PONDERANDO PANELES DE SNP PARA PREDICCIÓN GENÓMICA POR *SIMULATED ANNEALING*

Martín de Hijas-Villalba[1], M., Varona[2], L., Noguera[3], J.L., Ibáñez-Escriche[4], N., Rosas[5], J.P., y Casellas[1], J.
[1]Departament de Ciència Animal i dels Aliments, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona). [2]Departamento de Anatomía, Embriología y Genética Animal, Universidad de Zaragoza, 50013 Zaragoza. [3]Genètica i Millora Animal, Institut de Recerca i Tecnologia Agroalimentàries, 25198 Lleida. [4]Departament de Ciència Animal, Universitat Politècnica de València, 46071 València. [5]Programa de Mejora Genética "Castúa", INGA FOOD SA, 06200 Almendralejo, España.; Melani.MartinDeHijas@uab.cat

INTRODUCCIÓN

Los programas de selección son una herramienta esencial para el incremento de la producción de alimentos y habitualmente utilizan toda la información disponible (ej., fenotipos, genealogía e incluso datos genómicos) para calcular los valores mejorantes de los candidatos a la reproducción. En este contexto, varios estudios han demostrado que los programas de selección basados en datos genómicos son más precisos al predecir dichos valores mejorantes (**GEBV**) que los métodos convencionales basados únicamente en el fenotipo y la genealogía (Schaetter, 2006; Legarra et al., 2008). Por esta razón, los modelos de predicción genómica han sido implementados en programas de selección de animales y plantas en todo el mundo (Meuwissen et al., 2001; Gianola, 2013).

Actualmente, el método para evaluación genética se basa principalmente en gBLUP (*genomic Best Linear Unbiased Prediction*, Henderson, 1973; Gianola et al., 2003). El gBLUP utiliza una matriz de relaciones genómicas (**G**), que se calcula a partir de la información de miles de SNPs (*Single Nucleotide Polymorphism*), cada uno de ellos con el mismo nivel de importancia. Sin embargo, estudios previos han obtenido mejoras en predicción genómica cuando no todos los SNPs tienen el mismo peso en la construcción de **G**. En modelos Bayesianos, el modelo BayesRC de MacLeod et al. (2016) que incorpora conocimiento biológico previo al modelo sería un ejemplo. En él, las diferentes regiones de DNA se distribuyen en diferentes clases dependiendo de sus probabilidades de contener genes causales del carácter bajo estudio. En cuanto a modelos basados en gBLUP, el modelo iterativo WssGBLUP de Wang et al. (2012) propone ponderar los SNPs a partir de los GEBV obtenidos tras un ssGBLUP previo tradicional. Pero el presente estudio está basado en un modelo previo de Casellas (2012) que utilizó *simulated annealing* (SA) para decidir qué SNPs debían contribuir en la creación de **G** (ponderación 1) y cuáles no (ponderación 0). En él, se obtuvieron resultados positivos, pero la eliminación de SNPs puede conllevar la pérdida de información, y por ello, en este estudio se propone un nuevo mecanismo para la construcción de **G**, donde todos los SNPs pueden participar en ella, pero con diferente ponderación. En el nuevo modelo (**modelo W**), los SNPs toman valores entre 0 y 1 y su rendimiento fue comparado con el modelo estándar gBLUP (**modelo G**) en función a su precisión al calcular los valores genéticos predichos en poblaciones simuladas.

MATERIAL Y MÉTODOS

Los datos simulados utilizados pertenecieron a dos tipos de poblaciones. Ambos tipos consistían en poblaciones de individuos diploides con cromosomas autosómicos de 100 cM con 6.000 SNPs y 250 QTL (Quantitative Trait Loci) cada uno, que evolucionaron durante 1.000 generaciones (Ne=100) más 5 generaciones finales (Ne=200) no superpuestas y con apareamiento aleatorio. La diferencia entre las poblaciones consistió básicamente en el número de cromosomas. Mientras las poblaciones P1 contenían un único cromosoma autosómico, las poblaciones P2 tenían 30. Los genotipos de las generaciones descendientes fueron simulados teniendo en cuenta los fenómenos de recombinación, mutación y desequilibrio genético. Una vez creados los valores genéticos, los fenotipos se obtuvieron para cada individuo añadiendo un residuo aleatorio (Lillehammer et al., 2013) obtenido de una distribución normal y estandarizado para cada heredabilidad.

Los datos simulados fueron analizados en el contexto de la evaluación genética implementando un gBLUP bajo el modelo jerárquico:

$$\mathbf{y=\mu+Za+e}$$

donde **y** era el vector de observaciones (fenotipos), **μ** la media poblacional, **a** los efectos genéticos aditivos del animal, **e** el vector de efectos residuales y **Z** la matriz de incidencia que relacionaba los fenotipos con los efectos genéticos del animal (Mrode y Thompson, 2005). Se construyeron las ecuaciones de modelo mixto (MME, Henderson, 1950).

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\alpha \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{a} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

donde **A** era la matriz de parentescos (Wright, 1922) y **X** un vector columna de unos, capturando la media poblacional. Como los datos simulados produjeron no solo datos fenotípicos sino también genómicos, la matriz **A** fue generalizada a una matriz de relaciones genómicas **G** (Legarra et al., 2009) calculada siguiendo el primer método de VanRaden (2008). Finalmente, para evitar singularidades durante la inversión de la matriz **G**, se asumió una reparametrización estándar del modelo (Henderson, 1984), obteniendo:

$$\mathbf{y}=\boldsymbol{\mu}+\mathbf{Z}(\mathbf{G}\mathbf{i})+\mathbf{e}$$

donde $\mathbf{i}=\mathbf{G}^{-1}\mathbf{a}$ seguía una distribución normal $N(\mathbf{0},\mathbf{G}^{-1}\mathbf{G}\mathbf{G}^{-1}\sigma_g^2)$.
El modelo W fue resuelto por el método iterativo de Gauss-Seidel (Mrode & Thompson, 2005) bajo *simulated annealing* basado en el error cuadrático medio (MSE) de los fenotipos como parámetro estadístico de comparación. La correlación entre valores mejorantes simulados (TBV) y predichos (GEBV) fue también calculada y los resultados de ambos estadísticos se compararon con el modelo G.
En total se evaluaron 100 poblaciones P1 diferentes, con heredabilidades entre 0,1 y 0,5, y tres poblaciones P2, con heredabilidades 0,1, 0,25 y 0,4.

RESULTADOS Y DISCUSIÓN

Para evaluar el comportamiento del modelo W, los resultados fueron comparados con los valores obtenidos por el modelo tradicional G (donde todos los SNP tienen la misma ponderación). Para cada población, se obtuvo el porcentaje de mejora de correlación y MSE. Los resultados para poblaciones P1 (Figura 1), revelan una reducción del MSE que oscila entre un 7,7 y un 26,7%. Estos resultados sugieren además una correlación positiva con la heredabilidad, (**r** = 0,827). Cuando la heredabilidad del carácter es superior, el modelo consigue un mejor ajuste, ya que detecta mejor el efecto genético. En cuanto a la correlación entre TBV y GEBV, el modelo W consigue aumentarla entre un 4 y un 11% respecto al modelo tradicional, y aunque su relación con la heredabilidad del carácter en estudio no es tan clara, hay una tendencia a aumentar la mejora de correlación en bajas heredabilidades (**r** = -0,625). Esto puede estar relacionado con la contribución del componente genético al determinar el fenotipo. Con heredabilidades más bajas, el incremento de información que se produce al disponer de una matriz de relaciones más ajustada a la covarianza real entre los individuos emparentados proporciona un mayor incremento relativo de precisión medida como correlación ente TBV y GEBV. Hay más probabilidades de que un efecto genético pequeño (h^2=0,1) no sea percibido por el modelo tradicional G, y por lo tanto sea más fácil de incrementar con un mejor ajuste de los GEBV.
Las poblaciones P2 con 30 cromosomas muestran resultados similares. Tras tan solo 6.000 iteraciones, las 3 poblaciones muestran una reducción de MSE entre un 10% y un 15%; y un aumento en la correlación de TBV y GEBV entre un 1,5 y un 3%.
Comparando con los estudios previos, según Lourenco et al. (2017), el modelo WssGBLUP sólo necesita dos iteraciones para maximizar la precisión genómica, pero tras estas iteraciones su precisión decrece en genes poligénicos. Además, se llegó a la conclusión de que el modelo WssGBLUP no presenta ninguna mejora respecto al modelo gBLUP cuando el número de QTL es superior a 500, mientras que el modelo propuesto en el presente estudio ha obtenido mejoras en precisión con números de QTL superiores a 1700. En cuando al modelo BayesRC, cuando los QTL están distribuidos de manera aleatoria por todo el genoma, el aumento en precisión respecto al modelo gBLUP tradicional no es muy elevado. Por lo tanto, sería interesante comparar el aumento en precisión de ese modelo con el modelo W propuesto bajo esas condiciones, ya que, el hecho de disponer de suficiente información

biológica sobre posición de genes y lugares propensos a afectar a un carácter en particular puede suponer una gran limitación para utilizar el modelo BayesRC.
Podemos concluir pues, que la implementación de pesos en los valores de los SNPs en el modelo W, para la construcción de la matriz **G**, supone una mejora en cuanto a ajuste de los valores mejorantes predichos de manera sencilla y sin causar pérdida de información.

REFERENCIAS BIBLIOGRÁFICAS

• Casellas, J. 2012. 4th International Conference on Quantitative Genetics. Edimburgh, Scotland (poster). • Gianola, D. 2013. Genetics 194: 573–596. • Gianola, D. et al. 2003. Genetics 163: 347-365. • Henderson, C. R. 1950. Annals of Mathematical Statistics 21, 309. • Henderson, C. R. 1973. Proceedings of the Animal Breeding and Genetics Symposium in honor of Dr. J. L. Lush. ASAS-ADSA, Champaign, Illinois, pp 10-41. • Legarra, A. et al. 2008. Journal of Dairy Science 91: 360-366. • Legarra, A. et al, 2009. Journal of Dairy Science 92: 4656–4663. • Lourenco, D. A. L. et al. 2017. Journal of Animal Breedding and Genetics 134: 463–471. • MacLeod, I. M. et al. 2016. BMC Genomics 17: 144 • Meuwissen, T. H. E. et al. 2001. Genetics 157: 1819-1829. • Mrode, R. A. et al. 2005. Wallingford, Oxfordshire, UK. CABI. • Schaeffer, L. R. 2006. Journal of Animal Breedding and Genetics 123: 218-223. • Van Raden, P. M. 2008. Journal of Dairy Science 91: 4414–4423. • Wang, H. et al. 2012. Genetics Research 94: 73–83. • Wright, S. 1922. American Naturalist 56: 330–338.

Agradecimientos: Investigación encuadrada en el proyecto CGL 2016-80155-R.

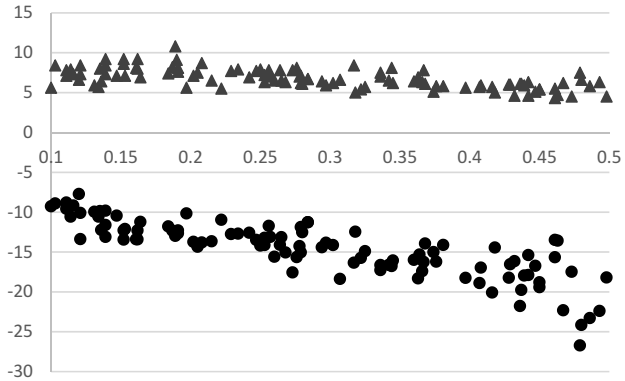


Figura 1. *Diferencias entre los resultados obtenidos para MSE (círculo) y correlación entre TBV y GEBV (triángulo) en los modelos G y W en poblaciones P1.*

WEIGHTING SNP PANELS FOR GENOMIC PREDICTION BY SIMULATED ANNEALING

ABSTRACT: The aim of this research was to propose a new approach for genomic evaluation where SNPs were weighted by a value ranging between 0 and 1 (**model W**) and compared against standard gBLUP models (**model G**) by analyzing simulated datasets.
Mixed model equations were solved by the iterative Gauss-Seidel procedure. The performance of model W was evaluated under simulated annealing (SA) based on the mean squared error (MSE). Two kinds of populations of 1000 individuals were tested; populations P1 with 1 autosomal chromosome of 100 cM (P1) and heritability ranging from 0.1 to 0.5, and populations P2 with 30 chromosomes and heritability of 0.1, 0.25 or 0.4. Model W with weighted SNPs have reported better fit to simulated data than Model G for both MSE and correlation. As expected, the implementation of weights for SNP when constructing the genomic relationship matrix provided higher accuracies than former approaches where all SNPs have the same influence (weight equal 1).

Keywords: SNP panels, valores mejorantes, programa de selección