MÉTODOS DE ESTIMAÇÃO E VALIDAÇÃO NA SELEÇÃO GENÔMICA

Ísis Fernanda de Almeida[1], Cosme Damião Cruz[2], Marcos Deon Vilela de Resende[3], Ramon Vinícius de Almeida[4], Fernanda Rodrigues Mendes[5]

Resumo: A seleção genômica ampla (GWS) 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. Este trabalho teve por objetivo avaliar a acurácia ao se aplicar diferentes métodos estatísticos de seleção genômica (RR-BLUP e BLASSO) e diferentes formas de validação. Foram simulados genomas com dez grupos de ligação contendo 100 marcas bi-alélicas e co-dominantes por grupo, totalizando 1010 marcadores apresentando intervalos entre marcas adjacentes equidistantes. Esses genomas foram utilizados para a geração da população simulada de famílias de meios-irmãos com 1.000 indivíduos. As diferentes características quantitativas, uniformes e binomiais, de herdabilidade 0,40, foram avaliadas. Para um tipo de validação criou-se uma população de validação independente com 1000 indivíduos e para outro tipo, aplicou-se a validação cruzada de Jacknife, usando a mesma população para estimação e validação. Os programas estatísticos utilizados foram GENES, SELEGEN GENÔMICA e R. As acurácias apresentaram valores entre 0,55 e 0,92. De forma geral, a validação feita através da população independente apresentou valores mais elevados de acurácia. Para famílias de meios-irmãos e com distribuição binomial dos efeitos genéticos, o método BLASSO foi ligeiramente superior ao RR-BLUP. Para a distribuição uniforme, os dois métodos de análise apresentaram valores superiores de acurácia a depender da forma de validação.

PALAVRAS-CHAVE: análise genômica, genética quantitativa, marcadores moleculares.

[1]Professora Mestre da Universidade Estadual de Goiás. Unidade Palmeiras de Goiás. Rua 07, S/N, Setor Sul - Palmeiras de Goiás- Goiás- 76.190.000 - Fone-Fax: 643571-1198. Email: isisagro@gmail.com
[2]Professor Doutor da Universidade Federal de Viçosa. Av. Peter Henry Rolfs, S/N, Campus Universitário. Email: cdcruz@ufv.br
[3]Professor Doutor da Universidade Federal de Viçosa/Embrapa Florestas. Av. Peter Henry Rolfs, S/N, Campus Universitário. Email: marcos.deon@ufv
[3]Professor Doutor. Instituto Federal Triângulo Mineiro- Rua João Batista Ribeiro n. 4000 – Distrito Industrial II. Uberaba-MG. Email: ramon@iftm.edu.br
[4]Professora Mestre da Universidade Estadual de Goiás. Unidade Palmeiras de Goiás. Email: mendesfr@yahoo.com.br

METHODS OF ESTIMATION AND VALIDATION IN GENOMIC SELECTION

Abstract: A genome-wide selection (GWS) is the simultaneous use of hundreds or thousands of markers, which cover the genome full, so that all genes of a quantitative trait are in linkage disequilibrium with at least a portion of the markers. The objective this work to evaluate the accuracy by applying different statistical methods of genomic selection (RR-BLUP and BLASSO) and different forms of validation. Genomes were simulated with ten linkage groups containing 100 bi-allelic and co-dominant marks per group, totaling 1010 markers and presenting equidistant intervals between adjacent marks. These genomes were used for the generation of the simulated population of half-sib families with 1000 individuals. Different quantitative traits, uniform and binomial, heritability 0.40, were evaluated. For a type validation, an independent validation population of 1000 individuals was created and for another type, Jacknife cross-validation was applied, using the same population for estimation and validation. The statistical programs used were GENES, SELEGEN GENOMICA and R. The accuracies had values between 0.55 and 0.92. Generally, the validation done through validation population showed values higher accuracy. For half sib families and binomial distribution of genetic effects, BLASSO was slightly superior to RR-BLUP. For the uniform distribution, the two methods provided superior accuracy according to the validation scheme adopted.

.

KEYWORDS: genomic analysis, quantitative genetics, molecular markers.

INTRODUÇÃO

A seleção genômica ampla (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. Por atuar em todo o genoma é considerada ampla, capturando todos os genes que afetam um caráter quantitativo, sem a necessidade de identificar previamente os marcadores com efeitos significativos, ou seja, sem o uso de testes de significância para marcas individuais, e

de mapear QTL (Quantitative Trait Loci), como no caso da MAS (Seleção Assistida por Marcadores). Os métodos tradicionais de seleção usam o fenótipo para inferir sobre o efeito do genótipo e a GWS usa o genótipo, com efeito genético pré-estimado em uma amostra da população, para inferir sobre o fenótipo a ser expresso nos candidatos à seleção (RESENDE, 2008).

A implementação da seleção genômica impõe vários desafios estatísticos e computacionais como a dimensionalidade do modelo, colinearidade entre marcas e a complexidade das características quantitativas. Para isso, vários métodos têm sido propostos, que diferem entre si pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo, entre eles o RR-BLUP e o BLASSO.

O método RR-BLUP (ridge regression-BLUP) estima simultaneamente os efeitos de todas as marcas (MEUWISSEN et al., 2001; WHITTAKER et al., 2000), sendo estas consideradas efeitos aleatórios com variância comum, ou seja, assumem que todos os marcadores contribuem igualmente para a variação genética (ausência de genes de efeitos maiores). No entanto, assumir que as marcas individuais possuem a mesma variância pode não ser apropriado para algumas marcas localizadas em regiões não associadas à variância genética, enquanto outras estão localizadas em regiões associadas a QTL (GODDARD e HAYES, 2007). Para contornar esta questão, muitos autores sugeriram metodologias que utilizam efeito shrinkange específico para cada marcador, que é o caso da metodologia BLASSO de Park e Casella (2008).

O estimador LASSO Bayesiano (BLASSO) ajusta uma variância separada para cada marca e força os estimadores para zero, como no caso do RR-BLUP. No entanto, efetivamente permite que alguns estimadores sejam identicamente iguais a zero, realizando simultaneamente o procedimento de shrinkage e seleção de covariáveis (DE LOS CAMPOS et al., 2009).

Baseando-se nas variadas abordagens associadas a cada metodologia de análise, o objetivo deste trabalho foi avaliar a acurácia ao se aplicar diferentes métodos estatísticos de estimação (RR-BLUP e BLASSO) sob diferentes formas de validação, utilizando uma população, simulada, de famílias de meios-irmãos.

MATERIAIS E MÉTODOS

Simulação dos dados fenotípicos e genotípicos

Foram simulados genomas com dez grupos de ligação, contendo 100 marcas bi-alélicas e co-dominantes por grupo, com intervalos entre marcas adjacentes equidistantes, o que gerou um total de 1010 marcadores. Dentre esses marcadores, 100 estariam associados a características quantitativas. Os genomas gerados foram utilizados para a simulação de uma família de meios-irmãos composta por 1.000 indivíduos. Esse tipo de família foi escolhido porque é muito utilizada no melhoramento genético vegetal e animal.

Diferentes características quantitativas também foram simuladas; uma controlada por genes de iguais e pequenos efeitos e outra apresentando genes de maior efeito, com base numa distribuição binomial. A herdabilidade aplicada no estudo foi de 0,40.

Metodologias de Análise

A partir dos dados fenotípicos gerados, foram estimados os efeitos de cada um dos locos marcadores que somados, compõem o valor genético genômico predito de cada indivíduo. Uma das formas de estimação foi através da metodologia RR-BLUP.

Esse método é denominado regressão aleatória ou regressão de cumeeira (*Ridge Regression*-BLUP) . Os coeficientes de regressão *ridge* são definidos como aqueles que minimizam a seguinte função:

$$(1/N)\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 + \lambda\sum_{i=1}^{n}\beta_i{}^2$$

em que λ é o parâmetro de penalização (ou shrinkage) ou parâmetro *ridge*, *n* é o número de marcadores e *N* é o número de indivíduos. O primeiro termo da equação é a soma de quadrados dos resíduos da regressão e o segundo termo é a penalização, a qual depende da magnitude dos coeficientes de regressão via $\sum_{i=1}^{n}\beta_i{}^2$. Quando λ não é conhecido, a escolha arbitrária do mesmo leva ao método de regressão *ridge regression* (RR). Se o parâmetro de regressão for associado a $\lambda = \sigma_e^2/\sigma_{gi}^2 = \sigma_e^2/(\sigma_g^2/n)$ tem-se a regressão aleatória BLUP para o efeito do segmento cromossômico *i*, em que σ_{gi}^2 é a variância genética associada ao loco ou segmento *i* e σ_g^2 e σ_e^2 são a variância genética do caráter e variância residual, respectivamente. O seguinte modelo linear misto geral foi usado para estimar os

efeitos dos marcadores, conforme Resende et al. (2008).

$$y = Xb + Zh + e;$$

em que:

y é o vetor de observações fenotípicas,

b é o vetor de efeitos fixos,

h é o vetor dos efeitos aleatórios dos marcadores e

e refere-se ao vetor de resíduos aleatórios.

X e *Z* são as matrizes de incidência para *b* e *h*.

O outro método, LASSO, combina shrinkage (regularização) com seleção de covariáveis e envolve o seguinte problema de otimização, via minimização de:

$$(1/N)\sum_{j}^{N}(y_j - \sum_{i=1}^{n} x_{ij}\beta_i)^2 + \lambda\sum_{i=1}^{n}|\beta_i|$$

em que $\sum_{i=1}^{n}|\beta_i|$ é a soma dos valores absolutos dos coeficientes de regressão. As soluções em que os coeficientes de regressão se distanciam de zero sofrem regularização através do parâmetro λ. No LASSO Bayesiano (BLASSO), esse parâmetro controla a precisão da distribuição *a priori* atribuída aos coeficientes de regressão. Para isso, pode ser implementado via analise Bayesiana do LASSO (DE LOS CAMPOS et al., 2009). Essa implementação impõe como distribuição *a priori* dos "p" coeficientes de regressão um produto de densidades exponenciais duplas:

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

Para construção da distribuição conjunta *a priori* dos parâmetros, os autores exploram o fato de a distribuição exponencial dupla poder ser representada como uma mistura de densidades normais com parâmetro de escala, com o processo de mistura de variâncias controlado por distribuição exponencial. Na distribuição *a priori* conjunta construída, a densidade atribuída aos coeficientes de regressão regularizados por LASSO será: $\prod_{j=1}^{p} N(\beta_{j1}|0, \sigma_e^2\tau_j^2)$, resultando em variâncias específicas para cada coeficiente de regressão. Por sua vez, a distribuição *a priori* para o parâmetro de escala τ_j^2 será representada por: $\prod_{j=1}^{p} exp(\tau_j^2|\lambda)$, no qual o parâmetro de suavização λ influencia o ajuste dos coeficientes de regressão. A informação *a priori* para esse parâmetro é dada por uma distribuição com hiperparâmetros conhecidos. Caso a distribuição escolhida seja conjugada, é

possível de obter amostras da distribuição, a posteriori conjunta, por meio de um amostrador de Gibbs.

Validação

A validação cruzada foi realizada pela reamostragem de um grupo de indivíduos via procedimento Jacknife (HELTSHE e FORRESTER, 1983). A metodologia generalizada do Jacknife baseia-se 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 caso, tomando-se k=1, a população de validação usa todos os C indivíduos e a população de estimação C-1 indivíduos e, em C repetições. Em cada repetição, 200 indivíduos foram removidos da população e utilizados para a formação da população de validação, sendo os outros indivíduos restantes utilizados na população de estimação dos efeitos dos marcadores. Além da validação de Jacknife foi utilizada uma população de validação distinta da população de estimação composta por mil indivíduos.

Como na validação dos resultados o valor fenotípico é conhecido, a seleção genômica em cada subgrupo analisado foi avaliada ao calcular a correlação do valor genético predito com o fenótipo observado nos indivíduos. Esta correlação é conhecida como capacidade preditiva ($r_{y\hat{y}}$) da seleção genômica em estimar os fenótipos, sendo dada teoricamente pela acurácia de seleção ($r_{g\hat{g}}$) multiplicada pela raiz quadrada da herdabilidade individual (*h*) ou, em outras palavras $r_{y\hat{y}} = r_{g\hat{g}}.h$ (RESENDE, 2008). Por outro lado, os coeficientes de correlação envolvendo valores genéticos genômicos verdadeiros (*GBV* fixados na simulação) e preditos ($G\hat{B}V$) fornecem diretamente a acurácia seletiva e foram utilizados a fim de medir a capacidade do método em predizer de forma acurada.

Análises estatísticas

As simulações foram implementadas através dos softwares GQMOL e GENES (Cruz, 2006) e as análises estatísticas através dos softwares SELEGEN GENÔMICA (Resende, 2007) para o método RR-BLUP com validação Jacknife e R (R Development Core Team 2011) para o método BLASSO com validação independente.

RESULTADOS E DISCUSSÃO

A característica quantitativa com distribuição uniforme apresentou efeitos fixos de QTL dentro dos 100 locos simulados, por outro lado, a binomial apresentou efeitos variados dentro de cada loco, com alguns QTLs de maior efeito (Figura 1).

Para a validação de Jacknife, um percentual de indivíduos é retirado da amostra a fim de se realizar a estimação dos valores genômicos das marcas usando os indivíduos restantes. Não houve variação acentuada nos valores de acurácia entre as diferentes características quantitativas, exceto para o método RR-BLUP na população de validação independente, onde houve uma amplitude de variação de 0,10 (Tabela 1).

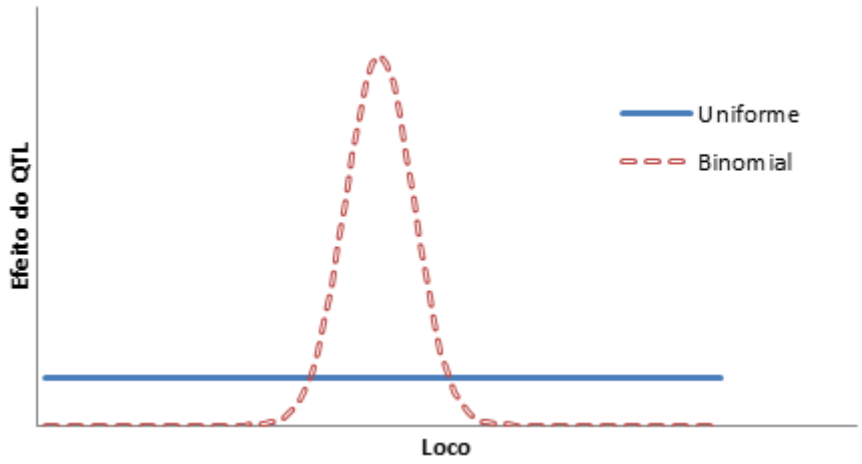

Figura 1 - Distribuição de efeitos de QTL para cada loco

Na validação de Jacknife a acurácia apresentou valores em torno 0,55 e 0,69 para o RR-BLUP e o BLASSO, respectivamente (Tabela 1). Em contrapartida, a população de validação apresentou acurácias mais elevadas, variando de 0,82-0,92 para o RR-BLUP e de 0,85-0,86 para o BLASSO. Em simulações feitas por Meuwissen et al. (2001) o RR-BLUP apresentou uma acurácia de 0,732. Em adição a esses estudos, Muir (2007) simulou 512 genótipos para um caráter com uma baixa herdabilidade (h^2=0,1) que resultou numa acurácia ainda maior, de 0,83. Crossa et al.(2010) genotipando 264 linhas de milho com 1135 marcadores obtiveram uma acurácia de 0,53 para produção de grãos utilizando o BLASSO.

Tabela 1 - Acurácia dos métodos RR-BLUP e BLASSO

		Métodos	
	Característica	RR-BLUP	BLASSO
Validação de Jacknife	Uniforme	0,55	0,68
	Binomial	0,55	0,69
População de Validação Independente	Uniforme	0,92	0,85
	Binomial	0,82	0,86

Alguns autores apontam para a superioridade de métodos bayesianos em relação o método BLUP (MEUWISSEN et al. 2001; USAI et al.2009; CROSSA et al. 2010). Entretanto, há relatos quanto a inversão deste comportamento esperado (ZHONG et al.2009; HABIER et al.2010). Zhong et al. (2009) obtiveram valores de 0,62 e 0,61 de acurácia para RR-BLUP e Bayes B, respectivamente. Segundo estes autores, uma distribuição de efeitos aleatórios mais complexa, como a utilizada em métodos bayesianos, só é útil quando as marcas estão fortemente associadas com o QTL de grande efeito. Tal associação ocorre quando o efeito do QTL é elevado e quando as marcas estão em alto desequilíbrio de ligação com o QTL de importância. Como as duas distribuições (uniforme e binomial) envolvidas na análise não apresentam QTLs de grandes efeitos, as acurácias apresentaram valores aproximados.

Na validação de Jacknife, oitocentos indivíduos foram utilizados em cada processo de estimação, duzentos a menos que na população de validação. Ou seja, o menor valor de acurácia observado para a validação de Jacknife pode ser resultado do menor número de indivíduos da população de estimação. O contrário também seria verdadeiro, de forma que quando utilizou-se a população de validação independente, um maior número de indivíduos contribuiu para a elevação do valor da acurácia.

Ao verificar a influência do parentesco e do tamanho da população de estimação na acurácia dos valores genômicos, Habier et al. (2010) observaram que em populações com maior parentesco o RR-BLUP apresentou valores mais acurados que o Bayes B, e que, para

ambos os métodos, os valores de acurácia decaíram com a diminuição da população de estimação. Corroborando com esses resultados, Van Raden et al. (2009) avaliando uma população de 3500 touros, genotipados com 38416 SNPs, obtiveram acurácias de 0,44 a 0,79 para características com herdabilidades entre 0,04 e 0,50, sendo que tanto o RR-BLUP quanto o Bayes B apresentaram valores semelhantes. Ao diminuir o número de marcas em 75% a acurácia decaiu de 0,53 para 0,50, ao passo que a diminuição na população de estimação em 68% a acurácia caiu de 0,53 para 0,35. De acordo com esses autores, aumentos na acurácia são lineares ao incremento do número de indivíduos.

CONCLUSÕES

Para famílias de meios-irmãos e com distribuição binomial dos efeitos genéticos, o método BLASSO foi ligeiramente superior ao RR-BLUP.

Para a distribuição uniforme, os dois métodos de análise apresentaram valores superiores de acurácia a depender da forma de validação.

Na comparação entre as formas de validação, a diferença entre os valores de acurácias observados está relacionada ao tamanho diferenciado da população de estimação.

REFERÊNCIAS

CROSSA, J.;DE LOS CAMPOS, G.;PEREZ, P.;GIANOLA, D.;BURGUENO, J.;ARAUS, J. L.;MAKUMBI, D.;SINGH, R. P.;DREISIGACKER, S.;YAN, J. B.;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, n.2, p.713-U406, 2010.

CRUZ, C.D. **Programa Genes: Biometria**. Editora UFV. Viçosa (MG), 2006. 382p.

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, n.1, p.375-385, 2009.

GODDARD, M. E.;HAYES, B. J. Genomic selection. **Journal of Animal**

Breeding and Genetics, v.124, n.6, p.323-330, 2007.

HABIER, D.;TETENS, J.;SEEFRIED, F. R.;LICHTNER, P.;THALLER, G. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. **Genet Sel Evol**, v.42, p.5, 2010.

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

MEUWISSEN, T. H. E.;HAYES, B. J.;GODDARD, M. E. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, v.157, n.4, p.1819-1829, 2001.

MUIR, W. M. Comparasion of genomic and traditional BLUP-estimated breeding values accuracy and selection response under alternative trait and genomic parameters. **Journal of Animal Breeding and Genetics**, v.124, p.342-355, 2007.

PARK, T.;CASELLA, G. The Bayesian Lasso. **Journal of the American Statistical Association**, v.103, n.482, p.681-686, 2008.

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. 2007. Selegen Genômica RR-BLUP. Sistema de Seleção Genômica Ampla (GWS) computadorizada via modelos lineares mistos. (CD-ROM).

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

RESENDE, M. D.;LOPES, P. S.;SILVA, R. L.;PIRES, I. E. Seleção genômica ampla (GWS) e maximização da eficiência do melhoramento genético. **Pesquisa Florestal Brasileira**, v.56, p.63-78, 2008.

USAI, M. G.;GODDARD, M. E.;HAYES, B. J. LASSO with cross-validation for genomic selection. **Genetics Research**, v.91, n.6, p.427-436, 2009.

VAN RADEN PM, TASSELL CP, WIGGANS GR, SONSTEGARD TS, SCHANABEL RD, TAYLOR JF,

SCHENKEL FS. Reliability of genomic predictions for North American Holstein bulls. **Jornal Dairy Science**, v. 92, p.16–24, 2009.

WHITTAKER, J. C.;THOMPSON, R.;DENHAM, M. C. Marker-assisted selection using ridge regression. **Genetical Research**, v.75, n.2, p.249-252, 2000.

ZHONG, S, DEKKERS, JCM., FERNANDO, RL., JANNINK, JL. Factors Affecting Accuracy From Genomic Selection in Populations Derived From Multiple Inbred Lines: A Barley Case Study. **Genetics**, v. 182, p.355-364, 2009.