DETECCION DE SNPs REDUNDANTES EN GENOTIPADOS DE ALTA DENSIDAD Y USO DE *BAGGING* PARA INCREMENTAR LA PRECISIÓN EN LA PRE-SELECCIÓN DE SNP PARA LA SELECCION GENOMICA

González- Recio, O. [1*], Naya, H. [2], Weigel, K. A. [3], Gianola, D. [4] y Rosa, G.J.M. [3]
[1]Departamento de Mejora Genética Animal. Instituto Nacional de Tecnología Agraria y Alimentaria. Ctra. La Coruña km 7,5. 28040 Madrid.
[2]Unidad de Bioinformática, Institut Pasteur de Montevideo, Uruguay.
[3]Dairy Science department. University of Wisconsin-Madison. 53076 WI,USA.
[4]Animal Science department. University of Wisconsin-Madison. 53076 WI,USA.
Correo electrónico: gonzalez.oscar@inia.es

INTRODUCCIÓN

La posibilidad de genotipar animales para una gran cantidad de SNPs a lo largo del genoma ha generado un gran interés para su aplicación en la selección animal. Sin embargo, estas nuevas tecnologías genotipan una gran cantidad de SNPs, que sobrepasan el número de individuos genotipados, lo que hace que los modelos tradicionales no sean efectivos bajo estas circunstancias. En los últimos años se han propuesto diferentes estrategias para abordar la sobreparametrización de estos modelos. Caben destacar los modelos Bayesianos jerarquizados en los que se asumen varianzas especificas a priori para los efectos o varianzas de los SNPs (e.g. Meuwissen et al., 2001; Gianola et al., 2003). Sin embargo, la mayoría de estos modelos adolecen de supuestos demasiado estrictos, o de severas distribuciones a priori que no permiten el aprendizaje Bayesiano. Los métodos no paramétricos y de machine learning (Gianola et al., 2006) permiten analizar datos con ruido, redundancia, e inconsistencias, ya que la función que relaciona los datos con las covariables es inespecífica.

En este trabajo se evalúan dos nuevas estrategias de machine learning aplicadas a la selección genómica para reducir la dimensionalidad y colinearidad de los genotipos, y para mejorar la habilidad predictiva del método.

MATERIAL Y MÉTODOS

Se usaron las evaluaciones genéticas de vida productiva (VP) y el genotipo (Illumina® BovineSNP50 BeadChip) de 3305 sementales Holstein con prueba de progenie en Estados Unidos, que fueron cedidos por el Laboratorio de Genómica functional bovina y los programas de mejora animal del USDA-ARS Beltsville Agricultural Research Center (Beltsville, MD). Tras la edición de los genotipos y la imputación de los genotipos faltantes, se usó un total de 32,611 SNPs en los análisis. Se seleccionaron los sementales transmisores de baja o alta VP, correspondientes a los machos por debajo del percentil α y por encima del percentil (1- α), respectivamente, para sus PTA de VP, simulando estudios caso-control. El grupo transmisor de alta VP es susceptible de poseer genes favorables para los caracteres de producción, morfológicos y resistencia a enfermedades.

La detección de SNPs se realizó usando machine learning, en concreto bagging con criterios de teoría de la información.

Bagging

El método bagging (de las siglas en inglés bootstrap aggregating) fue descrito inicialmente por Breiman (1996). Supongamos $\Psi = (\mathbf{c}, \mathbf{X})$ es el grupo de machos pertenecientes a las clases (**c**) de alta y baja VP con sus respectivos genotipos (**X**). Se realizó un muestreo aleatorio con reposición (bootstrapping) para cada una de las clases por separado, de manera que cada $(c_i, \mathbf{x}_i)$ puede aparecer más de una vez o ninguna en cada subset $\Psi^{(B)}$. Se realizaron 150 réplicas de bootstrapping utilizando los percentiles α=(0.10, 0.15, 0.20) en iguales proporciones, y se calculó la ganancia de información (IG) de cada SNP en cada $\Psi^{(B)}$, de tal

forma que se obtuvo $\phi_B = (IG_{b1}, IG_{b2}, ..., IG_{bp})$, B=1-150, donde IG_{bi} es la IG del SNP *i* en la muestra de bootstrapping *b*.
Detalles sobre el cálculo de IG pueden encontrarse en Long et al. (2007).
La IG para cada SNP procedente del bagging se calculó como la media de las 150 réplicas de bootstrapping.

Eliminación de redundancia

La redundancia entre pares de SNPs se estimó a través de la información mutua, que mide la cantidad de información que comparten dos variables aleatorias (Cover y Thomas, 1991). La información mutua entre dos SNPs, $MI(S_i, S_j)$, se define como:

$$MI(S_i; S_j) = \sum_{s_i, s_j} P_{S_i S_j}(s_i, s_j) \log \frac{P_{S_i S_j}(s_i, s_j)}{P_{S_i}(s_i) P_{S_j}(s_j)},$$

donde $P_{S_i}(s_i)$ y $P_{S_j}(s_j)$ son las distribuciones marginales y $P_{S_i S_j}(s_i, s_j)$ es la distribución de probabilidad conjunta. A menor información mutual, mayor independencia entre las variables (SNPs). Se consideró que dos SNPs son redundantes si su MI > 0.37 (percentil 95).

Se seleccionaron los 2000 SNPs con mayor IG. A partir de estos, se seleccionó otro set de 2000 SNPs, procediendo de la siguiente forma: si entre los 2000 SNPs existía un par redundante, se eliminó el SNP con menor IG, sustituyéndolo por el siguiente SNP de la lista completa con mayor IG. Se procedió iterativamente hasta que se eliminaron todos los SNPs redundantes en los 2000 seleccionados.

RESULTADOS Y DISCUSIÓN

La figura 1 muestra la MI de pares de SNPs en el mismo cromosoma, en función de su distancia en pares de bases. Como era de esperar, la mayor redundancia se observa entre SNPs más próximos, aunque como se observa en la figura, proximidad no necesariamente conlleva redundancia, probablemente debido a diferentes frecuencias alélicas o mayor desequilibrio de ligamiento. La tabla 1 muestra el número de SNPs por cromosoma entre los 2000 con mayor IG, antes y después de aplicar la eliminación de redundancia. Se observa una reducción importante del número de SNPs en aquellos cromosomas más informativos, puesto que contienen SNPs informativos pero redundantes entre si. Esta eliminación de redundancia provoca la entrada de nuevos SNPs en cromosomas menos informativos, pero con SNPs que confieren una IG suficiente como para ser seleccionados. Las diferencias son más notables cuando se reduce el número de SNPs seleccionados. La figura 2 muestra el ejemplo de dos importantes cromosomas (BTA2 y BTA14) en los que previamente se han detectado genes relacionados con la vida productiva (Grisart et al., 2004; Schnabel et al., 2005). El 21 % de los SNP seleccionados después de eliminar redundancia fue diferente a los 2000 SNPs iniciales.
La pre-selección de SNPs aporta ventajas en los estudios de asociación y la selección genómica, puesto que permite reducir la dimensionalidad del problema y detectar aquellas regiones de interés en la expresión del carácter. Además permite el diseño de SNPs de baja densidad, que permiten genotipar individuos a un menor coste y con una buena capacidad para predecir observaciones futuras (Long et al.; 2007; González-Recio et al., 2008; Weigel et al., 2009). Los resultados obtenidos en este trabajo animan a un estudio más profundo sobre el comportamiento y habilidad predictiva de estos métodos en estudios de asociación con genoma completo y valoraciones genómicas.

REFERENCIAS BIBLIOGRÁFICAS

●Breiman, L, 1996. Machine Learning 24: 123–140. ●Gianola, D et al. 2006. Genetics 173: 1761-1776.●Gianola D. et al. 2006. Genetics 173:1761-1776. ●Gonzalez-Recio O. et al. 2008.

Genetics 178: 2305-2313. ● Grisart B. et al. 2004. Proc. Natl. Acad. Sci. U.S.A.,101:2398-403. ● Long, N. et al. 2007. J. Anim. Breed. Genet., 124 (6): 377-389. ●Meuwissen, T. H. E. et al. 2001. Genetics 157: 1819-1829. ● Schnabel et al. 2005. Animal Genetics, doi:10.1111/j.1365-2052.2005.01337.x ● Cover, T.M. & J.A. Thomas. 1991. Elements of information theory. John Wiley and sons, New York. ● Weigel K. A. et al. J. Dairy Sci. (submitted).

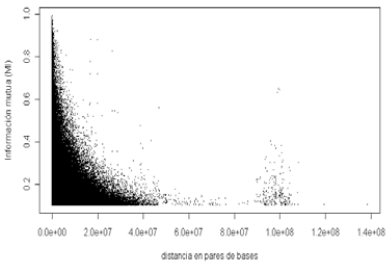

***Figura 1**. Información mutua entre pares de SNPs en función de la distancia en pares de bases*

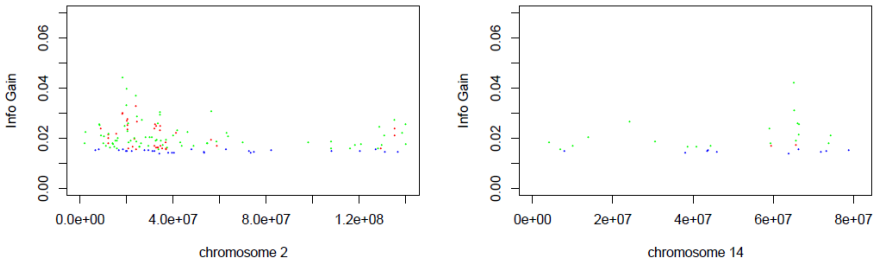

***Figura 2.** SNPs seleccionados antes y después de considerar redundancia (verde), SNPs redundantes eliminados (rojo) y SNPs seleccionados después de eliminar redundancia (azul) para los cromosomas bovinos 2 y 14.*

REDUCTION OF COLINEARITY IN HIGH-DENSITY SNPs GENOTYPES, AND USE OF BAGGING TO INCREASE ACCURACY OF PRE-SELECTION OF SNPs IN GENOMIC SELECTION

ABSTRACT. Machine learning provides tools to deal with crude, noisy, inconsistent and redundancy genomic data from high throughput assays for dense genotyping. This work provides preliminary results from two novel methods to reduce redundancy and increase accuracy in genome-wide association studies and genomic selection. Mutual information theory detects redundant SNPs from high-density SNP assay. The bootstrap aggregating method (bagging) increases accuracy of feature selection. The combination of these to methods might create a proper scenario for a higher predictive ability using low dense SNP assays, and should be investigated further in the future.
***Keywords:** theory information, bagging, snp, genomic selection.*