# VARIANTES RARAS: ERRORES DE SECUENCIACIÓN Y HEREDABILIDAD FALTANTE

González-Recio[1], O., Daetwyler, H.D., MacLeod, I.M., Pryce, J.E., Bowman, P.J., Hayes, B.J. y Goddard, M.E.
[1]Department of Environment and Primary Industries, Bundoora, Victoria 3083, Australia.
ogrecio@gmail.com

## INTRODUCCIÓN

El uso masivo de polimorfismos de un sólo nucleótido (SNP) en los modelos de análisis de varianza no han sido capaces, en la mayoría de las ocasiones, de recuperar toda la varianza aditiva genética que explican las relaciones de parentesco genealógico (Manolio et al., 2009). Este efecto es conocido como heredabilidad faltante, y ocurre incluso utilizando todos los marcadores simultáneamente en poblaciones ganaderas donde la genealogía es amplia y fiable (Jensen et al., 2012; Haile-Mariam et al., 2013, Roman_Ponce et al., 2014).

Una de las hipótesis para explicar la heredabilidad faltante es el desequilibrio de ligamiento incompleto entre los SNPs y las mutaciones causales, especialmente aquellas que aparecen en frecuencias alélicas bajas en las poblaciones (Gibson, 2012). La teoría conocida como variantes raras-enfermedades complejas ha sido respaldada por algunos autores, aunque aún no ha podido ser comprobada o refutada. El projecto "1000 Bull Genome Project" (Daetwyler et al., 2014) ofrece la posibilidad de utilizar las variantes provenientes de la secuenciación para estudiar la problemática de la heredabilidad faltante y mejorar la precisión de las evaluaciones genéticas. Sin embargo, la secuenciación tiene asociada un error estimado en el 1% (trabajos previos no publicados), pero este valor es una media global, y no tiene en cuenta la frecuencia alélica.

El objetivo de este estudio es validar las variantes raras en dúos de padre-hijo para diferenciarlas de errores de secuenciación, y estimar que proporción de heredabilidad faltante explican las variantes raras en la población Holstein, así como su contribución a la mejora de la precisión de las evaluaciones genéticas.

## MATERIAL Y MÉTODOS

Para este estudio se usaron 429 secuencias de individuos provenientes de 15 razas del proyecto "1000 Bull Genomes Project". Tras el control de calidad (McKenna et al., 2010; Grant et al., 2011; Purcell et al., 2007), se seleccionaron 2.785.440 variantes (SNP e indels) en regiones codificantes y aquellas en regiones proximales ($\pm$2000 pb "up-" y "down-stream") a las regiones codificantes. De ellas se filtraron aquellas en desequilibrio de ligamiento menor a 0,9999 (675.062). Las secuencias de los individuos de raza Holstein (122) y Jersey (26) fueron utillizadas como conjunto de referencia para la imputación (Browning y Browning, 2009) de 3.311 toros Holstein con genotipo basado en el chip "Bovine HD SNP" (632.002 SNP después de aplicar control de calidad y filtrado).

Las variantes genómicas se clasificaron como comunes (MAF>0,05), infrecuentes (0,01<MAF<0,05) y raras (MAF<0,01). Con el propósito de diferenciar variantes raras de errores de secuenciación se seleccionaron sólo aquellas variantes que al menos aparecieran en 2 de los animales secuenciados, y se evaluó el porcentaje de variantes raras que son errores de secuenciación utilizando 38 dúos de padres e hijos. En total se usaron 83.856 variantes raras en regiones codificantes y proximales, de las 4.442.216 encontradas en el total del genoma. El número de variantes infrecuentes fue de 102.549.

Los caracteres analizados fueron las desviaciones fenotípicas de las hijas corregidas por efectos ambientales y maternos para litros de leche (KL), kg de grasa (KG), kg de proteína (KP), e intervalo entre partos (IP) como indicador de fertilidad. Los componentes de varianzas se estimaron utilizando modelos mixtos con matriz de relaciones genómicas (GBLUP), incorporando las variantes comunes, infrecuentes y raras de manera sucesiva, y realizando una comparación de modelos por medio del test del cociente de la verosimilitud. Los análisis se implementaron con el programa ASReml 3 (Gilmour et al., 2009). La capacidad predictiva de las secuencias y las variantes raras se evaluó en un escenario de

validación cruzada con 2.832 toros Holstein como conjunto de referencia, y 465 toros descendientes de estos como conjunto de validación.

## RESULTADOS Y DISCUSIÓN

Más de la mitad de las variantes genómicas detectadas en la secuenciación presentaron MAF<0,02 (Figura 1, izquierda). La Figura 1 (derecha) muestra la proporción de variantes alélicas para el alelo de menor frecuencia, que estando en heterocigosis en el padre, aparecen también en el hijo para 38 dúos, en función de la frecuencia de los alelos menores. Para los alelos raros y de menor frecuencia, la probabilidad es cercana a 0,50. Sin embargo, la proporción observada en los dúos fue la mitad (0,25). Esto indica que aproximadamente la mitad de las variantes raras detectadas durante los procesos de identificación de variantes en la secuenciación del genoma son errores de secuenciación. Para los loci con MAF entre 0,02 y 0,05, los errores de secuenciación son todavía altos, y a partir de MAF>0,10, las proporciones observadas en la progenie se aproximan a lo esperado por herencia mendeliana.

En cuanto a la varianza genética explicada por los marcadores, el genotipado de alta densidad capturó entre el 81 y el 93% de la varianza genética aditiva en comparación con el pedigrí (heredabilidad faltante de entre 7 (IP) y 19 (KG) %). Las variantes comunes provenientes de secuencias tampoco capturaron la misma varianza genética que el pedigrí, (heredabilidad faltante de 10% (KL), 17% (KG), 15% (KP) y 2% (IP)). Para los caracteres de producción esta heredabilidad faltante es mayor a la esperada cuando no se cuenta con la secuenciación de los individuos de la población base, nuestro caso fue del 5%. Para el IP, nuestros análisis indican que la secuenciación si recupera toda la heredabilidad esperada.

La Tabla 1 muestra el porcentaje de varianza explicada por las variantes comunes, el pedigrí, variantes infrecuentes y las variantes raras cuando se introducen simultáneamente en el modelo. Las variantes comunes capturaron la mayor parte de la varianza aditiva total (entre el 76 y el 84%). Las relaciones de parentesco capturaron el 23% (KL y KP) y 14% (KG) en los caracteres productivos ($P$>0,01). Las variantes infrecuentes y raras no capturaron varianza aditiva de forma relevante ni significativa, tan sólo un 3% de ésta fue explicada por las variantes raras para KG. Sin embargo, las relaciones de parentesco no capturaron más del 2% de la varianza, acorde con lo esperado ya que las variantes comunes fueron la única fuente de información en el modelo, tal como se explica más arriba. Las variantes raras en cambio explicaron el 14% de la varianza genética aditiva total.

A pesar de las diferencias en la proporción de varianza aditiva capturada, las predicciones en validación cruzada no mejoraron sustancialmente utilizando datos de secuenciación con respecto a los genotipados de SNPs. Las variantes comunes mejoraron entre un 2 y un 3 % la capacidad predictiva de los chips de SNPs sólo para los caracteres KL y KG. Las variantes raras sólo mejoraron un punto (2%) la precisión de las predicciones de IP. Cabe destacar que éstas son estimas puntuales y no se estimó su incertidumbre.

Estos resultados implican que aunque las variantes genómicas comunes pueden explicar la mayor parte de la varianza genética aditiva, existe una pequeña proporción que no es posible capturar, al menos contando sólo con las regiones codificantes y sus regiones proximales. Las variantes raras no son la causa de la heredabilidad faltante en caracteres que han sido objeto de una selección intensa, pero si en aquellos sujetos a menor intensidad de selección o caracteres de "fitness". Es necesario desarrollar estrategias que sean capaces de estimar de forma precisa el efecto de estas variantes raras para implementar su selección o purga en los programas de mejora.

## REFERENCIAS BIBLIOGRÁFICAS

• Browning, B.L. & Browning, S.R. 2009. Am. J. Hum. Genet. 84: 210-223 • Daetwyler, H.D., et al. 2014. Nat. Genet. doi:10.1038/ng.3034 •Gibson, G. 2012. Nat. Rev. Genet. 13: 135-145. • Gilmour, A.R., et al. 2009. ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, HP1 1ES, UK www.vsni.co.uk • Grant, J.R., et al. 2011 Bioinformatics 27: 2300-2301. • Haile-Mariam, M., et al. 2013. J. Anim. Breed. Genet. 130(1): 20-31. • Jensen, J., et al. 2012. Genet. Sel. Evol. 13: 44. • Manolio, T.A., et al. 2009. Nature 461: 747-753.• McKenna, A., et al. 2010. Genome Res. 20: 1297-303. • Purcell, S., et al. 2007.

Am. J. Hum. Genet. 81(3): 559-575. • Román-Ponce, S.I., et al. 2014. Genet. Sel. Evol. 46: 36.

**Agradecimientos:** Los autores agradecen la financiación de ”CRC dairy futures”, así como al proyecto ”1000 bull genomes” y a ADHIS por la cesión de los datos utilizados.

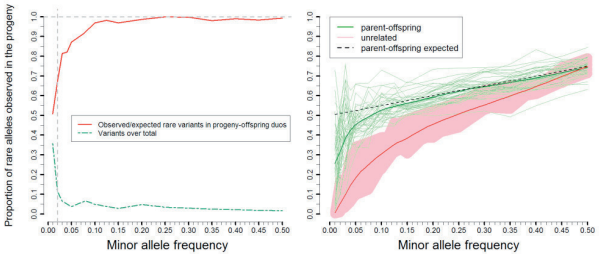
***Tabla 1**. Proporción de heredabilidad explicada por cada fuente de información. Se indica el nivel de significación del test del cociente de la verosimilitud de los modelos.*

| | Variantes comunes | Pedigrí | Variantes infrecuentes | Variantes raras |
|---|---|---|---|---|
| **Grasa (kg)** | 83% | 14%** | 0% | 3%† |
| **Leche (L)** | 77% | 23%** | 0% | 0% |
| **Proteína (kg)** | 76% | 23%** | 0% | 1% |
| **Fertilidad (días)** | 84% | 2% | 0% | 14%** |

***Tabla 2**. Precisión (cor) y error cuadrático medio (EMC) de la predicción genómica. El set de referencia fueron 2832 machos Holstein, y el set de validación 465 toros descendientes de aquellos.*

| Carácter | SNP chip Bovine HD[1] | | Variantes comunes | | Variantes comunes y raras | |
|---|---|---|---|---|---|---|
| | *cor* | *EMC* | *cor* | *EMC* | *cor* | *EMC* |
| **Grasa (kg)** | 0,60 | 146 | 0,57 | 143 | 0,57 | 143 |
| **Leche (L)** | 0,61 | 114261 | 0,63 | 110916 | 0,63 | 110923 |
| **Prot (kg)** | 0,65 | 81 | 0,65 | 81 | 0,65 | 81 |
| **Fertilidad (días)** | 0,42 | 182 | 0,42 | 182 | 0,43 | 182 |

***Figura 1**. Proporción observada de alelos transmitidos a la progenie y proporción de variantes sobre el total en función de la frecuencia alélica (izquierda), y proporción de variantes transmitidas a la progenie en 38 dúos de padre-hijo y dúos de animales no emparentados y la proporción de transmisión alélica esperada (derecha).*




## RARE VARIANTS: SEQUENCING ERRORS AND MISSING HERITABILITY

**ABSTRACT:** Sequence variants in coding regions from 429 sequenced animals were used to impute high density SNP genotypes of 3311 Holstein sires to sequence. There were 675,062 common variants (MAF>0.05), 102,549 uncommon variants (0.01<MAF<0.05), and 83,856 rare variants (MAF<0.01). We estimated that ~ 50% of variants with MAF<0.01 are sequencing errors. Common sequence variants captured 83%, 77%, 76% and 84% of the total genetic variance for fat, milk, and protein yields and fertility, respectively, using GBLUP. Rare variants captured 3%, 0%, 1% and 14% of the genetic variance for fat, milk and protein yields, and fertility respectively, whereas pedigree explained the remaining amount of genetic variance (none for fertility). Using common sequence variants slightly improved accuracy of genomic predictions. However, rare variants only increase the predictive ability of fertility by 2%. These results suggest that rare variants recover a small percentage of the missing heritability for complex traits, however very large reference sets will be required to exploit this in genomic evaluations for fitness traits like fertility.

**Keywords*:*** Rare variants, next generation sequencing, missing heritability, genomic prediction.