45ª Reunião Anual da Sociedade Brasileira de Zootecnia
Lavras, MG - UFLA - 22 a 25 de julho de 2008
45ª Reunião Anual SBZ2008 Lavras-MG

Predição do valor genético total de touros da raça Holandesa por meio de mapas densos de marcadores

Marcos Vinicius B. Silva[1,2], Curtis Van Tassell[1], Tad Sonstegard[1], Lakshmi Matukumalli[3], Steven Schroeder[1], Paul VanRaden[4], George Wiggans[4]

[1]Bovine Functional Genomics Laboratory, Agricultural Research Service, USDA Beltsville, MD, 20705, USA
[2]Embrapa Gado de Leite, Juiz de Fora, MG, 36038-330, Brasil – Bolsista do CNPq. e-mail: marcos@ars.usda.gov
[3]Bioinformatics and Computational Biology, George Mason University, Manassas, VA 20110, USA
[4]Animal Improvement Programs Laboratory, Agricultural Research Service, USDA Beltsville, MD

Resumo: A seleção genômica é a forma de seleção assistida em que os marcadores genéticos distribuídos ao longo de todo o genoma são usados na avaliação genética, esperando-se que os *quantitative trait loci* (QTL) estejam em desequilíbrio de ligação com pelo menos um marcador. O uso da seleção genômica tornou-se possível graças ao grande número de SNPs descorbertos durante o seqüenciamento do genoma bovino e aos novos métodos de genotipagem de SNPs em larga escala, como o novo *Illumina iSelect Bovine 50K Chip*. Na seleção genômica, a escolha dos indivíduos é baseada na soma das estimativas dos efeitos de todos os marcadores ao longo do genoma, ajustados como efeitos aleatórios. O objetivo desse trabalho foi usar a *ridge regression* para predizer o valor genético genômico de 474 touros jovens holandeses nascidos após 2000, a partir de grande número de marcadores simples em todo o genoma. Para tanto, foram usados os valores genéticos de 2.378 touros da raça Holandesa nascidos antes de 2000, publicados em Agosto/2003. Os efeitos dos marcadores foram estimados por *ridge regression* (solução BLUP) por meio de programas escritos nas linguagens Fortran e R *environment*. Pode-se concluir que a *ridge regression* não é metodologia mais adequada para estimação dos valores genômicos e que, provavelmente, métodos bayesianos que assumam distribuição *a priori* das variâncias associadas com os segmentos cromossômicos possam fornecer estimativas mais confiáveis dos valores genômicos.

Palavras–chave: bovinos, marcador genético, seleção genômica, SNP

Prediction of total genetic value in Holstein breed using dense marker maps

Abstract: Genomic selection is a form of marker-assisted selection in which genetic markers covering the whole genome are used, in which all quantitative trait loci (QTL) are in linkage disequilibrium with at least one marker. The use of genomic selection approach has become feasible thanks to the large number of single nucleotide polymorphisms (SNP) discovered by genome sequencing and new methods to efficiently genotype large number of SNPs as the new Illumina iSelect Bovine 50K Chip. With genomic selection, selection can be based on the sum of estimates of effects of all markers across the genome, fitted as random effects. The objective of this paper was estimate the genomic breeding value of 474 Holstein young bulls, borne after 2000, predicted from a large number of estimated single marker effects across the whole genome using the PTA milk estimated from 2,378 bulls were borne before 2000. The marker effects were estimated by BLUP estimation (ridge regression) using scripts in R software and Fortran. It was concluded that the ridge regression it is not so accurate for estimating the genomic breeding value. Bayesian methods that assumed a prior distribution for the variance associated with each chromosome segment probably can give more accurate predictions of breeding values than ridge regression.

Keywords: bovine, genetic marker, genomic selection, SNP

Introdução

Por volta de 1990, o foco das pesquisas em genética quantitativa passou a ser direcionado também para a genética molecular. Uma das justificativas é que o uso de informações do DNA possibilite acelerar o ganho genético em relação ao método tradicional, o qual utiliza somente os registros fenotípicos (Meuwissen et al., 2001). Sob esse novo enfoque, os estudos foram divididos em duas etapas: a detecção de marcadores associados a *quantitative trait loci* (QTL) e o seu uso na seleção assistida por marcadores (SAM) (Misztal, 2006). Quando marcadores de DNA estão disponíveis, eles

podem ser usados para determinar se a variação alélica dos locos dos marcadores, ao longo do mapa de ligação, está associada à variação na característica quantitativa.

Até recentemente, o número de marcadores de DNA identificados no genoma de bovinos era limitado e o custo de genotipagem era alto, o que restringia os experimentos delineados para o mapeamento de QTL. Todavia, segundo Van Tassell et al. (2007), a genômica de bovinos entrou em uma nova era pela disponibilidade de dados de seqüências de todo o genoma. Esses autores desenvolveram um chip de 50.000 *single-nucleotide polymorphism* (SNP) possibilitando assim a genotipagem de animais em larga escala. Desta forma, pode-se cobrir todo o genoma com marcadores possuindo distância média de 0,05 cM, assumindo-se que os mesmos são informativos e estão em desequilíbrio de ligação (DL). Objetivou-se, com esse trabalho, usar a *ridge regression* para predizer o valor genético genômico de 474 touros jovens holandeses em avaliação genética nos EUA, a partir de dados genotípicos de aproximadamente 50.000 SNP, gerados por meio da tecnologia *iSelect™ Infinium Custom Genotyping* da Illumina.

Material e Métodos

Neste estudo, foram utilizadas informações dos genótipos de 39.835 *single-nucleotide polymorphism* (SNP), cuja menor frequência alélica era superior a 5%, geradas pelo *Bovine Functional Genomics Lab*, do *Agricultural Research Service/ United States Department of Agriculture(USDA)*, em Beltsville, EUA, por meio da tecnologia iSelect™ Infinium Custom Genotyping da Illumina (*Illumina Bovine SNP50 chip*).

O conjunto de dados inicial consistia das informações dos valores genéticos para produção de leite da população referência (2.138 touros da raça Holandesa, nascidos antes do ano 2000) provenientes da avaliação genética realizada em Agosto/2003, pelo USDA, nos EUA. Usando-se esse conjunto de dados, as predições genômicas foram feitas por meio da *ridge regression*, conforme proposto por Meuwissen et al. (2001), o qual considera as soluções geradas por essa metodologia como sendo BLUP. Nesse caso, assume-se que os efeitos dos haplótipos em cada segmento cromossômico são iguais. Segundo Hayes (2007), a *ridge regression* pode ser aplicada à seleção genômica da seguinte forma:

$$\begin{bmatrix} 1_n'1_n & 1_n'X \\ X'1_n & X'X + I\lambda \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} 1_n'y \\ X'y \end{bmatrix} \text{ e } \hat{g} = (X'X + \lambda I)^{-1} X'y$$

onde $\mathbf{1_n}$ é um vetor onde todos os elementos são iguais a 1, **X** é a matriz incidência ligada aos genótipos dos 39.835 marcadores SNP, **y** é o vetor contendos os valores genéticos dos 2.138 animais, $\hat{g}$ é o vetor contendo as soluções para todos os efeitos dos marcadores e $\lambda = \sigma_e^2 / \sigma_g^2$. O valor de λ usado foi igual a 30, considerando-se herdabilidade da produção de leite como 0,25.

Após a estimação do vetor de soluções na população referência, multiplicou-se a matriz de genótipos dos 474 touros candidatos à seleção por esse vetor, obtendo-se as estimativas dos valores genômicos para esses animais.

De modo a verificar a precisão das estimativas obtidas foram estimados os coeficientes de regressão e as correlações entre os valores genéticos e os valores genômicos em ambas as populações. Ressalte-se que, para a população de candidatos à seleção, os valores genéticos usados foram aqueles publicados pelo USDA em Agosto/2007.

Nas análises estatísticas foram usados programas em ambiente R (R Development Core Team, 2007) e Fortran.

Resultados e Discussão

As relações entre os valores genéticos e os valores genômicos estimados, nas populações referência e de candidatos à seleção, podem ser observadas na Figura 1. Vê-se que, na população de referência, a relação entre os valores foi linear, como esperado. O coeficiente de regressão ± erro-padrão foi igual a 1,00000 ± 0,00008 e a correlação, 0,99.

Na população de candidatos à seleção, pode ser observada a dispersão dos valores, sem nenhuma tendência clara. Apesar do coeficiente de regressão ter sido significativo (P<0,01), ele foi muito reduzido (0,19 ± 0,05). A correlação estimada entre os valores genéticos e genômicos, nessa população, também foi baixo (0,16).

Provavelmente, a razão da reduzida acurácia obtida pode ser explicada pela igual ponderação dada pela *ridge regression* a todos os marcadores distribuídos pelo genoma, enquanto, de fato, muitos dos marcadores têm efeito muito pequeno ou simplesmente não possuem efeito nenhum sobre a característica.

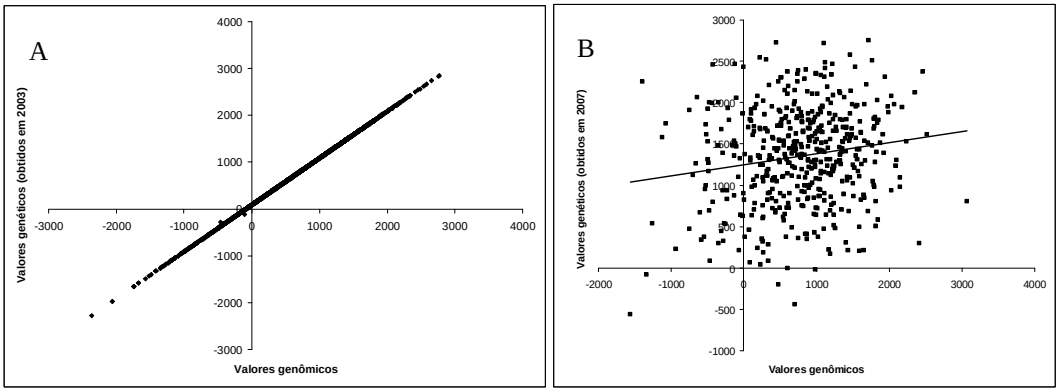

Figura 1 Relações entre os valores genéticos e os valores genômicos nas populações referência (A) e de candidatos à seleção (B)

Um ponto importante nessa discussão, é qual o valor de σ_g^2 (ou seja, da variância dos efeitos dos marcadores no segmento cromossômico) deve ser usado no BLUP. Um modo, Segundo Meuwissen et al. (2001), é o uso de modelo de deriva genética-mutação. Outra maneira de estimar σ_g^2 seria primeiro estimar a variância genética aditiva total (usando-se REML, por exemplo), dividindo-se posteriormente pelo número de segmentos cromossômicos. Melhores estimativas dos valores genéticos podem ser obtidas, de acordo com Hayes (2007), por métodos que permitam que as variâncias dos segmentos cromossômicos possam ser diferentes, como os métodos bayesianos, por exemplo.

Outra questão é que, apesar da distância média entre os marcadores ser de, aproximadamente, 0,05 cM, ela poderá na verdade ser muito grande entre alguns marcadores, fazendo com que alguns QTL não sejam incluídos na avaliação. Uma provável solução para isso pode ser o uso de blocos de haplótipos de marcadores e não o de marcadores simples.

A seleção genômica aumenta a taxa de ganho genético e reduz o custos de testes de progênie tradicionais em até 92% (Schaeffer, 2006), pois permite aos criadores selecionarem precocemente os animais, de ambos os sexos, que tenham herdados segmentos cromossômicos de maior mérito. Para tanto, o desenvolvimento de metodologias, de algoritmos computacionais e de programas é necessário para incorporar dados genômicos em programas de avaliação genética de modo adequado.

Conclusões

A *ridge regression* não parecer ser a metodologia mais adequada para estimação dos valores genéticos genômicos, em mapas densos de marcadores. Métodos que permitam que as variâncias dos segmentos cromossômicos sejam diferentes, como os bayesianos, podem ser mais apropriados.

Literatura citada

HAYES, B. QTL mapping, MAS, and Genomic Selection - A Short course. Ames, Iowa, 2007.

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.

MISZTAL, I. Challenges of application of marker assisted selection – a review. **Animal Science Papers and Reports**, v. 24, p. 5-10, 2006.

VAN TASSELL, C.P., MATUKUMALLI, L., TAYLOR, C., SMITH, T., SONSTEGARD, T., SCHANABEL, R., SILVA, M.V.G.B., WIGGANS, G., LIU, G., MOORE, S., TAYLOR, J. 2007. **Construction and application of a bovine high-density SNP assay**. American Dairy Science and Society of Animal Science. San Antonio, USA.(CD-ROM)

R DEVELOPMENT CORE TEAM. 2007. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. (URL http://www.R-project.org)

SCHAEFFER, L. Strategy for applying genome-wide selection in dairy cattle. **Journal of Animal Breeding and Genet**ics , v.123, p. 218-223.