PREDICCIÓN GENÓMICA MEDIANTE UNA APROXIMACIÓN SEMI-PARAMÉTRICA EN VACUNO LECHERO

S. T. Rodríguez-Ramilo[1,3], J. A. Jiménez-Montero[2] y Ó. González-Recio[3]
[1]CONAFE. Ctra. de Andalucía, Km. 23,6. 28340, Madrid
[2]Dpto. de Producción Animal. ETSI Agrónomos. UPM. 28040, Madrid
[3]Dpto. de Mejora Genética Animal. INIA. Ctra. La Coruña Km. 7,5. 28040, Madrid
E-mail: rodriguez.silvia@inia.es

INTRODUCCIÓN

La selección simultánea en función de miles de SNPs (Single Nucleotide Polymorphisms) está siendo una alternativa para incrementar la ganancia genética en vacuno lechero. Los valores genómicos directos pueden ser estimados reemplazando la matriz tradicional de relaciones de pedigrí por la matriz de relaciones genómicas (VanRaden, 2008; Yang et al., 2010). Generalmente, se emplea una población de referencia para predecir los valores mejorantes en los nuevos candidatos a la selección (Hayes et al., 2009).

Existen distintas metodologías para llevar a cabo predicciones genómicas (e. g. Meuwissen et al., 2001; de los Campos et al., 2009). Dentro de los algoritmos de aprendizaje automático se engloban los métodos no paramétricos o semi-paraméticos, los cuales se pueden implementar mediante regresiones sobre los marcadores (González-Recio et al., 2010) o mediante la construcción de estructuras de (co)varianza apropiadas. Un ejemplo de estas últimas metodologías es la Regresión de espacios de Hilbert con núcleo reproductor (RKHS de sus siglas en inglés), la cual se ha sugerido como una alternativa para la predicción de valores genómicos, ya que requiere menos asunciones a la hora de modelar caracteres cuantitativos complejos (Gianola y van Kaam, 2008). En este trabajo se propone la utilización de la metodología RKHS empleando como estructura de kernels la matriz de relaciones genómicas para realizar predicciones genómicas en siete caracteres evaluados en vacuno lechero.

MATERIAL Y MÉTODOS

Datos

Los 18446 toros de la población de EuroGenomics fueron genotipados con el Bovine 50K chip (Illumina inc., San Diego), aunque algunos machos de la población holandesa fueron genotipados con el CRV 60K chip y posteriormente imputados. Se descartaron los SNPs con un porcentaje de datos faltantes por encima del 5% y con una frecuencia alélica menor de 5%. Por tanto, el análisis se realizó con 36971 SNPs.

Los machos nacidos antes del 2005 (N = 14487) fueron empleados como población de referencia (R) y los nacidos con posterioridad al año 2005 (N = 3959) se usaron como población de validación (V) para evaluar la habilidad predictiva del algoritmo. Como datos fenotípicos se emplearon las pruebas deregresadas de tres caracteres de producción: kilos de leche (KL), kilos de grasa (KG), y kilos de proteína (KP); dos caracteres de tipo: estatura (EST) y anchura de pecho (ANPE); recuento de células somáticas (RCS) y días abiertos (DA). Las pruebas deregresadas de la población de referencia y de la población de validación fueron las de enero del 2009 y diciembre del 2011, respectivamente.

Modelo

La aproximación semi-paramética RKHS permite inferir valores genómicos mejorantes sin realizar estrictas asunciones a priori. En el contexto de la selección genómica (Gianola y van Kaam, 2008; González-Recio et al., 2008) el modelo se puede formular como sigue:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{K}\boldsymbol{\alpha} + \mathbf{e}$$

donde $\mathbf{y}$ (14487×1) es la prueba deregresada de los machos en la población de referencia. El primer término ($\mathbf{X\beta}$) es el término paramétrico con $\boldsymbol{\beta}$ como vector de efectos sistemáticos (solamente μ se ajustó en este caso porque los datos fueron precorregidos), y $\mathbf{X}$ es la matriz de incidencias. El término no paramétrico viene dado por $\mathbf{K\alpha}$, donde $\mathbf{K}$ es una matriz definida positiva de kernels, y $\boldsymbol{\alpha}$ es un vector de coeficientes no paramétricos que se asume que siguen una distribución $\boldsymbol{\alpha} \sim N(0, \mathbf{K}^{-1}\sigma_\alpha^2)$, donde σ_α^2 representa la inversa del parámetro de sintonización $(\sigma_\alpha^2 = \lambda^{-1})$. Los residuos $\mathbf{e}$ se distribuyen como $N(0, \mathbf{R} = \mathbf{N}^{-1}\sigma_e^2)$ donde $\mathbf{N} = \{n_i\}$ es la diagonal de la matriz donde n_i es el número de descendientes del macho i y σ_e^2 es la varianza residual. Se realizó un análisis Bayesiano del modelo implementado a través del algoritmo de muestreo de Gibbs.

Matriz de kernels

La idea que subyace en cualquier matriz de kernels es medir la distancia que existe entre los genotipos de los distintos individuos (González-Recio et al., 2009). En este estudio se implementa como distancia la matriz de relaciones genómicas (VanRaden, 2008; Yang et al., 2010). La relación genómica (G_{ij}) entre los individuos *i* y *j* se calcula como:

$$G_{ij} = \frac{1}{L}\sum_{k=1}^{L}\frac{(g_{ik} - \hat{p}_k)(g_{jk} - \hat{p}_k)}{\hat{p}_k(1-\hat{p}_k)}$$

donde g_{ik} se refiere a las frecuencias génicas del valor de los genotipos *AA*, *Aa* y *aa*, codificado como 1, 1/2 y 0, respectivamente, para el individuo *i* en el locus *k*, donde *i* = 1, *N* y *k* = 1, *L* (número de loci). La frecuencia génica es la mitad del número de copias del alelo de referencia *A*. La estima de la frecuencia alélica para el locus *k* en la población actual se designa como $\hat{p}_k$.

Estimación de ratios de varianza y valores genómicos directos (VGD)

A partir de distintos tamaños muestrales de la población de referencia (*N* = 4992, 6981, 10957 y 14487) se obtuvieron los ratios de la varianza genómica y residual. Los valores genómicos directos de cada carácter se calcularon para cada individuo de la población de validación de la siguiente forma:

$$\mathbf{VGD} = \mu + \mathbf{G}_{\mathbf{V,R}} \times \boldsymbol{\alpha}$$

Criterios de comparación

La precisión de las predicciones genómicas se estimó mediante el coeficiente de correlación de Pearson (*r*), pendiente de regresión (*b*) y el error cuadrático medio (*ECM*) entre los VGD predichos en los toros de la población de validación y las pruebas deregresadas de diciembre del 2011. También se evaluó el coeficiente de correlación de Pearson entre el índice de pedigrí (IP) y las pruebas deregresadas de diciembre del 2011.

RESULTADOS Y DISCUSIÓN

La Figura 1 muestra los ratios de la varianza genómica y residual para los siete caracteres evaluados en distintos tamaños de muestra de la población de referencia. Las estimas son más elevadas en los caracteres de tipo, seguidos estos de los caracteres de producción y finalmente RCS y DA. En general, a medida que se incrementa el tamaño de muestra el

ratio se reduce, llegando a ser inferior a la heredabilidad real. Este descenso es más acusado en los caracteres de tipo y producción que en RCS y DA. Consecuentemente, las predicciones genómicas se llevaron a cabo empleando las varianzas obtenidas a partir de *N* = 4992.

En la Tabla 1 se muestra el coeficiente de correlación de Pearson, la pendiente de la recta de regresión y el error cuadrático medio entre los VGD y las pruebas deregresadas. También se muestra la correlación entre el IP y las pruebas deregresadas. En todos los caracteres evaluados (excepto en DA) la correlación con IP fue menor que la correlación con los VGD. Los caracteres EST y DA mostraron la mayor (0,78) y menor (0,49) correlación con los VGD, respectivamente. Por el contrario, la mayor pendiente de la recta de regresión se detectó en DA (1,02) y la menor en EST (0,81). El error cuadrático medio varió entre 176002,20 (KL) y 0,51 (EST).

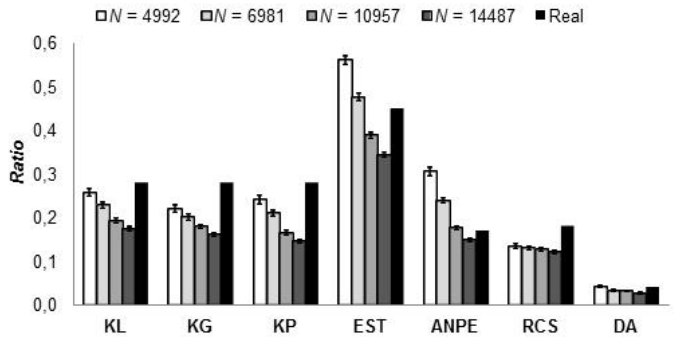

Figura 1. *Estima de los ratios entre varianza genómica y residual en los caracteres evaluados en distintos tamaños de muestra (N) de la población de referencia. Real: heredabilidad real.*

Tabla 1. *Coeficiente de correlación de Pearson, pendiente de la recta de regresión, error cuadrático medio y correlación con el índice de pedigrí en los caracteres evaluados.*

Carácter	*r*	*b*	*ECM*	*IP*
KL	0,74	0,94	176002,20	0,54
KG	0,71	0,90	299,55	0,51
KP	0,69	0,91	205,87	0,54
EST	0,78	0,81	0,51	0,61
ANPE	0,64	0,87	1,36	0,58
RCS	0,70	0,94	93,74	0,55
DA	0,49	1,02	382,92	0,53

En resumen, los resultados indican que la aproximación semi-paramética RKHS permite inferir valores genómicos mejorantes en distintos caracteres en vacuno lechero de forma adecuada cuando se emplea la matriz de relaciones genómicas.

REFERENCIAS BIBLIOGRÁFICAS

• de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K.A. & Cotes, J.M. 2009. Genetics 182:375-385.• Gianola, D. & van Kaam, J.B.C.H.M. 2008. Genetics 178:2289-2303.• González-Recio, O., Gianola, D., Long, N., Weigel, K.A., Rosa, G.J.M. & Avendaño, S. 2008. Genetics 178:2305-2313.• González-Recio, O., Gianola, D., Rosa, G.J.M., Weigel, K.A. & Kranis, A. 2009. Genet. Sel. Evol. 41:3.• González-Recio, O., Weigel, K.A., Gianola, D., Naya, H. & Rosa, G.J.M. 2010. Genet. Res. (Camb.) 92:227-237.• Hayes, B.J., Bowman, P.J., Chamberlain, A.C., Verbyla, K. & Goddard, M.E. 2009. Genet. Sel. Evol. 41:51.• Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Genetics 157:1819-1829.• VanRaden, P.M. 2008. J. Dairy Sci. 91:4414-4423.• Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., Visscher, P.M. 2010. Nat. Genet. 42:565-569.

SEMI-PARAMETRIC APPROACH FOR GENOME WIDE PREDICTION IN DAIRY CATTLE

ABSTRACT: This study aims to implement a semi-parametric approach in dairy cattle within the context of genome-wide prediction. The Reproducing Kernel Hilbert Spaces Regression model with the genomic matrix as kernel seems to perform properly when inferring direct genomic values in seven phenotypic traits.

Keywords: Reproducing Kernel Hilbert Spaces Regression, non-parametric methods, genome-assisted prediction.