PONDERACIÓN DE 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

INTRODUCCIÓN

Los modelos actuales para la evaluación genómica se basan principalmente en gBLUP (genomic Best Linear Unbiased Prediction, Henderson, 1973; Gianola et al., 2003), que compila la información de miles de SNPs (Single Nucleotide Polymorphism) en la matriz genómica de parentescos **G**. Sin embargo, algunos estudios han sugerido la necesidad de ponderar la contribución de los distintos SNP en la construcción de **G** (Casellas. 2012). En el presente trabajo se propone un nuevo mecanismo donde todos los SNPs participan en la construcción de **G** pero con distintas ponderaciones, con valores entre 0 y 1 (**modelo W**) y lo compara con el modelo standard gBLUP (**modelo G**) a partir del análisis de datos simulados.

MATERIAL Y MÉTODOS

Los datos simulados consisten en poblaciones de individuos diploides con 2 cromosomas autosómicos (100 cM cada uno) con más de 1.500 SNPs polimórficos y 300 QTL (Quantitative Trait Loci) también polimórficos tras 1.001 generaciones no superpuestas de apareamiento aleatorio (Ne = 100). Para cada generación, los genotipos de la descendencia fueron simulados teniendo en cuenta los fenómenos de recombinación, mutación y desequilibrio genético. Para los análisis genómicos se usaron los genotipos de la generación 1,001 (1000 individuos) y se simuló un fenotipo para cada individuo con heredabilidad 0.1, 0.25 o 0.4.

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

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

donde **y** era el vector de observaciones (fenotipos), **μ** la media poblacional, **a** los efectos aleatorios del animal, **e** el vector de efectos residuales y **Z** la matriz de incidencia que relaciona los datos fenotípicos con los efectos aleatorios del animal (Mrode & Thompson, 2005). Para resolverlo se construyeron ecuaciones de modelo mixtas (MME, Henderson, 1950)

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

donde **A** era la matriz de parentescos (Wright, 1922) y **X** un vector columna de unos. Al disponer de información genotípica para todos los individuos de la generación 1.001, 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 $\mathbf{i}=\mathbf{G}^{-1}\mathbf{a}$ sigue una distribución normal $N(\mathbf{0},\mathbf{G}^{-1}\mathbf{G}\mathbf{G}^{-1}\sigma_g^2)$.

El modelo fue resulto por el método iterativo de Gauss-Seidel (Mrode & Thompson, 2005) bajo *simulated annealing*.

El ajuste del modelo W fue evaluado de manera independiente bajo dos parámetros estadísticos distintos, el error cuadrático medio (MSE) de los fenotipos y la correlación entre valores mejorantes simulados y predichos. Por lo tanto, se examinaron dos variantes del modelo, una variante que decide si el cambio propuesto en el peso del SNP durante el alineamiento simulado era aceptado o no si la correlación total entre valores mejorantes estimados y simulados aumentaba (modelo W_{Cor}); y otra variante que aceptaba los cambios si el MSE total disminuía (modelo W_{MSE}). Ambos modelos obtenían resultados tanto para correlación como para MSE.
En total se evaluaron 9 poblaciones diferentes bajo ambas variantes del modelo (obteniendo un total de 18 réplicas), 3 poblaciones para cada una de las 3 diferentes heredabilidades del carácter fenotípico bajo estudio (0,1, 0,25 y 0,4).

RESULTADOS Y DISCUSIÓN

Los resultados obtenidos por las diferentes poblaciones fueron comparados tras 80.000 iteraciones de *simulated annealing*. Al evaluar las poblaciones según sus heredabilidades con test t Student, únicamente se encontraron diferencias significativas en una de las comparaciones. Existen diferencias entre poblaciones con heredabilidad 0,1 y 0,4 para el parámetro MSE bajo la variante del modelo W_{MSE}, pero no hay diferencias entre el resto de parámetros. Por lo tanto, parece que el modelo W está poco influenciado por la heredabilidad del carácter en estudio (Figura 1).
Para evaluar la actuación del modelo W, los resultados fueron comparados con los valores obtenidos en las mismas poblaciones por el modelo tradicional G (donde todos los SNP contabilizan al mismo nivel). Para cada población, se obtuvo el porcentaje de mejora de ambos parámetros estadísticos, correlación y MSE. Según el test t Student, se encontraron diferencias significativas tanto para correlación (p=0,015) como para MSE (p=2,0×10^{-7}) dependiendo del parámetro en el que se basaba el *simulated annealing* para resolver el sistema de ecuaciones. Cuando el modelo W fue resuelto tomando la correlación como referencia (modelo W_{Cor}), ésta incrementó un mínimo de un 17%, y hasta un 60% en una de las simulaciones. El MSE en cambio, aunque en la mayoría de poblaciones se redujo hasta en un 10%, en dos de las poblaciones sufrió incluso un pequeño incremento. Por otro lado, cuando el modelo W se basaba en el MSE (modelo W_{MSE}), los valores de MSE se redujeron entre un 15 y un 24,5%, mientras que los resultados para correlación oscilaron entre valores ligeramente negativos hasta una mejora del 20%. Cabe destacar que los valores más extremos, tanto para correlación como para MSE en ambas variantes del modelo pertenecen a la misma población, por lo que esto podría ser artefacto de los datos simulados para esa población concreta (población 3 en el gráfico). De hecho, debemos considerar que los resultados obtenidos estarán fuertemente influenciados por el diseño de la simulación, el reducido número de cromosomas, y la distribución aleatoria tanto de SNPs como de QTLs.
Ambas variantes del modelo obtuvieron mejores resultados tanto para MSE como para correlación que el modelo tradicional. Escoger una variante del modelo u otra dependerá de los objetivos del análisis, si estos se centran más en la precisión del modelo (modelo W_{Cor}) o la bondad de ajuste con los datos analizados (modelo W_{MSE}).

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. Jay 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. • 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. • Wright, S. 1922. American Naturalist 56, 330–338.

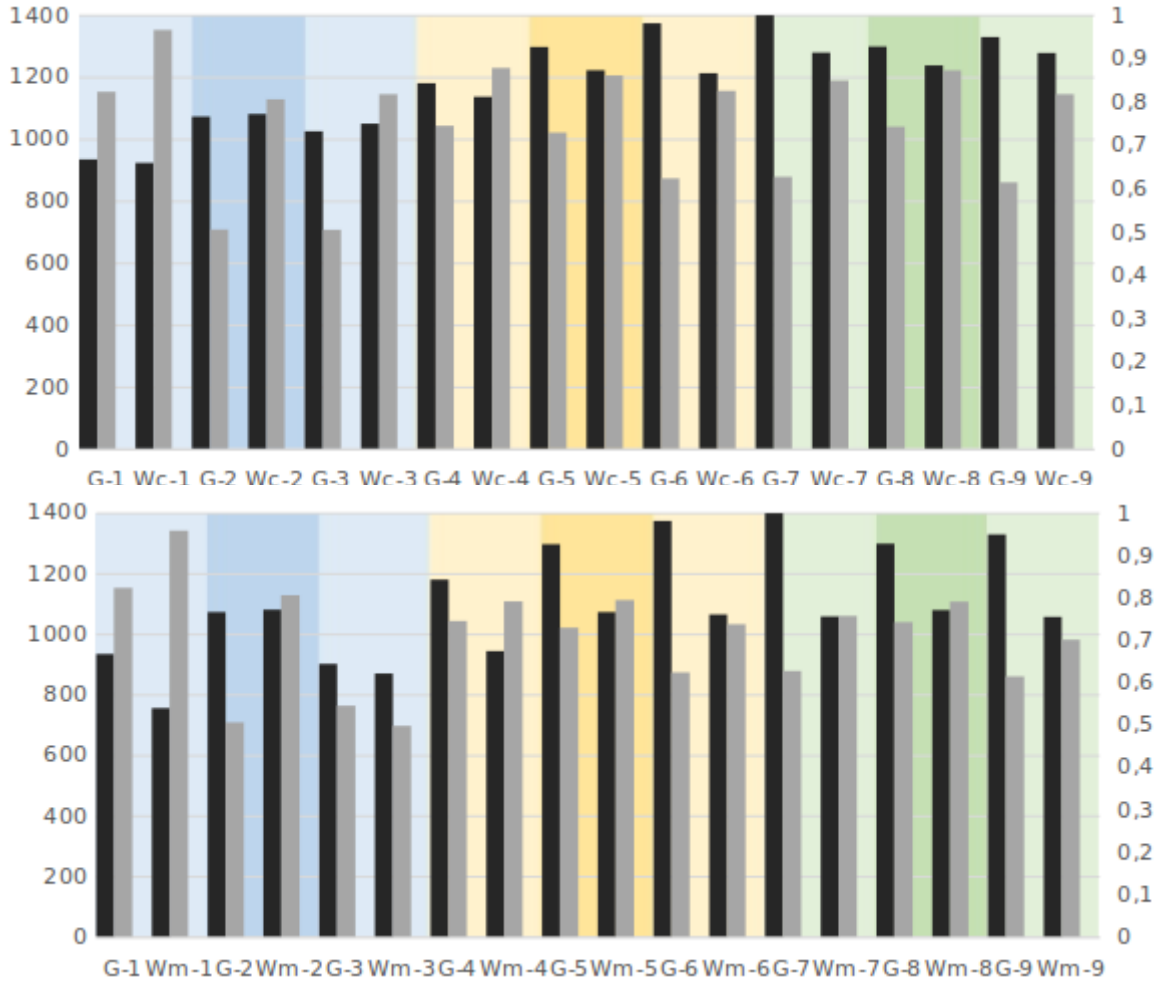

Figura 1. *Comparación de los valores de MSE (negro) y correlación entre valores mejorantes estimados y reales (gris) obtenidos por el modelo tradicional G y las variantes del modelo W_{Cor} (superior) y W_{MSE} (inferior). En sombreado azul se encuentran las poblaciones de baja heredabilidad (0,1); en naranja las poblaciones con heredabilidad intermedia (0,25) y en verde las poblaciones con elevada heredabilidad (0,4).*

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 simulated annealing on simulated data sets.
The performance of model W was evaluated under two different statistical parameters, the mean squared error (MSE, **model W_{MSE}**) and the correlation between simulated and predicted breeding values (**model W_{Cor}**); and trait heritabilities (0.1, 0.25 and 0.4). Model W with weighted SNPs have reported better fit to simulated data than Model G for both MSE and correlation. When using correlation as reference statistic (**model W_{Cor}**), the accuracy of the model improved more than 17% for all simulations. When MSE was the statistical parameter (**model W_{MSE}**), the fit of the model improved by at least 15%. The fit of the model when using correlation as statistical reference and accuracy under MSE criteria were also improved, but in a more variable way. 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: genomic prediction, single nucleotide polymorphism, simulated annealing