PREDICCIÓN GENÓMICA EN PRESENCIA DE ESTRATIFICACIÓN DE POBLACIONES

González-Recio, O.[1*], Forni, S. [2], Gianola, D. [3,4], Rosa, G.J.M. [4] y Weigel, K.A.[3]
[1]Departamento de Mejora Genética Animal. Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria. Ctra. La Coruña km 7,5. 28040 Madrid.
[2]PIC/GENUS Plc. 100 Bluegrass Commons Blvd ste 2200, Hendersonville, TN 37075 (USA).
[3]Dairy Sci. Dpt. University of Wisconsin-Madison. 1675 Observatory Dr., WI 53706 (USA)
[4]Animal Sci. Dpt. University of Wisconsin-Madison. 1675 Observatory Dr., WI 53706 (USA)
[*]Correo electrónico: gonzalez.oscar@inia.es

INTRODUCCIÓN

La información genómica ha permitido incrementar la precisión de las evaluaciones genéticas en poblaciones domésticas, tanto vegetales como animales, así como disminuir el intervalo generacional. Esta precisión aumenta cuanto mayor es el tamaño de la muestra de referencia sobre la que se entrenan los modelos estadísticos usados para predecir el mérito genético de los individuos de la población (Consortium, WTCC. 2007; Van Raden et al., 2009). Esto ha llevado a que, en ocasiones, se unan muestras de poblaciones heterogéneas de menores tamaños para aumentar el tamaño de la muestra de entrenamiento y así aumentar la precisión de las predicciones. Hasta la fecha los resultados no han sido del todo convincentes, sobre todo cuanto más alejadas genéticamente son las poblaciones que componen la muestra de referencia (Goddard y Hayes, 2009; Hayes et al., 2009; Ibañez-Escriche et al., 2009; Toosi et al., 2010).

Es posible que esto se deba a la interacción genotipo ambiente o a la diferente base genética de las poblaciones. Por ejemplo, dos poblaciones con diferente nivel productivo seguramente diferirán en las frecuencias alélicas de algunos marcadores genotipados en los chips actuales. Estos marcadores con diferente frecuencia alélica pueden estar o no en desequilibrio de ligamiento con QTL asociados al carácter de interés. Las técnicas estadísticas tradicionales pueden conducir a asociaciones espurias en el caso de que estos marcadores no estén asociados al QTL, lo que afectará en mayor o menor medida en función de la cantidad de marcadores que se encuentren en estas circunstancias, y de la magnitud de las diferencias de sus frecuencias alélicas. También, es factible que la frecuencia de algunas combinaciones genotípicas sea marcadamente diferente en las muestras de entrenamiento y de predicción; en el lenguaje de la regresión, esto implica que habrían partes de la "región experimental" que no son cubiertas por el modelo de entrenamiento, y esto puede ser crucial cuando hay variabilidad de tipo no-aditiva. En consecuencia, es esperable que la estratificación genética de las poblaciones tenga un importante efecto sobre la precisión de las predicciones en poblaciones heterogéneas. Es importante detectar dicha estratificación de poblaciones y desarrollar estrategias para corregir, en la medida de lo posible, la estratificación de poblaciones en multipoblaciones.

Los métodos no paramétricos han mostrado ciertas ventajas en relación a los métodos tradicionales a la hora de predecir el valor genético de animales y plantas usando información del genoma completo (de los Campos et al., 2010). En este trabajo se propone la utilización del algoritmo de bosques aleatorios (Random Forest, **RF**; Breiman, 2001) para detectar la estratificación de poblaciones, y clasificar individuos dentro de multipoblaciones, construyendo eventualmente una matriz de similitudes genéticas que pueda ser usada en como matriz de kernels en regresiones no paramétricas.

MATERIAL Y MÉTODOS

Datos

Se utilizaron datos de tres líneas (poblaciones) porcinas cedidos por la empresa PIC North America, Genus Plc. Dos de estas líneas (Línea A y Línea B) representan poblaciones de animales puros del núcleo de selección (923 y 919 individuos respectivamente), mientras que la tercera línea (C), formada por 700 individuos, fue una población cruce de la línea A con otras líneas no especificadas (sin contener Línea B). Todos los individuos fueron genotipados por 6742 SNP identificados previamente en regiones candidatas asociadas a hernia escrotal. Tras

una edición de genotipos se utilizaron en los análisis finales 5302 SNPs codificados como 0, 1 o 2, dependiendo del numero de un determinado alelo que presentaban para cada SNP.
Para poder evaluar la capacidad predictiva de los métodos, los análisis se realizaron bajo un escenario de validación cruzada dejando el 15% de los individuos más jóvenes como muestra de predicción, que no fue usada en absoluto para entrenar los algoritmos.

Detección de estratificación de poblaciones
Se usó el algoritmo de bosques aleatorios para clasificar los animales en los diferentes estratos de población, y se evaluó la capacidad predictiva de clasificación para nuevos animales de los cuales no se conocería a priori a que población pertenecen. Se construyeron 2000 árboles utilizando la multi-población compuesta por las 3 líneas. Cada árbol $h_t(\mathbf{X})$ se construye utilizando un subconjunto de la muestra de aprendizaje muestreado al azar con repetición (bootstrapping) y mediante sencillas reglas basadas en diferentes combinaciones de genotipos asigna cada individuo a un estrato. Cada individuo es asignado a un nodo terminal dentro de cada árbol en el que coincide con otros individuos. Estos nodo terminales predicen la población a la que pertenecen sus individuos en función de la población mayoritaria de individuos en dicho nodo. Se promediaron los 2000 árboles para obtener una estima final de cada individuo. Los individuos del conjunto de datos de predicción son pasados a posteriori por cada uno de los árboles, y son asignados a la población mayoritaria del nodo terminal al que llegan, y promediados a lo largo del RF.

Matriz de similitudes genéticas
A partir del resultado del RF se construyó una matriz (**K**) de similitudes genéticas de dimensiones 2542x2542, el numero total de animales. Cada elemento de la matriz ($k_{i,j}$) refleja el parecido genómico ($\mathbf{x}_i - \mathbf{x}_j$) entre pares de individuos, y se calcula como el porcentaje de árboles en el que los individuos *i* y *j* coinciden en el mismo nodo terminal. Los elementos $k_{i,j}$ se expresan en un rango entre 0 y 1, siendo 1 si los individuos *i* y *j* coinciden en el mismo nodo terminal en todos los árboles, y 0 en el caso de que no coincidan en ningún árbol. Tanto los individuos de la muestra de aprendizaje como los de la muestra de validación están representados en esta matriz, ya que como se ha comentado anteriormente, los individuos de la muestra de validación son pasados a través del árbol a posteriori, y terminan en un nodo, en el cual coinciden con otros individuos del conjunto total de datos.
Se realizó un análisis de componentes principales, calculando los autovectores y autovalores sobre está matriz para identificar los posibles estratos de poblaciones.

RESULTADOS Y DISCUSIÓN

La Figura 1 muestra un gráfico de los tres primeros autovectores de la matriz **K**. Se muestran los individuos de las líneas A, B y C en color negro, rojo y verde, respectivamente. Los individuos de la muestra de aprendizaje aparecen representados con puntos, mientras que los individuos en la muestra de validación se representan como triángulos. En esta figura se puede ver como el algoritmo RF separa los individuos en la muestra de aprendizaje (puntos) de cada una de las poblaciones, y además es capaz de predecir y asignar correctamente la población a la que pertenecen los individuos del la muestra de validación (triángulos). También es posible ver como los individuos de las poblaciones A y C, están más próximos entre si, que con respecto a los individuos de la línea B, lo que es esperable ya que la línea C es un producto del cruce de individuos de la línea A. La Tabla 1 muestra el valor medio de los elementos $k_{i,j}$ de la matriz de similitudes para cada una de las líneas. Se observa que la población B es más homogénea, mientras que la población C es la más heterogénea (menores valores de similitud entre individuos), lo cual es esperable debido a que la población de la línea C proviene de un cruce de diferentes líneas.

Este trabajo muestra algunos resultados preliminares para el tratamiento de la estratificación de poblaciones en la predicción de mérito genético usando información genómica..

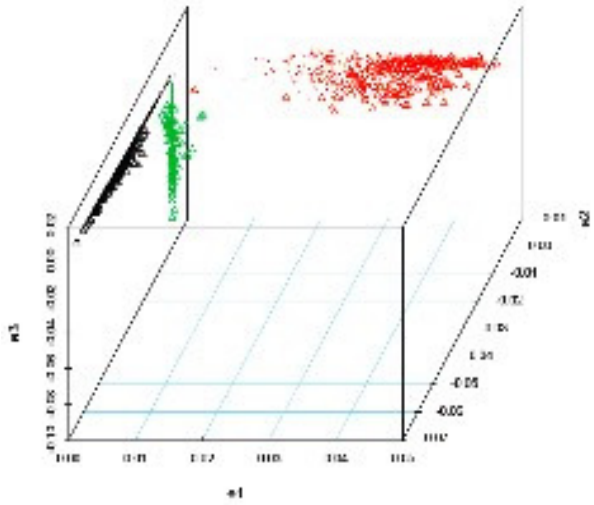

Figura 1. *Representa los 3 primeros autovectores de la matriz de similitud resultante del algoritmo RF. Cada color representa una población. Los individuos del set de aprendizaje y del set de validación se representan como puntos y triángulos respectivamente.*

El algoritmo RF permite detectar estratificación de poblaciones en muestras de poblaciones heterogéneas, y a su vez produce una matriz de similitudes genómicas que puede ser usada como matriz de kernels en regresiones no paramétricas del tipo RKHS (Gianola et al., 2006). Además, el algoritmo RF permite calcular la importancia de cada SNP en la estratificación de poblaciones, pudiendo así detectar aquellos SNPs que pueden provocar asociaciones espurias o magnitudes sesgadas de las estimas de su efecto sobre el carácter. Es necesario que trabajos futuros se centren en el desarrollo de estos métodos, para poder paliar el efecto de la estratificación de poblaciones en las predicciones genómicas usando poblaciones heterogéneas, y comparar su comportamiento con métodos usados actualmente.

REFERENCIAS BIBLIOGRÁFICAS

• Breiman L. Machine Learning 45(1):5-32, 2001. • Consortium, WTCC. 2007. Nature Rev. Genet. 447:661-678. • De los Campos, G., Gianola, D. & Allison, D.B. 2010. Nature Rev. Genet. 11: 880-886. • Gianola, D., Fernando, R.L. & Stella, A. 2006 Genetics 173: 1761-1776. • Goddard, M.E. & Hayes B.J. 2009. Nature Rev. Genet. 10:381-391. • Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K. & Goddard, M.E. 2009. Genet. Sel. Evol. 41:51. • Ibañez-Escriche, N., Fernando, R.L., Toosi, A. & Dekkers, J.C.M. 2009. Genet. Sel. Evol. 41:12-22. • Toosi, A., Fernando, R.L. & Dekkers, J.C.M. 2010. J. Anim. Sci. 88(1): 32-46. • VanRaden, P.M., Van Tasell, C.P., Wiggands, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F. 2009. J. Dairy Sci. 92:16-24.

GENOME WIDE PREDICTION OF GENETIC MERIT IN THE PRESENCE OF POPULATION ESTRATIFICATION

ABSTRACT: This study aims to account for population stratification when several populations are used in a genome-wide prediction context. The Random Forest algorithm is used to create a similarity matrix to cluster individuals according to sub-populations. At the same time, such matrix can be used as a kernel for future non-parametric genome-wide prediction methods.

Keywords: population stratification, non-parametric methods, genomic selection.