a capacidade preditiva do GWS por meio da correlação dos valores fenotípicos observados com os preditos; e, (iii) população de seleção, que comtempla indivíduos apenas genotipados, que serão avaliados por meio da predição dos valores genéticos genômicos ou fenótipos futuros.

Na prática, segundo Resende et al. (2012) essas populações podem ser fisicamente distintas (três populações diferentes) ou não. Neste caso, uma mesma população é usada consecutivamente para estimação e validação, utilizando um esquema *Jackknife* de validação cruzada.

Dessa forma, o objetivo deste trabalho foi avaliar a capacidade preditiva e o viés da GWS na validação independente e validação cruzada, quando números diferentes de indivíduos são utilizados nas equações de estimação, bem como a eficiência computacional das análises.

Material e Métodos

Para o presente estudo foram utilizados dados de eucalipto, pertencentes ao projeto de seleção genômica desenvolvido pela Embrapa em conjunto com a Veracel Celulose, contendo 1000 clones fenotipados para diâmetro altura do peito (DAP), altura (ALT) e volume (VOL), e, 936 clones fenotipados para densidade básica pelo pilodyn® (PIL). Todos os clones foram genotipados com 2668 marcadores DArT (*Diversity Arrays Technology*) (Sansaloni et al. 2010).

As análises foram realizadas no *software* R, versão 2.14.1 (R Development Core Team 2011). A matriz de incidência genotípica foi parametrizada conforme recomendado por Resende et al. (2010), em que cada coluna de marcas foi centrada e padronizada.

O computador utilizado possui um processador Intel Core i5 – 2410M, que opera à frequência de 2.30 Ghz, com 4 MB de cache L2, 4 GB de memória RAM, sistema operacional Windows 7 64 bits e placa de vídeo Intel (R) HD *Graphics Family*.

Os valores genéticos genômicos () foram preditos via *Ridge Regression* com auxílio do pacote rrBLUP (Endelman 2011). A capacidade preditiva () foi calculada pela correlação entre os e os fenótipos observados (*f*). Para verificar o viés da predição, a regressão de *f* em foi obtida por , em que, representa a variância dos valores genéticos genômicos preditos. Dessa forma, a melhor predição será aquela com igual a 1.

A validação independente consistiu em dividir a população em duas, uma para estimação e outra para validação. A divisão foi feita aleatoriamente e o número de indivíduos na população de estimação foi de 50%, 60%, 70%, 75%, 80%, 85%, 90% e 95% do número total de indivíduos.

A metodologia generalizada do *Jackknife* é baseada na divisão do conjunto de N dados amostrais em g grupos de tamanho igual a k, sendo que, a estimação da variância do estimador θ de interesse consiste na omissão de k observações em cada reamostragem (Resende 2008). Neste estudo foi utilizado k igual a 500, 250, 125, 100, 50, 25, 20, 10, 5, 4, 2 e 1 para as variáveis DAP, ALT e VOL. E k igual a 468, 234, 156, 117, 52, 24, 18, 12, 6, 4, 2 e 1 para PIL, uma vez que, os diferentes k tem que ser múltiplo do número de indivíduos.

Resultados e Discussão

Na Figura 1, observa-se o tempo, em minutos, do processamento de análise da GWS (eficiência computacional) para todos os diferentes tamanhos de grupos utilizados na validação *Jackknife*. Nota-se que, para k igual a 1, levou-se de 150 (PIL) a 198 (DAP, ALT e VOL) minutos para conclusão da análise, para cada característica. Em contrapartida quando k foi igual a 10 e 12, o tempo gasto ficou abaixo dos 20 minutos e com k igual a 125 e 117, a análise levou de 0.69 a 1.12 minutos para ser realizada. É importante ressaltar que, para o presente estudo, os dados são relativamente pequenos (1000, 936 indivíduos com 2668 marcadores), porém, para situações com números maiores de marcas e mais indivíduos, o tempo pode ser fator essencial.

As capacidades preditivas () e os para os diferentes tamanhos (k) de grupos da validação *Jackknife* podem ser visualizadas na Figura 2. Para DAP, variou de 0.85 a 0.88, sendo que, o valor de 0.88 foi obtido pela maioria dos tamanhos (k=1, 2, 4, 5, 10, 20, 125). Em relação a , os valores foram iguais a 0.98 e 0.99, podendo concluir que, para todos os tamanhos de grupos, o viés foi praticamente igual para essa característica. Para o caráter ALT, quando k foi igual a 500 e a 125, os valores de foram iguais a 0.43 e 0.45, respectivamente, para todos os outros, foi igual a 0.46. A regressão variou de 0.95 a 1.02, sendo que o valor ideal (1.00) foi

obtido quando k foi igual a 20 e 125 (Figura 1). Dessa forma, para DAP e ALT, presando pela eficiência computacional, k igual a 125 obteve resultados eficientes (Figura 1 e 2).

A capacidade preditiva para VOL obteve valor máximo (0.45) quando k foi igual a 100 e valor mínimo (0.42) para k igual a 500. Para os demais tamanhos, foi igual a 0.44. Os betas das regressões para esta variável apresentaram valores de 1.01 (k=10 e 20) a 1.07 (k=500) (Figura 1). Portanto, para essa variável, k=20 ou igual a 100 (,=1.04), pode ser eficientemente aplicável (Figura 2).

Em relação a variável PIL, variou de 0.36 (k=468) a 0.43 (k=1, 2, 12 e 24), quando k foi igual a 4, 6, 18, 52 e 117 foi igual a 0.42. Na regressão, beta foi igual a 0.92 (k=468), 0.98 (k=156 e 234), 0.99 (k=52 e 117), 1.00 (k=1, 2, 4, 6 e 18) e 1.01 (k=12 e 24) (Figura 1). Assim, seguindo o mesmo raciocínio, para PIL, k=117, apresentou predições satisfatórias.

Portanto, pôde-se perceber que, com k igual a 125 (DAP e ALT), 100 (VOL) e 117 (PIL) – 10 a 12.5% do total de indivíduos –, as predições foram praticamente idênticas as obtidas com k=1, com a vantagem de necessitarem de aproximadamente 1% do tempo de processamento computacional em relação a k=1.

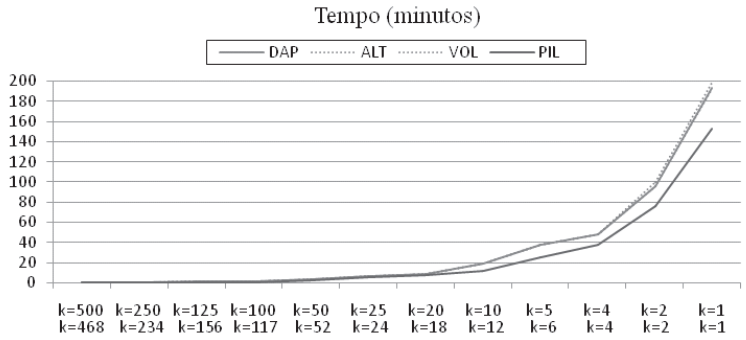

Figura 1. Tempo de execução da função RR-BLUP para os diferentes tamanhos (k) de grupos utilizados na validação *Jackknife*.

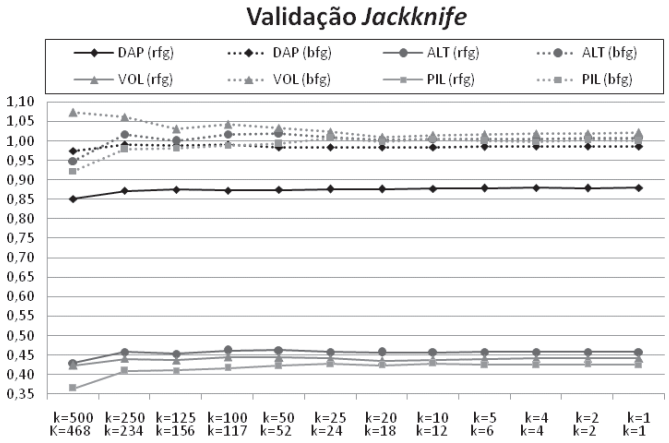

Figura 2. Capacidade preditiva (rfg) e regressão de *f* em (bfg) para os vários números de grupos usados na validação por *Jackknife* para as variáveis DAP, ALT, VOL e PIL.

Na validação independente, o tempo de análise computacional foi muito rápido nesse estudo, para analisar todos os grupos foi preciso um tempo médio de 51 segundos para cada variável. Verificou-se que aumenta com o incremento de indivíduos usados na população de estimação até 85% para DAP, ALT e VOL, e, até 90% para PIL. Foram observados valores de maiores que na validação cruzada, porém, esses valores foram

acompanhados de alto viés (Figura 3). Dessa forma, visando uma predição com menor viés possível e boa capacidade preditiva, as melhores porcentagens utilizadas foram: 80% para DAP, ALT e PIL e 70% para VOL (Figura 3).

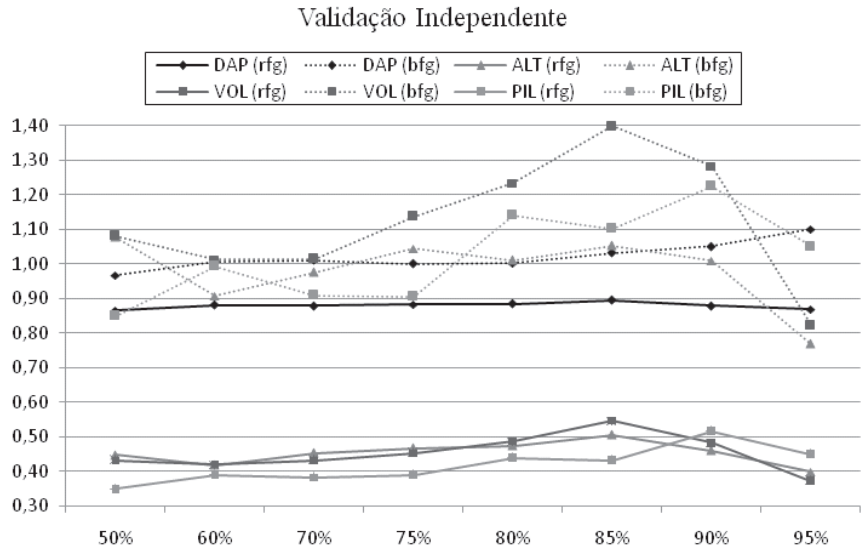

Figura 3. Capacidade preditiva (rfg) e beta da regressão (bfg) obtidos na validação independente para as diferentes porcentagens de indivíduos deixados na população de estimação.

Agradecimentos

Os autores agradecem a Embrapa e a Veracel Celulose pela disponibilização dos dados. A Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pela concessão das bolsas de estudos. E ao professor Luiz Alexandre Peternelli, pelas valiosas considerações.

Referências

Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. **Plant Genome** 4:250-255.

R Development Core Team (2011) **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Resende MDV (2008) **Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, 330p.

Resende MDV et al. (2010) **Computação da seleção genômica ampla (GWS)**. Colombo: Embrapa Florestas, 79p.

Resende MDV et al. (2012) **Seleção genômica ampla (GWS) via modelos mistos (REML/BLUP), inferência bayesiana (MCMC), regressão aleatória multivariada (RRM) e estatística espacial**. Viçosa: UFV, 291p. Disponível em: http://www.det.ufv.br/ppestbio/corpo_docente.php.

Sansaloni CP et al. (2010) A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in *Eucalyptus*. **Plant Methods** 6:6-16.