Uso de información genómica en evaluaciones genéticas

I. Aguilar[1,2] , I. Misztal[2], D. L. Johnson[3], A. Legarra[4], S. Tsuruta[2], T. J. Lawlor[5]

[1]*Instituto Nacional de Investigación Agropecuaria, INIA Las Brujas, 90200, Uruguay.*
[2]*Animal & Dairy Science Department, University of Georgia, Athens, GA, 30602, EE.UU.*
[3]*LIC, Private Bag 3016, Hamilton 3240, NZ.*
[4]*INRA, UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, Francia*
[5]*Holstein Association USA Inc., Brattleboro, VT 05302, EE.UU.*

Introducción

Las evaluaciones genéticas tradicionales usan información fenotípica y de pedigrí para predecir los valores de cría de las características de relevancia económica para la selección de los animales candidatos a ser utilizados como progenitores. En una reciente revisión, Hill (2008) ha mostrado el éxito de la mejora genética en varias especies domesticas.

Aunque se han descubierto genes con polimorfismos conocidos que afectan a caracteres cuantitativos, en general no han sido ampliamente utilizados dado el poco aporte relativo comparado con la selección basada en valores de cría estimados usando registros fenotipícos y relaciones de parentesco (Goddard, 2009). Basado en varios estudios, el autor presenta cuatros razones que apoyan dichas conclusiones. En primer lugar, la selección tradicional basada en valores de cría estimados, es eficaz. En segundo lugar hay muchos genes que influyen en la expresión de las características. En tercer lugar, dado que las características están controladas por muchos genes, los efectos estimados son pequeños y por lo tanto es difícil obtener estimaciones precisas. Por último, son pocos los genes conocidos responsables de explicar una parte importante de la variación observada de las características cuantitativas.

Selección Genómica

En los últimos años, la disponibilidad de marcadores moleculares de alta densidad del tipo polimorfismos de un solo nucleótido (SNP del inglés single-nucleotide polymorphisms) y la disponibilidad de plataformas para el genotipado de individuos a costos rentables, ha llevado al desarrollo de métodos de selección llamados selección genómica o selección sobre todo el genoma (Meuwissen *et al.,* 2001). La selección genómica puede ser definida como una forma de selección asistida por marcadores, donde un gran número de marcadores genéticos distribuidos sobre todo el genoma, se encuentran en desequilibrio de ligamiento con sectores del cromosoma asociados a características cuantitativas (Meuwissen *et al.*, 2001).

En humanos, resultados del Consorcio Internacional HapMap han identificado más de 3.1 millones de SNPs (Frazer *et al.*, 2007); y chips de 500,000 SNPs están siendo utilizados en estudios de asociación sobre todo el genoma (GWAS, del inglés genome-wide association studies; por ejemplo Weedon *et al.* (2008)).

En ganado vacuno, se ha desarrollado un chip de aproximadamente 57.000 SNPs (Van Tassell *et al.,* 2008; Matukumalli *et al*., 2009) y esta disponible comercialmente mediante el chip Illumina BovineSNP50 BeadChip (Illumina Inc., San Diego, CA). Un subconjunto de SNPs de este chip ha sido seleccionado y se esta utilizando en la evaluación genética nacional de ganado lechero de Estados Unidos (Wiggans *et al.*, 2009).

Resultados de estudios de simulación (Meuwissen *et al*., 2001; Habier *et al*., 2007; Van Raden, 2008) han demostrado que se podría obtener un aumento sustancial en la precisión de la estimación de los valores de cría de animales sin registros (propios o de sus hijos/as), en comparación con las evaluaciones genéticas tradicionales. Esto puede mejorar la ganancia genética mediante la reducción del intervalo generacional, reduciendo los costos de la prueba de progenie de toros en aproximadamente 90% (Schaeffer, 2006). Por otro lado, Konig *et al.* (2009) sostiene que los beneficios de los programas de mejoramiento que usan información genómica son debidos a la sustancial reducción en el

intervalo de generación, al aumento de la precisión de los valores de cría estimados y al aumento de la intensidad de selección de los padres de vacas.

Varios estudios con datos reales se han llevado a cabo para evaluar la exactitud de la selección genómica en diferentes especies animales. Estos estudios incluyen por ejemplo: ratones (Legarra *et al.*, 2008; de los Campos *et al.*, 2009), pollos (Gonzalez-Recio *et al.*, 2009; Chen *et al.*, 2010), ganado de carne (Garrick, 2010), entre otros. Los estudios de selección genómica realizados en ganado lechero incluyen varias poblaciones: Estados Unidos (VanRaden *et al*., 2009), Australia (Hayes *et al*., 2009), Canadá (Van Doormaal *et al.*, 2009), Nueva Zelanda (Harris and Johnson, 2010), Noruega (Luan *et al*., 2009) y Dinamarca (Su *et al*., 2010).

Por lo general, los valores de cría genómicos se obtienen mediante la estimación del efecto de cada uno de los SNPs y posteriormente acumulado el efecto de cada uno ellos, basándose en todos los marcadores genéticos del genoma (Meuwissen *et al*., 2001). El efecto de cada SNP se puede estimar utilizando diferentes supuestos sobre la distribución a priori de sus efectos. Meuwissen *et al.* (2001) definen dos métodos Bayesianos utilizando diferentes tipos de distribución *a priori*. El primer método (BayesA) utiliza una distribución chi-cuadrado invertida para la varianza del marcador, y el segundo método (BayesB) utiliza una prior que tiene una alta densidad de ceros, permitiendo que algunos marcadores tengan un efecto nulo. Por otro lado, asumiendo una distribución normal con igual varianza para los efectos de los marcadores, resulta en el método conocido como GBLUP (Meuwissen *et al*., 2001; Habier et al., 2007; VanRaden, 2008). Los valores de cría genómicos también se pueden estimar con un modelo simple que incluye una matriz de relaciones genómicas (Nejati-Javaremi *et al*., 1997), las cuales pueden ser derivadas de la información de los SNPs, y resultan en un modelo equivalente al GBLUP (Habier *et al.*, 2007; VanRaden, 2008). Mediante el uso de la matriz de relaciones observadas, la selección genómica utiliza los desvíos debidos a el muestreo Mendeliano (Goddard, 2009). Es así que podemos diferenciar matrices de relaciones 'esperadas' o 'promedio' la cuales están basadas en la información de pedigrí, de matrices de relaciones 'realizadas', basadas en información de marcadores moleculares.

Experimentos con datos reales en ganado lechero (Hayes *et al.*, 2009; VanRaden *et al*., 2009) han indicado que el uso de un gran número de marcadores, con igual varianza, es apropiado para la mayoría de las características. Poca (o ninguna) pérdida de precisión se produjo para la mayoría de las características asumiendo igual varianza para cada marcador SNP (Cole *et al*., 2009; VanRaden *et al*., 2009).

Los resultados de estudios de asociación genómica (GWAS) en humanos, muestran que una pequeña fracción de la variación total es explicada por las variantes genéticas de gran efecto (Maher, 2008). Los estudios realizados en humanos en estatura, que tiene estimaciones de heredabilidad en torno a un 80-90%, encontraron que las variantes genéticas sólo explican aproximadamente un 5% de la varianza total (Weedon *et al.*, 2008). En este sentido Goldstein (2009) estimó que se requerirían aproximadamente 93.000 SNP para explicar el 80% de la variación de la altura en humanos. Yang *et al*. (2010) han demostrado que el 45% de la varianza de estatura puede explicarse al considerar todos los SNPs simultáneamente (294.831 SNPs) mediante el uso de un modelo lineal basado en una matriz de relaciones genómicas.

Selección Genómica en Una Etapa

En general, no todos los animales de una población son genotipados, y múltiples valores de cría pueden ser calculados. Así es que los valores de cría tradicionales son calculados mediante el uso de datos fenotípicos e información de pedigrí, en tanto que valores de cría genómicos son estimados para los animales genotipados. Actualmente, las evaluaciones genómicas en ganado lechero se calculan mediante un procedimiento que involucra varias etapas. Una evaluación típica requiere: 1) evaluación tradicional, con un modelo animal por ejemplo, 2) la extracción de pseudo-observaciones tales como valores de cría de-regresados, 3) la estimación de los efectos de los efectos genómicos, y 4) la combinación de los valores de cría genómicos con los valores de cría tradicionales mediante el uso de índices de selección (Hayes *et al.,* 2009; VanRaden *et al.,* 2009; Harris and Johnson, 2010).

El uso de múltiples pasos en la implementación de la selección genómica en las evaluaciones genéticas implica la utilización de varios parámetros y asunciones, resultando en una metodología compleja y por lo tanto sujeta a errores. Misztal *et al.* (2009) han sugerido integrar la información genómica en una evaluación genética de un solo paso (SGUE: Selección Genómica en Una Etapa) en el cual la matriz de relaciones basada en el pedigrí (A) es combinada con una matriz de relaciones genómicas (G) basada en la información de marcadores moleculares (SNPs). En este sentido, han propuesto que la matriz de relaciones entre animales basadas en información de pedigrí (A) puede ser modificada a una matriz (H), que incluye tanto las relaciones basadas en genealogía así como las diferencias basada en la información genómica ($A_{\ddot{A}}$): $H = A + A_{\ddot{A}}$, donde $A_{\ddot{A}}$ es una matriz que contiene las desviaciones debido a la información genómica.

Considerando un modelo animal, unicaracter como el usado generalmente en las evaluaciones genéticas tradicionales:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

donde:

y corresponde al vector de registros,

b al vector de efectos fijos, y

u al vector del efecto animal;

X y Z las correspondientes matrices de incidencia. Bajo el modelo de herencia infinitesimal se asume que var (u) = A σ^2_u , donde A es la matriz de relaciones basadas en el pedigrí. Finalmente se asume var (e) = Iσ^2_e.

Las ecuaciones de los modelos mixtos pueden ser modificadas para considerar la matriz H:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + H^{-1}\alpha \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix}$$

donde α es la relación de varianzas $\left(\alpha = \sigma^2_e / \sigma^2_u\right)$

En casos donde la H^{-1} es imposible de obtener, dicho sistema de ecuaciones simétrico, puede ser transformado a un sistema de ecuaciones no-simétricos (Henderson, 1984):

$$\begin{bmatrix} X'X & X'Z \\ HZ'X & HZ'Z + I\alpha \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ HZ'y \end{bmatrix}$$

Estos sistemas de ecuaciones no simétricos, requieren de algoritmos especiales para estimar las soluciones de los efectos. Misztal *et al.* (2009) presentaron algoritmos similares a los utilizados en las evaluaciones genéticas, basado en el método de gradiente conjugado precondicionado (PCG, (Tsuruta *et al.,* 2001)), pero con soporte para sistemas de ecuaciones no-simétricos (BiCGSTAB, (Vorst, 1992)).

Legarra *et al.* (2009) presentan formulas para derivar la matriz de relaciones H. Las mismas se basan en asumir la distribución conjunta de los valores de cría de animales sin información genómica y genotipados. Particionando el vector de valores de cría (u) en animales sin información genómica (u1) y con información genómica (u2); la densidad conjunta se puede expresar como p(u1,u2) = p(u1|u2) p(u2). La distribución condicional p(u1|u2) se basa en las propiedades de la distribución normal multivariada, y p(u2) se basa únicamente en la información genómica (relaciones genómicas). La distribución conjunta de u1 y u2 es:

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \begin{bmatrix} A_{11} + A_{12} A_{22}^{-1}(G - A_{22})A_{22}^{-1}A_{21} & A_{12} A_{22}^{-1}G \\ GA_{22}^{-1}A_{21} & G \end{bmatrix}$$

donde:

G es la matriz de relaciones genómicas. La matriz G generalmente es definida como:

$$G = \frac{MM'}{2\sum p(1-p)}$$

donde:

M es la matriz de incidencia de los SNPs, con dimensión igual al número de animales por el número de SNPs. Los valores de M corresponden a 0-2p, 1-2p, o 2-2p, si el locus es homocigoto para el

primer alelo, heterocigoto u homocigoto para el segundo alelo, respectivamente; siendo p la frecuencia alélicas del segundo alelo. Legarra *et al.* (2009) presentan formulas alternativas para su implementación en evaluaciones genéticas utilizando sistemas de ecuaciones no-simétricas:

$$H = A + \begin{bmatrix} A_{12}A_{22}^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I \\ I \end{bmatrix} (G - A_{22}) \begin{bmatrix} I & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & A_{22}^{-1}A_{21} \end{bmatrix}$$

Dichas operaciones resultan en operaciones secuénciales de multiplicaciones y sumas de matrices. Las multiplicaciones que involucran particiones de la matriz A, pueden ser implementadas sin tener explícitamente la matriz A en forma densa, utilizando la metodología desarrollada por Colleau (2002). Formulas para su implementación fueron presentada por Misztal *et al.* (2009), en tanto que Aguilar et al. (2010b) presentan un pseudo-código para la implemenacion del método de Colleau.

Evaluaciones genéticas utilizando métodos no-simétricos han sido implementados para poblaciones de miles e incluso algunos millones de animales, sin embargo en poblaciones mas grandes han existido problemas de convergencia.

La inversa de la matriz H ha sido derivada por Aguilar *et al.* (2010a) y por Christensen & Lund (2010):

$$H^{-1} = A^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{bmatrix}$$

Esta formula disminuye drásticamente las operaciones computacionales involucradas con los sistemas de ecuaciones no-simétricos y permite su inclusión en forma directa en las ecuaciones de los modelos mixtos utilizados en las evaluaciones genéticas tradicionales. Generalmente los programas de evaluaciones genéticas utilizan A^{-1} para la resolución de los valores de cría o en la estimación de componentes de varianza. Reemplazando A^{-1} por H^{-1} permite utilizar la selección genómica de forma muy simplificada en los actuales programas computacionales usados en las evaluaciones genéticas. Por otro lado, permite el uso de modelos mas sofisticados: modelos con efectos maternos, multi-carácter, de regresión aleatoria, modelos umbral, etc.

Una posible limitante en la implementación utilizando H^{-1} es la necesidad de la creación de la matriz G y A_{22} y de la inversión de ambas matrices. Para esto, se han propuesto métodos computacionales eficientes (Aguilar *et al*., 2010b). La creación de la matriz G, la cual involucra una multiplicación de matrices puede ser eficientemente realizada utilizando subrutinas que realizan operaciones en bloques para optimizar el uso de la memoria cache y la memoria principal. Por otro lado dichas subrutinas pueden ser fácilmente paralelizadas para su uso en computadores de mas de un procesador.

La creación de la matriz de relaciones en base a pedigrí para los animales no genotipados (A_{22}), puede ser creada en forma simple utilizando la metodología propuesta por Colleau (2002). El algoritmo realiza la multiplicación de la matriz A por un vector, sin tener almacenada explícitamente dicha matriz en forma densa, mediante dos lecturas del pedigrí en forma secuencial. Utilizando vectores con ceros y unos, en la posición de cada animal genotipado, podemos obtener las relaciones de parentesco basadas en pedigrí de los animales genotipados.

La inversión de las matrices G y A_{22} representan un costo computacional importante, pudiendo ser una limitante para su aplicación, aunque el uso eficientes subrutinas y procesamiento en paralelo ha permitido su utilización. Las diferentes metodologías para la creación e inversión de las matrices necesarias para la implementación de SGUE, fueron testeadas utilizando diferentes números de animales genotipados, variando desde 1.000 hasta 30.000 para un panel de 40.000 SNPs (Aguilar *et al.*, 2010b). El tiempo de procesamiento para la creación de la matriz (G) que considero el mayor número de animales fue de aproximadamente una hora, siendo similar el tiempo requerido para la inversión de cada matriz.

Aplicaciones

La metodología SGUE ha sido implementada y validada en la evaluación genética nacional de calificación final en ganado Holando de Estados Unidos (Aguilar *et al.*, 2010a). Información del BovineSNP50 Bead chip (Illumina, San Diego, CA) para 6508 pa-

dres fue utilizada junto con registros de calificación final de 6 millones de vacas, y aproximadamente 9 millones de animales. En comparación a métodos de selección genómica de varios pasos (VanRaden *et al.*, 2009), el método SGUE resultó en valores similares en términos de precisión y sesgo (Aguilar *et al.*, 2010a).

La extensión del método de SGUE a modelos multi-carácter es simple y sencilla. Tsuruta et al. (2010) han aplicado dicha metodología en la evaluación de características asociadas al tipo usando modelos multi-carácter. El uso en modelos multicarácter resultó mas preciso que en los modelos unicaracter.

Estudios de aumento de precisión en características de baja heredabilidad tales como fertilidad, analizada como el resultado de cada inseminación artificial, fueron realizado usando el método SGUE (Aguilar *et al*., 2010c). La aplicación de modelos multicarácter para tasa de preñez en una evaluación genética nacional demostró tener importantes incrementos en precisión mediante el uso de información genómica y la utilización de modelos multi-carácter.

Conclusiones

Las precisiones logradas por el uso de información genómica mediante SGUE son comparables con las obtenidas por los métodos que involucran varias etapas. Las principales ventajas del método de selección genómica de una sola etapa son su simplicidad y la derivación automática de pesos para la combinación de las diversas fuentes de información en la estimación de valores de cría genómicos. Por otra parte, la generalización para el uso de estructuras de datos complejas o modelos más complicado es sencilla y de fácil implementación en los programas computacionales utilizados comúnmente en las evaluaciones genéticas.

Referencias

AGUILAR, I.; MISZTAL, I.; JOHNSON, D. L. *et al*. 2010a. J. Dairy Sci. 93:743–752.

AGUILAR, I.; MISZTAL, I.; LEGARRA, A. *et al*. 2010b. J. Anim. Breed. Genet. Accepted.

AGUILAR, I., MISZTAL, I.; TSURUTA, S. *et al*. 2010c. J. Dairy Sci. Submitted.

CHEN, C.Y.; MISZTAL, I. AGUILAR, I. *et al.* 2010. J. Anim. Sci. Accepted: doi:10.2527/jas.2010–3071.

CHRISTENSEN, O.; LUND, O. 2010. Genet. Sel. Evol. 42:2.

COLE, J. B.; VANRADEN, P. M.; O'CONNELL, J. R. *et al*. 2009. J. Dairy Sci. 92:2931–2946.

COLLEAU, J. J. 2002. Genet. Sel. Evol. 34:409–421.

DE LOS CAMPOS, G.; NAYA, H.; GIANOLA, D. *et al.* 2009. Genetics 182:375–385.

FRAZER, K. A.; BALLINGER, D. G.; COX, D. R. 2007. Nature 449:851–861.

GARRICK, D. J. 2010. 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany.

GODDARD, M. 2009. Genetica 136:245–257.

GOLDSTEIN, D. B. 2009. N Engl J Med 360:1696–1698.

GONZALEZ-RECIO, O.; GIANOLA, D.; ROSA,G. *et al.* 2009. Genet. Sel. Evol. 41:3.

HABIER, D.; FERNANDO, R. L.; DEKKERS, J. C. M. 2007. Genetics 177:2389–2397.

HARRIS, B. L.; JOHNSON D. L. 2010. J. Dairy Sci. 93:1243–1252.

HAYES, B. J.; BOWMAN, P. J.; CHAMBERLAIN, A. J. et al. 2009. J. Dairy Sci. 92:433–443.

HENDERSON, C. 1984. Application of linear models in animal breeding. University of Guelph, Ontario.

HILL, W. G. 2008. Lohmann Information 43:3–20.

KONIG, S.; SIMIANER, H.; WILLAM, A. 2009. J. Dairy Sci. 92:382–391.

LEGARRA, A.; AGUILAR, I.; . MISZTAL, I. 2009. J. Dairy Sci. 92:4656–4663.

LEGARRA, A.; ROBERT-GRANIE, C.; MANFREDI, E. et al. 2008. Genetics 180:611–618.

LUAN, T.; WOOLLIAMS, J. A.; LIEN, S. *et al.* 2009. Genetics 183:1119–1126.

MAHER, B. 2008. Nature 456:18–21.

MATUKUMALLI, L. K.; LAWLEY, C. T.; SCHNABEI, R. D. et al. 2009. PLoS ONE 4:e5350.

MEUWISSEN, T. H. E., B. J. HAYES, AND M. E. GODDARD. 2001. Genetics 157:1819–1829.

MISZTAL, I.; LEGARRA, A.; AGUILAR, I. 2009. J. Dairy Sci. 92:4648–4655.

NEJATI-JAVAREMI, A.; SMITH, C.; GIBSON, J. P. 1997. J. Anim. Sci. 75:1738–1745.

SCHAEFFER, L. R. 2006. J. Anim. Breed. Genet. 123:218–223.

SU, G.; GULDBRANDTSEN, B.; GREGERSEN, V. R. *et al.* 2010. J. Dairy Sci. 93:1175–1183.

TSURUTA, S.; AGUILAR, I.; MISZTAL, I. *et al*. 2010. Commun. No. 489 in 9th World Congress on Genetics Applied to Livestock Production, Leipzig, Germany.

TSURUTA, S.; MISZTAL, I.; STRANDEN, I. 2001. J. Anim. Sci. 79:1166–1172.

VAN DOORMAAL, J.; KISTEMAKER, G.; SULLIVAN, P. G. *et al*. 2009. Interbull Bull. 40:214–218.

VAN TASSELL, C. P.; SMITH, T. P. L.; MATUKUMALLI, L. K. *et al.* 2008. Nature Methods 5:247–252.

VANRADEN, P. M. 2008. J. Dairy Sci. 91:4414–4423.

VANRADEN, P. M.; VAN TASSELL, C. P. WIGGANS, G. R. *et al.* 2009. J. Dairy Sci. 92:16–24.

VORST, H. A. V. D. 1992. SIAM J. Sci. Stat. Comput. 13:631–644.

WEEDON, M. N.; LANGO, H.; LINDGREN, C. M. *et al.* 2008. Nat. Genet. 40:575–583.

WIGGANS, G. R.; SONSTEGARD, T. S.; VANRADEN, P. M. *et al.* 2009. J. Dairy Sci. 92:3431–3436.

YANG, J.; BENYAMIN, B.; MCEVOY, B. P. *et al.* 2010. Nat Genet 42:565–569.