Revista da Estatística UFOP, Vol III(3), 2014, ISSN 2237-8111
Edição especial: 59ª Reunião Anual da Regional Brasileira da Sociedade Internacional de Biometria - RBRAS
Trabalhos apresentados na versão poster.

Correção para estrutura de população baseada em análise de covariância via autovetores da matriz de parentesco genômico

Camila Ferreira Azevedo[1,6]
Marcos Deon Vilela de Resende[2]
Fabyano Fonseca e Silva[3]
José Marcelo Soriano Viana[4]
Magno Sávio Valente[5]
Moyses Nascimento[1]

1 Introdução

A correção para estrutura de população permite eliminar dos valores genéticos totais os efeitos dos genitores ou de famílias, visando trabalhar apenas com o efeito da segregação mendeliana, que consiste em uma estimativa para o mérito genético de indivíduos não aparentados. Em genética de associação (GWAS) essa correção é de extrema importância, pois permite uma análise de associação entre alelos e QTLs (*Quantitative Trait Loci*) devida apenas ao desequilíbrio de ligação (LD) livre de genealogia e evitando os falsos positivos (DAETWYLER et al., 2012). Em contrapartida, na seleção genômica ampla (GWS) ao longo prazo, é inapropriado utilizar o fenótipo bruto ou o valor genético total, sem a correção para efeitos de família, com o objetivo de prever o valor genético dos indivíduos em gerações futuras (RESENDE et al., 2012).

A análise de covariância via autovetores (ajustados como covariáveis de efeitos fixos) associados aos componentes principais da matriz de parentesco genômico (EVG – Eigen Vectors of G) foi apresentada por Price et al. (2006) e Patterson et al. (2006) como uma metodologia estatística para corrigir a estrutura genética de população. Tal técnica pode ser usada como uma alternativa aos procedimentos de correção descritos por Garrick et al. (2009) e Resende et al. (2012), uma vez que nem sempre se tem conhecimento das informações sobre o parentesco genealógico entre os indivíduos que este procedimento requer.

A correção via EVG consiste em incluir no modelo, como covariáveis de efeitos fixos, os autovetores da matriz de parentesco genômica G que estão associados aos maiores autovalores e primeiros componentes principais de G, com o objetivo de capturar a variância genética devida à estrutura de população. Assim, como G está associada aos indivíduos, e não

[1]DET – UFV. e-mail: camila.azevedo@ufv.br
[2] DEF – UFV /EMBRAPA Florestas.
[3]DZO– UFV
[4]DBG – UFV
[5]Programa de Pós-Graduação em Genética e Melhoramento – UFV.
[6] Agradecimentos: Fapemig, Capes e CNPq pelo apoio financeiro.

aos marcadores, os autovetores informam sobre os indivíduos que dominam as relações de parentesco e os agrupam em subgrupos estruturados. Desse modo, o ajuste dos autovetores de G como covariáveis do modelo fornece uma correção para essa estruturação.

Diante do exposto, o presente trabalho teve por objetivo realizar um estudo para avaliar o comportamento da análise de covariância via EVG para a correção de estrutura de população quanto à eficiência na estimação dos valores genômicos livres de parentesco utilizando dados simulados de 1000 indivíduos e 2000 marcadores SNPs em quatro cenários com dois tipos de arquitetura genética e dois níveis de herdabilidade.

2 Material e métodos

Os dados genotípicos simulados são provenientes de um genoma diplóide de comprimento igual a 200 centimorgans (CM), assumindo cromossomos de tamanho igual, com dois alelos cada. O número de marcadores eqüidistantes é de 2000 SNPs (*Single Nucleotide Polymorphisms*), sendo que 100 destes marcadores são considerados realmente genes (QTLs). Um total de 1000 indivíduos pertencentes a 20 famílias de irmãos completos (com 50 indivíduos cada) foram genotipados e fenotipados.

Os dados fenotípicos foram simulados considerando dois tipos de arquitetura genética, a primeira atribuindo um efeito de pequena magnitude no fenótipo para cada um dos 100 QTLs (modelo infinitesimal) e a segunda com três genes de maior efeito responsável por 50% da variabilidade genética e efeitos de pequena magnitude atribuídos aos demais. Além disso, foram definidos dois níveis de herdabilidade no sentido amplo, 0,30 e 0,50, associados à herdabilidades no sentido restrito em torno de 0,20 e 0,35, respectivamente.

Os cenários foram definidos como: Cenário 1, característica controlada por genes de pequeno efeito com herdabilidade 0,3; Cenário 2, característica controlada por genes de pequeno efeito com herdabilidade 0,5; Cenário 3, característica controlada por genes de pequeno e maior efeito com herdabilidade 0,3; Cenário 4, característica controlada por genes de pequeno e maior efeito com herdabilidade 0,5.

O método GBLUP (*Genomic Best Linear Unbiased Predictor*) foi utilizado incluindo no modelo os autovetores da matriz de parentesco genômica G, como pode ser visto a seguir

$$y = Xb + \sum_{i=1}^{v} U_i \alpha_i + Zg + e \quad (1)$$

em que y é o vetor de dados fenotípicos; b é o vetor de efeitos fixos com matriz de incidência X; g é o vetor de efeitos genéticos aditivos (aleatórios) de indivíduos com matriz de

incidência Z=Wm sendo W a matriz de incidência de marcadores e m o vetor de efeitos genéticos aditivos (aleatórios) das marcas, $g \sim N(0, \sigma_g^2 G)$ sendo G a matriz de parentesco aditivo genômica ($G = (WW')/[tr(WW')/N]$, sendo N o número de indivíduos e tr o operador traço matricial) e σ_g^2 a variância genética aditiva do caráter; e é o vetor de resíduos, $e \sim N(0, I\sigma_e^2)$; v é o número (neste trabalho variou de 1 a 50) de autovetores U_i, associados aos componentes principais com os maiores autovalores. Os autovetores foram ajustados como efeitos fixos, em que α_i são os coeficientes da regressão de efeitos fixos.

Os quatros cenários foram simulados dez vezes, nove destas réplicas foram assumidas como populações de treinamento e a réplica restante como população de validação. Em cada repetição, efeitos de marcadores foram estimados via GBLUP com inclusão de autovetores e utilizados para estimar os valores genéticos dos indivíduos na décima população. Assim, a população de validação foi utilizada para avaliar a concordância entre os valores genéticos preditos e os valores paramétricos dos efeitos da segregação mendeliana (fenótipos sem efeito de família) por meio da capacidade preditiva (correlação entre os dois valores, a qual é também equivalente à acurácia da predição, quando se simulam os valores paramétricos) e do viés de estimação. A capacidade preditiva, o viés e a herdabilidade foram obtidos para cada repetição da simulação e o resultado final se deu como a média entre esses valores.

Todas as rotinas computacionais das análises foram implementadas no software R (R Development Core Team, 2010) utilizando os pacotes *pedigreemm* e *rrBLUP* e as funções *pedigreemm* e *mixed.solve*, respectivamente.

3 Resultados e discussões

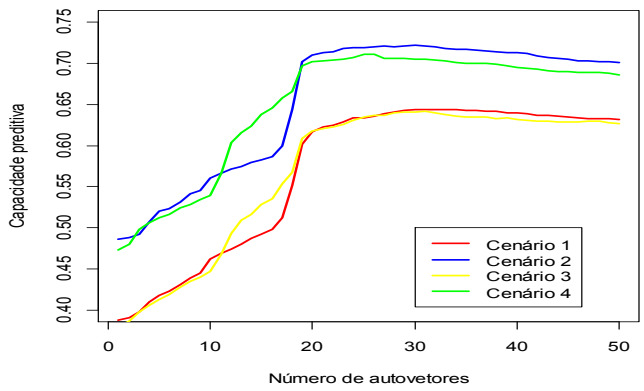

Figura 1 – Capacidade preditiva envolvendo os valores genéticos preditos e os valores paramétricos dos efeitos da segregação mendeliana para cada cenário.

As estimativas médias da capacidade preditiva dos efeitos da segregação mendeliana são apresentadas na Figura 1, considerando a inclusão de 1 a 50 autovetores no modelo. É fácil perceber que existe um padrão de comportamento entre as curvas dos cenários em que os níveis de herdabilidade são coincidentes, apesar da diferença na arquitetura genética. Além disso, as curvas atingem uma assíntota e a capacidade preditiva tende a se estabilizar a partir da inclusão de 20 autovetores no modelo, o que concorda com o reportado por Janss et al. (2012) que afirma que o número de autovetores utilizados nas análises devem ser 10 ou 20.

Os resultados em valores médios para a herdabilidade, a capacidade preditiva e o viés quando se considera 20 autovetores são apresentados na Tabela 1. As estimativas das herdabilidades, em todos os cenários, obtidas pelas análises se comparadas com a herdabilidade paramétrica no sentido restrito (aditiva) foram superestimadas e se comparadas com a herdabilidade no sentido amplo foram subestimadas, como era esperado, uma vez que apenas se considerou o modelo genético aditivo. A herdabilidade no sentido restrito foi superestimada devido ao uso de um modelo de estimação sem dominância, em presença de dominância na simulação.

Tabela 1 – Resultados médios da validação realizada para cada cenário.

Cenários	$^{2}_{a\,par,}$	$^{2}_{d\,par,}$	^{2}	$r_{y,smd}$	$b_{y,smd}$	$r_{g,smd}$	$b_{g,smd}$
Cenário1	0,21	0,10	0,27	0,21	0,20	0,62	1,15
Cenário2	0,35	0,17	0,46	0,32	0,27	0,71	1,89
Cenário3	0,20	0,13	0,24	0,25	0,25	0,62	1,17
Cenário4	0,33	0,21	0,42	0,34	0,32	0,70	1,36

$^{2}_{a\,par.}$é a herdabilidade paramétrica aditiva; $^{2}_{d\,par.}$é a herdabilidade paramétrica devido a dominância; h^2é a herdabilidade obtida pelo método EVG; $r_{y,smd}$ e $b_{y,smd}$ correlação e viés obtidos entre o valor genético predito sem a inclusão de autovetores e a segregação mendeliana; $r_{g,smd}$ e $b_{g,smd}$correlação e viés obtidos entre a segregação mendeliana e o valor genético predito com a inclusão de autovetores.

Os resultados de capacidade preditiva entre o valor genético predito e a segregação mendeliana, obtidos na análise em que não há a inclusão de autovetores foram inferiores aos resultados obtidos quando se inclui tais covariáveis. Os baixos valores são consequências da inadequada estimação do valor genético quando não se corrige para efeito de estrutura de população, pois para capturar bem os valores genéticos não corrigidos, não se deve corrigir para estrutura de população (Powell et al., 2010). Além disso, os vieses de predição foram inferiores a 1 quando não houve a inclusão de autovetores e superiores a 1 quando houve, indicando que os valores genéticos preditos foram superestimados e subestimados, respectivamente (RESENDE et al. 2012). Mas os vieses foram menores com o método EVG.

Com a inclusão dos autovetores, a predição torna-se, em média, 2,5 vezes mais eficiente do que no modelo tradicional visto o aumento da capacidade preditiva. Outra vantagem dessa correção na GWS, é que o catálogo de efeitos de marcas estimado é válido para a predição em geração futuras, sem a necessidade de uma re-estimação de novos efeitos de marcas.

4 Conclusão

De forma geral, a análise com EVG pode ser uma alternativa viável e eficiente para a correção de estrutura da população, principalmente quando não se tem conhecimento das relações de parentesco genealógico. O método GBLUP com EVG propicia uma adequada GWAS e eficiente GWS em gerações avançadas de melhoramento.

5 Bibliografia

[1]DAETWYLER, H.D.; KEMPER, K.E.; VAN DER WERF, J.H.J.; HAYES, B.J. Components of the accuracy of genomic prediction in a multi-breed sheep population. **Journal of Animal Science**. v. 90, n. 10, p. 3375-3384,2012.

[2]GARRICK, D. J.; TAYLOR, J. F.; FERNANDO, R. L. Deregressing estimated breeding values and weighting information for genomic regression analyses. **Genetics Selection Evolution**, v. 41, p. 55, 2009.

[3]JANSS, L.; DE LOS CAMPOS, G.; SHEEHAN, N., SORENSE, D. Inferences from Genomic Modelsin Stratified Populations. **Genetics**. v. 192, p. 693-704, 2012.

[4]PATTERSON, N. J.; PRICE, A.L.; REICH, D. Population Structure and Eigenanalysis. **Plos Genetics**. v. 2, n. 12, p. 2074 - 2093, 2006.

[5]POWELL, J. E., VISSCHER, P. M., GODDARD, M. E. Goddard. Reconciling the analysis of IBD and IBS in complex trait studies. **Nature Reviews Genetics**, v. 11, p.800-805, 2010.

[6]PRICE, A.L., PATTERSON, N. J., PLENGE, R.M., WEINBLATT, M.E., SHADICK, N.A., REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics**.v. 38, n. 8, p. 904-909,2006.

[7]R DEVELOPMENT CORE TEAM. *R*: A **language and environment for statistical computing**. R Foundation for Statistica Computing, Vienna, Austria, 2010.Disponivel em: http://www.R–project.org.

[8]RESENDE, M.D.V.; SILVA, F.F.; LOPES, P.S.; AZEVEDO, C.F. **Seleção Genômica Ampla (GWS) via Modelos Mistos (REML/BLUP), Inferência Bayesiana (MCMC), Regressão Aleatória Multivariada e Estatística Espacial**. 1. ed. , 2012.