Genomic selection in small dairy cattle populations

Ph.D Thesis by José Antonio Jiménez Montero

Under supervision of:

Advisors:

Dr. Oscar González Recio and Dr. Rafael Alenda Jiménez

Department advisor:

Prof. Agustín Blasco Mateu

Valencia, March 2013

RESUMEN

La selección genómica está cambiando profundamente el mercado del vacuno de leche. En la actualidad, es posible obtener una alta precisión en las valoraciones genéticas de animales muy jóvenes sin la necesidad del fenotipo propio o el de sus hijas. Por tanto, la respuesta genética de un programa genómico bien diseñado supera netamente a la selección tradicional. Esta mejora está modificando uno de los principios tradicionales del mercado de vacuno de leche como era la preferencia de uso de toros con altas fiabilidades frente a otros animales con valores genéticos *a priori* superiores.

Esta tesis contiene seis capítulos en los cuales se estudian de las bases para la implementación del programa de selección genómica en el vacuno de leche español. Para ello se realizaron estudios de simulación y valoraciones genómicas con datos reales de la primera población nacional de referencia.

El objetivo principal de esta tesis es contribuir a la implementación de la selección genómica en el vacuno de leche español. Los objetivos específicos son: (1) Estudiar alternativas de genotipado en poblaciones reducidas de vacuno lechero. (2) Desarrollar y validar metodología para la evaluación de grandes cantidades de genotipos. (3) Estudiar el efecto de los procesos de imputación de genotipos en la capacidad predictiva de los genotipos resultantes.

Las principales cuestiones relacionadas con la selección genómica en vacuno lechero fueron discutidas en el capítulo 1 incluyendo: aspectos estadísticos y genéticos en los que se basa la selección genómica, diseño de poblaciones de referencia, revisión del estado del arte en cuanto a la metodología desarrollada para evaluación genómica, diseño y métodos de los algoritmos de imputación, e implementación de la selección genómica en vacuno de leche a nivel de programa de selección, centro de inseminación y de granja comercial.

En el capítulo 2 se realizó un estudio de simulación comparando estrategias de genotipado selectivo en poblaciones de hembras frente al uso de selección tradicional o selección genómica con una población de referencia de machos. La población de referencia española estaba formada en principio por algo más de 1,600 toros con prueba de progenie. Este tamaño no es, en principio, suficiente para obtener predicciones genómicas de alta fiabilidad. Por tanto, debían evaluarse diferentes alternativas para incrementar la habilidad predictiva de las evaluaciones. Las estrategias que consisten en usar como población de referencia los animales en los extremos de la distribución fenotípica permitían mejorar la precisión de la evaluación. Los resultados usando 1,000 genotipos fueron 0.50 para el carácter de baja heredabilidad y 0.63 para el de heredabilidad media cuando la variable dependiente fue el fenotipo ajustado. Cuando se usaron valores genéticos como variable dependiente las correlaciones fueron 0.48 y 0.63 respectivamente. Para los mismos caracteres, una población de 996 machos obtuvo correlaciones de 0.48 y 0.55 en las predicciones posteriores. El estudio concluye que la

estrategia de genotipado que proporciona la mayor correlación es la que incluye las hembras de ambas colas de la distribución de fenotipos. Por otro lado se pone de manifiesto que la mera inclusión de las hembras élite que son las habitualmente genotipadas en las poblaciones reales produce resultados no satisfactorios en la predicción de valores genómicos.

En el capítulo 3, el Random Boosting (**R-Boost**) es comparado con otros métodos de evaluación genómica como Bayes-A, LASSO Bayesiano y G-BLUP. La población de referencia española y caracteres incluidos en las evaluaciones genéticas tradicionales de vacuno lechero fueron usados para comparar estos métodos en términos de precisión y sesgo. Las predicciones genómicas fueron más precisas que el índice de pedigrí tradicional a la hora de predecir los resultados de futuros test de progenie como era de esperar. Las ganancias en precisión debidas al empleo de la selección genómica dependen del carácter evaluado y variaron entre 0.04 (Profundidad de ubre) y 0.42 (Porcentaje de grasa) unidades de correlación de Pearson. Los resultados promediados entre caracteres mostraron que el LASSO Bayesiano obtuvo mayores correlaciones superando al R-Boost, Bayes-A y G-BLUP en 0.01, 0.03 y 0.03 unidades respectivamente. Las predicciones obtenidas con el LASSO Bayesiano también mostraron menos desviaciones en la media, 0.02, 0.03 y 0.10 menos que Bayes-A, R-Boost y G-BLUP, respectivamente. Las predicciones usando R-Boost obtuvieron coeficientes de regresión más próximos a la unidad que el resto de métodos y los errores medios cuadráticos fueron un 2%, 10% y 12% inferiores a los obtenidos a partir del B-LASSO, Bayes-A y G-BLUP, respectivamente. El estudio concluye que R- Boost es una metodología aplicable a selección genómica y competitiva en términos de capacidad predictiva.

En el capítulo 4, el algoritmo de machine learning R-Boost evaluado en el capítulo 3 es descrito e implementado para selección genómica adaptado a la evaluación de grandes bases de datos de una forma eficiente. Tras la incorporación en el consorcio Eurogenomics, el programa genómico español pasó a disponer de más de 22,000 toros probados como población de referencia, por tanto era necesario implementar un método capaz de evaluar éste gran conjunto de datos en un tiempo razonable. El nuevo algoritmo denominado R-Boost realiza de forma secuencial un muestreo aleatorio de SNPs en cada iteración sobre los cuales se aplica un predictor débil. El algoritmo fue evaluado sobre datos reales de vacuno de leche empleados en el capítulo 3 estudiando más en profundidad el comportamiento de los parámetros de sintonización. Esta propuesta de modificación del Boosting puede obtener predicciones sin perdida de precisión o incrementos de sesgo empleando tan solo un 1% del tiempo de computación original.

En el capítulo 5 se evalúa el efecto de usar genotipos de baja densidad imputados con el software *Beagle* en cuanto a su posterior habilidad predictiva cuando son incorporados a la población de referencia. Para ello se emplearon dos métodos de evaluación R-Boost y un BLUP con matriz genómica. Animales de los que se conocían los SNPs incluidos en los chips

GoldenGate Bovine 3K y BovineLD BeadChip, fueron imputados hasta conocer los SNPs incluidos en el BovineSNP50v2 BeadChip. Posteriormente, un segundo proceso de imputación obtuvo los SNPs incluidos en el BovineHD BeadChip. Tras imputatar desde dos genotipados a baja densidad, se obtuvo similar capacidad predictiva a la obtenida empleando los originales en densidad 50K. Sin embargo, sólo se obtuvo una pequeña mejora (0.002 unidades de Pearson) al imputar a HD. El mayor incremento se obtuvo para el carácter días abiertos donde las correlaciones en el grupo de validación aumentaron en 0.06 unidades de Pearson las correlaciones en el grupo de validación cuando se emplearon los genotipos imputados a HD. En función de la densidad de genotipado, el algoritmo R-Boost mostró mayores diferencias que el G-BLUP. Ambos métodos obtuvieron resultados similares salvo en el caso de porcentaje de grasa, donde las predicciones obtenidas con el R-Boost fueron superiores a las del G-BLUP en 0.20 unidades de correlación de Pearson. El estudio concluye que la capacidad predictiva para algunos caracteres puede mejorar imputando la población de referencia a HD así como empleando métodos de evaluación capaces de adaptarse a las distintas arquitecturas genéticas posibles.

Finalmente en el capitulo 6 se desarrolla una discusión general de los estudios presentados en los capítulos anteriores y se enlazan con la implementación de la selección genómica en el vacuno lechero español, que se ha desarrollado en paralelo a esta tesis doctoral. La primera población de referencia con unos 1.600 toros fue evaluada en el capítulo 4 y fue usada para comparar los distintos métodos y escenarios propuestos en los capítulos 3, 4 y 5. La primera evaluación genómica obtenida para los caracteres incluidos en el capítulo 4 de esta tesis estuvo disponible para los centros de inseminación incluidos en el programa en septiembre de 2011. La población de Eurogenomics se incorporó en Noviembre de dicho año, completándose la primera evaluación para los caracteres incluidos en el índice de selección ICO en Febrero de 2012 empleando el R-Boost descrito en el capítulo 3. En mayo de 2012 las evaluaciones del carácter proteína fueron validadas por Interbull y finalmente el 30 de Noviembre del 2012 las primeras evaluaciones genómicas oficiales fueron publicadas on-line por la federación de ganaderos CONAFE (http://www.conafe.com/noticias/20121130a.htm).

.