Controle de Qualidade de Dados Genotípicos para Estudos Genômicos em Clones de Eucalipto

Gislâyne Maira Pereira de Faria[1], Camila Santana Pereira[2] Italo Stefanine Correia Granato[3], Marcos Deon Vilela de Resende[4], Fabyano Fonseca e Silva[5], Karina Carnielli Zamprogno Ferreira[6], Leonardo Novaes Rosse[7], Carolina Paola Sansaloni[8], César Daniel Petroli[9], Dario Grattapaglia[10]

Resumo

Este trabalho foi realizado com o objetivo de testar diferentes porcentagens de MAF e Call Rate na avaliação da qualidade dos dados genotipados, e suas implicações nas predições dos valores genéticos genômicos dos indivíduos. Para tanto, foram utilizados dados de eucalipto, contendo 1000 clones fenotipados para diâmetro da 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 ($\hat{g}\hat{g}$) foram preditos via *Ridge Regression* e a capacidade preditiva ($r_{\hat{g}f} r_{\hat{g}f}$) foi calculada pela correlação entre os $\hat{g}\hat{g}$ e os fenótipos observados (*f*). Para verificar o viés da predição, $b_{f\hat{g}} b_{f\hat{g}}$ foi obtido pela regressão de *f* em $\hat{g}\hat{g}$. A qualidade da genotipagem foi satisfatória, uma vez que a maioria das marcas apresentaram elevado Call Rate e a maioria dos marcadores apresentou MAF acima de 10%. Verificou- se em todos os casos, que as predições dos efeitos dos marcadores não foram viesadas. Dessa forma, é importante levar em consideração o controle de qualidade. Em uma análise de seleção genômica, recomenda-se averiguar diferentes valores de controle de qualidade, visando adotar o melhor corte para cada situação.

Introdução

O uso de marcadores moleculares tem contribuído de forma importantíssima nos programas de melhoramento, isto porque possibilita a aplicação direta das informações de DNA na seleção, consequentemente ajuda a aumentar a eficiência seletiva e a ter uma maior rapidez na obtenção de ganhos genéticos com a seleção. Ademais, possui baixo custo, em comparação com a tradicional seleção baseada em dados fenotípicos. De acordo com Resende et al., (2008), a seleção genômica ampla (*genome wide selection*– GWS) proposta por Meuwissen et al. (2001), tornou-se um método muito atrativo, sendo que, geneticistas e melhoristas renomados, adeptos de métodos tradicionais, têm demonstrado e confirmado a superioridade e exequibilidade prática do método em benefício do melhoramento animal e vegetal.

A seleção genômica ampla consiste na predição simultânea dos efeitos genéticos de inúmeros marcadores de DNA dispersos em todo o genoma de um organismo, de forma a capturar os efeitos de todos os locos, de grandes e pequenos efeitos, e explicar toda a variação genética de um caráter quantitativo. O procedimento RR-BLUP (RR - *Random Regression*; BLUP - *Best Linear Unbiased Prediction*) usa preditores do tipo BLUP e ajusta os efeitos dos marcadores como variáveis explicativas, permitindo assim, a predição de valores genéticos genômicos dos indivíduos, o que demonstra grande utilidade no melhoramento genético.

Contudo, uma das fases da seleção genômica ampla consiste na avaliação da qualidade dos dados genotipados. O Call Rate (taxa de atendimento, qualificação ou repetibilidade) é uma medida de qualidade

[1] Doutoranda do Programa de Pós-graduação em Genética e Melhoramento – UFV/Viçosa. E-mail: gislaynemaira@yahoo.com.br
[2] Doutoranda do Programa de Pós-graduação em Genética e Melhoramento – UFV/Viçosa. E-mail: camilajsc@yahoo.com.br
[3] Mestrando do Programa de Pós-graduação em Genética e Melhoramento – UFV/Viçosa. E-mail: italo.granato@gmail.com
[4] Pesquisador Pós-doutor da Embrapa Floresta e Professor Credenciado do Departamento de Estatística – UFV/Viçosa. E-mail: marcos.deon@gmail.com
[5] Professor Pós-doutor do Departamento de Estatística – UFV/Viçosa. E-mail: fabyanofonseca@ufv.br
[6] Pesquisadora Doutora da Empresa Veracel Celulose S.A. Salvador, BA. E-mail: karina.zamprogno@veracel.com.br;
[7] Pesquisador Doutor da Empresa Veracel Celulose S.A. Salvador, BA. E-mail: leonardo.rosse@veracel.com.br
[8] Doutoranda do Programa de Pós-graduação em Biologia Molecular – UNB/Brasília. E-mail: carosansaloni@hotmail.com
[9] Doutorando do Programa de Pós-graduação em Biologia Molecular – UNB/Brasília. E-mail: petrolic@hotmail.com
[10] Pesquisador Pós-doutor da Embrapa Recursos Genéticos e Biotecnologia e Professor do curso de Pós-graduação em Ciências Genômicas e Biotecnologia – UCB/Brasília. E-mail: dario.grattapaglia@embrapa.br

utilizada para eliminar marcadores com grande quantidade de valores perdidos, normalmente, opta-se por trabalhar com Call Rate maior que 95%. Outra medida de qualidade é a MAF (frequência do alelo menos comum), que está relacionada com o polimorfismo dos locos marcadores na população, quando estes são pouco variáveis não apresentam relevância genética na população. Geralmente utiliza-se MAF entre 1% ou 5%.

Caso a quantidade de dados perdidos por marcador não seja elevada, a imputação de marcas pode ser uma alternativa eficiente para evitar a necessidade de eliminar marcadores, que por ventura, podem ser importantes às características quantitativas. Para imputação existem várias metodologias disponíveis (Gengler et al., 2011; Poland et al., 2012). A eliminação de marcadores pouco polimórficos pode não ser uma estratégia interessante, pois nesse caso poderia excluir marcadores que apresentam alelos raros associados à característica quantitativa. Para melhorar a capacidade dos métodos estatísticos de seleção genômica em detectar o efeito em alelos raros, Meuwissen et al. (2011) recomendam a padronização (divisão pela raiz quadrada da heterozigose de cada loco) dos valores de incidência dos marcadores, com isso marcadores com baixa variabilidade tende a aumentar seus valores na matriz de incidência genotípica.

Dessa forma, o objetivo deste trabalho foi testar diferentes porcentagens de MAF e Call Rate na avaliação da qualidade dos dados genotipados, e suas implicações nas predições dos valores genéticos genômicos dos indivíduos.

Material e Métodos

Para elaboração deste trabalho, foram utilizados dados de Eucalipto, pertencentes ao projeto de seleção genômica desenvolvido pela Embrapa em conjunto com a Veracel Celulose. Foram cedidos dados de 1000 clones fenotipados para volume (VOL), altura (ALT) e diâmetro da altura do peito (DAP) 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).

Foi realizado analises para todas as características com o modelo RR-BLUP (Meuwissen et al., 2001), utilizando o pacote rrBLUP (Endelman 2011) do software R, versão 2.14.1 (R Development Core Team 2011), considerando diferentes níveis de eliminação para Call Rate e MAF. Foram utilizados call rate iguais a 0.7, 0.8, 0.9 e 0.95, em combinação com MAF iguais a 0.01, 0.05 e 0.1, totalizando 12 diferentes cenários. Para todas as combinações foram calculados os seguintes parâmetros: capacidade preditiva ($r_{\hat{g}f} r_{\hat{g}f}$) calculada pela correlação entre os $\hat{g}\hat{g}$ e os fenótipos observados (*f*); viés dado pela regressão de *f* em $\hat{g}\hat{g}$ por $b_{f\hat{g}} = Cov(\hat{g},f)/\sigma^2_{\hat{g}} b_{f\hat{g}} = Cov(\hat{g},f)/\sigma^2_{\hat{g}}$, em que, $\sigma^2_{\hat{g}}\sigma^2_{\hat{g}}$ representa a variância dos valores genéticos genômicos preditos. Dessa forma, a melhor predição será aquela com $b_{f\hat{g}} b_{f\hat{g}}$ igual a 1.

Resultados e Discussão

De acordo com a Figura 1, observa-se que a maioria dos marcadores apresentou MAF acima de 10%, esse elevado número de marcadores polimórficos, é coerente com a estrutura de uma população alógama, como é o caso do Eucalipto. A qualidade da genotipagem foi satisfatória, uma vez que a maioria das marcas apresentaram elevado Call Rate.

Tabela 1. Porcentagem de marcas excluídas considerando diferentes valores de MAF e Call Rate (CR), para as características ADV (Altura, Diâmetro da altura do peito e Volume) e PIL.

Característica	CR	MAF		
		0,01	0,05	0,1
ADV	0,7	0,22	6,78	16,34
	0,8	1,54	8,09	17,65
	0,9	11,58	18,14	27,13
	0,95	33,4	39,06	46,62

PIL	0,7	0,22	6,59	16,27
	0,8	1,31	7,68	17,35
	0,9	9,71	16,08	25,45
	0,95	30,02	36,21	44,64

Para todas as características ouve grande variação no número de marcas eliminadas, sendo que o numero de marcadores eliminados ao considerar menores níveis de eliminação por Call Rate (CR) e MAF, foi relativamente baixo, enquanto que, com maiores valores, o numero de marcadores excluídos aumentou gradativamente, com destaque para CR igual a 0,95, onde a eliminação de marcas foi acima de 30% em todos os níveis (Tabela 1).

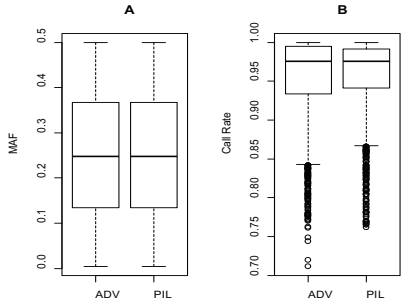

Figura 1. Boxplot dos valores de MAF (A) e Call Rate (B) para as características ADV (Altura, Diâmetro na altura do peito e Volume) e PIL (densidade básica pelo pilodyn®).

Pelos resultados obtidos na Tabela 2, verifica-se, em todos os casos, que as predições dos efeitos dos marcadores não foram viesadas. Quanto à eliminação, ocorreu um pequeno decréscimo na capacidade preditiva e na herdabilidade capturada com aumento da intensidade de marcas excluídas, com isto, pode-se supor que a eliminação considerando o MAF e Call Rate, pode ter levado a perda de marcadores importantes para as características em questão.

Vale ressaltar que, os valores de incidência dos marcadores foram centrados na média e padronizados, com isto, o método de seleção genômica fica mais eficaz para a detecção de QTLs, associados a alelos raros (Meuwissen et al. , 2011; Resende et al., 2012), destacando os efeitos da eliminação de marcadores considerando o MAF, que é um critério de eliminação de marcadores pouco polimórficos.

Tabela 2. Resultados da predição dos parâmetros capacidade preditiva (r_{fg}), regressão do GEBV nos valores fenotípicos (b_{fg}) e herdabilidade capturada (h^2) pelo método de seleção genômica considerando a eliminação de marcadores para diferentes MAF e Call Rate (CR), sem a eliminação dos marcadores (S.E.), para as variáveis Altura (ALT), Diâmetro na altura do peito (DAP), densidade básica pelo pilodyn® (PIL) e Volume (VOL).

Caracteres	CR	0,7			0,8			0,9			0,95			S.E.
	MAF	0,01	0,05	0,1	0,01	0,05	0,1	0,01	0,05	0,1	0,01	0,05	0,1	
DAP	r_{fg}	0,87	0,87	0,87	0,88	0,87	0,87	0,87	0,86	0,86	0,84	0,83	0,83	0,88
	b_{fg}	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98	0,98
	h^2	0,87	0,86	0,85	0,87	0,86	0,85	0,85	0,84	0,83	0,78	0,77	0,77	0,87
PIL	r_{fg}	0,42	0,42	0,42	0,42	0,42	0,42	0,41	0,41	0,41	0,41	0,39	0,39	0,43
	b_{fg}	1,00	1,00	0,99	1,00	1,00	0,99	0,99	0,99	0,99	0,99	0,99	0,99	1,00
	h^2	0,38	0,37	0,36	0,38	0,37	0,37	0,36	0,35	0,34	0,33	0,31	0,31	0,38

ALT	r_{fg}	0,45	0,45	0,44	0,45	0,45	0,44	0,46	0,45	0,44	0,46	0,45	0,44	0,46
	b_{fg}	1,00	1,0	0,99	1,00	1,00	1,00	1,00	1,00	0,99	1,00	1,00	0,99	1,01
	h^2	0,26	0,25	0,23	0,25	0,25	0,23	0,25	0,24	0,23	0,24	0,22	0,21	0,26
VOL	r_{fg}	0,43	0,43	0,43	0,43	0,43	0,42	0,43	0,43	0,42	0,42	0,42	0,41	0,44
	b_{fg}	1,01	1,01	1,00	1,01	1,01	1,00	1,01	1,01	1,00	1,01	1,01	1,00	1,01
	h^2	0,30	0,30	0,29	0,30	0,30	0,29	0,29	0,29	0,28	0,26	0,26	0,24	0,31

Quando se utiliza o critério de eliminação pelo Call Rate, são excluídas marcas com elevado número de dados perdidos. Como foi observado o decréscimo da capacidade preditiva e da herdabilidade capturada, pode-se concluir que, a imputação dos dados perdidos foi eficiente, uma vez que vários marcadores que provavelmente estão associados a características foram mantidos na analise.

Desse modo, conclui-se que, é importante levar em consideração o controle de qualidade, entretanto, deve-se considerar que este controle pode levar a exclusão de marcadores importantes para determinadas características, como foi observado neste trabalho. Dessa forma, em uma análise de seleção genômica, recomenda-se averiguar diferentes valores de controle de qualidade, visando adotar o melhor corte para cada situação.

Agradecimentos

Os autores agradecem a empresa de papel e celulose VERACEL. A Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) pela concessão da bolsa de estudo, a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico – CNPq pela concessão das bolsas de estudos. E ao professor Cosme Damião Cruz, pelas valiosas considerações.

Referências

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

Gengler, N.; Mayeres, P.; Szydlowski, M. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. **Animal, Cambrige**, v.1, n. 1, p. 21-28, 2007. DOI: 10.1017/S1751731107392628

Meuwissen, T.H. E; Hayes B. J; Goddard, M. E (2001) Prediction of total genetic value using genome-wide dense marker maps. **Genetics** 57:1819-1829.

Meuwissen, T. H. E.; Luan, T.; Woolliams, J. A. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. **Journal of Animal Breeding and Genetics**, v. 128, n. 6, p.429-39, 2011.

Poland, J., J. Endelman et al. 2012. Genomic selection in wheat breeding using genotyping-bysequencing. **Plant Genome** 5:103-113. doi: 10.3835/plantgenome2012.06.0006

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

Resende, M.D.V 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, C. P.; Petroli, C. D.; Carling, J.; Hudson, C. J.; Steane, D. A.; Myburg, A. A.; Grattapaglia, D.; Vaillancourt, R. E. and Kilian, A. A high-density Diversity Arrays Technology (DArT) microarray for genome-wide genotyping in Eucalyptus**. Plant Methods** 2010.