UNIVERSIDAD POLITECNICA DE VALENCIA

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

Genomic selection in small dairy cattle populations

A thesis submitted to the Polytechnic University of Valencia in fulfillment of the requirements for the degree of doctor of philosophy with international Doctor Mention

By

José Antonio Jiménez Montero

Sig.

Thesis Advisors

Dr. Oscar González Recio **Dr. Rafael Alenda Jiménez**

Sig. Sig.

Department advisor

Prof. Agustín Blasco Mateu

Sig.

Agradecimientos

Me gustaría dar las gracias al profesor Rafael Alenda por la confianza que ha tenido en mi durante estos años y especialmente por inculcarme el interés por la mejora genética animal durante los años de estudiante de ingeniería. Muchas gracias por esas charlas tan estimulantes para mí.

Gracias Oscar por todo lo que me has enseñado (y espero sigas enseñándome) y por permitirme invadir tu despacho durante tanto tiempo. Me siento un privilegiado al haber tenído un director de tesis de tu nivel, tan cercano y tan involucrado en mi formación.

Muchas gracias Profesor Agustín Blasco por su ayuda con los trámites burocráticos así como por la formación de calidad recibida en los cursos del Master en Mejora Genética Animal junto al resto de profesores que me hicieron comprender las bases de la mejora genética. Muchas gracias también a los buenos compañeros del Master.

Quiero agradecer a CONAFE y a los centros de inseminación por permitirme realizar los trabajos de esta tesis con los genotipos de la población española y hacerme partícipe de las reuniones del programa genómico.

Muchas gracias Aberekin que junto a la UPM y el INIA a través del proyecto CDTI-P080250866 ha financiado el periodo de investigación y me ha incluido durante este tiempo en sus mesas técnicas donde tanto he podido aprender sobre implementación de la mejora genética en vacuno de leche.

También quiero agradecer al Departamento de Producción Animal de la ETSI Agrónomos de la UPM y al Departamento de mejora genética animal del INIA por la buena acogida recibida durante los periodos de desarrollo de la tesis en sus instalaciones. Especialmente quiero agradecer al Dr. García Cortes por el tiempo dedicado en mi formación en programación y a Cristina Meneses por su paciencia con mis preocupaciones sobre el vacuno de carne.

I would like to thank to Pr. Daniel Gianola and Pr. Kent Weigel who made possible my stay at the University of Wisconsin-Madison. Thanks to the Animal Science department for making me feel at home during that period.

Gracias a la profesora Soledad Álvarez de la USAL por iniciarme en el conocimiento académico de la producción animal y a Ana Díaz por su ayuda en mis años de estudiante.

Muchas gracias al C.R.D.O. Guijuelo que me dio la oportunidad de tener un primer trabajo relacionado con mis estudios, a Javier Martín Tereso que me inició en el mundo de la investigación en el RRC de Nutreco y por supuesto a Hypor España y a Javier Santamartina por darme la oportunidad de conocer el funcionamiento de una empresa de mejora genética porcina durante los años que trabajé con ellos. Esta experiencia previa me ha permitido afrontar los estudios de doctorado con una perspectiva diferente.

Muchas gracias a la Banda Municipal de música de El Barco de Ávila y en especial a su director Alfonso Márquez, por todos los valores tanto musicales como sobre todo humanos y de trabajo en equipo que me han inculcado desde mi infancia y que tanto me están ayudando en mi carrera profesional.

Muchas gracias a todos aquellos amigos de El Barco de Ávila, del Tomás Luis de Victoria, de la universidad, del mundo ganadero y tantos otros que me hacen un poquito mejor persona.

Quiero agradecer a mi familia, Mari Cruz, José Antonio, David, Sofía, tíos, primos y también aquellos que no están, por la educación recibida, por los valores que me han enseñado, por ser un modelo ejemplar de conducta tanto en casa como en la calle, y durante los últimos años, muchas gracias por ser mi banco de pruebas particular para validar mi capacidad como divulgador

de la mejora genética animal. Siguiendo vuestros pasos sé que voy por el buen camino.

Finalmente, aunque lo más importante, muchas gracias María Elena por todos estos años juntos, por tantas cosas positivas que me aportas y por que cada día me haces más feliz.

Contents

Index of Tables

Index of figures

Summary

Genomic selection is producing profound changes in dairy cattle market since reliable breeding values, which double the reliability of the pedigree index, can be obtained earlier in an animal's life. As a result, genetic gains of properly designed genomic programs are considerably larger than genetic gains obtained with traditional approaches. The industry has introduced this new tool all around the world faster than any other previous improvement.

This thesis contains six chapters, in which initial stages for the implementation of genomic selection program in Spanish Holstein population were studied using simulations and real data. The initial interest began in 2008 (González-Recio et al., 2008), when the results obtained by VanRaden, (2008) were used to involve the Spanish industry in genomic selection. This research has been used to obtain the official genomic breeding values and implement the imputation of genotypes.

The global aim of this thesis was to contribute with practical recommendations for implementing genomic selection in the Spanish dairy cattle. The specific objectives were: (1) To study alternative genotyping strategies for small populations, (2) to develop and validate methods for the evaluation of large data sets of genotypes, and (3) to study the effect of imputation on predictive ability.

The main topics with respect to genomic selection in dairy cattle were discussed in chapter 1 including: genetic and statistical aspects underlying genomic selection, design of proper reference populations (**RP**), review of methodology for genome-assisted evaluation, imputation, and implementation of genomic selection in dairy cattle breeding programs. Breeding values with medium high accuracies are now available early in the life of the animals. This is modifying one of the traditional principles of dairy market: the strong preference for highly reliable bulls.

In chapter 2, a simulation study was carried out comparing female-selective genotyping strategies with traditional pedigree index and a bull RP. The Spanish male RP has 1,600 genotypes, which is not large enough to provide reliable predictions. Alternatives should be evaluated to improve predictive ability. The accuracy of predicted genomic breeding values using the two-tailed strategies was better than the accuracy obtained using other strategies (0.50 and 0.63 using yield deviations as phenotype and 0.48 and 0.63 using breeding values in low- and medium-heritability scenarios, respectively, using 1,000 genotyped cows). When 996 genotyped bulls were used as the training population, the sire' strategy led to accuracies of 0.48 and 0.55 for low- and medium-heritability traits, respectively. The most informative strategy involved genotyping of females that exhibited upper and lower extreme values within the distribution. Including just top animals resulted in poor results.

Several methods for implementing genome assisted evaluations were compared in Chapter 3. Methods including marker regression included Bayesian methods (Bayes-A, Bayesian LASSO and Random Boosting (R-Boost). G-BLUP was also utilized using the genomic relationship matrix. The Spanish RP was used to compare those methods in terms of predictive ability and bias. Genomic predictions were more accurate than traditional pedigree indices for predicting future progeny test results of young bulls. The gain in accuracy, due to inclusion of genomic data, varied by trait and ranged from 0.04 to 0.42 Pearson correlation units. Results averaged across traits showed that Bayesian LASSO had the highest accuracy with an advantage of 0.01, 0.03 and 0.03 points in Pearson correlation compared with R-Boost, Bayes-A, and G-BLUP, respectively. The B-LASSO predictions also showed the least biased predictions (0.02, 0.03 and 0.10 SD units less than Bayes-A, R-Boost and G-BLUP, respectively), measured as the mean difference between genomic predictions and progeny test results.

The R-Boost algorithm provided genomic predictions with regression coefficients closer to unity, for four out of five traits and also resulted in mean squared error estimates that were 2%, 10%, and 12% smaller than B-LASSO, Bayes-A, and G-BLUP, respectively. R-Boost seemed to be a competitive marker regression methodology in terms of predictive ability.

Chapter 4 describes the R-Boost algorithm tested in Chapter 3 for genomic evaluations in large data sets. After joining the Eurogenomics consortium with more than 22,000 bulls in the RP, a feasible method with reasonable computation times, and no impaired predictive ability was required. The random boosting uses a random selection of markers to add a subsequent weak learner to the predictive model. Optimization of the algorithm and behavior of tuning parameters was tested in real dairy cattle data. Those tuning parameters control the percentage of single nucleotide polymorphisms (**SNP**) sampled per iteration and the level of shrinkage over the regression coefficient estimation. The proposed modification of the original boosting algorithm can be run in 1% of the time used with the original algorithm, and with negligible differences in accuracy and bias.

In Chapter 5, genotypes from the GoldenGate Bovine 3K and BovineLD BeadChip for 834 animals were imputed to a BovineSNP50v2 BeadChip using *Beagle*. Those genotypes were subsequently imputed to the BovineHD BeadChip. Predictive ability of imputed and native genotypes as RP in genome-assisted evaluations was compared using G-BLUP and R-Boost. Imputed low density genotypes achieved similar predictive ability than native genotypes. However, marginal better selection efficiency was obtained after imputation to HD (0.002 greater Pearson correlation units). The largest improvements were found for Days Open after imputation to HD genotypes (up to 0.06 greater Pearson correlation units). R-Boost was more sensitive to marker density than G-BLUP. Both methods performed similar except for Fat Percentage, where R-Boost outperformed G-BLUP with up to

0.20 Pearson correlation units. The predictive ability of certain traits may be improved either by imputing genotypes to HD or by utilizing a method that takes into account the genetic architecture of the trait.

Finally, in chapter 6 a general discussion links the studies previously covered with the implementation of genomic selection in the Spanish dairy cattle is reported. The first Spanish RP with above 1,600 progeny tested bulls was tested as a proper source of genomic information in chapter 4 and was used for comparing methods and scenarios in chapters 3, 4 and 5. First genomic evaluation was carried out for those traits included in Chapter 4 of this thesis and results were used for AI centers in September 2011. The Eurogenomics population was included on November 2011. First complete genomic evaluation for the 26 traits included in the Spanish index (ICO) was carried out in February 2012 using Random Boosting as described in chapter 4. In May 2012 Spanish genomic evaluation for protein yield was validated by Interbull. Finally, on November 30th 2012, first official genomic evaluations were published on-line by CONAFE (http://www.conafe.com/noticias/20121130a.htm).

Resum

La selecció genòmica està canviant profundament el mercat del boví de llet. Actualment, és possible obtenir valoracions genètiques fiables d'animals molt joves sense necessitat de disposar del fenotip propi o el de les seves filles. Per tant, la resposta genètica d'un programa genòmic ben dissenyat supera netament a la selecció tradicional.

Aquesta tesi es composa de sis capítols en els que s'estudia l'establiment de les bases per a implementar un programa de selecció genòmica en el boví de llet espanyol. Amb aquesta finalitat, s'han realitzat estudis de simulació i valoracions genòmiques amb dades reals de la primera població de referència nacional.

L'objectiu principal d'aquesta tesi és contribuir a la implementació de la selecció genòmica en el boví de llet espanyol. Els objectius específics són: (1) Estudiar alternatives de genotipat en poblacions reduïdes de boví lleter. (2) Desenvolupar i validar metodologia per a l'avaluació de grans quantitats de genotips. (3) Estudiar l'efecte dels processos d'imputació de genotips en l'habilitat predictiva dels genotips resultants.

Les principals qüestions relacionades amb la selecció genòmica en boví lleter van ser discutides el capítol 1 incloent: aspectes estadístics i genètics en què es basa la selecció genòmica, disseny de poblacions de referència adequades, revisió de la metodologia desenvolupada per a l'avaluació, disseny i metodologia de programes d'imputació i implementació de la selecció genòmica en boví de llet a nivell de programa de selecció, centre d'inseminació i granja comercial. La selecció genòmica està revolucionant el mercat del boví de llet, ja que és possible aconseguir valors genètics molt més precisos d'animals joves, en comparació amb els obtinguts mitjançant índexs de pedigrí tradicionals. Aquesta millora està modificant un dels principis tradicionals del mercat de boví de llet com era la preferència d'ús

de toros amb altes fiabilitats respecte animals amb valors genètics *a priori* superiors.

En el capítol 2 es va realitzar un estudi de simulació comparant estratègies de genotipat selectiu en poblacions de femelles enfront de l'ús de selecció tradicional o selecció genòmica amb una població de referència de mascles. La població espanyola estava formada per una mica més de 1,600 toros amb prova de progènie. Aquest mida no és, en principi, suficient per obtenir prediccions genòmiques d'alta fiabilitat. Per tant, calia avaluar diferents alternatives per incrementar l'habilitat predictiva de les avaluacions. Les estratègies que inclouen el genotipat com a població de referència dels animals en ambdós extrems de la distribució permetien millorar la precisió de l'avaluació. Els resultats usant 1,000 genotips van ser 0.50 per al caràcter de baixa heretabilitat i 0.63 per al d'heretabilitat mitjana quan la variable dependent fou el fenotip ajustat. Quan varen usar-se valors genètics com a variable dependent, les correlacions van ser 0.48 i 0.63, respectivament. Per als mateixos caràcters, una població de 996 mascles va obtenir correlacions de 0.48 i 0.55 en les prediccions posteriors. L'estudi conclou que l'estratègia de genotipat que proporciona la major correlació és la que inclou les femelles de les dues cues de la distribució de fenotips. D'altra banda es fa evident que la mera inclusió de les femelles d'èlit, que són les habitualment genotipades, produeix resultats molt pobres en la predicció de valors genòmics.

En el capítol 3, el Random Boosting és comparat amb altres mètodes d'avaluació genòmica utilitzant metodologia Bayessiana (Bayes-A i LASSO Bayessià) i amb un G-BLUP usant la matriu genòmica. La població de referència espanyola va ser utilitzada per comparar aquests mètodes en termes de precisió i biaix. Les prediccions genòmiques van ser més precises que l'índex de pedigrí tradicional a l'hora de predir els resultats de futurs test de progènie. Els guanys obtinguts en precisió derivats de l'ús de la selecció

genòmica depenen del caràcter avaluat i varien entre 0.04 i 0.42 unitats de correlació de Pearson. Els resultats promig entre caràcters demostraren que el LASSO Bayessià va obtenir majors correlacions superant al Random Boosting, Bayes-A i BLUP genòmic en 0.01, 0.03 i 0.03 unitats, respectivament. Les prediccions obtingudes amb el LASSO també van mostrar menys desviacions respecte la mitja, 0.02, 0.03 i 0.10 menys que Bayes-A, R Boost i G-BLUP, respectivament. Les prediccions usant Random Boosting van obtenir coeficients de regressió més propers a la unitat que la resta de mètodes i els errors mitjans quadràtics van ser un 2%, 10% i 12% inferiors als obtinguts a partir del B-LASSO, Bayes-A i G - BLUP, respectivament. L'estudi conclou que el Random Boosting és una metodologia aplicable en selecció genòmica i competitiva en termes d'habilitat predictiva.

En el capítol 4 l'algoritme de machine learning Random Boosting avaluat en el capítol 3 és descrit i implementat per a selecció genòmica i adaptat a l'avaluació eficient de grans bases de dades. Després de la incorporació al consorci Eurogenomics, el programa genòmic espanyol va passar a disposar de més de 22,000 toros provats com a població de referència. Es va fer necessària doncs, l'implementació d'un mètode capaç d'avaluar aquest gran conjunt de dades en un temps raonable. El nou algoritme anomenat Random Boosting realitza de forma seqüencial una selecció aleatòria d'SNPs a cada iteració sobre els quals s'aplica un predictor feble. L'algoritme va ser avaluat sobre les dades reals de boví de llet emprades en el capítol 3 i van estudiar-se més en profunditat el comportament dels paràmetres de sintonització. Aquesta proposta de modificació del Boosting permet obtenir prediccions sense pèrdua de precisió ni increments de biaix emprant només un 1% del temps de computació original.

En el capítol 5 s'avalua l'efecte d'usar genotips de baixa densitat imputats mitjançant el programari Beagle pel que fa a la seva posterior habilitat

predictiva quan aquests són incorporats a la població de referència. Amb aquesta finalitat, es varen utilitzar dos mètodes d'avaluació: Random Boosting i un BLUP amb matriu genòmica. Animals dels que s'en coneixia els SNPs inclosos en els xips GoldenGate Bovine 3K i BovineLD BeadChip varen ser imputats fins a conèixer els SNP's inclosos en el BovineSNP50v2 BeadChip. Posteriorment, un segon procés d'imputació va permetre obtenir els SNP's inclosos en el BovineHD BeadChip. Els genotipats a baixa densitat després de ser imputats, van obtenir similar capacitat predictiva que els originals en densitat 50K. Tanmateix, només és va obtenir una petita millora (en 0.002 unitats de Pearson) a l'imputar HD. El major increment es va obtenir per a dies oberts on les correlacions en el grup de validació varen augmentar en 0.06 unitats de Pearson quan es van emprar els genotips imputats a HD. En funció de la densitat de genotipat, l'algoritme Random Boosting mostra més diferències que el BLUP genòmic. Ambdós mètodes varen obtenir resultats similars tret del cas d' percentatge de greix, on les prediccions obtingudes amb el Random Boosting varen ser superiors a les del G-BLUP en 0.20 unitats de correlació de Pearson. L'estudi conclou que la capacitat predictiva d'alguns caràcters pot millorar imputant la població de referència a HD i usant mètodes d'avaluació que siguin capaços d'adaptar-se a les diferents arquitectures genètiques possibles.

Finalment, en el capítol 6 es duu a terme una discussió general dels estudis presentats en els capítols anteriors que s'enllacen amb la implementació de la selecció genòmica en el boví lleter espanyol, desenvolupada paral·lelament a aquesta tesi doctoral. La primera població de referència, amb uns 1,600 toros, va ser avaluada en el capítol 4 i va ser usada per comparar els diferents mètodes i escenaris proposats en els capítols 3, 4 i 5. La primera avaluació genòmica obtinguda per als caràcters inclosos en el capítol 4 d'aquesta tesi va estar disponible per als centres d'inseminació inclosos en el programa al mes de setembre del 2011. La població d'Eurogenomics es va incorporar al

novembre del mateix any, completant la primera avaluació per als caràcters inclosos en l'índex de selecció ICO al febrer de 2012 emprant el Random Boosting descrit en el capítol 3. El maig de 2012 les avaluacions del caràcter Proteïna van ser validades per INTERBULL i finalment el 30 novembre 2012 les primeres avaluacions genòmiques oficials van ser publicades on-line per la federació de ramaders CONAFE (http://www.conafe.com/noticias/ 20121130a.htm).

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).

.

1

General Introduction:

Infinitesimal model and genomic selection

Animal breeding aims to improve economic productivity of future generations of domestic species through selection under a changing cost and income scenario. Most of the traits of economic interest in livestock have a complex and quantitative expression i.e. influenced by a large number of genes and affected by environmental factors. Statistical analysis of phenotypes and pedigree information allows estimating the genetic merit (breeding values) of the selection candidates based on Fisher's infinitesimal model (Fisher, 1919). The infinitesimal model assumes that quantitative traits are determined by an infinite number of loci with very small effects on these characters. It is assumed that these quantitative trait loci (**QTL**) are homogeneously distributed throughout the genome. The population mean of quantitative traits is modified choosing the best genotypes in the population using the predicted breeding values obtained with the Best Linear Unbiased Predictor (**BLUP**) methodology (Henderson, 1975). The genetic improvement obtained with this traditional quantitative method is due to the average probability of sharing certain variants of genes between relatives.

Nowadays, thanks to the advances in the molecular techniques, a large number of genetic markers are known and can be individually genotyped. Different authors have proposed strategies to use and integrate these new sources of information (Dekkers, 2010; Fernando and Grossman, 1989; Lande and Thompson, 1990). Marker-assisted selection provided options for extra gains by increasing selection accuracy when a sufficiently large number of markers are used (Villanueva et al., 2005).

The sequencing of the human genome, completed in 2003, followed by those of several animal species as cattle (Elsik et al., 2009), have paved the way to a new tool that uses genomic information of each animal. These modern sequencing techniques allow genotyping thousands of sources of variation

throughout the genome. Some of them may relate to the productive performance of the animals, their morphology or resistance to diseases. Some markers represent differences in chemist bases (Adenine, Cytosine, Guanine and Thymine) in certain positions of the DNA sequence. Those markers are known as SNP when differ in a single base. It is expected that some of those variations will be close to QTLs of interest. Therefore, SNPs are used as markers under the assumption that they will be inherited jointly to QTLs due to the existing linkage disequilibrium (**LDQ**) in the genome. Breeding values can be estimated through marker effects estimation considering all QTLs simultaneously. Those marker effects are expected consistent across families. Selection based on this genomic predictions was named Genomic Selection (**GS**) (Haley and Visscher, 1998) and is becoming a new paradigm for genetics.

The breeder equation under GS

Genetic response

The main advantage that genomics provides is the increment of the selection accuracy at an early age of the animal compared to traditional pedigree index (**PI**) when the own phenotype and pedigree is not available (Goddard, 2009). This development is of great importance because it changes the reliability of the information available at the key moment selection decisions have to be taken, such as: bulls to be progeny tested, replacement of heifers, cow culling and mating.

Given the genetic response equation:

$$\partial G = \frac{i \rho \sigma_a}{L}$$

where δG is the expected genetic gain, i is the selection intensity applied, ρ is the accuracy of the evaluation, σ_a is the additive genetic standard deviation, and L the considered generation interval.

The use of genomics has some effect on all terms of the equation:

Accuracy and generation interval

Genomic selection provides a greater ρ in comparison with the PI and a reduction of L as a higher ρ can be obtained at earlier ages of individuals (Dekkers, 2010). This improvement is of special interest for those programs based on the selection of highly reliable individuals, as is the case in dairy cattle.

The accuracy of GS depends, among other factors, on the LDQ between SNPs (Calus et al., 2008). LDQ is defined as the non-random association between the alleles at two different genome loci. LDQ can be caused by migration, mutation, selection or genetic drift in small populations, or any other event that may affect the genetic structure of the population. Sargolzaei et al. (2008) found averaged values for LDQ of 0.31 calculated as r^2 (Hill and Robertson, 1968). Afterwards, Banos and Coffey, (2010) reported levels around 0.30 for r^2 in a Holstein population, and concluded that this is the minimum level required for reliable prediction of genomic breeding values.

High density arrays may provide enough LDQ between genome segments to trace all QTL affecting the traits of interest (Hayes, 2007). However, LDQ is decreasing if a recombination occurs in the meiosis previous to the development of the gametes of each new generation (Habier et al., 2009). To maintain the reliability of genomic predictions, new genotyped and phenotyped individuals should be included in the RP. The estimation of the chromosomal segment effects should be re-evaluated at least every three generations according to Hayes (2007).

The first results obtained with simulations were overly optimistic, showing accuracy of predictions higher than 0.80 (Meuwissen et al., 2001). A revolution in testing programs was proposed with the optimistic expectation that progeny testing cost could be reduced with 92% (Schaeffer, 2006). Besides this overoptimistic background, other factors should be taken into account for the field implementation of GS. An initial large investment in genotyping would be necessary. It is essential to maintain commercial farms involved in the data recording, which is a requirement for a correct estimation of SNPs effects. It must be pointed out also that many dairy farmers tend to use semen from proven bulls with high reliability for mating designs, for compensating weak points and deficiencies of their cows.

Selection intensity

Selection intensity could be incremented in the sire-sire or dam-sire paths due to more reliable information about the individuals or discriminating between full sibs.

Genetic variability

Finally, GS can reduce the emphasis on the family information in comparison with traditional breeding, which is related to a lower increase of inbreeding (Daetwyler et al., 2007). Inbreeding increments are related to a reduction in genetic variability and therefore σ_a, which may negatively affect the genetic gain. However, results are contradictory (Pedersen et al., 2010), and there is a risk of dangerously speeding up inbreeding by sampling only the apparently best families and promoting the most profitable matings in the short term (Dürr and Philipsson, 2012).

The state of the art genomic evaluations have not achieved the initially expected reliabilities (Pryce and Daetwyler, 2012). Genomic information improved the accuracies of genetic values equivalent to 11 daughters in a

traditional progeny test (VanRaden et al., 2009b). Under these assumptions, dairy cattle genomics is focused on pre-selection of candidates to be tested based on their genomic values (Lillehammer et al., 2010). Nevertheless, some young bulls evaluated based on their genomic information have been marketed due to outstanding genomic values, and some aggressive breeding strategies based on genomics have been proposed.

Reference population and genotyping strategies

The dairy cattle breeding market is highly competitive. Breeding programs not implementing these new tools may be gone in disadvantage regarding other competitors in a few years. The leading countries in dairy cattle sector have developed their genomic programs with different alternatives regarding methods and genotyping strategies.

The first step in the implementation of GS is to create a RP of genotyped animals. It is not straightforward to determine what animals should be genotyped first. Most of the countries have genotyped proven bulls for that purpose. Other type of genotyped animals are not yet or less useful for the RP, for instance: young bulls waiting for progeny proof, elite females (bull dams or candidates), top ranking heifers, and production cows from the entire population (see cases of countries such as Denmark, Canada, Finland, Sweden, Unites States Ireland or Holland).

¿What is a reference population?

The fundamental step in genomic selection is the collection of phenotypes (own or from progeny) and DNA samples from those genotyped animals. The "RP" is used to train a statistical model that estimates the effects of each SNP or genomic combinations thereof, on phenotypes. These estimates allow to predict of genomic breeding values for new individuals with the only source of information of their DNA (Dekkers, 2010).

Features of the reference population

The characteristics of the RP, such as the population size or the type of animals, determine the accuracy of predicted genetic values of young animals (Hayes et al., 2009a; VanRaden et al., 2009b). Other aspects determining the predictive accuracy are: the reliability of the phenotypic information, the genetic relatedness of the population, both in the training and validation sets, or the genotyping density. The establishment of an appropriate RP is one of the key aspects in a genomic program. The strategy to include animals in the RP depends on the goal (Pérez-Cabal et al., 2012). For dairy cattle RPs, Saatchi et al, (2010) recommended to use reliable (>90%) progeny tested sires from recent generations rather than older bulls.

Size of the reference population

The required size of the RP is inversely proportional to the heritability of the trait and directly proportional to the effective population size (Goddard, 2009). National populations could be enough for some traits with high h^2, but not for traits with low h^2. One of the challenges in small populations, and especially for low heritability traits is to increase the predictive accuracy obtained with genomic evaluations (VanRaden et al., 2009a).

Different international collaboration frameworks have emerged to increase the accuracy of genome-enhance predictions for a successful implementation of genomic selection. The first association appeared between Canada and the United States to share genotypes and technical knowledge in 2008 with an initial population of around 17,000 genotypes (VanRaden et al., 2009a; Wiggans et al., 2011, 2008). Recently, Great Britain and Italy joined the consortium. The current size of the whole population is above 50,000 genotypes including cows. Other European countries created the Eurogenomics consortium in 2009 (Holland, Germany, France, Finland, Sweden, Denmark). It incorporated Spain in 2011, and Poland in 2012.

Around 22,000 50K genotypes and more than 1,000 high density (**HD**) genotypes are currently shared between these countries. Other countries are currently working in cooperation programs (Cromie et al., 2012; de Roos et al., 2009; Lund et al., 2011). This implies sharing both genotypes and phenotypic records while maintaining the estimation procedures separately at the national centers or research partners.

The increment in reliability due to the RP size was estimated to range between 8% and 11% above the reliability obtained with the national RP using 15,996 sires from the Eurogenomics consortium (Lund et al., 2011). Larger RP will be necessary for multi-breed evaluations, as well as higher density arrays, to compensate for the greater genetic diversity. Although this should be confirmed in real population and develop strategies and methods applied to these particular cases.

Predictive ability of genomics depends mainly in the size of the RP. To raise the number of animals used as reference has been the main objective of many programs. As progeny tested bulls are the most accurate source of information, to share their genomic information is the most successful strategy to enlarge the RP. However, for small populations without many reliable bulls or for those traits not routinely recorded, different alternatives should be considered.

Reference populations closed or dynamics

As mentioned above, the LDQ decays with the passing of generations. We should consider changes in allele frequencies, estimates of spurious effects and the possibility of emergence of new mutations in the population. The reliability of genomic evaluations is enhanced when the parents of the animals to be evaluated are included in the RP (Weigel et al., 2009).

For these reasons, the RP should be dynamic and remain open to the entry of new animals, thus it is important to maintain or create a good data recording scheme. Collaboration contracts with commercial farms have been suggested to optimize the volume and quality of this information (König., 2010). However, progeny testing schemes will continue to be of great importance in dairy cattle to reach high reliability, at least for the next years.

Genotyping Strategies

Most countries only have sires genotyped as RP (Loberg and Dürr, 2009). They are a good and easy representation of the genetic structure of the population, and achieve high reliabilities due to the large amount of information generated by their daughters. They are important for the AI centers and spread most of the genetic improvement. Many programs have a limited number of highly reliable sires that avoids good genomic predictions (VanRaden et al., 2009a). The efficiency of the program could improve considering alternative selective genotyping. The inclusion of the most informative females should be evaluated for this purpose (Sen et al., 2009; Spangler et al., 2008).

Currently, the best females (bull dams candidates) have been genotyped in some countries (Loberg and Dürr, 2009), and some genomic evaluation systems (e.g., USA or Australia) now include cows in their training population. (Pryce et al., 2012; Wiggans et al., 2011) However, preferential treatment of particular cows in the genomic predictions via their performance records has to be somehow corrected.

Nieuwhof et al, (2012) showed that the inclusion of cows in a RP produced slight differences regarding bulls in terms of correlations with daughter performance. The main advantage of including sires and dams in the reference set was an improvement in regression coefficients for many traits, compared with both the PI and the genomic breeding values (**DGV**) from a

population composed only by sires. While it was expected that bias might increase with the inclusion of cows in the reference set. Including cows with good quality records in the reference set resulted in better selection decisions for some traits.

In a few years, several populations will have more cows than bulls genotyped. It can be hypothesized that the genotypes of large female populations are an alternative to improve the accuracy of the genomic evaluations. Females represent the larger portion within the Holsteins population and most of the traits of economic interest are measured in them. The association between phenotype and genotype of the same animal should be greater than the association between a genotype and the averaged phenotypes of their progeny. A massive genotyping of females would allow capturing more genomic associations between markers and phenotypes. This genotyping could be combined with other strategies such as the genotyping at different densities and imputation of missing SNPs. Those designs allows to increase the size of the RP at a low cost (Habier et al., 2009; Weigel et al., 2009, 2010b).

Methods applied to genomic prediction

The knowledge of the genome of an animal brings a new and complementary source of information to that previously available for selection. In such a situation, information is obtained for a large number of markers. However, only few thousand of individuals are genotyped leading to the curse of dimensionality problem also known as the large p small n problem. This scenario generates an over-parameterization in traditional methods. Therefore it is necessary to develop, adapt and implement new methods in the genome-enhance evaluations.

Different approaches are currently used for estimating genomic values, and it is important to assess the performance of diverse methodologies and identify

the methods that provide the greatest predictive ability in a given population. Genomic prediction methods can be categorized as: 1) methods that regress phenotypic records on SNP markers directly, and 2) methods that compute genomic values as a function of the genomic relationship using a (co)variance structure between subjects (De los Campos et al., 2009).

Methods based on marker regression, approximate genomic values as a linear regression of phenotypes on marker genotype codes as

$$\mathbf{y} = 1'\mu + \mathbf{X}\beta + \mathbf{e}$$

Where y is a vector of dependent variables, 1 is a vector of ones, μ is the population mean or intercept, **X** is a $n \times p$ matrix of codes (e.g., -1, 0, 1 for aa, aA, and AA genotypes, respectively) of n samples and p markers, **β** is a vector of allelic substitution effects for each marker, and **e** is a vector of residuals.

As said previously, GS is carried out in $n << p$ scenarios. Methods based on marker regression need to introduce some shrinkage on the estimation of marker effects. On the contrary, methods based on genomic matrix do not suffer of large p small n problems, as the amount of unknown effects is generally not larger than in the traditional BLUP models (González-Recio et al., 2008).

Below is an overview on the methods that have been proposed for genomic-enhanced evaluations.

Least squares

This approach does not make assumptions about the distribution of the markers as their effects are treated as fixed. It simply deals with more parameters to estimate than available data. Therefore, SNPs pre-selection through ordinary least squares regressions is commonly applied prior to the analyses. Those SNPs with larger effect are selected, assuming that the others do not have any effect on the considered trait (Meuwissen et al., 2001). This methodology applied to genomic selection showed unsatisfactory results (Goddard and Hayes, 2007).

BLUP (Ridge Regression)

Here, a normal prior $N(0, \mathbf{I}\sigma_\beta^2)$ is assigned to the marker effects (β). It must be noted this is not the traditional BLUP described by Henderson (1975). Usually, it is considered that the shrinkage on markers effect is homogeneous; however, this shrinkage is allelic frequency dependant with less shrinkage on markers that have intermediate allelic frequency (Gianola, personal communication). Regarding prediction ability, these methods do not fit well for those cases where genes with large effect are involved, such in the case of DGAT1 in content of fat in milk (Meuwissen et al., 2001).

Bayesian Alphabet

Bayes A

This model proposes Bayesian regressions on the genomic markers. It was originally proposed by Meuwissen et al. (2001). Bayes A assumes a normal prior distribution on the SNPs effects, with zero mean and variance σ_j^2 associated to each marker. This variance is assumed to be distributed as a scaled inversed Chi-squared with 4.012 degrees of freedom and scale parameter 0.002. The choice of these hyperparameters fitted the simulations

used by the authors, but they have been extended to many cases where these values have not been justified. Furthermore, these hyperparameters do not allow Bayesian learning, as evidenced in Gianola et al. (2009). Contrary to what argued initially, this method does not assume different variances for each SNPs, because the prior distribution for the variance of the effects is the same for all markers. The shrinkage on the marker coefficient estimate depends on the estimated marker effect and the allelic frequency of such marker (Gianola, personal communication). Bayes A provides accurate predictions, although it seldom outperforms the methods described below.

Bayes B

This statistical approach was also described in the study of Meuwissen et al. (2001). Bayes B is likely the most accepted model, despite the flaws on its formulation. Bayes B assumes a normal prior distribution on the SNPs effects with zero mean and variance σ_j^2. Then, a mixture of distributions is assumed on this variance being equal to zero with probability π and distributed as in Bayes A with probability $1-\pi$. This formulation is ill-posed from a Bayesian point of view, as assuming a zero variance implies absence of uncertainty about the marker effect, and therefore the inference lacks Bayesian sense. Furthermore, the election of π is arbitrary with no justification and the choice of the hyperparameters in the inversed chi-squared distribution suffer the same drawbacks as in Bayes A. However, Bayes B is one of the most used methods and provides high accurate predictions, especially for those traits regulated by large effect genes as fat percentage. In the original article of Meuwissen et al. (2001), averaged accuracies (5 replicates) between predicted and simulated values resulted 0.318 ±0.018 for least squares, 0.732 ±0.030 for BLUP, 0.798 for Bayes A (1 replicate) and 0.848 ±0.012 for Bayes B.

Bayes C

Bayes C was proposed to amend some of the defects of Bayes B, as the estimation of the probability π or the distribution of mixtures, which in Bayes C is applied on the SNPs effects instead of the variances. In a comparison using simulated data, Bayesian BLUP, Bayes A, Bayes B and Bayes C achieved similar predictive ability and over 0.85 in terms of Pearson correlation (Verbyla et al., 2010).

Bayes SSVS

Verbyla et al. (2009) proposed Bayes-SSVS, adding stochastic search variable selection for the selection of SNP included in the model of prediction and those set at zero variance. The main advantage of the method is the reduction in computational time when compared with the original Bayes B algorithm.

Bayes Cπ & Dπ

Habier et al. (2011), described the Bayes Cπ and Bayes Dπ methods. To address the drawback of BayesA and BayesB regarding the impact of prior hyper-parameters exposed by Gianola et al. (2009) and the prior probability on π. The former applies a single variance common to all SNPs instead of locus specific variance for those 1-π non-zero markers. The second proposes a prior on the scale parameter of the marker effect variance, which follows an inverse chi-square prior. In addition, the proportion π of SNP is also considered unknown and thereby estimated from data. However, accuracies of these alternative bayesian methods were similar than original methods. None of them were preferred when they were compared in terms of prediction accuracy. Bayes Cπ was competitive in terms of computational time as its Gibbs algorithm is faster than the Metropolis-Hastings algorithm of the other methods.

Bayesian LASSO

The Bayesian counterpart of the LASSO method (Park and Casella, 2008) has been proposed for its implementation in genomic selection. This methodology considers a Laplace (double exponential, DE) prior distribution on the markers effects. This method performs larger shrinkage on the marker coefficients estimates through zero than methods such as BLUP or Bayes A. A large number of markers are estimated with a very small effect, almost null, while a small proportion of marker effects are allowed to have large effects. This produces an effect similar to the pre-selection of covariates (De Los Campos et al., 2010).

The Bayesian LASSO depends on a shrinkage parameter over the distribution of the marker effects. Several alternatives have been proposed for the estimation of this parameter. Bayesian estimation is perhaps preferred for the philosophy of the method. Legarra et al. (2010) proposed a modification of the method considering two different variances, one for the conditional distribution of SNPs effects, and another for the residuals. Note that, SNP effect posterior distribution was conditional on the residual variances in the original version.

Bayesian LASSO has been widely applied in the genomic evaluations. Usai et al. (2009) reported better results using Bayesian LASSO than G-BLUP and Bayes-A. They concluded that it provides accurate predictions, especially for low density genotyping. Cleveland et al. (2010) reported similar results when compared Bayesian LASSO and two variants of Bayes A, but they found better predictions using Bayesian LASSO for those traits regulated by a larger number of genes with a small effect.

Bayes R

Bayes R has been recently proposed. It uses a mixture of four zero-mean normal distributions as prior distribution on the SNP effects. The first

distribution assumes zero variance zero effect, and the last one with approximately 1% of the total genetic variance. The prior of the proportions of SNP in each distribution was the Dirichlet distribution. This method also includes a polygenic effect estimated using the average relationship matrix. As in Bayes B, this method assumes no uncertainty for those SNP assigned in the zero mean zero variance class, and the total genetic variance is assumed known without uncertainty.

Elastic Net and SNP pre-selection

The elastic net algorithm, as implemented for genomic selection by Croiseau et al. (2011), corresponds to a combination of the ridge regression BLUP and Bayesian LASSO, with an additional parameter α, taking a value in [0, 1], to weight the RR and LASSO penalties. With a=1, a LASSO model is defined, whereas with a=0 a full ridge regression model is chosen. The objective of this method is to provide a more flexible tool to deal with n << p scenarios. It has resulted in encouraging results especially for small populations (Croiseau et al., 2011; Sánchez et al. 2011). The authors also include SNP pre-selection that can be implemented before carrying out a genomic evaluation. Markers included in evaluation were selected following QTL detection procedures using a combined linkage disequilibrium and linkage analysis (LDLA) (Druet et al., 2008; Meuwissen and Goddard, 2001). From this LDLA, a value of the likelihood ratio test (**LRT**) was obtained for each haplotypes. Then, the 50 SNPs around each detected LRT peak (±25) were included in a pre-selected set of SNPs used for genomic evaluation. This marker pre-selection did not clearly improve original methods in terms of prediction accuracy but reduced the computation time of marker regression algorithms.

G-BLUP

This method is similar to the traditional BLUP evaluations described by Henderson (1975). However, it uses a genomic relationship matrix built from molecular information instead of traditional pedigree relationship matrix. Those individuals sharing identical by state genotype for a larger number of markers are expected to be genetically more similar and will have larger values in the corresponding cells of the matrix.

This method has gained acceptance by the scientific community and is used in the official evaluations of several countries such as U.S.A or Canada. First evaluations performed with real data showed reliabilities for the combined trait Net Merit of 63% compared to 32% resulted from pedigree index (VanRaden, 2008). Luan et al. (2009) founded higher accuracies using G-BLUP than using Bayes B. However, Mrode et al. (2010) in a comparison between two G-BLUP, two Bayes-A and two Bayes-B reported similar results for the considered methods with some variations depending on the evaluated trait.

Single-Step Genomic Selection

Most countries combine genomic-enhanced breeding values obtained from genomic models with traditional proofs (Hayes et al., 2009b; VanRaden et al., 2009b). However, there is not consensus on what is the best approach for blending these predictions that are obtained from different sources of information, different animals, and different model assumptions. To address this problem, Misztal et al. (2009) proposed an evaluation where pedigree relationship is reinforced with contributions from the genomic relationship matrix. This procedure is expected to improve the evaluation of not genotyped animals. This method was tested using the combined morphological trait Final Score of U.S.A Holstein records by Aguilar et al. (2010). Coefficient of determination resulted 24 % for PI, 40% for a G-

BLUP combined with PI in a multiple step procedures and ranged from 37% to 41% for six different single-step approaches.

This methodology avoid the *ad-hoc* combination of genomic and traditional predictions, however is necessary to include a weighting parameter between the genomic and pedigree relationship matrix without any theoretical justification. Parameter estimates may be biased if the genomic relationship coefficients are in a different scale than pedigree-based coefficients. Forni et al. (2011) suggested re-scaling the genomic relationship matrix using the observed allele frequencies to obtain average diagonal elements of 1. Vitezica et al. (2011) concluded that Single Step resulted less biased than multiple step approach. Recent results of Nordic Red cattle (Su et al., 2012b) showed that this method outperformed PI accuracies for the whole data set of animals included in the evaluation (genotyped and non genotyped). Slightly greater accuracy was also reported compared with DGV using G-BLUP and blended genomic values (**GEBV**) (2.2 % and 1.3 % respectively).

The "two-step" approach is undertaken in most dairy cattle populations, although research on a single-step approach for genomic predictions is in an advanced stage of research in some countries as new Zealand (Harris et al., 2012).

Machine Learning algorithms

These methodologies have emerged recently, and aim to optimize predictive ability in a set of data without necessity of adjusting a specific pattern of inheritance Many algorithms have been developed in the machine learning field (Long et al., 2007). Some of them are discussed below.

*Reproducing Kernel Hilbert Spaces Regression (**RKHS**)*

Gianola et al. (2006), proposed a semi-parametric methods in the genomic evaluations as an alternative to SNPs regressions. These methods are more

attractive for multiple and complex interactions that may exist in the biological and metabolic systems. Those methods use to shown better predictive ability than those previously described. The results obtained so far show that they are not worse than the Bayesian regression and in many cases over-performed them in predictive ability (González-Recio et al., 2009; Konstantinov and Hayes, 2010; Long et al., 2010).

For example, in a study about feed conversion rate in broilers, correlations between observed and predicted phenotypes are similar to those obtained with Bayes A (0.27), while correlation with the PI is 0.11 (González-Recio et al., 2009). However, the correlation increases if a pre-selection of SNPs was carried out and RKHS was subsequently implemented. Several authors showed statistical details of the theory underlying these methods (Gianola and de los Campos, 2008; Gianola and Kaam, 2008; Wahba, 1999). The main disadvantage of these methods is the necessity of tuning internal parameters, and the fact that interpretation of results does not respond to a traditional genetic model.

Random Forest

The Random Forest (**RF**) algorithm builds classification or regression trees from genotypes and phenotypes of individuals using randomization of the sample. It considers all markers but also their possible interactions, environmental factors and even interactions between them. Those methods present a predictive ability equal or better than other parametric methods (González-Recio and Forni, 2010). The RF algorithm offers the possibility of capturing the effects of a large number of interactions gene-gene and gene-environment (Sun, 2010). This should be a major advantage in the study of complex diseases although it has been seldom uses in genome-assisted evaluations.

Support Vector Machine

Support Vector Machine (**SVP**) are supervised learning models with associated learning algorithms that analyze data and recognize patterns used for classification and regression analysis. These SVPs perform robust regression for quantitative response. This method exploits the relationships between observations through arraying predictors in observation space using a set of inner products. It can be considered as a specific learning algorithm within the general reproducing kernel Hilbert spaces (RKHS) regression. In a study by Moser et al. (2009), SVR gave the highest accuracy when compared with RR-BLUP or Bayesian regression.

Neural Networks

Those machine learning algorithms can accommodate complex genotype-phenotype relationships including additivity but also dominant or epistatic effects. Bayesian Radial Basis functions models, as described by Long et al. (2010), outperformed Bayes-A when different epistasis and dominant scenarios were simulated. Similarly, predictive ability was improved using Neural Networks on dairy cows and wheat genomic data regarding models that used using only pedigree. Gianola et al. (2011) concluded that Neural Networks may be useful for predicting complex traits using high-dimensional genomic information, where the number of unknowns exceeds sample size. Neural Networks can capture non-linear dependencies in an adaptive manner. This may be useful for prediction of phenotypes.

Boosting

This algorithm has shown competitive behavior in prediction studies in multiple domains. In a multi-specie study (dairy cattle and broilers), Pearson correlations between predicted and observed responses for productive life were 0.65, 0.53, 0.66, y 0.63 using two Boosting approaches, Bayesian-LASSO and Bayes-A respectively. In the broiler example, outcomes for

those methods were 0.33, 0.37, 0.26 and 0.27 respectively, showing a notable advantage of the Boosting algorithm over Bayesian models. Based on these results, machine learning algorithms are a suitable alternative to other methods used for genomic evaluations at the expense of a lower interpretability of results (González-Recio et al., 2010). In a comparison between the three methods, Ogutu et al. (2011) concluded that Pearson correlation was greater for boosting, intermediate for SVMs and lowest for RF but differed little among the three methods.

It must be noted that results obtained with different methods depend on many factors including genetic architecture of the trait, the RP size, the dependant variable used or marker density among others (Calus, 2010). It is necessary to have enough information about the real performance of the methods in order to decide which is the most suitable for each case. It seems inappropriate to give a single recommendation.

Several methods have been proposed but no one shows clear advantages over the others in terms of prediction ability and almost every country is following their own developments. Some convergence should be expected in the future, if any methodology over-performs the others.

Implementation of genomic selection in dairy cattle

Dependent variable

Accurate phenotyping is still the main pre-requisite for a successful breeding program, and is even more important within a genomic context. The reduction of testing schemes and phenotyping data collection could squander any potential advantage of GS.

It is also possible to use Predictive Transmitting Ability (**PTA**) as dependent variable in the genomic evaluations. PTA is the predicted genetic merit that an animal transmits to his offspring for a given trait, including information

coming from its relatives. However it is convenient to use daughter yield deviation (**DYD**) as dependent variable to avoid that information from relatives influence the genomic breeding value of a given animal. The DYD for sires is the average of the phenotypes of the offspring adjusted by the genetic value of the mating and the environmental effects. In the extreme case, phenotypes can be used as response variable, but this has not yet been studied, and further research is necessary on this context. There is not an agreement about the most suitable response variable and depends on the problem to be solved or the data available. For instance, in dairy cattle each country uses different measures: DYD, de-regressed national and international EBV or phenotypic records as the case of Israel (Loberg and Dürr, 2009). Currently, most of the genomic evaluations include bulls from other countries in the RP. Genotypes of foreign bulls are only useful if they have available phenotypic information. In these cases, the only source of information is the sire's deregressed multi across country evaluations (**MACE**) EBV expressed on each national scale (Liu, 2011).

Dependant variable used for most genomic evaluations of dairy cattle is still a traditional breeding value or another kind of pseudo-phenotype. To achieve the maximum benefits of GS real phenotypes should be the desired dependent variable in the future. Prediction of future phenotypes should be the goal of GS.

Chips

Chips that contain SNPs specifically chosen for their large effects, gave high accuracy of genomic breeding values (De los Campos et al., 2009). However, designed chips based on evenly spaced SNPs along the genome, produce more reliable predictions (Kong et al., 2008), and they may be used for multiple traits evaluations. If the cost of genotyping limit SNP panels to 750 loci or fewer, assays based on selected SNP with large estimated effects

on the trait may be preferred (Weigel et al., 2010a). The Bovine SNP50 Chip (Illumina, San Diego, CA, USA) genotyping assay with 54001 SNPs (Matukumalli et al., 2009) has been, so far, the preferred and most commonly used in dairy cattle. It is usually referred as the 50K density chip, and was replaced by a second version enhanced to 54609 markers in 2011.

Low density assays include the Golden Gate Illumina Bovine 3K with 2,900 SNPs. This technology was recently re-designed to improve imputation accuracy in multiple breeds (Boichard et al., 2012), and has been replaced by the Illumina BovineLD BeadChip using 6909 SNP's. Recently, a customized Genseek Genomic Profiler assay is becoming popular. This chip includes 10K SNPs including those in the BovineLD, some for parentage verification and some markers related to diseases. The Affymetrix MegAllele GeneChip Bovine Mapping offers similar density and has been also used in some genotyping process (Sargolzaei et al., 2008). Ultra High density assay chips include 777,962 SNP's for the Illumina BovineHD BeadChip or 640,000 for the Affymetrix Axiom Genome-Wide BOS 1 Array. For most of those purposes, high density platforms (>500K) offers new expectations. The use of those "ultra-high density" assays provide larger linkage disequilibrium between SNPs and QTLs, and therefore, higher reliability of the estimations (VanRaden et al., 2013). This increase may be of particular relevance in situations in which using the current chips cannot obtain sufficient accuracy.

In addition, other customized SNP arrays are currently used within some genomic programs, as the CRV 60,000-marker chip (De Roos et al., 2009).

International collaboration

In addition to the aforementioned joint RPs, a major international consortium has now been established to pool records for dry matter intake, and feed efficiency from Ireland, Australia, U.S.A. , The Netherlands, U.K. and Germany (De Haas et al., 2012; Veerkamp et al., 2012). Similar strategies

are expected in the near future. A similar approach was chosen by the most important Brown Swiss populations (Austria, France, Germany, Slovenia, Switzerland, and the United States), implementing a common estimation framework known as Intergenomics (Santus, 2011).

An important issue for the development of genomic selection is the management of property rights on the information required for carrying genomic evaluations out (Genotypes, phenotypes and pedigrees). This a key point for marketing strategies and exchange of genotypes. Each country has different policies regarding the ownership of genomic data. For instances, genotypes belong to private companies, to farmers, or to different organizations within the dairy cattle market as research centers, AI companies, breeder associations, herd books, research projects or different govern departments (Loberg and Dürr, 2009).

Blending traditional and genomic information

Different sources of information have to be taken into account for the publication of genetic merit. Dairy cattle programs are overwhelmed with genetic evaluations for several production traits (kg of milk, fat and protein, percentages of fat and protein, lactoglobulines…), functional traits (fertility, somatic cell count, longevity…), and more than 16 linear type traits, plus their respective genome-enhanced breeding values.

Researchers and industry have tried to provide a blended genomic evaluation combining DGV and traditional proofs in different manners regarding the country. Currently, there are phenotyped and genotyped animals but also phenotyped but non-genotyped, genotyped but non-phenotyped and animals without any of this information. A whole joint evaluation is becoming another challenge nowadays.

Currently, the most common option is running traditional evaluations and genomic enhanced separately and then combine both results (Hayes et al., 2009b; VanRaden, 2008). Usually, pseudo-data (DYD or breeding values) are used as dependant variable for sire genomic evaluation which do not have own phenotypes for the traits of interest, but do have reliable progeny proofs. The blended genomic value contains information from DGV and traditional proofs. Both values were initially weighted according to the estimated reliability of each (VanRaden et al., 2009b), although some modifications have been proposed recently.

DGV and traditional breeding values are still two different sources of information and the way to blend them is not completely clear yet. In addition, reliabilities of genomic values and therefore reliabilities of blending values have not a standard way of measurement. Even within a GS program, way of estimate reliability depends on animal amount of information. For bulls with daughter information, VanRaden et al. (2009) proposed a selection index for the predictor bulls that included: 1) direct genomic prediction; 2) subset PTA; and 3) published PTA. where subset PTA refers to a traditional genetic evaluation considering only genotyped animals. However, the computation of weights based on reliabilities of the three sources of information is not clearly justified.

Blending DGV and PI could mask those individuals with large differences between both sources of information. These cases are not evident for the breeder. In practice, genomics is used as a source of additional information to the traditional evaluations. Young candidates with different traditional and genomic breeding values should be addressed based on the knowledge of both evaluations independently.

To obtain reliabilities of blended genomic values, the R^2 from PI and from the nonlinear model were divided by mean reliability of daughter deviations.

Then, the difference between the published and observed PI reliability was added to the adjusted genomic R^2 to obtain the realized genomic reliability. Following the same approach for blending genomic values, Su et al. (2012b), computed expected reliabilities as the weighted average of the original reliabilities, using the weights previously computed for genomic values.

Some methods include a polygenic effect in the genomic prediction model instead of posterior blending predictions. Incorporating pedigree information doesn't improve prediction accuracy if genotyping is dense enough (Calus and Veerkamp, 2007). Genomic evaluations are more consistent but decreases the correlation between DGV and EBV of sires in the RP and prediction accuracy is reduced as the polygenic variance increases (Liu et al., 2011). However, proportion of variance explained by markers is trait dependant (Jensen et al., 2012). They conclude that for all traits analyzed more than 92% of all additive genetic variance could be explained using 44K SNP markers. Also, that further increases in marker density will have limited effects on predictive accuracy, unless better methods are used to distinguishing between markers with real effects and markers with no effect. With full sequencing for a substantial number of animals, SNP that are the causative mutation or are closely linked to it may be identified. Identification of those SNP may enable an increase in evaluation accuracy and a decreased number of SNP needed for evaluation.

Results of this thesis are based on DGV using different evaluation methodologies. However, the end products in actual implementations of GS are GEBV blending DGV with traditional EBV. The aim of the thesis is the study and comparison of genotypes, and methods. Using DGVs is justified because they are less influenced by other sources of information.

Genomic selection across populations

Combining data sets from different populations has been proposed as a way to increases accuracy for small populations. First simulation studies showed that reliability improvement in joint populations depends on marker density and genetic distance between populations (Roos et al., 2009). However, some results with real data showed that no improvement was found when genomic matrix was used and only slightly improvement when Bayesian regression were used as the method of evaluation (Hayes et al., 2009a; Pryce et al., 2011). In a similar study, Erbe et al. (2012a), found greater accuracies for the smaller population in an across-breed evaluation when the method of evaluation was Bayes-R, and the SNP's were a subset of the HD array including only those 58,532 SNP's in the transcripted part of the bovine genome. However, no improvement was found for the larger population. Accounting for breed-specific SNP allele effects as suggested by Ibáñez-Escriche et al., (2009) is an alternative to increase DGV reliability, however it was not clearly demonstrated with real data using G-BLUP (Makgahlela et al., 2012).

International evaluations

The traditional genetic evaluations from each country are combined in an international evaluation, carried out by Interbull throughout the MACE procedure. However, the complexity of genomic evaluation, the different methods used in the respective national genomic evaluations and the different dependent variable between countries, limit the implementation of such a methodology on genome-enhanced breeding values. Greater efforts in research are required to respond to this problem. Sullivan and VanRaden (2009) proposed the G-MACE method that can deal with genomic data and is no longer based on independency of data sets across countries, as far as the group of involved countries is sharing data and genotypes to get better

predictions. The first official results for international genomic validations (GEBV Test) were published in the Interbull website in August 2010 for protein yield. Data came from Canada, Denmark, Sweden, Finland, France, Germany, Poland, New Zealand, Netherlands and United States. The European commission has accepted genomic evaluations from those countries validates through "GEBV test" as valid procedures within the European Union. First Spanish genomic evaluations using the entire Eurogenomics RP passed the GEBV test in May 2012.

Future developments

New Traits

GS offers additional benefits in those traits that are difficult to measure (e.g. disease resistance, feed efficiency or meat quality), those traits of low heritability (e.g. Related to fertility), sex linked (e.g. Milk production), expressed at late ages (e.g. Longevity) or even those measured after slaughtering (e.g. Carcass traits). An appropriate choice of individuals to be phenotyped and genotyped might be favor the implementation of GS for those traits. Genomic values can be provided for the rest of the genotyped animals, when marker effects are estimated in a correctly phenotyped population. The phenotype of the candidates or their closest relatives is no longer required to provide accurate predictions (Dekkers, 2010). However, genomic selection reliability is expected to be greater for animals related with those individuals used as RP (Pérez-Cabal et al., 2012). Phenotypic information of an indicator trait genetically correlated with those new traits and recorded on a large scale can be integrated in the genomic evaluation model to improve the accuracy of predictions for those traits (Calus and Veerkamp, 2007)

Gene introgression

New genes introgression could be managed in a more efficiently way using marker information (Amador et al., 2012). These techniques might be of interest for further progress on the disease resistance, adaptation to hard environmental conditions, quality productions or increments on productive efficiency (Odegard et al., 2009).

Sources of genetic variability

Most of the methods applied in genomic selection exploit additive markers variance (Gianola et al., 2009). However, other sources of variability as dominant or epistatic effects included in models of whole-genome evaluation could increases the accuracy of predictions (Toro and Varona, 2010).

Full sequencing for a substantial number of animals, should provide SNP that are the causative mutation or are closely linked to it. Identification of those SNP may enable an increase in evaluation accuracy and a decreased number of SNP needed for evaluation (Wiggans et al., 2011). Simulation studies shown that, current methods used in genomic selection could be not able to identify recent mutations affecting traits of interest (Casellas and Varona, 2011). In fact, those SNP's with low minor allele frequency (**MAF**) are usually removed during the quality control process before genomic evaluations.

In addition, copy number variations (**CNV**s), which represent a significant source of genetic diversity in mammals, have been shown to be associated with phenotypes. Cattle genome is copy number variable within and between breeds (Fadista et al., 2010).Other structural variants or signals that are identified through SNP, as epigenetic effects, may play also an important role that is not included yet in current evaluations of breeding values

(González-Recio, 2012). Taking into account those sources of variation could increase reliability of breeding values in the future.

Imputation

Despite the improvement in reliability of young selected candidates, genomic selection may be economically unfeasible in commercial farms due to an unaffordable chip price. Genotyping was initially restricted to males and elite females in most dairy cattle population. So, a key point in a genomic selection program is the optimization of genomic information in breeding programs (Pryce and Daetwyler, 2012). Low density SNPs panels have been developed with the objective to reduce genotyping costs. Those less expensive low density genotyping platforms have increased the number of genotyped animals. However, performance of low density panels, in terms of predictive ability, is not competitive for most cases. Imputation methods have been developed to solve that problem. Accurate genotype imputations of predictions about those SNP not included in the low density assay may be obtained using high density genotypes as reference. Imputation methods combine a "reference panel" of individuals genotyped at a dense set of polymorphic sites (usually SNP's) with a sample from a genetically similar population genotyped at a subset of sites out of the dense set of polymorphisms (Howie et al., 2009). Imputation utilizes on the linkage disequilibrium between SNPs in the high density panel with the premises that SNPs with large linkage disequilibrium are inherited jointly.

One of the first steps in the imputation process is phasing haplotypes. Usually, genotypes does not provide phase of the haplotypes. In a biallelic locus, phase is unknown for heterozygote individuals; considering a couple of SNP's undefined animals are double heterozygous. Older phasing methods often use linkage information (Sobel and Lange, 1996), and provide the most probable phase between SNPs according to haplotype frequencies.

Phasing methods that solely rely on LDQ tend to mistakenly introduce recombinations when applied to genotypes covering long genetic distances (Kong et al., 2008). For such a reason, some methods introduce family relationships. Family based algorithms could increase imputation accuracy (e.g., Albers et al., 2007; Ding et al., 2007).

Imputation could be useful for incorporating animals genotyped at low density into the genetic evaluations; e.g commercial females. The inclusion of some of these animals could increase the size of the RP for SNP effects estimation in low h^2 traits, preselection of progeny testing candidates or genomic mating design. In addition, these chips with lower prices would allow implementing genomic selection in species or breeds in which current cost is not affordable, for instances, breeds with a reduced individual value, like poultry, sheep or swine. Combination of the information from different SNPs platforms is also available after imputation process (Druet et al., 2010).

Genotyping a large RP at extra large high density could be cost prohibitive. Therefore, it would be possible to genotype a subset of the RP and then impute the remaining genotypes. The predictive accuracy of a posterior genomic evaluation should be checked to ensure that it out-performs the results obtained before the imputation.

Reference population for imputation

Currently, different density platforms are marketed. Those genomic programs including genotyping strategies need a RP for imputing missing SNPs. The RP must include a representative sample of the genetic background of the whole population, with similar allelic frequencies as the population to be imputed (Hao et al., 2009). Phenotypic information of reference animals is not needed for imputation. The imputation accuracy is function of the relatedness between animals in the RP and those to be

imputed (Meuwissen and Goddard, 2010). For instance calves genotyped at high density are a good source of information in order to impute their dams or sisters genotyped at lower density. However*, a priori,* the bulls highly represented in the population or animals from most common matings (sire x maternal grand sire) would be good candidates to be genotyped as RP. The RP should also be large. The larger the population size, the larger the imputation accuracy. Animals in the RP should be genotyped with the higher density SNP platform possible (Weigel et al., 2010c). In a genomic program integrating different SNPs panels an additional RP for imputation purposes should be considered.

Methods, imputation accuracy and reliability of imputed genotypes

Several methods have been developed for imputation, and different software is currently available (Kong et al., 2008). Imputation methods could be compared by the error rate of imputation, which is the percentage of SNPs incorrectly imputed. In a comparison study of different software carried out by (Biernacka et al., 2009) *Mach* (Li and Abecasis, 2006) and *Impute* (Marchini et al., 2007) produced lower imputation error than *Plink* (Purcell et al., 2007) and *Fastphase* (Scheet and Stephens, 2006) on a rheumatoid arthritis case-control data set. For all methods, imputation is more reliable for SNP genotypes that are in strong LD and those with lower MAF (Pei et al., 2008). In a similar comparison between *Impute*, *Mach*, *Beagle* (Browning and Browning, 2009) and *Plink*, the latter performed consistently poorer than the other three. Based on those results, Nothnagel et al. (2009) recommended *Mach* or *Beagle* because these programs are more user-friendly and require less memory than *Impute*. Pei et al. (2008) found better results when imputation is carried out by *Mach* and *Impute* instead of *Fastphase*, *Plink* or *Beagle*. While in another study, *Beagle* resulted similar or more accurate than *Fastphase*, *Impute* or *Mach* for SNP imputation from

different assays, and was also competitive in computational efficiency (Howie et al., 2009).

Regarding dairy cattle genotypes, low density genotypes (2-4K) could be accurately imputed to high density genotype (50K) (accuracy above 90%). Weigel et al., (2010b) reported accuracies of 0.869, 0.758, 0.709 or 0.687 using Fastphase1.2 and 0.926, 0.887, 0.758 or 0.7 12 with *Impute* 2.0 when 90, 95, 98, or 99% of BovineSNP50 genotypes were masked in a population of 3,146 North America Jersey cattle.

Recently, other software has been developed for animal breeding program. Those methods are designed to combine population and pedigree haplotyping such as *Findhap.f90* (VanRaden et al., 2011), *FImpute* (Sargolzaei et al., 2011), *AlphaImpute* (Hickey et al., 2012) or *Phasebook* (Druet and Georges, 2010), the latter based on *Beagle*. When those new approaches are included in comparisons, *Beagle* was shown to be about twice as accurate as *Findhap* (Segelke et al., 2012). Johnston et al. (2011) concluded that: *FImpute* was the fastest program and was the most accurate software program for animals with family information, while *Beagle* was the most accurate software for animals with limited family information.

Some of the aforementioned software *Beagle* (v3.3), *Impute* (v2.0), *Fastphase* (v1.4), *AlphaImpute*, *Findhap*(v2), and *FImpute*(v2), were included in and ensemble-based system considering each method as a classifier. *Beagle* and *FImpute* had the greatest accuracy among the six imputation packages, the best imputation accuracies were those that had *Beagle* as first classifier in the proposed ensemble (Sun et al., 2012). A different approach was proposed by Calus et al. (2011) in this case using a multivariate mixed model framework. This new approach over-performed *Fastphase* and *Beagle* when genotyping density was low, but *Beagle* outperformed the other methods at high SNP density.

Once low density genotypes have been imputed to high density, it is possible to estimate the genomic values with similar accuracy than that obtained with high density genotyping (Berry and Kearney, 2011). Accuracy of DGV of young selection candidates could be increased after imputation compared to those form pure low density typed SNP. Weigel et al. (2010a) showed that animals genotyped at low density but with enough phenotypic information, could be included in the RP after imputation to higher density panels. The overall accuracy of SNPs effects estimation was increased.

Low density SNP panels could be designed using those SNP more informative in terms of predictive ability. Young Holstein Bulls genotyped for 300 to 2000 highly selected SNP could provide DGV for lifetime merit with correlations of 0.43 to 0.57 with future PTA, while correlation using the BovineSNP50 Beadchip achieved a correlation of 0.61 (De los Campos et al., 2009). For individual traits platforms with 500 to 1000 selected SNP, where selection was based on largest estimated effects, resulted in correlations of 0.55 to 0.65 with PTA for progeny testing.

However, dairy cattle genetic programs include several traits with different informative SNPs for each of them. Under this scenario, low density platforms designs based on evenly spaced SNPs are preferred to obtain good predictive accuracy across traits. In a Jersey cattle study based on 1446 sires genotyped with a 42556 SNPs, genomic values were estimated showing a correlation of 70,6% with sires PTAs from traditional evaluations. After removing 93% of the SNPs based on equidistant physical location and high minor allele frequency of still typed SNP, (equivalent to 3K chip) and posterior imputation of these SNPs, average correlation with PTAs was 68.5% (Weigel et al., 2010c). A critical issue with imputed genotypes is how to integrate them effectively into the genomic evaluation system. One can use these posterior probabilities directly or pick up the ‘‘best-guess’’ genotype to perform the subsequent evaluation (Pei et al., 2010).

Objectives

The first step in the implementation of genomic selection is to create a RP of genotyped animals. The RP is used to train a statistical model that estimates the effects of each SNP or genomic combinations between phenotypes and SNPs (or combination of SNPs). The estimates obtained from the RP allow the prediction of genomic breeding values for new individuals with the only source of information of their DNA (Dekkers, 2010). The characteristics of these RP, like the size or the animals included, is relevant to increase the accuracy of future predicted DGV (Hayes et al., 2009a; VanRaden et al., 2009b). Most countries only have sires genotyped as RP (Loberg and Dürr, 2009). They are a good representation of the genetic structure of the population, and achieve high reliabilities due to the large amount of information generated by their daughters. However, many programs have a limited number of highly reliable sires that precludes high reliabilities of genomic predictions (VanRaden et al., 2009a). The efficiency of those programs could improve considering international agreements for joining several RP or alternative selective genotyping. The inclusion of the most informative females should be evaluated for this purpose (Spangler et al, 2008; Sen et al., 2009). It can be hypothesized that the genotypes of large female populations are an alternative to improve the accuracy of the genomic evaluations.

After genotyping a RP, different approaches are currently used for estimating genomic values, and it is important to assess the performance of diverse methodologies and identify the methods that provide the greatest predictive ability in a given population. Genomic prediction methods can be categorized as: 1) methods that regress phenotypic records on SNP markers directly, and 2) methods that compute genomic values as a function of the genomic relationship using a (co)variance structure between subjects (De los Campos et al., 2009). Several methods have been developed but no one shows clear advantages over the others in terms of prediction ability and

almost every country is following their owns developments. If some methodology over-performs the others some convergence should be expected in the future. Machine Learning algorithms are an appealing alternative to Bayesian regressions and G-BLUP. These methodologies have emerged recently, and are aim to optimize predictive ability in a data set without adjusting a specific pattern of inheritance. Boosting is one of the machine learning algorithms implemented for genomic selection with great predictive ability (Ogutu et al. 2011). This is also a suitable alternative to other methods used for genomic evaluations (González-Recio et al., 2010).

Despite the improvement in reliability of young selected candidates, genomic selection may be economically unfeasible in commercial farms due to an unaffordable chip price. So, next key point in a genomic selection program was the optimization of genotype density in candidate animals (Pryce and Daetwyler, 2012). Low density SNPs panels have been developed for this purpose. Low density genotyping platforms have increase the number of genotyped animals due to their low prices. However, performance of low density panels in terms of predictive ability is not competitive for most cases (Weigel et al., 2009). Imputation methods have been proposed with the aim to solve that problem. These methods combine a "reference set" of individuals genotyped at a dense panel of polymorphic sites (usually SNP's) with a set from a genetically similar population genotyped at a subset of those sites in the dense panel (Howie et al., 2009).

Several methods have been proposed for imputation, and different software are currently available (Kong et al., 2008). Among them, *Beagle* (Browning and Browning, 2009) is reported as competitive when compared to other approaches (Johnston et al., 2011; Segelke et al., 2012; Sun et al., 2012).

The main objective of this thesis was to contribute with knowledge to the implementation of genomic selection in the Spanish dairy cattle, which occurred in parallel to the development of this Doctoral thesis.

The specific objectives were

1) Creation of a reference population sufficiently informative when the progeny tested sire population is limited.

2) Development of a competitive and reliable genomic evaluation in terms of prediction accuracy, computationally efficient and flexible for further future developments.

3) Implement a flexible and efficient imputation design for different density genotypes.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Albers, C.A., Heskes, T., Kappen, H.J., 2007. Haplotype Inference in General Pedigrees Using the Cluster Variation Method. Genetics 177, 1101–1116.

Amador, C., Toro, M.Á., Fernández, J., 2012. Molecular Markers Allow to Remove Introgressed Genetic Background: A Simulation Study. PLoS ONE 7, e49409.

Banos, G., Coffey, M.P., 2010. Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. Journal of Dairy Science 93, 2775–2778.

Berry, D.P., Kearney, J.F., 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5, 1162–1169.

Biernacka, J., Tang, R., Li, J., McDonnell, S., Rabe, K., Sinnwell, J., Rider, D., Andrade, M. de, Goode, E., Fridley, B., 2009. Assessment of genotype imputation methods. BMC Proceedings 3, S5.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., Van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A., Wiggans, G.R., For the Bovine LD Consortium, 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE 7, e34130.

Browning, B.L., Browning, S.R., 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of

Trios and Unrelated Individuals. The American Journal of Human Genetics 84, 210–223.

Calus, M. p. l., Veerkamp, R. f., 2007. Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. Journal of Animal Breeding and Genetics 124, 362–368.

Calus, M.P.L., 2010. Genomic breeding value prediction: methods and procedures. animal 4, 157–164.

Calus, M.P.L., Meuwissen, T.H.E., De Roos, A.P.W., Veerkamp, R.F., 2008. Accuracy of Genomic Selection Using Different Methods to Define Haplotypes. Genetics 178, 553–561.

Calus, M.P.L., Veerkamp, R.F., Mulder, H.A., 2011. Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. J ANIM SCI 89, 2042–2049.

Casellas, J., Varona, L., 2011. Short communication: Effect of mutation age on genomic predictions. Journal of Dairy Science 94, 4224–4229.

Cleveland, M., Forni, S., Deeb, N., Maltecca, C., 2010. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. BMC Proceedings 4, S6.

Croiseau, P., Legarra, A., Guillaume, F., Fritz, S., Baur, A., Colombani, C., Robert-Granié, C., Boichard, D., Ducrocq, V., 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. Genetics Research 93, 409–417.

Cromie, A.R., Berry, D.P., Wickham, J.F., Kearney, J.F., Pena, J., vanKaam, J.B., Gengler, N., 2012. International Genomic Co-operation; Who, what, when, where, why and how?. Interbull Bulletin 42, 72.

Daetwyler, H. d., Villanueva, B., Bijma, P., Woolliams, J. a., 2007. Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics 124, 369–376.

De Haas, Y., Calus, M.P.L., Veerkamp, R.F., Wall, E., Coffey, M.P., Daetwyler, H.D., Hayes, B.J., Pryce, J.E., 2012. Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. Journal of Dairy Science 95, 6103–6112.

De los Campos, G. de los, Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182, 375–385.

De Los Campos, G., Gianola, D., Rosa, G.J.M., Weigel, K.A., Crossa, J., 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92, 295–308.

De Roos, A.P.W., Schrooten, C., Mullaart, E., Van der Beek, S., De Jong, G., Voskamp, W., 2009. Genomic selection at CRV. Interbull Bulletin 39, 47.

Dekkers, J.C.M., 2010. Animal genomics and genomic selection. Adapting Animal Production to Changes for a Growing Human Population. Presented at the International Conference, Lleida.

Ding, X.D., Simianer, H., Zhang, Q., 2007. A New Method for Haplotype Inference Including Full-Sib Information. Genetics 177, 1929–1940.

Druet, T., Fritz, S., Boussaha, M., Ben-Jemaa, S., Guillaume, F., Derbala, D., Zelenika, D., Lechner, D., Charon, C., Boichard, D., Gut, I.G., Eggen, A., Gautier, M., 2008. Fine Mapping of Quantitative Trait Loci Affecting Female Fertility in Dairy Cattle on BTA03 Using a Dense Single-Nucleotide Polymorphism Map. Genetics 178, 2227–2235.

Druet, T., Georges, M., 2010. A Hidden Markov Model Combining Linkage and Linkage Disequilibrium Information for Haplotype

Reconstruction and Quantitative Trait Locus Fine Mapping. Genetics 184, 789–798.

Druet, T., Schrooten, C., De Roos, A.P.W., 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. Journal of Dairy Science 93, 5443–5454.

Dürr, J., Philipsson, J., 2012. International cooperation: The pathway for cattle genomics. Animal Frontiers 2, 16–21.

Elsik, C.G., Tellam, R.L., Worley, K.C., 2009. The Genome Sequence of Taurine Cattle: A Window to Ruminant Biology and Evolution. Science 324, 522–528.

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., Goddard, M.E., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95, 4114–4129.

Fadista, J., Thomsen, B., Holm, L.-E., Bendixen, C., 2010. Copy number variation in the bovine genome. BMC Genomics 11, 284.

Fernando, R., Grossman, M., 1989. Marker assisted selection using best linear unbiased prediction. Genetics Selection Evolution 21, 1–11.

Fisher, R.A., 1919. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh 52, 399–433.

Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Genetics Selection Evolution 43, 1.

Gianola, D., Campos, G. de los, Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183, 347–363.

Gianola, D., De los CAMPOS, G., 2008. Inferring genetic values for quantitative traits non-parametrically. Genetics Research 90, 525–540.

Gianola, D., Fernando, R.L., Stella, A., 2006. Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics 173, 1761–1776.

Gianola, D., Kaam, J.B.C.H.M. van, 2008. Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics 178, 2289–2303.

Gianola, D., Okut, H., Weigel, K., Rosa, G., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics 12, 87.

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.

Goddard, M. e., Hayes, B. j., 2007. Genomic selection. Journal of Animal Breeding and Genetics 124, 323–330.

González-Recio, O., 2012. Epigenetics: A New Challenge in the Post-Genomic Era of Livestock. Front Genet 2.

GonzáLez-Recio, O., Forni, S., 2010. RanFoG: Random Forest in a java package to analyze disease resistance using genomic information. Presented at the XV Reunión Nacional de Mejora Genética Animal, Vigo (Spain).

González-Recio, O., Gianola, D., Long, N., Weigel, K.A., Rosa, G.J.M., Avendaño, S., 2008. Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers. Genetics 178, 2305–2313.

González-Recio, O., Gianola, D., Rosa, G., Weigel, K., Kranis, A., 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genetics Selection Evolution 41, 3.

González-Recio, O., Weigel, K.A., Gianola, D., Naya, H., Rosa, G.J.M., 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. Genetics Research 92, 227–237.

Habier, D., Fernando, R., Kizilkaya, K., Garrick, D., 2011. Extension of the bayesian alphabet for genomic selection. BMC Bioinformatics 12, 186.

Habier, D., Fernando, R.L., Dekkers, J.C.M., 2009. Genomic Selection Using Low-Density Marker Panels. Genetics 182, 343–353.

Haley, C.S., Visscher, P.M., 1998. Strategies to Utilize Marker-Quantitative Trait Loci Associations. Journal of Dairy Science 81, Supplement 2, 85–97.

Harris, B.L., Winkelman, A.M., Johnson, D.L., 2012. large-scale single-step genomic evaluation for milk production. Interbull Bulletin 46, 20–24.

Hayes, B., Bowman, P., Chamberlain, A., Verbyla, K., Goddard, M., 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41, 51.

Hayes, B.J., 2007. QTL, Mapping, MAS, and genomic selection. short course. Iowa State University (USA).

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009b. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92, 433–443.

Henderson, C.R., 1975. Best linear unbiased estimation and prediction under a selection model. Biometrics 31, 423–447.

Hickey, J.M., Kinghorn, B.P., Tier, B., Van der Werf, J.H., Cleveland, M.A., 2012. A phasing and imputation method for pedigreed populations that results in a single-stage genomic evaluation. Genet Sel Evol 44, 9.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics 38, 226–231.

Howie, B.N., Donnelly, P., Marchini, J., 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5, e1000529.

Ibánẽz-Escriche, N., Fernando, R.L., Toosi, A., Dekkers, J.C., 2009. Genomic selection of purebreds for crossbred performance. Genetics Selection Evolution 41, 12.

Jensen, J., Su, G., Madsen, P., 2012. Partitioning additive genetic variance into genomic and remaining polygenic components for complex traits in dairy cattle. BMC Genetics 13, 44.

Johnston, J., Kistemaker, G., Sullivan, P.G., 2011. Comparison of different imputation methods. Interbull Bulletin 0.

Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D.F., Stefansson, H., Stefansson, K., 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nature Genetics 40, 1068–1075.

Konstantinov, K., Hayes, B., 2010. Comparison of BLUP and Reproducing kernel Hilbert spaces methods for genomic prediction of breeding values in Australian Holstein Friesian cattle., in: Proc. 9th World Cong. Genet. Appl. Livest. Prod. Presented at the 9th World Cong. Genet. Appl. Livest. Prod., Leipzig (Germany).

Lande, R., Thompson, R., 1990. Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124, 743–756.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2010. Lasso Bayesiano Mejorado para selección genómica. Presented at the XV Reunión Nacional de Mejora Genética Animal., Vigo (Spain).

Li, Y., Abecasis, G., 2006. Mach 1.0: rapid haplotype reconstruction and missing genotype inference. Am J Hum Genet S 79, 2290.

Lillehammer, M., Meuwissen, T.H.E., Sonesson, A.K., 2010. Effects of alternative genomic selection breeding schemes on genetic gain in dairy cattle., in: Proc. 9th World Cong. Genet. Appl. Livest. Prod. Presented at the 9th World Cong. Genet. Appl. Livest. Prod., Leipzig (Germany).

Liu, Z., 2011. Use of MACE results as input for genomic models. Interbull Bulletin 0.

Liu, Z., Seefried, F., Reinhardt, F., Rensing, S., Thaller, G., Reents, R., 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genetics Selection Evolution 43, 19.

Loberg, A., Dürr, J.W., 2009. Interbull survey on the use of genomic information. Proceedings of the Interbull technical workshop. Proceedings of the Interbull technical workshop 39, 3–14.

Long, N., Gianola, D., Rosa, G. j. m., Weigel, K. a., Avendaño, S., 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics 124, 377–389.

Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A., Kranis, A., GonzáLez-Recio, O., 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics Research 92, 209–225.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M., Meuwissen, T.H.E., 2009. The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics 183, 1119–1126.

Lund, M., De Roos, A., De Vries, A., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, G., 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43, 1–8.

Makgahlela, M. l., Mäntysaari, E. a., Strandén, I., Koivula, M., Nielsen, U. s., Sillanpää, M. j., Juga, J., 2012. Across breed multi-trait random regression genomic predictions in the Nordic Red dairy cattle. Journal of Animal Breeding and Genetics n/a–n/a.

Marchini, J., Howie, B., Myers, S., McVean, G., Donnelly, P., 2007. A new multipoint method for genome-wide association studies by imputation of genotypes. Nature Genetics 39, 906–913.

Matukumalli, L.K., Lawley, C.T., Schnabel, R.D., Taylor, J.F., Allan, M.F., Heaton, M.P., O'Connell, J., Moore, S.S., Smith, T.P.L., Sonstegard, T.S., Van Tassell, C.P., 2009. Development and Characterization of a High Density SNP Genotyping Assay for Cattle. PLoS ONE 4, e5350.

Meuwissen, T., Goddard, M., 2010. The Use of Family Relationships and Linkage Disequilibrium to Impute Phase and Missing Genotypes in Up to Whole-Genome Sequence Density Genotypic Data. Genetics 185, 1441–1449.

Meuwissen, T.H., Goddard, M.E., 2001. Prediction of identity by descent probabilities from marker-haplotypes. Genetics Selection Evolution 33, 605.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–4655.

Moser, G., Tier, B., Crump, R., Khatkar, M., Raadsma, H., 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution 41, 56.

Mrode, R., Coffey, M.P., Straden, I., Meuwissen, T.H.E., vanKaam, J.B., Kearney, J.F., Berry, D.P., 2010. A comparison of various methods for the computation of genomic breeding values of dairy bulls using software at genomicselection.net, in: Proc. 9th World Cong. Genet. Appl. Livest. Prod. Presented at the 9th World Cong. Genet. Appl. Livest. Prod., Leipzig (Germany), p. Abstract 518.

Nieuwhof, G.J., Beard, K.T., Konstantinov, K.V., Reich, C.M., Mason, B.A., Hayes, B.J., 2012. validation of genomic evaluations in Australian Jersey cattle using a reference set that includes cows, in: AUSTRALASIAN DAIRY SC I ENCE SYMPOSIUM 2012 Proceedings. Presented at the AUSTRALASIAN DAIRY SC I ENCE SYMPOSIUM 2012, Melbourne, Australia, pp. 31–33.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., Franke, A., 2009. A comprehensive evaluation of SNP genotype imputation. Human Genetics 125, 163–171.

Odegard, J., Yazdi, M.H., Sonesson, A.K., Meuwissen, T.H.E., 2009. Incorporating Desirable Genetic Characteristics From an Inferior Into a Superior Population Using Genomic Selection. Genetics 181, 737–745.

Ogutu, J., Piepho, H.-P., Schulz-Streeck, T., 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings 5, S11.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association 103, 681–686.

Pedersen, L.D., Sørensen, A.C., Berg, P., 2010. Marker-assisted selection reduces expected inbreeding but can result in large effects of hitchhiking. J. Anim. Breed. Genet. 127, 189–198.

Pei, Y.-F., Li, J., Zhang, L., Papasian, C.J., Deng, H.-W., 2008. Analyses and Comparison of Accuracy of Different Genotype Imputation Methods. PLoS ONE 3, e3551.

Pei, Y.-F., Zhang, L., Li, J., Deng, H.-W., 2010. Analyses and Comparison of Imputation-Based Association Methods. PLoS ONE 5, e10827.

Pérez-Cabal, M.A., Vazquez, A.I., Gianola, D., Rosa, G.J.M., Weigel, K.A., 2012. Accuracy of Genome-Enabled Prediction in a Dairy Cattle Population using Different Cross-Validation Layouts. Front Genet 3.

Pryce, J.E., Arias, J., Bowman, P.J., Davis, S.R., Macdonald, K.A., Waghorn, G.C., Wales, W.J., Williams, Y.J., Spelman, R.J., Hayes, B.J., 2012. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. Journal of Dairy Science 95, 2108–2119.

Pryce, J.E., Daetwyler, H.D., 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim. Prod. Sci. 52, 107–114.

Pryce, J.E., Gredler, B., Bolormaa, S., Bowman, P.J., Egger-Danner, C., Fuerst, C., Emmerling, R., Sölkner, J., Goddard, M.E., Hayes, B.J., 2011. Short communication: Genomic selection using a multi-breed, across-country reference population. Journal of Dairy Science 94, 2625–2630.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., De Bakker, P.I.W., Daly, M.J., Sham, P.C., 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. The American Journal of Human Genetics 81, 559–575.

Roos, A.P.W. de, Hayes, B.J., Goddard, M.E., 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183, 1545–1553.

Saatchi, M.A., Netaji-Javaremi, A., Miraei-Ashtiani, S.R., Mehrabani-Yeganeh, H., Moradi-Shahrebabak, M., 2010. The impact of using individuals from different generations with different accuracy of BLUP EBVs on accuracy of Genomic EBVs, in: Proc. 9th World

Cong. Genet. Appl. Livest. Prod. Presented at the World Cong. Genet. Appl. Livest. Prod., Leipzig (Germany).

Sánchez, J.P., García-Gámez, E., Gutiérrez-Gil, B., Arranz, J.J., 2011. Preliminary evaluation of genomic selection procedures in the Churra dairy population. Presented at the XIV Jordanas Sobre Produccion Animal, Zaragoza, Espana, 17 y 18 de Mayo de 2011., Asociación Interprofesional para el Desarrollo Agrario, pp. 551–553.

Santus, E., 2011. Intergenomics: business rules and transition into services. Interbull Bulletin 0.

Sargolzaei, M., Chesnais, J.P., Schenkel, F.S., 2011. FImpute—An efficient imputation algorithm for dairy cattle populations. J. Dairy Sci. 94(E-Suppl. 1), 421(Abstr.).

Sargolzaei, M., Schenkel, F.S., Jansen, G.B., Schaeffer, L.R., 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91, 2106–2117.

Schaeffer, L. r., 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123, 218–223.

Scheet, P., Stephens, M., 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics 78, 629–644.

Segelke, D., Chen, J., Liu, Z., Reinhardt, F., Thaller, G., Reents, R., 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. Journal of Dairy Science 95, 5403–5411.

Sen, Ś., Johannes, F., Broman, K.W., 2009. Selective Genotyping and Phenotyping Strategies in a Complex Trait Context. Genetics 181, 1613–1626.

Sobel, E., Lange, K., 1996. Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am J Hum Genet 58, 1323–1337.

Spangler, M.L., Sapp, R.L., Bertrand, J.K., MacNeil, M.D., Rekaya, R., 2008. Different methods of selecting animals for genotyping to maximize the amount of genetic information known in the population. J ANIM SCI 86, 2471–2479.

Su, G., Madsen, P., Nielsen, U.S., Mäntysaari, E.A., Aamand, G.P., Christensen, O.F., Lund, M.S., 2012. Genomic prediction for Nordic Red Cattle using one-step and selection index blending. Journal of Dairy Science 95, 909–917.

Sullivan, P.G., VanRaden, P.M., 2009. Development of Genomic GMACE. Interbull Bulletin 40, 157–161.

Sun, C., Wu, X.-L., Weigel, K.A., Rosa, G.J.M., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F., Gianola, D., 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genetics Research 94, 133–150.

Sun, Y.V., 2010. Multigenic modeling of complex disease by random forests. Adv. Genet. 72, 73–99.

Toro, M.A., Varona, L., 2010. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol 42, 33.

Usai, M.G., Goddard, M.E., Hayes, B.J., 2009. LASSO with cross-validation for genomic selection. Genetics Research 91, 427–436.

VanRaden, P., O'Connell, J., Wiggans, G., Weigel, K., 2011. Genomic evaluations with many more genotypes. Genetics Selection Evolution 43, 1–11.

VanRaden, P., Wiggans, G.R., Van Tassell, C.P., Sonstegard, T.S., Schenkel, F.S., 2009a. Benefits from collaboration in genomics. Interbull Bulletin 40, 67–72.

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., Van Kaam, J.B.C.H.M., Valentini, A., Van Doormaal, B.J., Faust, M.A., Doak, G.A., 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96, 668–678.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., 2009b. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16–24.

Veerkamp, R.F., Coffey, M.P., Berry, D.P., De Haas, Y., Strandberg, E., Bovenhuis, H., Calus, M.P.L., Wall, E., 2012. Genome-wide associations for feed utilisation complex in primiparous Holstein–Friesian dairy cows from experimental research herds in four European countries. animal 6, 1738–1749.

Verbyla, K., Bowman, P., Hayes, B., Goddard, M., 2010. Sensitivity of genomic selection to using different prior distributions. BMC Proceedings 4, S5.

Verbyla, K.L., Hayes, B.J., Bowman, P.J., Goddard, M.E., 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet Res (Camb) 91, 307–311.

Villanueva, B., Pong-Wong, R., Fernández, J., Toro, M.A., 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J ANIM SCI 83, 1747–1752.

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genet Res (Camb) 93, 357–366.

Wahba, G., 1999. Advances in Kernel Methods: Support vector machines, reproducing kernel Hilbert spaces and the randomized GAVC. B. Scholkopf, C. Burges and A. Smola. MIT Press, Cambridge, MA.

Weigel, K.A., De los Campos, G., González-Recio, O., Naya, H., Wu, X.L., Long, N., Rosa, G.J.M., Gianola, D., 2009. Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. Journal of Dairy Science 92, 5248–5257.

Weigel, K.A., De los Campos, G., Vazquez, A.I., Rosa, G.J.M., Gianola, D., Van Tassell, C.P., 2010a. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science 93, 5423–5435.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M., Wiggans, G.R., 2010b. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science 93, 2229–2238.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M., Wiggans, G.R., 2010c. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93, 2229–2238.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Chesnais, J.P., Schenkel, F.S., Van Tassell, C.P., 2008. Genomic evaluations in the United States and Canada: A collaboration, in: Proc. Inte. Comm. Anim. Recording. Presented at the International Commitee Animal Recording, Niagara Falls NY, p. 6.

Wiggans, G.R., VanRaden, P.M., Cooper, T.A., 2011. The genomic evaluation system in the United States: Past, present, future. Journal of Dairy Science 94, 3202–3211.

2

Genotyping strategies for genomic selection in small dairy cattle populations

J.A. Jiménez-Montero

O. González-Recio

R. Alenda

Published in Animal,(2012) 6:1216-1224.

The design of the reference population is fundamental to maximizing the benefits of genomic selection. Currently, most of the genotyped animals are sires; however, the number of sires available in some populations might not be enough to make an appropriate genomic evaluation. This study presents an optimal genotyping design that includes females in the reference population, suggesting that two-tailed strategies are preferable to increase the reliability of genomic selection in small cattle populations.

Abstract

This study evaluated different female-selective genotyping strategies to increase the predictive accuracy of genomic breeding values in populations that have a limited number of sires with a large number of progeny. A simulated dairy population was utilized to address the aims of the study. The following selection strategies were used: random selection, two-tailed selection by yield deviations, two-tailed selection by breeding value, top yield deviation selection, and top breeding value selection. For comparison, two other strategies: genotyping of sires and pedigree indexes from traditional genetic evaluation were included in the analysis. Two scenarios were simulated, low heritability (h^2=0.10) and medium heritability (h^2=0.30). Genomic breeding values were estimated using Bayesian Lasso. The accuracy of predicted genomic breeding values using the two-tailed strategies was better than the accuracy obtained using other strategies (0.50 and 0.63 for the two-tailed by yield deviations strategy and 0.48 and 0.63 for the two-tailed by breeding values strategy in low- and medium-heritability scenarios, respectively, using 1000 genotypes cows). When 996 genotyped bulls were used as the training population, the sire' strategy led to accuracies of 0.48 and 0.55 for low- and medium-heritability traits, respectively. The random strategies required larger training populations to outperform the accuracies of the pedigree index, but selecting females from the top of the yield deviations or breeding values of the population did not improve accuracy relative to that of the pedigree index. Bias was found for all genotyping strategies considered, although the top strategies produced the most biased predictions.

Strategies that involve genotyping cows can be implemented in breeding programs that have a limited number of sires with a reliable progeny test.

The results of this study showed that females that exhibited upper and lower extreme values within the distribution of yield deviations may be included as training population to increase reliability in small reference populations. The strategies that selected only the females that had high estimated breeding values or yield deviations produced suboptimal results.

Introduction

Genomic selection (**GS**) is the most promising tool that has emerged for increasing the genetic gain rate in livestock (Weigel et al., 2010). Genetic evaluations that use genomic information aim to increase the accuracy of breeding value predictions. Genomic evaluations have focused mainly on sire breeding value predictions (**EBV)** that use daughter yield deviations (**DYD**) as the response variable in reference populations because sires have a larger impact on breeding programs than cows, and their DYDs are more accurate than cow phenotypes (Calus, 2009).

In genomically assisted evaluations, a reference population is needed to estimate marker effects that account for linkage disequilibrium between markers and quantitative trait loci (**QTL**). The characteristics of the training population, e.g., its size or the selection of animal genotyped, are important for increasing the accuracy of genomic predictions (Hayes et al., 2009; VanRaden et al., 2009a). There are challenges in reaching sufficiently high predictive accuracy, especially in small populations and particularly for low heritability traits (VanRaden et al., 2009b). In most countries, only sires have been genotyped and included in the reference population (Loberg and Dürr, 2009) because bulls drive the genetic structure of the population and provide high predictive accuracy due to the large amount of information from their daughters' averages. In some countries, however, there are a limited number of sires that have been progeny-tested, and this hampers the accuracy of the predictions in test populations (VanRaden et al., 2009b); thus, alternative strategies are required.

For instances, international collaborations for joining different populations have helped to increase population sizes (Wiggans et al., 2008; Cromie et al., 2010; Lund et al., 2010).

GS can be enhanced using female genotypes because economically important traits are measured in the female population, and cows comprise the largest proportion of the Holstein population. In addition, increasing attention has been directed at recording functional traits, particularly health traits. Female reference populations for genomic selection of those new phenotypes could be feasible (Ducrocq and Santus, 2011). Dominant and epistatic effects can be captured and exploited. The relationship between the genotype and phenotype of a cow is expected to be stronger than the relationship between a sire's genotype and his daughter phenotypes. Currently, in some genomic programs, the best females, which are candidates for bull dams, are being genotyped (Loberg and Dürr, 2009). However, with large numbers of selection candidates in the female population, a pre-selection of genotyped animals is needed to optimize genotyping costs (Blonk et al., 2010). Selective genotyping of the most informative individuals might increase genotyping efficiency (Spangler et al., 2008; Sen et al., 2009). However, there has been very limited research as to which animals are most informative in terms of single nucleotide polymorphism (**SNP**) effects and genomic predictions when females are used in the reference population.

The aim of this study was to evaluate female-selective genotyping strategies using simulation and to increase the predictive accuracy of genomic breeding values (GBVs) in populations that have a limited number of sires with large number of progeny.

Materials and methods

Simulations

Phenotypes and genotypes were simulated to mimic a dairy cattle population based on 996 progeny-tested sires and 40000 dams. These recorded and genotyped animals were used to select different training populations in genomic selection programs.

The simulation was performed with the QMSim software (Sargolzaei and Schenkel, 2009) using the following parameters: 1,000 historical generations were generated to produce a realistic level of linkage disequilibrium (**LD**) similar to that obtained for the currently used 50 K SNP chip.

LD, the non-random association of alleles between two loci, was measured using the r^2 parameter (Hill and Robertson, 1968). LD can be estimated using other measurements, such as D, D´ or different measures based on the chi-squared statistic (Zhao et al., 2005); however, r2 is the most common measure of LD for biallelic markers, and it is less sensitive to the effects of allelic frequency (Sargolzaei et al., 2008).

The first historical population was composed of 1000 females and 400 males. During the 1,000 historical generations, the population size decreased from 1,400 to 400 individuals with the same sex ratio, which mimicked a bottleneck and a decrease in the effective number ($\mathbf{N_e}$) to account for the evolution of the historical Holstein effective population size (Hayes *et al*., 2003; Sorensen *et al*., 2005). Previous simulation studies have used a similar effective number (e.g., Meuwissen et al., 2001). Following the bottleneck, historical population size was extended for 40 added generations. Then, 20,000 females and 300 males from the last historical generation were used as founders. Similar strategies (shrinkage and expansion of the population) have been used in simulations of dairy cattle populations (Habier et al., 2009, De Roos et al., 2009).

From this founder population and based on BLUP EBVs, 15 overlapping generations of selection were simulated as contemporary born animals. The population was under random mating between selected animals, and the average sex ratio was 0.5. During the 15 periods of selection, 51 out of 300 tested sires were selected as proven bulls (17%), and 9,000 out of 20,000 of the dams were culled (45%). Selection and culling criteria were based on EBVs. Individuals from the next offspring replaced culled animals. This overlapping active population was used to mimic a scaled representation of a dairy cattle population having higher selection intensity in males than in females. Individuals from progeny sets 10 through 15 were genotyped and used to create the training and validation sets. Genotyped animals had at least 10 generations of traditional selection and pedigree depth.

The simulated genome consisted of 30 chromosomes (100 cM each), and the recombination rate was adjusted to this distance. With the objective of obtaining a desired LD between adjacent SNPs, 9990 biallelic markers were equally spaced out over the genome. Additive genetic effects were determined by ninety quantitative trait loci (QTL) that were simulated and randomly distributed along the genome. QTL effects were generated based on a gamma distribution with a shape parameter equal to 0.4 (Hayes and Goddard, 2001; Meuwissen et al., 2001). QTL allelic effects were first sampled from the gamma distribution in such a way as to be positive or negative with a probability of 0.5. As expected, most of the QTL had a small effect, but others had a large effect. The mutation rate was fixed at 2.5e-5, and the number of crossovers was sampled from a Poisson distribution with positions randomly distributed. The new variants and the selection process as well as drift and Bulmer effects modified the genetic variance. True breeding values (TBVs) were calculated by summing all QTL effects and were

subsequently scaled to a realized genetic variance of 1. Distributions of the QTL effects of the traits are shown in Figure 2.1.

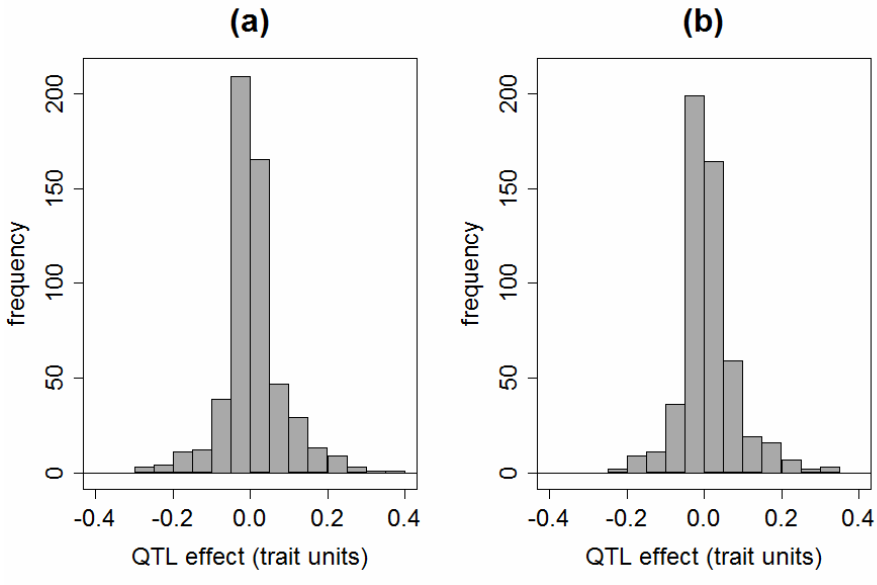

Figure 2.1 Distribution of simulated QTL effects: (a) 0.30 heritability trait scenario and (b) 0.10 heritability trait scenario.

The simulation study included two scenarios in terms of heritability (0.10 and 0.30). For each animal from sets 10 through 15, pedigrees, true breeding values, phenotypes, and genotypes were simulated, and breeding values were estimated. Analyses were performed on 10 replicates (five per trait), and the strategy and sizes of the training sets were designed to be sufficient for the aim of the study.

Selective Genotyping Strategies

Animals from progeny sets 11 through 14 represented a contemporary overlapping active population of 40,000 females. From them, 1000, 2000,

and 5000 females were selected and genotyped as training sets based on the following strategies:

1. At random (**RND**). - Females were randomly selected from the available population (generations 11 through 14).

2. Two-tailed Yield Deviation values (**TTYD**). - An equal number of females were selected from the lowest $\alpha/2$ and the highest $(1-\alpha/2)$ percentiles of the yield deviation distribution (for α=0.025, 0.05, and 0.125).

3. Two-tailed EBVs (**TTBV**). - This subset represented a selection of females that had estimated breeding values in the lowest $\alpha/2$ and highest $(1-\alpha/2)$ percentiles of the distribution (for α=0.025, 0.05, and 0.125).

4. Highest Yield Deviation values (**TopYD**). - In this strategy, cows that had yield deviations in the 1-α percentile (for α=0.025, 0.05, and 0.125) were selected.

5. Highest EBVs (**TopBV**). - Cows that had estimated breeding values in the 1-α percentile (for α=0.025, 0.05, and 0.125) were selected.

Genotyping strategies based on animals selected by their breeding values were included to evaluate the information provided by the pedigree and its effects on the accuracy of the genomic evaluation relative to the true breeding values.

As a reference, all sires (996) from progeny sets 10 through 13 were genotyped (**SiresDYD**). The distribution of family sizes showed values consistent with a dairy population of 40,000 contemporary cows (Figure 2.2).

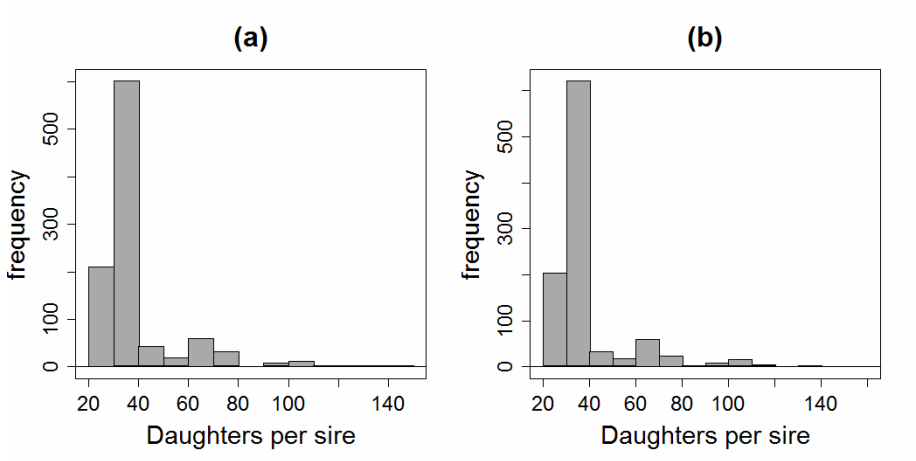

Figure 2.2. Distribution of the number of daughters per sire in (a) 0.30 heritability trait scenario and (b) 0.10 heritability trait scenario.

For each selection period, sires with higher EBVs were allowed to breed a new crop of progeny. Sires that had fewer than 40 daughters represented discarded progeny-tested bulls that were excluded after their first crop of daughters. In each period, 17% were proven bulls; González-Recio et al. (2005) reported a similar value for successful progeny-tested bulls in a progeny test program in Spain.

Daughter yield deviation was used as a dependent variable in the analysis of the SiresDYD strategy. When training and testing datasets overlap, evaluations of realized accuracies for genomic predictions can result in overconfidence (Amer and Banos, 2010); therefore, to avoid overlap between training and testing subsets, males of progeny sets 14 and 15 were excluded from the analysis. In addition, the records from cows from the last crop of daughters (15) were excluded from the estimates of DYD, as these animals were included in the validation set. To account for different accuracies in the DYD estimations, these values were weighted by their

prediction error variances in terms of number of daughter equivalents (VanRaden and Wiggans, 1991).

Genomic Evaluation Model

The Bayesian version of the LASSO method (de los Campos *et al*., 2009) was used to estimate SNP coefficients in the training populations. The response variables in the females strategies were the yield deviations, which are a result of a combination of a cow's genetic and residual values. Fixed effects were not simulated. The yield deviation was used as a dependent variable in the evaluation of all of the female-based selective genotyping strategies, including the strategies in which selection was based on breeding values. A single chain of Gibbs sampling was run using 10,000 iterations and a burn-in period of 2,500. Convergence was checked visually.

Accuracy of genomic evaluations

Accuracy is a common measurement of predictive ability (Goddard and Hayes, 2007; Luan et al., 2009) in genetic prediction studies. Accuracy was quantified using Pearson correlations between the predicted GBV and true breeding values simulated for generation 15. Means and standard deviations after five replicates were calculated for each strategy and trait.

Bias and MSE

True breeding values were known in the simulation. The average difference between the true and the predicted GBVs in the testing population provided a measure of the bias in the genomic predictions for each selective genotyping strategy. In addition, regression coefficients of traditional on genomic breeding were calculated for the SiresDYD strategy. Mean square error (**MSE**) of the estimator was calculated as prediction error. MSE was

used as a risk function to quantify differences between the estimator and the true value.

In addition, regression coefficients of true on estimated genomic breeding values were estimated, averages and standard deviations of intercepts, slopes and coefficients of determination were calculated for all considered strategies and sample sizes. Regression coefficients are usually considered as bias predictors when true breeding values are not known.

Results

Simulated population

In both scenarios (heritability either 0.10 or 0.30), the average LD (r^2 between adjacent markers) in generations 11 through 15 (training and testing sets) was 0.31. High LD values were observed only at small distances between pairs of SNPs (Figure 2.3). All chromosomes were simulated using the same parameters, and therefore, differences between them were not expected.

In the medium- and low-heritability scenarios, the average inbreeding coefficients in the last generation were 0.03 and 0.05, respectively, and the average accuracies of the pedigree indices were 0.35 (sd=0.05) and 0.41 (sd=0.04), respectively.

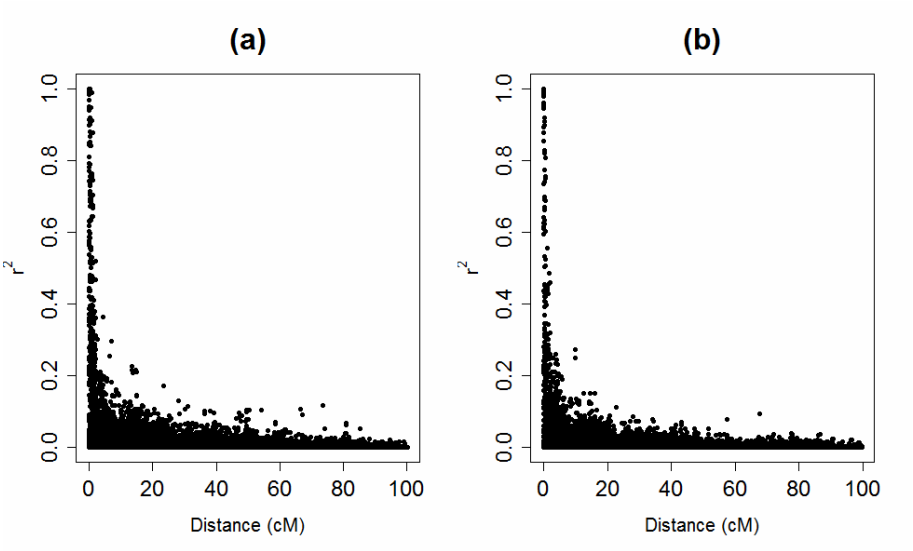

Figure 2.3. Distribution of r^2 between single-nucleotide polymorphism (SNP) pairs and physical distance: (a) chromosome 1 for the 0.10 heritability trait and (b) chromosome 7 for the 0.30 heritability trait.

Accuracy of genomic evaluations

The accuracy of genomic evaluations depended on the selective genotyping strategy used for the training set (Figure 2.4).

The predictive accuracy of the medium-heritability trait was greater than the accuracy of the low-heritability trait. As the size of the training populations increased, accuracies reached upper limits of approximately 0.75 (h^2=0.30) and approximately 0.70 (h^2=0.10). In the low- and medium-heritability traits, the accuracies of the SiresDYD strategy were 0.48 and 0.55, respectively, which indicated 37% and 34% increases, respectively, in accuracy relative to the accuracies of the pedigree indices (0.35 and 0.41, respectively).

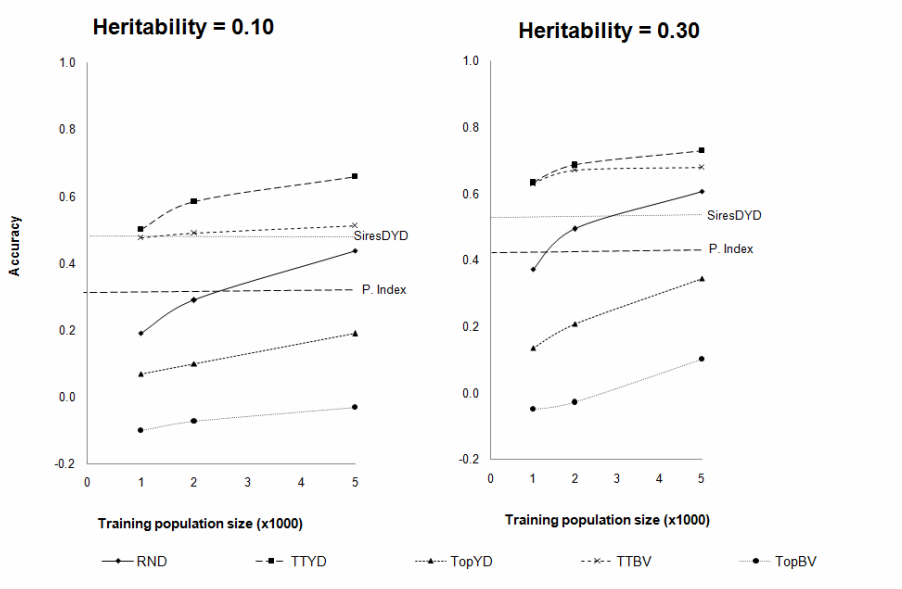

Figure 2.4 Estimated accuracies for genomic breeding values for two different heritabilities (0.10 and 0.30) in testing sets when 1000, 2000, or 5000 females in the training set were genotyped. The following genotyping strategies were used: cows at random (RND), top yield deviation cows (TopYD), top breeding value cows (TopBV), two-tailed yield deviation cows (TTYD), two-tailed breeding value cows (TTBV), all sires (SiresDYD), and pedigree index without GS.

Only the TTYD and TTBV strategies produced predictive accuracies that were better than those of the SiresDYD strategy (Table 2.1). When 1000 cows were genotyped as training set, in the TTYD strategy, the accuracies for low- and medium-heritability traits were 0.50 and 0.63, respectively. In the TTBV strategy, the corresponding values were 0.48 and 0.63.

Table 2.1. Average differences in the accuracy of predicted GBVs and standard deviations (in parenthesis) for each selective genotyping strategy[a] versus the SiresDYD[b] strategy based on the heritability and use of different female training sets and population sizes from a contemporary population of 40,000 animals

h^2	Size of	Two-Tailed Values		Top Values		Random
	Training set	Phen. (TTYD)[c]	EBV (TTBV)[d]	Phen. (TopYD)[e]	EBV (TopBV)[f]	RND[g]
0.3	1000	0.08(0.05)	0.08(0.06)	-0.42(0.11)	-0.60(0.12)	-0.18(0.08)
	2000	0.14(0.07)	0.12(0.07)	-0.34(0.08)	-0.58(0.13)	-0.06(0.05)
	5000	0.18(0.09)	0.13(0.08)	-0.21(0.07)	-0.45(0.11)	0.06(0.08)
0.1	1000	0.02(0.02)	0.00(0.08)	-0.41(0.09)	-0.58(0.09)	-0.29(0.07)
	2000	0.11(0.04)	0.01(0.08)	-0.38(0.11)	-0.55(0.09)	-0.16(0.05)
	5000	0.18(0.06)	0.04(0.07)	-0.29(0.07)	-0.51(0.06)	-0.04(0.06)

[a] Genotyping strategies for the training set
[b] Results are compared to a male genotyping strategy (SiresDYD), which genotypes all sires in the population as the training set (accuracies of the SiresDYD strategy were 0.48 and 0.55 for the 0.10 and 0.30 heritability traits, respectively).
[c] TTYD (Females with yield deviation in the $\alpha/2$ and $1-\alpha/2$ percentile)[h].
[d] TTBV (Females with EBVs in the $\alpha/2$ and $1-\alpha/2$ percentile)[h].
[e] TopYD (Females with yield deviation in the $1-\alpha$ percentile)[h].
[f] TopBV (Females with EBVs in the $1-\alpha$ percentile)[h].
[g] RND (Females selected at random)[h].
[h] (for α=0.025, 0.05, and 0.125).

In both the low- and medium-heritability scenarios, the use of two-tailed yield deviations data from generations 11 through 14 as criteria for the selection of animals in the training set produced the highest predictive accuracy regardless of the size of the training population. In all of the strategies, accuracy improved as the number of records in the training set increased. When the size of the training set increased from 1000 to 5000 genotyped cows, the RND strategy produced a greater increase in accuracy than the other strategies. Nevertheless, the accuracies of the RND strategy were always less than those produced by the two-tailed strategies. The accuracy of the RND strategy was greater than that of the SiresDYD strategy

only when 5000 cows were genotyped as the training set in the medium-heritability scenario.

Strategies based on the best females (TopYD and TopBV) produced the lowest accuracies, and at small training population sizes, the Top strategies produced negative values.

The heritability of the trait affected accuracy (Goddard and Hayes, 2009). The populations in our study required more than 5000 cows (12.5% of the simulated population) in the training set to achieve accuracy >0.66 in the low-heritability scenario.

Bias and MSE

Pedigree index predictions were biased in 0.01 trait units, which was lower than the values from the genomic predictions. The female-based selective genotyping strategies exhibited biases that were between those of the SiresDYD and the pedigree index (Table 2.2). Strategies that selected top animals only, including SiresDYD, produced more biased estimates than the other strategies (e.g., the SiresDYD strategy produced -0.97 and -2.23 for the low- and medium-heritability traits, respectively). The TopYD strategy produced the most biased estimate for the 0.10 heritability trait, and it gave the second-most biased prediction (after SiresDYD) for the 0.30 heritability trait (bias=1.74). In the Top and Random strategies, increases in the size of the training population reduced the bias. The RND strategy showed bias equal to 17% of that found in the SiresDYD strategy.

Table 2.2. Bias and mean square error (MSE) of genomic predictions in the testing set for different genotyping strategies, training set size and heritability

	h^2	Size of Genotyped Training set	Two-Tailed Values		Top Values		Random	SiresDYD[b]
			Phen. (TTYD)	EBV (TTBV)	Phen. (TopYD)	EBV (TopBV)	RND	
BIAS	0.10	1000	0.45 (0.18)	0.42 (0.06)	2.06 (0.03)	1.15 (0.09)	-0.15 (0.04)	-0.97 (0.12)
		2000	0.56 (0.18)	0.53 (0.09)	1.80 (0.03)	0.95 (0.09)	-0.09 (0.03)	
		5000	0.52 (0.11)	0.52 (0.08)	1.41 (0.04)	0.64 (0.08)	-0.04 (0.03)	
	0.30	1000	0.45 (0.10)	0.35 (0.06)	1.74 (0.02)	1.28 (0.06)	-0.36 (0.09)	-2.23 (0.32)
		2000	0.57 (0.10)	0.43 (0.08)	1.48 (0.03)	1.04 (0.07)	-0.26 (0.06)	
		5000	0.55 (0.06)	0.42 (0.06)	1.10 (0.05)	0.68 (0.09)	-0.16 (0.03)	
MSE	0.10	1000	0.51 (0.27)	0.38 (0.04)	4.32 (0.09)	1.41 (0.19)	0.10 (0.02)	1.00 (0.24)
		2000	0.66 (0.28)	0.48 (0.10)	3.32 (0.08)	1.01 (0.15)	0.08 (0.01)	
		5000	0.55 (0.16)	0.44 (0.09)	2.05 (0.09)	0.52 (0.09)	0.07 (0.01)	
	0.30	1000	0.64 (0.19)	0.42 (0.04)	1.84 (0.09)	1.41 (0.19)	0.29 (0.11)	5.20 (1.48)
		2000	0.75 (0.17)	0.45 (0.08)	1.29 (0.09)	1.01 (0.15)	0.20 (0.07)	
		5000	0.63 (0.09)	0.40 (0.06)	0.66 (0.08)	0.52 (0.09)	0.14 (0.01)	

[a]Bias measured as the difference between estimated and true breeding values (in genetic value units).

[b]Genotyped training set size = 996 animals for the SiresDYD strategy.

The biases in the two-tailed strategies were about 50% and 75% lower than those in the SiresDYD with the same training set size (Table 2.2). In the two-tailed strategies, an increase in the number of animals in the training set did not substantially reduce the bias. All but the RND and SiresDYD strategies overestimated the breeding values. To calculate DYD, the SiresDYD strategy used the records from the entire female population, whereas the Top strategies selected only the cows in the upper tail of the

distribution, which might be why the SiresDYD showed a bias that was more similar to the RND strategy. Similar patterns were apparent in the MSE; the RND and two-tailed strategies produced the lowest MSE. The MSE of the RND strategy was similar to that of the pedigree index and lower than that of the SiresDYD strategy with the same training set size.

In addition to bias results, regression coefficients of true on estimated genomic breeding values were calculated (Table 2.3, Table 2.4 and Table 2.5). Because of large differences between replicates for the "top" strategies, due to low accuracy, mean values across replicates are not informative.

Table 2.3. Averages and standard deviations of intercepts, of genomic predictions in the testing set, for different genotyping strategy, training set size and heritability regressions

		Size of Genotyped Training set	Two-tailed Values		Top Values		Random	SiresDYD[a]
	h^2		Phen. (TTYD)	EBV (TTBV)	Phen. (TopYD)	EBV (TopBV)	RND	
Intercept	0.10	1000	1.23	1.17	0.22	2.15	1.16	-1.83
			(0.09)	(0.07)	(0.95)	(0.15)	(0.10)	(0.10)
		2000	1.19	1.13	0.25	1.94	1.04	
			(0.09)	(0.06)	(1.00)	(0.17)	(0.10)	
		5000	1.11	1.05	-0.11	1.75	0.88	
			(0.09)	(0.06)	(0.73)	(0.14)	(0.07)	
	0.30	1000	2.42	2.22	-2.63	4.59	0.74	2.79
			(0.26)	(0.28)	(3.16)	(0.54)	(0.70)	(0.33)
		2000	2.32	2.06	-3.21	4.04	0.71	
			(0.19)	(0.24)	(2.63)	(0.51)	(0.57)	
		5000	2.12	1.88	-3.77	2.35	0.69	
			(0.17)	(0.22)	(2.77)	(1.22)	(0.39)	

[a]Genotyped training set size = 996 animals for the SiresDYD strategy.

Table 2.4. Averages and standard deviations of slopes of genomic predictions in the testing set, for different genotyping strategy, training set size and heritability regressions

		Size of Genotyped Training set	Two-tailed Values		Top Values		Random	SiresDYD[a]
	h^2		Phen. (TTYD)	EBV (TTBV)	Phen. (TopYD)	EBV (TopBV)	RND	
Slope	0.10	1000	0.21	0.24	0.38	-0.17	0.33	0.75
			(0.04)	(0.06)	(0.29)	(0.09)	(0.15)	(0.23)
		2000	0.22	0.25	0.41	-0.10	0.39	
			(0.03)	(0.06)	(0.30)	(0.08)	(0.15)	
		5000	0.26	0.28	0.57	-0.03	0.48	
			(0.04)	(0.08)	(0.27)	(0.10)	(0.11)	
	0.30	1000	0.32	0.37	1.14	-0.17	0.87	1.16
			(0.04)	(0.07)	(0.60)	(0.14)	(0.25)	(0.26)
		2000	0.33	0.40	1.30	-0.07	0.86	
			(0.05)	(0.08)	(0.54)	(0.19)	(0.20)	
		5000	0.38	0.45	1.52	0.30	0.84	
			(0.06)	(0.09)	(0.60)	(0.34)	(0.15)	

[a]Genotyped training set size = 996 animals for the SiresDYD strategy.

All strategies showed values different than 0 for the intercept and 1 for the slope regression as expected for unbiased predictions. In both cases, RND strategies were less deviated from the expected values than the TTBV and TTYD strategies. Finally, in the comparison between SiresDYD and two tailed strategies, averaged intercept estimation was closer to that expected for TTBV and TTYD, while the slopes of SiresDYD strategies were notably closer to 1 than the slopes of the two tailed strategies.

Table 2.5. Averages and standard deviations of coefficients of determination of genomic predictions in the testing set, for different genotyping strategy, training set size and heritability regressions

		Size of Genotyped Training set	Two-tailed Values		Top Values		Random	SiresDYD[a]
	h^2		Phen. (TTYD)	EBV (TTBV)	Phen. (TopYD)	EBV (TopBV)	RND	
r^2	0.10	1000	0.26 (0.11)	0.23 (0.07)	0.01 (0.01)	0.01 (0.01)	0.04 (0.03)	0.24 (0.12)
		2000	0.35 (0.11)	0.24 (0.06)	0.01 (0.02)	0.01 (0.01)	0.09 (0.05)	
		5000	0.44 (0.11)	0.27 (0.08)	0.04 (0.03)	0.01 (0.01)	0.20 (0.07)	
	0.30	1000	0.41 (0.09)	0.40 (0.08)	0.02 (0.02)	0.00 (0.00)	0.17 (0.10)	0.28 (0.14)
		2000	0.47	0.45	0.05	0.00	0.25	
	0.10	1000	0.26 (0.11)	0.23 (0.07)	0.01 (0.01)	0.01 (0.01)	0.04 (0.03)	0.24 (0.12)
		2000	0.35	0.24	0.01	0.01	0.09	

[a]Genotyped training set size = 996 animals for the SiresDYD strategy.

Discussion

Parameters of the simulated population

Quality control of the simulation before genomic evaluations was based on the LD between adjacent markers, the level of decay in LD with respect to the distance between SNPs, inbreeding values and the accuracy of traditional genetic evaluations. Simulated values were compared with Holstein real data. The average LD between adjacent markers in dairy cattle is related to the accuracy of genomic selection. Values of r2 between 0.20 and 0.31 have been reported for different populations (Banos and Coffey, 2010; Habier et al., 2010). LD values estimated in our simulation were similar to the values reported in Holstein cattle in North America (Sargolzaei et al., 2008). The level of decay in LD with respect to the distance between SNPs was also

similar to the results observed in real populations (Sargolzaei et al., 2008; De Roos et al., 2008).

Inbreeding values of the simulation were in the range of those reported in real dairy cattle populations (Kearney et al., 2004; González-Recio et al., 2006; González-Recio et al., 2007). Finally, the accuracies of the genetic evaluations were within the range of values reported for many traits in dairy cattle populations (González-Recio and Alenda, 2005; VanRaden et al., 2009a).

Accuracy of genomic evaluations

Differences between the pedigree index of traditional genetic evaluation and the SiresDYD genomic strategy were considered to be part of the simulation quality control. These results were similar to those reported in North American Holstein bulls (VanRaden et al., 2009a).

In the female-based strategies, the accuracies achieved using the two-tailed strategies (TTYD and TTBV) were greater than those obtained using the pedigree index, even at the smallest population size (1000). Compared to the SiresDYD strategy, the accuracies for the low- and medium-heritability traits derived from TTYD were 38% and 55% higher, respectively, but these increases were at the expense of an increase in the training population size from 1000 to 5000 animals. Two-tailed selections could be compared with the use of divergent lines in QTL detection and genome-wide association studies. Use of extreme samples appears to enhance the ability of selection procedures to select influential SNPs in genetic association studies. Higher accuracies reached by two-tailed strategies are consistent with a broiler mortality study by Long et al. (2009), who achieved similar conclusions.

The TopYD and TopBV strategies required a large number of animals to produce accuracies similar to those produced by the other strategies, which suggests that the TopYD and TopBV strategies were the least informative and should not be used to create a training population. Lower accuracies of Top strategies compared to RND have also been found by Ehsani *et al.* (2010), who compared different selective genotyping strategies and concluded that the selection of the best individuals does not provide good predictions compared with random selection.

Accuracy increased with the reference population size. This phenomenon has also been observed in previous simulations (Goddard and Hayes, 2009). In real populations, Lund et al. (2010) reported average reliability (square of accuracy) increases of between 8% and 11%. These results were obtained when the number of bulls in the training set was increased from the size of national training sets to the 15,966 shared genotypes of the EUROGENOMICS consortium (Holland, Finland, Sweden, Denmark, France and Germany).

The strategies based on yield deviations were more accurate than those that used EBVs as the selection criteria, which might be due to the low accuracy of the EBVs in cows. In the presence of epistasis and dominant effects, the strategies based on yield deviations might produce better results for the commercial population if the method can identify these effects. The dairy cattle industry might be interested in exploiting these effects in commercial populations, although dominance is not inherited and only part of the epistatic variance is transmitted to progeny.

Our study was based on a single trait rather than on multiple breeding objectives. Genetic evaluations are carried out for several traits, but only some of these traits explain the success of sires in the breeding program; for

example, udder composite is the key trait in sires' dams (González-Recio et al., 2005). Selection of different breeding goals may be reduced to a productivity-functionality index selection for 2/3 traits. The extreme individuals exhibiting both traits should be genotyped. Nonetheless the genotyping cost for the least profitable individuals must be carefully considered

Bias and MSE

The results from this study show that genotyping random females in the population lead to smaller biased predictions and MSE in the genomic-assisted evaluations. Genotyping only the top animals of the population, including sires, may lead to greater bias and MSE. Regression coefficients of true on genomic breeding values were not equal to 1. However, SiresDYD coefficients were in the range of similar values reported by other authors with real data of small dairy cattle populations (Olson et al., 2011). Female strategies showed low values, which could represent a potential problem in the application of female base GEBV. To deal with that problem, larger reference population sizes produce less bias and MSE. The RND strategy achieved always the smallest bias estimates. It must be pointed out that strategies that produced the more accurate predictions (TTYD, TTBV) also showed larger bias than the RND strategy. This is an interesting result for numerically small populations or when the economic resources for genotyping are limited. The genotyping strategy would need to focus either on maximizing accuracy or minimizing bias. The best strategy would depend on the purpose and organization of the breeding program. For instance, if comparison between non contemporaneous animals has to be done, the two tailed strategies may have some drawbacks, but they will maximize the genetic gain. The two tailed genotyping strategies showed smaller bias and

MSE than the SiresDYD strategies, suggesting that they might be interesting genotyping designs in numerically small populations In addition Patry and Ducrocq (2009) detected bias using GS and an underestimation of the breeding values when they were estimated based on pre-selected genomic animals. That source of bias does not affect our results as the selection was based on traditional breeding values. The estimation methodology and the model could be a source of bias in this study.

Conclusions

In small cattle populations, Two-tailed selection of females might be an advantageous strategy to create the training population in a genomic program, in terms of predictive ability, although at the expense of larger bias, mainly with small reference population sizes.

Random selection may be advisable for larger populations due to lower bias estimations. In addition, selection based on yield deviations rather than on EBVs might be preferable. However, strategies based on genotyping only the best cows (e.g., sires' dams) performed poorly.

A combination of two-tailed strategies based on the female population and the current male genotyping strategy should be considered, although the method to combine the DYD from sires and the yield deviations of cows must be developed.

All genotyping strategies considered based on genotyping the best animals resulted in biased evaluations, but largest bias was found for the "siresDYD" strategy.

Acknowledgments

The authors acknowledge funds from the project CDTI-P080250866 UPM-INIA, and to Beatriz Villanueva, Miguel Angel Toro and Agustín Blasco for helpful comments and suggestions.

References

Amer, P.R. and Banos, G. 2010. Implications of avoiding overlap between training and testing data sets when evaluating genomic predictions of genetic merit. Journal of Dairy Science 93, 3320-3330.

Banos, G. and Coffey, M.P. 2010. Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. Journal of Dairy Science 93, 2775-2778.

Blonk, R.J.W., Komen, J., van Arendonk J.A.M. 2010. Minimizing Genotyping In Breeding Programs With Natural Mating. World congress on genetic applied to livestock production abstract nº 195, Leipzig, Germany, 2-7 August 2010.

Calus, M.P.L. 2009. Genomic breeding value prediction: methods and procedures. Animal 4, 157 – 164.

Cromie, A.R., Berry, D.P., Wickham, B., Kearney, J.F., Pena, J., Van Kaam, J.B.C.H., Gengler, N., Szyda, J., Schnyder, U., Coffey, M., Moster, B., Hagiya, K., Welle, J.I., Abernethy, D. and Spelman, R. 2010. International genomic co-operation: Who, what, when, where, why and how? Interbull meeting. Riga, Latvia, May 31-June 4, 2010.

De los Campos, G., Naya, H., Pianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K. and Cotes, J.M. 2009. Posterior predicting quantitative traits with regression models for dense molecular markers and pedigrees. Genetics 182, 375–385.

De Roos, A.P.W., Hayes, B.J., Spelman, R.J. and Goddard, M.E 2008. Linkage Disequilibrium and Persistence of Phase in Holstein-Friesian, Jersey and Angus Cattle. Genetics 179, 1503–1512.

De Roos, A.P., Hayes, B.J. and Goddard, M.E. 2009. Reliability of genomic predictions across multiple populations. Genetics 183, 1545-1553.

Ducrocq, V. and Santus, E. 2011. Moving away from progeny test schemes: Consequences on conventional (Inter)national evaluations. Interbull Bulletin 43 (http://www.interbull.org/images/stories/Ducrocq_copy.pdf).

Ehsani, A., Janss, L. and Christensen, O.F. 2010, Effects of Selective Genotyping on Genomic Prediction. World congress on genetic applied to livestock production abstract nº 444, Leipzig, Germany, 2-7 August 2010.

Goddard, M.E. and Hayes, B.J. 2007. Genomic selection. Journal of Animal Breeding and Genetics. 124, 323–330.

Goddard, M.E. and Hayes, B.J. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. Nature Reviews Genetics 10, 381–391.

González-Recio, O. and Alenda, R. 2005. Genetic parameters for female fertility traits and a fertility index in Spanish dairy cattle. Journal of Dairy Science 88, 3282-3289.

González-Recio, O., Ugarte, C., and Alenda, R. 2005. Genetic analysis of an artificial insemination progeny test program. Journal of Dairy Science 88, 783-789.

González-Recio, O., Alenda, R., Chang, Y.M., Weigel, K. and Pianola, D. 2006. Selection for female fertility using censored fertility traits and investigation of the relationship with milk production. Journal of Dairy Science 89, 4438–4444.

González-Recio, O., López de Maturana, E. and Gutiérrez, J.P. 2007. Inbreeding depression on female fertility and calving ease in Spanish dairy cattle. Journal of Dairy Science 90, 5744-5752.

Habier, D., Dekkers, J.C.M. and Fernando, R.L. 2009. Genomic selection using low-density marker panels. Genetics 182, 343-353.

Habier, D., Tetens, J., Seefried, F.R., Lichtner, P. and Thaller, G. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42: 5.

Hayes, B.J. and Goddard, M.E. 2001. The distribution of effects of genes affecting quantitative traits in livestock. Genetics Selection Evolution 33(3), 209-229.

Hayes, B.J., Visscher, P.M., Mcpartlan, H.C. and Goddard, M.E. 2003. Novel multilocus measure of linkage disequilibrium to estimate past effective population size. Genome Research 13, 635–643.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. 2009. Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92, 433-443.

Hill, W.G. and Robertson, A. 1968. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics 38, 226-231.

Kearney, J.F., Wall, E., Villanueva, B. and Coffey, M.P. 2004. Inbreeding trends and application of optimized selection in the uk holstein population. Journal of Dairy Science 87, 3503-3509.

Loberg, A. and Dürr, J.W. 2009. Interbull survey on the use of genomic information. In Proceedings of the Interbull technical workshop 39, 3-14.

Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A., and Avendaño, S 2009. Comparison of classification methods for detecting associations

between SNPs and chick mortality. Genetics Selection Evolution 41:18.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M. and Meuwissen, T.H.E. 2009. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. Genetics 183, 1119–1126.

Lund, M.S., de Roos, A.P.W., de Vries, A.G., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, M. and Su, G. 2010. Improving genomic prediction by EuroGenomics collaboration. World congress on genetic applied to livestock production abstract nº 880, Leipzig, Germany, 2-7 August 2010.

Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829.

Olson, K.M., VanRaden, P.M., Tooker, M.E. and Cooper, T.A. 2011, Differences among methods to validate genomic evaluations for dairy cattle. Journal of Dairy Science 94, 2613-2620.

Patry, C., and Ducrocq, V. 2009. Evidence of a bias in genetic evaluation due to genomic selection. Interbull Bulletin 40, 167-171.

Sargolzaei, M., Schenkel, F.S., Jansen, G.B., and Schaeffer L.R. 2008. Extent of linkage disequilibrium in Holstein cattle in North America. Journal of Dairy Science 91, 2106-2117.

Sargolzaei, M. and Schenkel, F.S. 2009. QMSim: a large-scale genome simulator for livestock. Bioinformatics 25, 680-681.

Sen, S., Johannes, F. and Broman, K.W. 2009. Selective genotyping and phenotyping strategies in a complex trait context. Genetics 181, 1613-1626.

Sørensen, A.C., Sørensen, M.K. and Berg, P. 2005. Inbreeding in Danish dairy cattle breeds. Journal of Dairy Science 88, 1865–1872.

Spangler, M.L., Sapp, R.L., Bertrand, J.K., Mac Neil, M.D. and Rekaya, R. 2008. Different methods of selecting animals for genotyping to maximize the amount of genetic information known in the population. Journal of Animal Science 86, 2471-2479.

VanRaden, P.M. and Wiggans, G.R. 1991. Derivation, Calculation and use of national animal model information. Journal of Dairy Science 74, 2737-2746.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sontegard, T.S.G., Schnabel, R.D., Taylor, J.F., and Schenkel, F.S. 2009a. Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16-24.

VanRaden, P.M., Wiggans, G.R., Van Tassell, C.P., Sontegard, T.S.G. and Schenkel, F.S. 2009b. Benefits from collaboration in genomics. Interbull Bulletin 40, 67-72.

Weigel, K.A., Van Tassell, C.P., O´Connel, J.R., VanRaden, P.M., and Wiggans, G.R. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. Journal of Dairy Science 93, 2229-2238.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Chesnais, J.P., Schenkel, F. and Van Tassell, C.P. 2008. Genomic evaluations in the United States and Canada: A collaboration. In procedings of International Commite of Animal Recording, Niagara Falls, NY, June 16-20, 6 pp., 2008.

Zhao, H., Nettleton, D., Soller, M. and Dekkers, J.C.M. 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as

predictors of linkage disequilibrium between markers and QTL. Genetical Research 86, 77-87.

3

Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle

J. A. Jiménez-Montero

O. González-Recio

R. Alenda

Published in Journal of Dairy Science, (2013) 96: 625-634.

The use of appropriate methodology is essential for maximizing the benefits of genomic selection. Advanced statistical methods for genomic selection were compared using data from Spanish Holsteins; these methods included SNP regression using Bayesian methods (Bayes-A, Bayesian LASSO), a machine learning algorithm for genomic prediction and G-BLUP using the genomic relationship matrix. This study compared the performance of these methods in terms of their predictive correlations, bias, and mean squared error.

Abstract

The aim of this study was to evaluate methods for genomic evaluation of the Spanish Holstein population as an initial step toward the implementation of routine genomic evaluations. This study provides a description of the population structure of progeny tested bulls in Spain at the genomic level and compares different genomic evaluation methods with regard to accuracy and bias.

Two Bayesian linear regression models, Bayes-A and Bayesian-LASSO (**B-LASSO**), as well as a machine learning algorithm, Random-Boosting (**R-Boost**), and BLUP using a realized genomic relationship matrix (**G-BLUP**), were compared. Five traits that are currently under selection in the Spanish Holstein population were used: milk yield, fat yield, protein yield, fat percentage, and udder depth. In total, genotypes from 1859 progeny tested bulls were used. The training sets were composed of bulls born before 2005; including 1601 bulls for production and 1574 bulls for type, whereas the testing sets contained 258 and 235 bulls born in 2005 or later for production and type, respectively.

Deregressed proofs (**DRP**) from January 2009 Interbull (Uppsala, Sweden) evaluation were used as the dependent variables for bulls in the training sets, whereas DRP from the December 2011 DRPs Interbull evaluation were used to compare genomic predictions with progeny test results for bulls in the testing set.

Genomic predictions were more accurate than traditional pedigree indices for predicting future progeny test results of young bulls. The gain in accuracy, due to inclusion of genomic data varied by trait and ranged from 0.04 to 0.42 Pearson correlation units. Results averaged across traits showed that B-LASSO had the highest accuracy with an advantage of 0.01, 0.03 and

0.03 points in Pearson correlation compared with R-Boost, Bayes-A, and G-BLUP, respectively. The B-LASSO predictions also showed the least bias (0.02, 0.03 and 0.10 SD units less than Bayes-A, R-Boost and G-BLUP, respectively) as measured by mean difference between genomic predictions and progeny test results. The R-Boosting algorithm provided genomic predictions with regression coefficients closer to unity, which is an alternative measure of bias, for four out of five traits and also resulted in mean squared errors estimates that were 2%, 10%, and 12% smaller than B-LASSO, Bayes-A, and G-BLUP, respectively.

The observed prediction accuracy obtained with these methods was within the range of values expected for a population of similar size, suggesting that the prediction method and reference population described herein are appropriate for implementation of routine genome-assisted evaluations in Spanish dairy cattle. R-Boost is a competitive marker regression methodology in terms of predictive ability that can accommodate large data sets.

Introduction

Genomic selection (**GS**) is the most promising new technology since progeny testing for increasing the rate of genetic gain in dairy cattle (Weigel et al., 2010). It is based on simultaneous selection for thousands of single nucleotide polymorphisms (**SNP**). Direct genomic breeding values (**DGVs**) can be calculated as the sum of the effects of individual SNPs across the entire genome or genome-enhanced predictions can be computed by augmentation or replacing the traditional pedigree relationship matrix with the realized genomic matrix (Goddard, 2009). Typically, SNP effects are first estimated in a training or reference population and then used to predict the breeding values of new selection candidates (Hayes et al., 2009b).

In dairy cattle, GS has caused profound changes in practical breeding programs, because nearly all young bulls acquired by major artificial-insemination (**AI**) centers are now selected based on such evaluations (Wiggans et al., 2011). In addition, females can be evaluated with cost-effective genotyping strategies (Weigel et al., 2012), leading to genomic predictions with a similar reliability to that of young bulls.

National and international genetic evaluations of dairy cattle consider nearly two dozen phenotypic traits (VanRaden and Sullivan, 2010) and the inclusion of additional, complex traits is expected within the next decade. These new traits may include measures of disease resistance and residual feed intake (González-Recio and Forni, 2011; Pryce et al., 2012), and evaluation may consider crossbreed performance (Toosi et al., 2010), and genotype by environment interaction effects (Hayes et al., 2009a).

Several different approaches are currently used for estimating genomic values, and it is important to assess the performance of diverse methodologies and identify the methods that can provide the greatest

predictive accuracy in a given population. Genomic prediction methods can be categorized as: 1) methods that regress phenotypic records on SNP markers directly, and 2) methods that view genetic values as a function of the subject and use marker information to build the (co)variance structure between subjects (De los Campos et al., 2009). The first group of methods includes several Bayesian regression approaches, such as Bayes-A, Bayes-B (Meuwissen et al., 2001), and the Bayesian least absolute shrinkage selection operator (**B-LASSO**), as described by Park and Casella (2008). These regression-based methods are usually implemented after traditional BLUP genetic evaluation of the reference population, and the resulting breeding value estimates are then used directly or deregressed prior to use as a dependent variable in the genomic evaluation (VanRaden, 2008). In general, these methods are computationally time-consuming if the number of SNPs is large, and this could preclude their utilization in routine evaluation programs in some countries, despite the fact are used in some countries as The Netherlands**.** The second group includes methods that compute a realized relationship matrix from the markers, such as G-BLUP (VanRaden et al., 2009b), or single step (Misztal et al., 2009) methods, to augment or replace the traditional pedigree based relationship matrix. The Single-step method includes all pedigree and genomic information and avoids the need to subsequently combine the genomic and traditional breeding values (Aguilar et al., 2010). The performance of alternative genomic evaluation methodologies can vary depending on the trait and population structure (Daetwyler et al., 2010).

In addition to the afore-mentioned approaches, an alternative for dealing with large data sets and complex interactions between SNPs is machine learning algorithms (Long et al., 2007). Machine learning methods usually compare favorably to Bayesian regression models in terms of predictive ability (e.g., González-Recio et al., 2008; Moser et al., 2009; González-

Recio and Forni, 2011). Non-parametric or semi-parametric methods of this type can be implemented by regressions on markers (e.g., Boosting as in González-Recio et al., 2010)) or by building appropriate (co)variance structures (e.g., Reproducing Kernel Hilbert Spaces regression, (Gianola and van Kaam, 2008). Boosting algorithms are among the most appealing machine learning methods for genomic-prediction problems (Ogutu et al., 2011), and in a recent study they provided greater predictive ability and smaller bias than other methods (González-Recio et al., 2010). Efficiency of DGV prediction in dairy cattle can be enhanced by modification to the algorithm, specifically Random Boosting (**R-Boost**), as is described in a companion paper(González-Recio et al., 2013.). These modifications allow prediction of genomic values with SNP regression methods in very large data sets.

Over the last decade the Spanish breeding program has provided competitive bulls for the national and international markets due to a robust milk-recording scheme. Special care has been taken in recording morphologic traits. GS has revolutionized dairy cattle breeding since 2009. Taking advantage of this technology is necessary to maintain the program's viability.

The objective of this study was to use genotypic and phenotypic data from the Spanish Holstein population to compare several popular genomic evaluation methodologies. Two different Bayesian linear regressions (Bayes-A and B- LASSO), G-BLUP, and a machine learning algorithm (R-Boost) were compared. Five phenotypic traits were considered, and methods were evaluated based on predictive correlations, bias, and mean squared error.

Material and methods

Genotypes

A total of 1859 progeny-tested sires were genotyped. The BovineSNP50.v2 Beadchip (Illumina, San Diego, CA), was used to genotype 54,609 SNPs of 1619 Sires, whereas the remaining 240 sires were genotyped for 54,001 SNPs using the BovineSNP50.v1 Beadchip.

SNPs with a greater than 5% incidence of missing genotypes across individuals and SNPs with minor allele frequency (**MAF**) less than 5% were discarded, leaving only 39,714 SNPs for the analysis. Some animals had missing genotypes for certain markers; after editing, 0.01% of the SNPs were missing. Missing genotypes were then imputed with BEAGLE 3.3.2(Browning and Browning, 2009). In a pilot study, known SNP's were masked mimicking missing marker rate of the population. Resulted imputation allele error ratio was 0.008.

Linkage Disequilibrium Estimation

The haplotypes obtained by Beagle prior imputation were used to estimate the degree of linkage disequilibrium (**LD**) between SNPs; all genotyped bulls were used in this calculation. LD, which refers to the non-random association of alleles between two loci, was measured using the r^2 parameter (Hill and Robertson, 1968). LD can be estimated using other methods, such as D, D' or measures based on the Chi-squared statistic (Zhao et al., 2005); however, r^2 is the most common measure of LD for biallelic markers and is less sensitive to the effects of allelic frequency than other methods (Sargolzaei et al., 2008).

Phenotypes

The January 2009 de-regressed MACE proofs (**DRP**) from progeny testing, as described in Jairath et al. (1998), were used as dependent variables and included 1859 bulls for production and 1810 for type. The production and type data were collected from progeny between 1980 and 2008. Sire's DRP for milk yield (**MY**), fat yield (**FY**), protein yield (**PY**), fat percentage (**FP**), and udder depth (**UD**) were used. The estimated heritability based on traditional genetic evaluations in Spain is 0.28 and 0.30 for production traits and udder depth, respectively.

Training and Validation Data Sets

Training and validation data sets were generated based on year of birth of the bulls. A total of 1601 bulls with DRP in January 2009 that were born before 2005 were used for the production training set, whereas 1574 bulls from the same period were used for the type training set. Bulls born between 2005 and 2007 were used as the validation set; 258 bulls were used for production traits and 236 were used for type. Effective daughter contributions (**EDC**) were used as weighting factors to account for differences in progeny group size when computing genomic predictions (Jairath et al., 1998). Bulls in the testing sets had their DGV in December 2011 that were based on 20 or more EDC. Design of the training and testing sets followed the recommendations of (Mäntysaari et al., 2010); although the recommended four years gap between training and testing sets was reduced to three years due to small size of the reference population, thereby leaving more training set bulls maximize the accuracy of estimated DGV.

Genomic Evaluation Model

The general structure for the models in linear form is:

$$\boldsymbol{y} = \mu \mathbf{1}_n + \sum_j \mathbf{X}_j g_j + \mathbf{e} \quad ,$$

where y is the vector of phenotypic records, μ is the overall mean, 1n is a vector of n ones, ∑j is a summation over all markers, gj is the coefficient of marker j denoting the allele substitution effect, Xj is a design matrix of genotype codes for the respective marker, and e is a vector of residuals.

The data were analyzed using four different approaches: two models based on marker regression (Bayes-A and B-LASSO), a method based on a realized relationship matrix from the markers (G-BLUP) and a machine learning algorithm (R-Boost), which is described in the companion paper (González-Recio et al., 2013).

Bayes-A

Bayes-A was introduced by (Meuwissen et al., 2001). This method assumes that marker effects (g_j) are normally and independently distributed a priori as $N = (0, \sigma^2_{g_j})$, where $\sigma^2_{g_j}$ is an unknown variance associated with marker j. The prior distribution of the variances of the SNPs was a scaled inverted chi-squared distribution, $\sigma^2_j \sim X^{-2}_{(df,ss)}$, where ss is the scale parameter and df represents the degrees of freedom. The parameters ss and df were considered as hyper-parameters and were fixed for each trait as in Gianola et al. (2009). An improper prior was assumed for μ. Following (González-Recio et al., 2009), the residuals, e, were assumed to be distributed as $N(0, R = N^{-1}\sigma^2_e)$, where N = {ni} is a diagonal matrix with elements ni representing the

corresponding EDC of sire i, and σ_e^2 is the residual variance. The prior distribution for the residual variance σ_e^2 was assumed to be an inverted chi-squared distribution with hyper-parameters df and ss. The Gibbs sampler was run for 10,000 cycles, with the first 1,000 cycles of burn-in discarded. Convergence of the chain was checked by visual inspection, and inferences on the parameters were made on the mean posterior estimates after burn-in.

Bayesian- LASSO

The Bayesian counterpart of the LASSO model (De los Campos et al., 2009; Park and Casella, 2008) was also used to estimate SNP coefficients in the training population. The B-LASSO can be viewed as an optimization problem, using the sum of the absolute values of the regression coefficients (*L1*-norm) as a penalty, in the following regression model (Tibshirani, 1994):

$$\min_{\beta}\left\{\sum(\boldsymbol{y}_j-\mathbf{X}_j'\boldsymbol{\beta})^2+\lambda(t)\sum_j|\boldsymbol{\beta}_j|\right\},$$

where X is a vector of covariates, β is the corresponding vector of regression coefficients and λ is a smoothing parameter controlling the shrinkage of the distribution.

The LASSO estimates can be interpreted as the posterior mode in a Bayesian model considering a double Laplace prior for the coefficient estimates, as:

$$\pi(\beta)=\prod_{j=1}^{p}\frac{\lambda}{2}e^{-\lambda|\beta_j|},$$

as put forth by (Park and Casella, 2008), the smoothing parameter λ was assigned a prior distribution gamma (a, b). Values of the hyper-parameters of the prior distribution were set at 5.0 and 1.0, respectively for convenience.

The Laplace distribution results in stronger shrinkage of marker coefficients towards zero than Bayes-A. This prior gives each coefficient βj a high probability of being near zero while simultaneously giving some coefficients a chance to have large effect (Yi and Xu, 2008). In practice, this produces a similar outcome to variable selection (De Los Campos et al., 2010). Flat prior was assumed for μ. The prior distribution for the residual variance σ_e^2, was assumed to be an inverted chi-squared distribution with hyper-parameters df and ss and was weighted by the number of progeny of each bull, as detailed for the previous method. A single chain of Gibbs sampling was run using 25,000 iterations and a burn-in period of 15,000. The convergence of the chain was checked by visual inspection, and inferences on the parameters were made on the mean posterior estimates after burn-in.

G-BLUP

The G-BLUP is the most similar to traditional BLUP evaluations of the four alternatives considered herein. If many QTL exist with effects that are normally distributed with constant variance, the pedigree relationship matrix can be replaced with the genomic relationship matrix (**G**) where the latter is built from molecular information. Pairs of individuals sharing the same genotype for a large number of markers will be more similar genomically, and will have higher values in the corresponding off diagonal cells of the matrix, as is the case for pairs of related animals in a pedigree-based relationship matrix. The genomic relationship matrix was computed as

$G = \frac{ZZ'}{2\sum p_i(1-p_i)}$, following (VanRaden, 2008), where a more detailed description of this model is provided.

Random-Boosting

The boosting algorithm is a machine learning technique that combines several different predictors and a shrinkage factor (Friedman, 2000). Boosting iteratively adds basis functions, such that each addition further reduces the selected loss function (Hastie et al., 2005). In this study, ordinary least squares was chosen as the basis function, and it was successively applied to the residuals of the previous iteration in a sequential manner. The mean squared error (**MSE**) of prediction was used as the loss function to be minimized. Details of the algorithm are described in the companion paper (González-Recio et al., 2013). Following their results, *v* was set to 0.10 for production traits and 0.20 for type, while the percentage of SNPs selected at each iteration (*mtry*) was set to 0.50, 0.10, 0.01, 1.00 and 0.10 for MY, FY, PY, FP and UD, respectively.

Estimation of Direct Genomic Values

The DGVs for each trait were calculated for individuals in the testing set as:

$$DGV = \mu + \mathbf{X}_j\hat{\boldsymbol{\beta}}_j,$$

where $\boldsymbol{\mu}$ is the overall mean, $\mathbf{X}_j$ is a matrix of genotypes and $\hat{\boldsymbol{\beta}}_j$ is a vector of posterior means of SNP effects for each of the four methods. For the R-Boost method, $\hat{\boldsymbol{\beta}}_j$ represents the sum of the slope estimates from the model in which SNP *j* was selected.

Criterion for Comparisons

The accuracy of the genomic predictions was computed as the Pearson correlation between the predicted DGV of bulls in the testing set and their December 2011 DRP. Sire - maternal grandsire - maternal great grand sire index for sires in the testing set was used as a benchmark. It was calculated as 50% of the sires DRP, 25% of the maternal grand sires DRP and 12.5% of the maternal great-grand sires DRP. For simplicity we refer to these values as Sire-Pedigree Index (**Sire-PI**). Estimated accuracies were adjusted by EDC following the recommendations of (Mäntysaari et al., 2010).

The average difference between 2011 DRP and the predicted DGVs in the testing population provided a measure of bias in the genomic predictions; this bias estimate was standardized. Coefficients of regression of realized December 2011 DRP on estimated DGV were also calculated, because this parameter is also commonly used as a measure of prediction bias in genomic evaluations (Mäntysaari et al., 2010). Finally, MSE of prediction, which is linked to bias, slope and accuracy, was also estimated.

Results and discussion

Summary of genomic data

The distribution of genotyped bulls, by year of birth, is shown in Figure 3.1. Most of the bulls were born after 1990, thereby providing a recent reference population for prediction of genomic values of young animals. It is well known that GS results in higher responses for the generations closer to the reference population (Goddard, 2009).

After filtering, the distribution of MAF was nearly uniform with a mean of 0.28 (Figure 3.2). The average distance between adjacent SNPs was 0.06

Mb, and SNPs had average heterozygosity of 0.29. Linkage disequilibrium, between adjacent SNPs, measured as r2, was 0.24. All values were in the range reported previously values for other Holstein populations (Banos and Coffey, 2010; Habier et al., 2010; Wiggans et al., 2009a).

Figure 3.1. Number of genotyped bulls by year of birth.

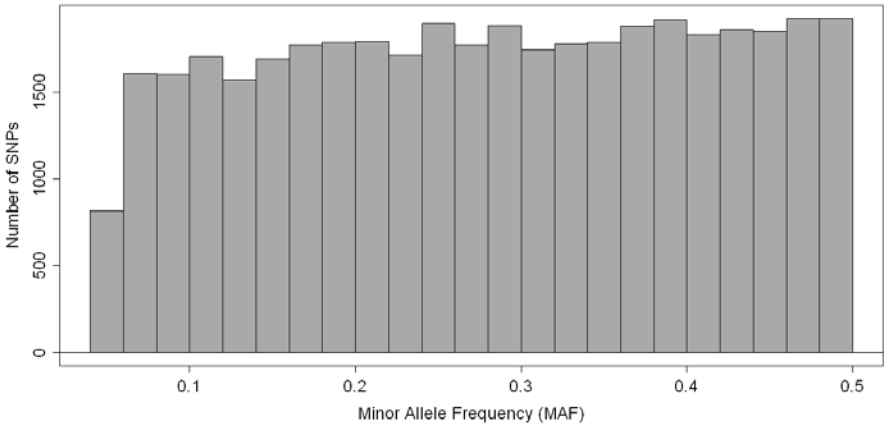

Figure 3.2. Distribution of minor allele frequencies (MAF) of the SNPs after quality control.

The Figure 3.3 shows the average r^2 between SNP pairs plotted against the map distance of up to 1 Mb and shows the standard deviations for the average r2 values across all 30 chromosomes. Average r^2 decreased exponentially with increasing distance between SNPs and was equal to 0.40, 0.24, 0.16 and 0.08 at distances of 0.01, 0.05, 0.1 and 1 Mb, respectively. The level of decay in LD with respect to the distance between SNPs was also similar to results from other populations (Habier et al., 2010; de Roos et al., 2009).

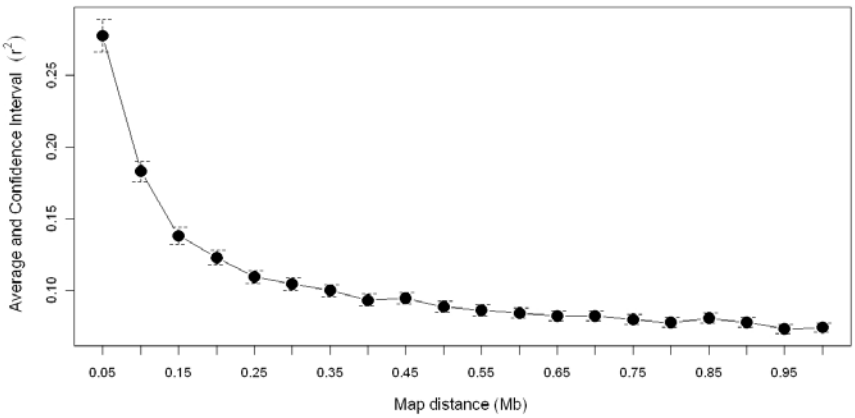

Figure 3.3. Average linkage disequilibrium (measured as r2) and confidence interval (estimated by R package gplots) between syntenic markers with respect to their physical distance.

Accuracy

Traditional Sire-PI accuracies ranged from 0.37 to 0.51 (Table 3.1). Predicted DGV showed higher accuracies than Sire-PIs, regardless of the genomic prediction model, with an average increment of 49%, ranging from 9% for UD to 83% for FP. Similar results have been reported previously in other Holstein populations (Moser et al., 2010; VanRaden et al., 2009b), indicating that selection of young animals based on genomic predictions is preferable to selection based on traditional pedigree information.

Table 3.1. Accuracy, standardized bias in means, bias in regression coefficients and mean squared error (MSE) of genomic predictions for different evaluation methodologies and five traits of economic interest in Spanish dairy cattle

Methods[1]	Milk Yield (MY)	Fat Yield (FY)	Protein Yield (PY)	Fat Percentage (FP)	Udder Depth (UD)
Accuracy					
Sire-PI	0.37	0.37	0.40	0.39	0.51
B- LASSO	**0.60**	**0.61**	**0.57**	0.74	0.55
Bayes-A	0.55	**0.61**	0.55	0.65	**0.56**
R-Boost	0.54	0.60	0.50	**0.81**	0.54
G-BLUP	0.58	0.59	**0.57**	0.62	0.55
Bias in means					
B- LASSO	0.04	**-0.05**	0.05	**0.01**	**-0.06**
Bayes-A	0.07	-0.07	**0.02**	-0.04	-0.11
R-Boost	**-0.01**	-0.09	**-0.02**	**0.01**	-0.22
G- BLUP	-0.19	-0.15	-0.16	0.14	**0.06**
Bias in Regression coefficients					
B- LASSO	0.73	0.80	0.70	1.06	0.63
Bayes-A	0.58	0.78	0.67	0.79	0.69
R-Boost	**0.87**	**0.99**	**0.80**	1.19	**0.82**
G- BLUP	0.71	0.80	0.70	**1.02**	0.64
MSE					
B- LASSO	**239992**	398	**206**	**0.03**	0.95
Bayes-A	289122	404	215	0.04	**0.90**
R-Boost	247593	**395**	216	**0.03**	0.92
G- BLUP	269619	423	219	0.04	0.95

In bold: The preferred method within trait and comparison criteria.
[1]Methods: Sire-PI (Traditional pedigree index), B- LASSO (Bayesian LASSO), Bayes-A, R-Boost (Random Boosting) and G-BLUP

Among methods, B-LASSO provided the highest Pearson correlations for MY (0.60), FY (0.61) and PY (0.57), as well as the highest Pearson correlation averaged across traits. Bayes-A provided greater accuracy for UD (0.56) and was equivalent in accuracy to B-LASSO for FY (0.61). R-

Boost achieved the greatest Pearson correlation for FP (0.81), whereas G-BLUP achieved the same accuracy as B-LASSO for PY (0.57). In general, differences in accuracy between methods were small for MY, FY, PY and UD but larger differences were found for FP. For instance, R-Boost outperformed GBLUP by 0.19 units of Pearson correlation for FP. Pearson correlation coefficients averaged across traits were 0.61, 0.60, 0.58, and 0.58 for B-LASSO, R-Boost, Bayes-A, and G-BLUP, respectively.

In a previous study based on simulated and growth data in mice, Usai et al. (2009) showed slightly greater accuracy with B-LASSO compared with G-BLUP and Bayes-A. Cleveland et al. (2010) reported a similar predictive ability for B-LASSO and two variants of Bayes-A in simulated data; however, the authors observed better performance of B-LASSO for traits that were regulated by many QTL with small effects. Legarra et al. (2011) reported slightly greater accuracies for B-LASSO than G-BLUP, but slightly better for B-LASSO. G-BLUP on real data showed reliabilities of 63% compared to 32% from pedigree index on the combined trait Net Merit(VanRaden, 2008). Others found higher or similar accuracies using GBLUP than using Bayes B (Luan et al., 2009; Mrode et al., 2010).

There were no relevant differences between R-Boost and the additive models based on marker regression, except for FP. Although machine learning techniques are expected to accommodate cryptic relationships in the data, the use of dependent variables that represent previously computed (additive, linear and smoothed), sire EBVs could mask such differences. R-Boost seems to provide some advantages over Bayesian regression when a small number of QTL regulate the trait under purely additive regulation (González-Recio and Forni, 2011). In the present study, genomic predictions

from R-Boost were more accurate for traits controlled by single genes that explain a large proportion of the genetic variance (e.g., DGAT1 for FP). Note that differences exist in accuracy for the R-Boost method between the results of this manuscript and the companion paper (González-Recio et al., 2013), presumably due to the adjustment for number of progeny in the present paper as suggested by (Mäntysaari et al., 2010).

Bias in the Mean

The DGV of bulls in the testing set showed an average deviation over the realized DRP of 0.08 genetic SD across methods and traits, with averages ranging from 0.05 (FP) to 0.11 (UD). Increasing size of the reference population may alleviate this problem (Liu et al., 2011; Lund et al., 2011). Standardized bias showed greater differences between methods than Pearson correlations. R-Boost resulted in nearly unbiased predictions for MY and FP and also produced the least bias for PY, whereas B-LASSO, produced the least bias in predictions for FY, FP and UD. Bayes-A showed a similar bias to R-Boost for PY. G-BLUP tended to provide, more biased predictions for all traits, with the exception of UD. The Methods with greater Pearson correlation can also produce more biased predictions; so both accuracy and bias should be considered when deciding which method has greater predictive ability. Therefore, MSE may be a more appropriate comparison criterion than the Pearson correlation, as it combines accuracy and bias.

When the genomic predictions of young bulls are compared with highly reliable, progeny-tested bulls, biases from genomic predictions must be taken into account. In addition, genomic predictions of future performance are expected to be biased when only genomically pre-selected bulls are allowed to produce offspring (Patry and Ducrocq, 2011). This was not the

case for bulls included in the present study, as they were genotyped after selection.

Bias in Regression Coefficients

The coefficients of regressing realized DRP on estimated DGV are commonly used as a measure of bias in genomic evaluations. The expected value for this slope coefficient is unity if evaluations predict the actual magnitude of differences between bulls, if the genotyped young bulls are a representative sample of the bulls in the population. However, the genotyped young bulls are typically pre-selected by the AI centers based on their EBV or Sire-PI (Mäntysaari et al., 2010). In our study, regression coefficients ranged between 0.58 for Bayes-A (MY) and 1.19 for the R-Boost (FP). R-Boost provided slope coefficient closest to unity for four of the five traits (0.87 for MY, 0.99 for FY, 0.80 for PY and 0.82 for UD). Bayes-A provided the smallest coefficients for all traits, except UD, whereas B-LASSO and GBLUP produced similar coefficients that exceeded unity only for FP.

These Regression coefficients were within the range reported in other studies in similar dairy cattle populations (Olson et al., 2011; Tsuruta et al., 2011). Some authors have suggested inclusion of a polygenic effect to address this problem (Liu et al., 2011), because this modification could reduce the overestimation of DGV. Low coefficients of regression for MY, PY and UD, could be possibly explained by higher selection on these traits compared with FY and FP.

MSE

The MSE can be viewed as a risk function that incorporates both the predictive variance and bias of an estimator. B-LASSO and R-Boost provided smallest MSE for all traits except UD, where Bayes-A

outperformed the other methods. For instance, regarding MY Bayes-A showed 20% and 17% greater MSE than B-LASSO and R-Boost, respectively. G-BLUP also showed greater MSE (from 5 to 12%) for MY and FY, as compared with B-LASSO and R-Boost. R-Boost was the preferred method across traits in terms of MSE providing the smallest MSE on average, followed by B-LASSO, Bayes-A, and G-BLUP respectively.

In a previous study, (Verbyla et al., 2009) showed similar differences in MSE between Bayesian regression models and G-BLUP. Their study reported larger MSE than the present results for the Spanish population, perhaps due to their smaller reference population (1098 progeny tested bulls). As stated previously, MSE reflects both bias and accuracy, but, it is often ignored when comparing genomic evaluation methods.

Conclusions

Implementation of GS in the Spanish Holstein breeding program will improve selection efficiency for both AI centers and commercial farms, and identification of superior animals at young ages will be more accurate than was previously believed possible.

The descriptors of the genomic structure of the population used in this study showed that the Spanish population is similar to other Holstein dairy cattle populations, as expected. Based on this similarity, genomic evaluations of genotyped animals for recorded traits included in the milk recording scheme should be feasible.

Different prediction methodologies, including non-parametric, implemented in this study showed similar predictive ability, and the optimal method was sometimes trait dependent. In general B-LASSO was preferable in terms of

Pearson correlations, and R-Boost provided regression coefficient estimates closest to unity. Both methods outperformed Bayes-A and G-BLUP in terms of predicted MSE. Methods that provided higher Pearson correlations also showed large biases, so MSE may be a more appropriate comparison criterion than Pearson correlations. Marker regression methods outperformed G-BLUP in terms of MSE due to larger bias in GBLUP estimates. Lastly the R-Boost method may provide computational advantages over B-LASSO and Bayes-A

Future collaborations with the EUROGENOMICS consortium, which has a reference population of more than over 20,000 progeny-tested bulls, is expected to substantially increase the accuracy of genomic predictions for Spanish Holsteins. Here the R-Boost method is expected to show some computationally advantages over B-LASSO and Bayes-A

Acknowledgments

The authors acknowledge funds from the project CDTI-P080250866 UPM and the agreement INIA-CC10-046, to CONAFE, ASCOL, ABEREKIN, XENETICA FONTAO and GENETICAL for providing biological samples and phenotypes used in this study and to “Dirección General de Producciones y Mercados Agrarios”, “Laboratorio Central de Veterinaria del Ministerio de Agricultura, Alimentación y Medio Ambiente” for support of genotyping process and especially Pr. Kent Weigel for helpful suggestions and comments.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full

pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Banos, G., Coffey, M.P., 2010. Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. Journal of Dairy Science 93, 2775–2778.

Browning, B.L., Browning, S.R., 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics 84, 210–223.

Cleveland, M., Forni, S., Deeb, N., Maltecca, C., 2010. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. BMC Proceedings 4, S6.

Daetwyler, H.D., Pong-Wong, R., Villanueva, B., Woolliams, J.A., 2010. The Impact of Genetic Architecture on Genome-Wide Evaluation Methods. Genetics 185, 1021–1031.

De los Campos, G. de los, Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182, 375–385.

De Los Campos, G., Gianola, D., Rosa, G.J.M., Weigel, K.A., Crossa, J., 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. Genetics Research 92, 295–308.

Friedman, J.H., 2000. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189–1232.

Gianola, D., Kaam, J.B.C.H.M. van, 2008. Reproducing Kernel Hilbert Spaces Regression Methods for Genomic Assisted Prediction of Quantitative Traits. Genetics 178, 2289–2303.

Goddard, M., 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136, 245–257.

González-Recio, O., Forni, S., 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genetics Selection Evolution 43, 1–12.

González-Recio, O., Gianola, D., Long, N., Weigel, K.A., Rosa, G.J.M., Avendaño, S., 2008. Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers. Genetics 178, 2305–2313.

González-Recio, O., Gianola, D., Rosa, G., Weigel, K., Kranis, A., 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genetics Selection Evolution 41, 3.

González-Recio, O., Jiménez-Montero, J.A., Alenda, R., n.d. The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. Journal of Dairy Science 96, 614–624.

González-Recio, O., Weigel, K.A., Gianola, D., Naya, H., Rosa, G.J.M., 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. Genetics Research 92, 227–237.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., Thaller, G., 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42, 5.

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27, 83–85.

Hayes, B., Bowman, P., Chamberlain, A., Verbyla, K., Goddard, M., 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genetics Selection Evolution 41, 51.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J., Goddard, M.E., 2009b. Invited review: Genomic selection in dairy cattle: progress and challenges. J. Dairy Sci. 92, 433–443.

Hill, W.G., Robertson, A., 1968. Linkage disequilibrium in finite populations. TAG Theoretical and Applied Genetics 38, 226–231.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B., Kolstad, B., 1998. Genetic Evaluation for Herd Life in Canada. Journal of Dairy Science 81, 550–562.

Legarra, A., Robert-Granié, C., Croiseau, P., Guillaume, F., Fritz, S., 2011. Improved Lasso for genomic selection. Genetics Research 93, 77–87.

Liu, Z., Seefried, F., Reinhardt, F., Rensing, S., Thaller, G., Reents, R., 2011. Impacts of both reference population size and inclusion of a residual polygenic effect on the accuracy of genomic prediction. Genetics Selection Evolution 43, 19.

Long, N., Gianola, D., Rosa, G. j. m., Weigel, K. a., Avendaño, S., 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics 124, 377–389.

Luan, T., Woolliams, J.A., Lien, S., Kent, M., Svendsen, M., Meuwissen, T.H.E., 2009. The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. Genetics 183, 1119–1126.

Lund, M., De Roos, A., De Vries, A., Druet, T., Ducrocq, V., Fritz, S., Guillaume, F., Guldbrandtsen, B., Liu, Z., Reents, R., Schrooten, C., Seefried, F., Su, G., 2011. A common reference population from

four European Holstein populations increases reliability of genomic predictions. Genetics Selection Evolution 43, 1–8.

Mäntysaari, E., Liu, Z., VanRaden, P., 2010. Validation Test for Genomic Evaluations. InterbullBull 41, 17–22.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Misztal, I., Legarra, A., Aguilar, I., 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science 92, 4648–4655.

Moser, G., Khatkar, M., Hayes, B., Raadsma, H., 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genetics Selection Evolution 42, 37.

Moser, G., Tier, B., Crump, R., Khatkar, M., Raadsma, H., 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genetics Selection Evolution 41, 56.

Mrode, R., Coffey, M.P., Straden, I., Meuwissen, T.H.E., vanKaam, J.B., Kearney, J.F., Berry, D.P., 2010. A comparison of various methods for the computation of genomic breeding values of dairy bulls using software at genomicselection.net, in: Proc. 9th World Cong. Genet. Appl. Livest. Prod. Presented at the 9th World Cong. Genet. Appl. Livest. Prod., Leipzig (Germany), p. Abstract 518.

Ogutu, J., Piepho, H.-P., Schulz-Streeck, T., 2011. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proceedings 5, S11.

Olson, K.M., VanRaden, P.M., Tooker, M.E., Cooper, T.A., 2011. Differences among methods to validate genomic evaluations for dairy cattle. Journal of Dairy Science 94, 2613–2620.

Park, T., Casella, G., 2008. The Bayesian Lasso. Journal of the American Statistical Association 103, 681–686.

Patry, C., Ducrocq, V., 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. Journal of Dairy Science 94, 1011–1020.

Pryce, J.E., Arias, J., Bowman, P.J., Davis, S.R., Macdonald, K.A., Waghorn, G.C., Wales, W.J., Williams, Y.J., Spelman, R.J., Hayes, B.J., 2012. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. Journal of Dairy Science 95, 2108–2119.

Roos, A.P.W. de, Hayes, B.J., Goddard, M.E., 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183, 1545–1553.

Sargolzaei, M., Schenkel, F.S., Jansen, G.B., Schaeffer, L.R., 2008. Extent of linkage disequilibrium in Holstein cattle in North America. J. Dairy Sci. 91, 2106–2117.

Tibshirani, R., 1994. Regression Shrinkage and Selection Via the Lasso. CiteSeerX.

Toosi, A., Fernando, R.L., Dekkers, J.C.M., 2010. Genomic selection in admixed and crossbred populations. J ANIM SCI 88, 32–46.

Tsuruta, S., Misztal, I., Aguilar, I., Lawlor, T.J., 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. Journal of Dairy Science 94, 4198–4204.

Usai, M.G., Goddard, M.E., Hayes, B.J., 2009. LASSO with cross-validation for genomic selection. Genetics Research 91, 427–436.

VanRaden, P., Sullivan, P., 2010. International genomic evaluation methods for dairy cattle. Genetics Selection Evolution 42, 1–9.

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., 2009. Invited Review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92, 16–24.

Verbyla, K.L., Hayes, B.J., Bowman, P.J., Goddard, M.E., 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. Genet Res (Camb) 91, 307–311.

Weigel, K.A., Hoffman, P.C., Herring, W., Lawlor Jr., T.J., 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. Journal of Dairy Science 95, 2215–2225.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., Van Tassell, C.P., 2009. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. Journal of Dairy Science 92, 3431–3436.

Wiggans, G.R., VanRaden, P.M., Cooper, T.A., 2011. The genomic evaluation system in the United States: Past, present, future. Journal of Dairy Science 94, 3202–3211.

Yi, N., Xu, S., 2008. Bayesian LASSO for Quantitative Trait Loci Mapping. Genetics 179, 1045–1055.

Zhao, H., Nettleton, D., Soller, M., Dekkers, J.C.M., 2005. Evaluation of linkage disequilibrium measures between multi-allelic markers as predictors of linkage disequilibrium between markers and QTL. Genetics Research 86, 77–87.

4

The gradient boosting algorithm and random Boosting for genome-assisted evaluation in large data sets

O. González-Recio

J.A. Jiménez-Montero

R. Alenda

Published in Journal of Dairy Science, (2013) 96:614-624

The expectations raised by genomic selection have caused that many more individuals have already available genotypes. The different consortia created worldwide have provided genetic evaluation units with several thousand of genotyped individuals. This study proposes a machine-learning algorithm to implement routine genome-assisted evaluation in a feasible manner with reasonable computation times, with no impaired predictive ability.

Abstract

In the next few years, with the advent of high density SNPs arrays and genome sequencing, genomic evaluation methods will need to deal with a large number of genetic variants and an increasing sample size. The boosting algorithm is a machine learning technique that may alleviate the drawbacks of dealing with such large data sets. This algorithm combines different predictors in a sequential manner with some shrinkage on them, each predictor is applied consecutively to the residuals from the committee formed by the previous ones, to form a final prediction based on some subset of covariates. Here, a detailed description is provided, and examples using data toy are included. A modification of the algorithm called "random boosting" was proposed to increase the predictive ability and speed up computation time of genome- assisted evaluation in large data sets. The random boosting uses a random selection of markers to add a subsequent weak learner to the predictive model. These modifications were applied to a real data set composed by 1797 bulls genotyped for 39,714 SNPs. De-regressed proofs of four yield traits and one type trait from January 2009 routine evaluations were used as dependent variables. A 2-fold cross validation scenario was implemented. Sires born before 2005 were used as a training sample (1576 and 1562 for production and type traits, respectively), whereas younger sires were used as a testing sample to evaluate predictive ability of the algorithm on yet to be observed phenotypes. Comparison with the original algorithm was provided. The predictive ability of the algorithm was measured as Pearson correlation between observed and predicted responses. Further, estimated bias was computed as the average difference between observed and predicted phenotype.

The results showed that the modification of the original boosting algorithm can be run in 1% of the time used with the original algorithm, and with negligible differences in accuracy and bias. This modification may be used to speed up the calculus of genome-assisted evaluation in large data sets such us those obtained from consortiums.

Introduction

In the last years, several methods have been proposed to incorporate high density marker information in the genetic evaluations (Aguilar et al., 2010; Gianola et al., 2006; González-Recio et al., 2008; Meuwissen et al., 2001). These methods are based on either linear regression on the marker effects (e.g. Bayes B, Bayesian LASSO) or in genomic covariance between genotyped individuals (e.g. GBLUP, Single Step GBLUP). These methods are supposed to deal with the curse of dimensionality problem, although some concerns have been raised about their convenience to analyze high-dimensional data (Gianola et al., 2009). Non-parametric model from the machine learning repository have been proposed as an alternative in genome–assisted evaluations because they are able to extract hidden relationships from large, noisy and redundant data and do not follow a particular parametric design. For instances, reproducing kernel Hilbert spaces (González-Recio et al., 2008), Radial basis functions (Long et al., 2010), random forest (González-Recio and Forni, 2011) neural networks (Gianola et al., 2011) or the boosting algorithm (González-Recio et al., 2010) have already been implemented in this context. In general, previous results showed that non-parametric methods have similar or better predictive accuracy than regression on SNPs and genomic relationship matrices. Further, machine-learning methods are attractive and flexible for the implementation of genome-assisted evaluation using high-density SNP arrays. SNP chips include more and more SNPs, and sequence data may soon be available increasing the computation requirements. Thus, new strategies need to be developed to deal with reference population samples with a larger number of genotyped individuals with chips including an increasing number of SNPs.

The gradient boosting algorithm (BOOST) is an interesting alternative in a genome-assisted evaluation context when many more animals and markers are genotyped or sequenced, because it performs variable selection, uses simple regression models in an additive fashion and is computationally fast and easy. BOOST is a machine learning algorithm classified as an ensemble method. It was first proposed by Freund and Schapire (1996) for classification problems and was known as AdaBoost. Since then, it has been utilizes in many fields showing similar or higher predictive accuracy than traditional methods both in classification and regression problems. The boosting algorithm has been previously used in the genome wide prediction of genetic merit and disease susceptibility in animal breeding (González-Recio et al., 2010; González-Recio and Forni, 2011), and also showed similar or higher accuracy than other methods such as Bayes A or Bayesian LASSO. The algorithm uses a reference data set to find a predictive model which, given some genotype markers (e.g. SNP), predicts the most likely genetic merit for individuals yet to be observed. It does not assume any particular mode of inheritance or parametric model, and as commented above, is suitable to analyze very high-dimensional, redundant and fuzzy data like high-density SNP chips.

Nonetheless, BOOST, just as any other method used in a genome-assisted evaluation context, has yet to deal with the estimation of regression equation on markers when several thousand genotyped animals are used in the reference population (VanRaden et al., 2011), such as in the case of the EuroGenomics consortium in which more than 22,000 genotypes are already available as a reference population. These methods need to be adapted or modified to be implemented in the new era of genomic evaluations with many more genotypes and phenotypes, to predict genetic merit of young sires and cows in an accurate manner with minimum computer requirements.

The objective of this article is to provide a comprehensive description of the boosting algorithm in a genome-assisted genetic evaluation context, and to propose modifications thereof to deal with the larger number of genotypes and phenotypes in genomic evaluations.

The manuscript is organized as follows: first a brief description of ensemble methods is provided, then the gradient boosting algorithm is detailed in a genome-assisted evaluation context. The implementation of gradient boosting is illustrated in a toy example using two different base regression functions (ordinary least square and reproducing kernel Hilbert space regression). A modification of the algorithm is proposed for its implementation in the genome-assisted evaluation with many more phenotypes and genotypes. Finally, this modification is applied to a real data set and compared to the original BOOST. Comparison with other methods commonly used in this context is provided in a companion paper (Jiménez-Montero et al., 2013) in a real genomic evaluation problem.

Methods

Brief description of ensemble methods

Ensemble methods are a linear combination of some models instead of using a single fit of the model (Hastie et al., 2005; Seni and Elder, 2010), that can be expressed in the form:

$$\mathbf{y} = c_0 + c_1 h_1(\mathbf{y};\mathbf{X}) + c_2 h_2(\mathbf{y};\mathbf{X}) + ... + c_m h_m(\mathbf{y};\mathbf{X}) + ... + c_M h_M(\mathbf{y};\mathbf{X}) + \mathbf{e}$$

$$\mathbf{y} = c_0 + \sum_{m=1}^{M} c_m h_m(\mathbf{y};\mathbf{X}) + \mathbf{e}$$

Where $h_m(\mathbf{y};\mathbf{X})$ ($m \in \{1,...,M\}$) is some sort of model or function implemented on the phenotypes and genotypes in some specified manner, c_0 is the population mean and c_m ($m \in \{1,...,M\}$) are the coefficients or weights for each

model. Each model $h_m(\mathbf{y};\mathbf{X})$is usually called 'weak learner' because they are simple models that are supposed to perform slightly better than random guess. It is important to point out that little improvement would be gained with a strong learner and computation time would increase significantly. The ensemble methods form a "committee" of predictors with potentially greater predictive ability than that of any of the individual predictors. They became popular as a relatively simple device to improve the predictive performance of a base procedure. Random Forest, Bagging or boosting are examples of ensemble methods. They have been used in different fields and may be implemented in studies using large amount of genomic information.

Gradient boosting

Gradient boosting is considered as an ensemble method (Hastie et al., 2005). This algorithm combines different predictors in a sequential manner with some shrinkage on them (Friedman, 2000). It also performs variable selection.

Gradient boosting, as an ensemble method, may be described as follows:

$$\mathbf{y} = \mu + \sum_{m=1}^{M} v h_m(\mathbf{y};\mathbf{X}) + \mathbf{e}$$

Each predictor ($h_m(\mathbf{y};\mathbf{X})$ for $m \in (1, M)$) is added in a sequential manner, and is applied consecutively to the residuals from the committee formed by the previous ones, weighted by $c_{i\neq 0}$=v. This algorithm can be calculated using importance sampling learning ensembles as described next:

(Initialization): Given data $(\mathbf{y}, \mathbf{X})$, let the prediction of phenotypes be $F_0 = \bar{\mathbf{y}}$.

Then, for m in {1 to M}, with M being large, calculate the loss function (L) for $(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i;\mathbf{x}_i, p_m))$

where p_m is the SNP (only one SNP is selected at each iteration) that minimizes $\sum_{i=1}^{n} L(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, p_m))$ at iteration m; $h(y_i; \mathbf{x}_i, p_m)$ is the prediction of the observation using learner $h(\cdot)$ on SNP p_m. Selection of SNP p_m may be based on the minimization of the loss function $L(\cdot)$ in the training set or in a tuning set previously put aside in an n-fold cross-validation scenario.

Next, update the predictions at iteration m in the form $F_m(\mathbf{x}_i) = F_{m-1}(\mathbf{x}_i) + v \cdot h(y_i; \mathbf{x}_i, p_m)$ with $v \in (0,1)$ being some shrinkage factor, e.g. v=0.01.

Each subsequent model is trained on the residuals of the previous one, which are actually residual estimates ($\hat{\mathbf{e}}$). These $\hat{\mathbf{e}}$ are expected to be identical and independently distributed as $\hat{\mathbf{e}} \sim N(0, \sigma^2_{e_m})$, where $\sigma^2_{e_m}$ is the residual variance for model m. Therefore, the larger M the smaller $\sigma^2_{e_m}$. This means that for larger m, the contribution of the selected SNP at m is expected to be smaller. The shrinkage parameter v aims to control this trade off between number of models and importance of the SNPs. The smaller v is, the smaller explained variance is subtracted at each iteration, and therefore new (or the same) SNPs are allowed to explain the remaining residual variance.

Note that a large variety of learners ($h_m(\mathbf{y}; \mathbf{X})$) and loss functions ($L(y_i, F_m(\mathbf{x}_i))$) may be proposed, each of them leading to different boosting model. For instances, classification and regression trees, generalized least squares regression or non-parametric kernel regression may be used as weak learners. A quadratic error term, the exponential L_1 loss function, the Gini index or the Huber loss function are some examples of loss functions that may be implemented within the algorithm.

The choice of the number of iterations, *M*, is a model comparison problem which may be overcome in many different ways (Friedman, 2000; González-Recio et al., 2010; Hastie et al., 2005). This parameter may control the complexity of the ensemble and the overfitting caused in the training set. A simple manner of choosing *M* is stopping the algorithm when the decrease in error rate or mean squared error in a tuning set is not relevant during a large enough number of iterations (*e.g.* 100). Once the coefficient and the weak learners have been estimated, predictions for yet-to-be observed records may be calculated as:

$$\hat{y}_i = \hat{F}_m(\mathbf{x}_i) = \hat{\mu} + \sum_{m=1}^{M} v\hat{h}_m(\mathbf{x}_i) \cdot$$

More details on the gradient boosting can be found in Freund and Schapire, (1996), (Friedman, 2000) and its implementation on genomic prediction in (González-Recio et al. (2010).

Following below is a toy data example to describe the procedure to compute predicted genomic merit of genotyped individuals using two different weak learners: ordinary least square and RKHS regression.

Illustrations

Let $\mathbf{y} = [\,21.08 \;\; 16.13 \;\; 18.41 \;\; 20.50 \;\; 12.95]$ be the vector of observed phenotypes for *n*=5 individuals. Each individual is genotyped for *p*=3 SNPs codified as 0, 1 or 2 if they share 0, 1 or 2 copies of the most frequent allele in the population (as an arbitrary coding example). Let the corresponding $\mathbf{X}$ matrix be:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 0 \\ 0 & 2 & 1 \\ 2 & 0 & 1 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix}$$

The mean estimate $\bar{\mathbf{y}}$ for trait **y** is 17.81. The algorithm is initialized setting $F_0 = 17.81$ for all individuals. Note that other environmental effects may be included to adjust the phenotype. Let the loss function be the mean squared error, and the shrinkage coefficient *v*=0.9. This value is used only for illustration purpose, and a smaller value (e.g. *v*=0.10) is usually desired.

Illustration 1: ordinary least square

Suppose that the weak learner ($h(\cdot)$) is the ordinary least squares regression and the mean squared error (**MSE**) was assumed as loss function $L(\cdot)$. The ensemble will be constructed adding the results of several of these regressions.

The first model, *m*=1, is estimated as follows:

The heuristic search begins by trying *p*=3 models in the form

$\mathbf{y} = F_0 + h_1(\mathbf{y}; \mathbf{X}_s, s) + \mathbf{e}$, with $s \in \{1,2,p=3\}$

In this case $h(\cdot)$ is a linear regression on SNP*s*, and the model becomes

$\mathbf{y} = F_0 + a_1 + b_1\mathbf{X}_s + \mathbf{e}$, where $\mathbf{X}_s$ is the column vector of the genotype codes for SNP *s*.

For simplicity, here the model was solved by least squares estimates, although other estimators like Bayesian regression may be used. The solutions for each SNP would be:

For *s*=1: $\hat{a}_1$ =-1.137 and $\hat{b}_1 = 1.137$ with MSE=8.44;

For *s*=2: $\hat{a}_1$ =0.593 and $\hat{b}_1$ =-0.370with MSE=8.87;

For *s*=3: $\hat{a}_1$ =2.336 and $\hat{b}_1$ =-1.946with MSE=11.07;

Hence, SNP *s*=1 is selected because it was the one minimizing the MSE. The new estimates are $\hat{\mathbf{F}}_1 = \bar{\mathbf{y}} + v\hat{h}_1(\mathbf{y}, \mathbf{X})$, with $\hat{h}_1(\mathbf{y}, \mathbf{X}) = -1.137 + 1.137\mathbf{x}_{\cdot,1}$. The prediction for animal *i* becomes now:

$\hat{y}_i = \bar{\mathbf{y}} + v\hat{a}_1 + v\hat{b}_1 x_{i,1}$; with $\hat{a}_1$ =-1.137, $\hat{b}_1 = 1.137$ and $x_{i,1}$ being the genotype code of individual *i* for SNP 1.

Note that, again for simplicity, the SNP minimizing the MSE in the same data set was selected and used to estimate *a* and *b*. In a real scenario, a tuning set should be kept apart and the selected SNP could be the one minimizing the MSE in the tuning set with $\hat{a}$ and $\hat{b}$ estimated in the training set.

A second model *m*=2 is then added to the ensemble as:

$\mathbf{y} = \hat{F}_1 + h_2(\mathbf{y}; \mathbf{X}_s, s) + \mathbf{e}$; with $s \in \{1,2, p = 3\}$. The model may be written as $\mathbf{r}_1 = \mathbf{y} - \hat{F}_1 = h_2(\mathbf{y}; \mathbf{X}_s, s) + \mathbf{e}$, and the dependent variable in the second model are the residuals obtained from *m*=1:

$$\begin{bmatrix} 3.27 \\ -0.66 \\ -0.43 \\ 2.69 \\ -4.87 \end{bmatrix} = \begin{bmatrix} 21.08 \\ 16.13 \\ 18.41 \\ 20.50 \\ 12.95 \end{bmatrix} - \begin{bmatrix} 17.81 + 0.9 \cdot (-1.137) + 0.9 \cdot 1.137 \cdot 1 \\ 17.81 + 0.9 \cdot (-1.137) + 0.9 \cdot 1.137 \cdot 0 \\ 17.81 + 0.9 \cdot (-1.137) + 0.9 \cdot 1.137 \cdot 2 \\ 17.81 + 0.9 \cdot (-1.137) + 0.9 \cdot 1.137 \cdot 1 \\ 17.81 + 0.9 \cdot (-1.137) + 0.9 \cdot 1.137 \cdot 1 \end{bmatrix}$$

$\mathbf{r}_1 \quad = \quad \mathbf{y} \quad - \quad \hat{F}_1$

The heuristic search begins again by trying *p*=3 models in the form

$\mathbf{r}_1 = h_2(\mathbf{y}; \mathbf{X}_s, s) + \mathbf{e}$, with $s \in \{1,2, p = 3\}$

As before $h(\cdot)$ is a linear regression on SNP*s*, and the model for the heuristic search is then

$\mathbf{r}_1 = a_2 + b_2\mathbf{X}_s + \mathbf{e}$, where $\mathbf{X}_s$ is the column vector of the genotype codes for SNP *s*.

The solutions for each SNP would be:

For *s*=1: $\hat{a}_2$ =-0.114 and $\hat{b}_2$ =0.114with MSE=8.44;

For *s*=2: $\hat{a}_2$ =-0.431 and $\hat{b}_2$ =0.269with MSE=8.40;

For *s*=3: $\hat{a}_2$ =2.336 and $\hat{b}_2$ =-1.946with MSE=6.33;

The SNP *s*=3 is selected in *m*=2 because it was the one minimizing the MSE. The new estimates are $\hat{\mathbf{F}}_2 = \hat{\mathbf{F}}_1 + v\hat{h}_2(\mathbf{y},\mathbf{X})$, with $\hat{h}_2(\mathbf{y},\mathbf{X}) = 2.336 - 1.946\mathbf{x}_{\cdot,3}$. The prediction model for animal *i* becomes now:

$\hat{y}_i = \bar{\mathbf{y}} + v\hat{a}_1 + v\hat{b}_1 x_{i,1} + v\hat{a}_2 + v\hat{b}_2 x_{i,3}$; with $\hat{a}_2$ =2.336, $\hat{b}_2$ =-1.946 and $x_{i,3}$ being the genotype code of individual *i* for SNP 3.

Subsequently, more models are added by selecting one SNP at each model *m* after the heuristic search is done on residuals $\mathbf{r}_{m-1} = \mathbf{y} - \hat{F}_{m-1}$ until MSE converges. In this case, the algorithm converged at the second decimal (MSE=5.71) for *M*=14, and final predictions were

$\hat{y}_i = \bar{\mathbf{y}} + v\hat{a}_1 + v\hat{b}_1 x_{i,1} + v\hat{a}_2 + v\hat{b}_2 x_{i,3} + \ldots + v\hat{a}_m + v\hat{b}_m x_{i,m} + \ldots + v\hat{a}_M + v\hat{b}_M x_{i,M}$. The SNP selected at each iteration were [1,3,3,2,1,2,1,3,2,1,2,1,2,1]. The predicted genomic merits of individuals in the toy data set were:

$$\hat{\mathbf{y}} = \hat{\mathbf{F}}_{14} = \begin{bmatrix} 21.08 \\ 16.13 \\ 18.41 \\ 16.73 \\ 16.73 \end{bmatrix}$$

For generalization, it can be shown that the non-parametric genomic merit of any individual using ordinary least square regression as weak learner is

$\hat{y}_{\cdot} = \bar{\mathbf{y}} + v(\hat{a}_1 + \hat{b}_1\mathbf{x}_{\cdot s_1} + \hat{a}_2 + \hat{b}_2\mathbf{x}_{\cdot s_2} + ... + \hat{a}_m + \hat{b}_m\mathbf{x}_{\cdot s_m} ... + \hat{a}_M + \hat{b}_M\mathbf{x}_{\cdot s_M})$, with $\hat{a}_m$ and $\hat{b}_m$ being the intercept and slope coefficient estimates in model *m*, and $\mathbf{x}_{\cdot s_m}$ is the vector for the corresponding genotypes codes for SNP selected at model *m*.

Here, the intercept estimates can be added to compute a global intercept ($\hat{a}_T$) that may be interpreted as a bias corrector.

$\hat{a}_T = v(\hat{a}_1 + \hat{a}_2 + ... + \hat{a}_m + ... + \hat{a}_M)$.

Then, SNP contribution to the genomic merit ($\hat{\mathbf{b}}_T\mathbf{x}_{\cdot}$) of the individual may be expressed as:

$$\hat{\mathbf{b}}_T\mathbf{x} = v\hat{\mathbf{b}}\boldsymbol{\Lambda}\mathbf{x}$$

where $\hat{\mathbf{b}} = \left(\hat{b}_1, \hat{b}_2, ..., \hat{b}_m, ..., \hat{b}_M\right)$ is a row vector of *M* dimensions containing the slope estimates at each model $m \in \{1,...,M\}$, $\mathbf{x}_{\cdot px1} = \left(x_{\cdot 1}, x_{\cdot 2}, ..., x_{\cdot j}, ..., x_{\cdot p}\right)'$ is the column vector with the genotype codes of the individual for the *p* SNPs, and $\boldsymbol{\Lambda}$ is an indicator matrix (*M* x *p*) with each row $m \in \{1,...,M\}$ indicating the SNP selected in model *m*. Each row *m* contains zero for those positions where the corresponding SNP is not included in the model *m*, and '1' in the position of the SNP included in model *m*. Hence, the non-parametric prediction of the genomic breeding value of a given individual would be:

$\hat{y}_{egbv} = \hat{a}_T + \hat{\mathbf{b}}_T\mathbf{x}_{\cdot}$

The global coefficient estimate ($\hat{b}_{T_j}$) for SNP *j* is the sum of the slope estimates in the model in which the SNP *j* was selected, as $\hat{\delta}_T = v\sum_{m=1}^{M} \Lambda_m \hat{\delta}_m$, where Λ_m is an indicator function equal to 1 if the SNP is selected at model *m* and 0 otherwise, and $\hat{b}_m$ is the slope estimate from model *m*.

It is clear that $\hat{\mathbf{b}}_T$ is a row vector containing the global coefficient estimates for each SNP in the form $\hat{\mathbf{b}}_T = v\hat{\mathbf{b}}\boldsymbol{\Lambda} = v\left(\hat{\delta}_{T_1}, \hat{\delta}_{T_2}, \ldots, \hat{\delta}_{T_j}, \ldots, \hat{\delta}_{T_p}\right)$.

It must be pointed out that although each $\hat{b}_m$ is calculated from a linear function, the sum of all $\hat{b}_m$ lacks of a linear interpretation as each of them is calculated from previously corrected phenotypes.

Predictions of new genomic breeding values for young genotyped individuals can be easily calculated using the regression equations obtained as described above.

Illustration 2: kernel regression or RKHS

Assume now that the weak learner ($h(\cdot)$) is a non-parametric regression (kernel or RKHS) as described in Kimeldorf and Wahba, (1971):

$$\mathbf{K}'\mathbf{y} = [\mathbf{K}'\mathbf{K} + \lambda\mathbf{K}]\boldsymbol{\alpha},$$

where $\mathbf{y}$ is the vector of phenotypes, $\mathbf{K} = \{k_{i,j}\}$ is a *n*x*n* matrix of kernels, λ is a smoothing parameter that may be interpreted as the variance explained by the kernel matrix, and $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_i, \ldots, \alpha_n]$ is a column vector of *n* non-parametric coefficients.

Following the reparametrization I in (De los Campos et al., 2009), the model equation can be written as follows:

$\left[\mathbf{K}+\frac{\sigma_e^2}{\lambda^{-1}}\mathbf{I}\right]\boldsymbol{\alpha}=\mathbf{y}$, with σ_e^2 being some residual variance. For equivalences between RKHS and BLUP see (De los Campos et al., 2009). Both the residual variance and λ must be estimated in a RKHS scenario. Maximum likelihood or Bayesian estimates from these parameters may be obtained using standard procedures. Further, the model may be simplified using a kernel regression model as that described in Gianola et al. (2006) without needing the estimation of these parameters, using the Nadaraya-Watson estimator (Nadaraya, 1964; Watson, 1964). Here, for convenience, a RKHS model is proposed but the ratio between σ_e^2 and λ was assumed equal to 1.

Then, a kernel matrix must be constructed for each SNP. Each matrix ($\mathbf{K}^s, s\in\{1,2,p=3\}$) must be semi-positive definite and contains the set of quantitative values representing genomic similarities between pairs of individuals ($k_{i,j}^s$) at a given locus *s*. A large variety of kernels have proved to be useful for genomic data (González-Recio et al., 2009, 2008; Schaid, 2010). Here, again for simplicity, the allele match kernel was used as illustration, the kernel score assays the number of common alleles between the locus *s*of two individuals *i* and *j*. The score is 4 if the genotypes of the individuals are the same; 2 if one is a heterozygote and the other is a homozygote, and 0 if they don't share any common allele (*i.e.* molecular relationship).

Therefore, the matrix $\mathbf{K}^s=\left\{k_{i,j}^s\right\}$ for each SNP $s\in\{1,2,p=3\}$ would be:

$$\mathbf{K}^1=\begin{bmatrix}4&2&2&4&4\\2&4&0&2&2\\2&0&4&2&2\\4&2&2&4&4\\4&2&2&4&4\end{bmatrix};\ \mathbf{K}^2=\begin{bmatrix}4&4&0&4&4\\4&4&0&4&4\\0&0&4&0&0\\4&4&0&4&4\\4&4&0&4&4\end{bmatrix};\ \mathbf{K}^3=\begin{bmatrix}4&2&2&0&0\\2&4&4&2&2\\2&4&4&2&2\\0&2&2&4&4\\0&2&2&4&4\end{bmatrix}$$

As described in Gianola et al. (2006), González-Recio et al. (2008) and De los Campos et al. (2009), the predicted genomic breeding merit of the individuals may be computed as $\hat{\mathbf{y}} = \mathbf{K}^{s*}\hat{\boldsymbol{\alpha}}$, where $\hat{\boldsymbol{\alpha}}$ are the non-parametric coefficient estimates, and $\mathbf{K}^{s*}$ is an *m*x*n* kernel matrix containing the genomic similarities at a given locus *s* between the *n* individuals with records and the *m* individuals whose genomic merit we aim to predict. For instance, if we aim to predict the genomic merit of the individuals with records $\mathbf{K}^{s*} = \mathbf{K}^{s}$, otherwise $\mathbf{K}^{s*} = \left\{k_{i,j}^{s*}\right\}$ with $k_{i,j}^{s*}$ being the genomic similarity between each individual without phenotype (*i*) and those with phenotypes (*j*).

Coming back to the toy data set, the predictions for yet-to-be observed records may be calculated as:

$\hat{y}_i = \hat{F}_m(\mathbf{x}_i) = \bar{\mathbf{y}} + \sum_{m=1}^{M} v\hat{h}_m(\mathbf{x}_i)$. In this case, the weak learner is a RKHS regression as described above ($[\mathbf{K}^{s*} + \mathbf{I}]\boldsymbol{\alpha} = \mathbf{y}$).

As before, $\mathbf{y} = [\,21.08 \;\; 16.13 \;\; 18.41 \;\; 20.50 \;\; 12.95]$ was the vector of observed phenotypes for *n*=5 individuals. And the algorithm was initialized setting $F_0 = 17.81$ for all individuals. Again, assume that the loss function is the mean squared error, and the shrinkage coefficient *v*=0.9.

The first model, *m*=1, is estimated as follows:

The heuristic search begins by trying *p*=3 models in the form

$\mathbf{y} = F_0 + h_1(\mathbf{y}; \mathbf{K}^{s}, s) + \mathbf{e}$, with $s \in \{1,2, p = 3\}$

With $h(\cdot)$ being the RKHS with $\mathbf{K}^{s}$ as kernel matrix. The model is

$\mathbf{y} = \mathbf{F}_0 + \mathbf{K}^s\boldsymbol{\alpha} + \mathbf{e}$, where $\mathbf{K}^s$ is the kernel matrix corresponding to locus *s*, and $\boldsymbol{\alpha}$ is the vector of non parametric coefficients for model *m*=1.

The solutions for each SNP would be:

For *s*=1: $\hat{\boldsymbol{\alpha}}_1 = [3.00 \; -0.45 \; 0.00 \; 2.42 \; -5.13]$ with a mean squared error (MSE)=8.29;

For *s*=2: $\hat{\boldsymbol{\alpha}}_1 = [3.41 \; -1.54 \; 0.12 \; 2.83 \; -4.73]$ with a mean squared error (MSE)=8.88;

For *s*=3: $\hat{\boldsymbol{\alpha}}_1 = [0.85 \; -1.38 \; 0.89 \; 3.77 \; -3.79]$ with a mean squared error (MSE)=6.40;

The SNP *s*=3 has produced the smallest MSE and was therefore selected in this case. The new estimates are $\hat{\mathbf{F}}_1 = \bar{\mathbf{y}} + v\hat{h}_1(\mathbf{y}, \mathbf{K}^3)$, with $\hat{h}_1(\mathbf{y}, \mathbf{K}^s) = \mathbf{K}^s\hat{\boldsymbol{\alpha}}_1$. The prediction for animal *i* now become:

$\hat{y}_i = \bar{\mathbf{y}} + v\mathbf{k}^3_{i,\cdot}\hat{\alpha}_1$; with *v* being the shrinkage coefficient, $\hat{\boldsymbol{\alpha}}_1 = [0.85 \; -1.38 \; 0.89 \; 3.77 \; -3.79]$ and $\mathbf{k}^3_{i,\cdot} = \{k^3_{i,\cdot}\}$ containing the vector with the genomic similarities between the individual *i* and each individual with record at locus 3.

As for the OLS learner, the SNP minimizing the MSE in the same data set as the one used to estimate *a* and *b* was selected, but a tuning set may be used as stated previously.

A second model *m*=2 is then added to the ensemble as:

$\mathbf{y} = \hat{F}_1 + h_2(\mathbf{y}; \mathbf{X}_s, s) + \mathbf{e}$; with $s \in \{1,2,p=3\}$. The model may be written as $\mathbf{r}_1 = \mathbf{y} - \hat{F}_1 = h_2(\mathbf{y}; \mathbf{X}_s, s) + \mathbf{e}$, and the dependent variable in the second model are the residuals obtained from *m*=1:

$$\begin{bmatrix} 1.09 \\ -1.41 \\ 0.86 \\ 3.66 \\ -3.90 \end{bmatrix} = \begin{bmatrix} 21.08 \\ 16.13 \\ 18.41 \\ 20.50 \\ 12.95 \end{bmatrix} - \begin{bmatrix} 17.81 + 0.9 \cdot 2.42 \\ 17.81 + 0.9 \cdot (-0.30) \\ 17.81 + 0.9 \cdot (-0.30) \\ 17.81 + 0.9 \cdot (-1.08) \\ 17.81 + 0.9 \cdot (-1.08) \end{bmatrix}$$

$$\mathbf{r}_1 = \mathbf{y} - (\hat{F}_0 + \hat{F}_1)$$

The heuristic search begins again by trying p=3 models in the form

$\mathbf{r}_1 = h_2(\mathbf{y}; \mathbf{K}^s, s) + \mathbf{e} = \mathbf{K}^s \alpha_1 + \mathbf{e}$, with $s \in \{1,2,p=3\}$

The solutions for each SNP would be:

For s=1: $\hat{\alpha}_2 = [0.87 \ -0.36 \ 0.10 \ 3.44 \ -4.12]$ with a mean squared error (MSE)=5.93;

For s=2: $\hat{\alpha}_2 = [1.22 \ -1.28 \ 0.17 \ 3.79 \ -3.77]$ with a mean squared error (MSE)=6.34;

For s=3: $\hat{\alpha}_2 = [0.31 \ -1.25 \ 1.02 \ 3.82 \ -3.74]$ with a mean squared error (MSE)=6.25;

The SNP s=1 is selected in m=2 because it was the one minimizing the MSE. The new estimates are $\hat{\mathbf{F}}_2 = \hat{\mathbf{F}}_1 + \nu \hat{h}_2(\mathbf{y}, \mathbf{K}^1)$, with $\hat{h}_2(\mathbf{y}, \mathbf{K}^1) = \mathbf{K}^1 \hat{\alpha}_2$. The prediction for animal i now become:

$\hat{y}_i = \bar{\mathbf{y}} + \nu \mathbf{k}^3_{i,\cdot} \hat{\alpha}_1 + \nu \mathbf{k}^1_{i,\cdot} \hat{\alpha}_2$; with $\hat{\alpha}_2 = [0.87 \ -0.36 \ 0.10 \ 3.44 \ -4.12]$ and $\mathbf{k}^1_{i,\cdot} = \{k^1_{i,\cdot}\}$ containing the vector with the genomic similarities between the individual i and each individual with record at locus 1.

As described previously, subsequent models are added to the residuals of the previous ensemble until a convergence criterion is reached. In this case, the algorithm converged at the second decimal in the MSE (=5.71) for M=7. The

SNP selected at each iteration were [3,1,3,1,1,3,1]. The predicted genomic merits of individuals in the toy data set were:

$$\hat{\mathbf{y}} = \hat{\mathbf{F}}_7 = \begin{bmatrix} 21.08 \\ 16.16 \\ 18.38 \\ 16.72 \\ 16.72 \end{bmatrix}.$$

For generalization, it can be shown that the non-parametric genomic merit of any individual using RKHS as weak learner is

$$\hat{y} = \bar{y} + v\left(\mathbf{k}_{i,\cdot}^{s_1}\hat{\alpha}_1 + \mathbf{k}_{i,\cdot}^{s_2}\hat{\alpha}_2 + \cdots + \mathbf{k}_{i,\cdot}^{s_m}\hat{\alpha}_m + \cdots + \mathbf{k}_{i,\cdot}^{s_M}\hat{\alpha}_M\right) = \bar{y} + v\sum_{m=1}^{M}\mathbf{k}_{i,\cdot}^{s_m}\hat{\alpha}_m,$$

with $\hat{\alpha}_m$ being the non parametric coefficient estimates at model *m*, and $\mathbf{k}_{i,\cdot}^{s_m} = \{k_{i,\cdot}^{s_m}\}$ the vector containing the genomic similarities between the individual *i* and each individual with record at the locus selected at model *m*. Hence, if the $\hat{\alpha}_m$ are estimated at each model using the residuals of the previous model they will differ between models, whereas the **K** matrix remains constant. Hence, if the phenotype of a new individual has to be predicted, the non-parametric coefficient estimates and the pairs of the genomic similarity between it and the individuals with observation should be computed once and electronically stored. A single text file may be stored for each individual containing the genomic similarity at each marker position with each individual in the reference population. The algorithm does not need to be run again, and the predictive equations can be computed in a straightforward manner, as with linear regression models.

Modification of the boosting algorithm

Randomboosting. The purpose of this modification is basically to speed up the algorithm for large data sets or too time consuming learners. We propose

to sample *mtry* covariates at random out of the *p* SNPs at each iteration, and select the SNP among the *mtry* that minimizes the given loss function. Therefore, computation time may be reduced in the order of $o\left(\frac{mtry}{p}\right)$ regarding the original algorithm, as only a small percentage of SNPs are tested for minimization of the loss function at each iteration. The parameter *mtry* may be tuned in the Random boosting modification. Studies of similar strategies used in the Random Forest algorithm showed that a value for *mtry* of 0.1**p* may achieve satisfactory results (Goldstein et al., 2010).

The boosting algorithm with this modification would flow as follows:

(Initialization): Given data $\Psi = (\mathbf{y}, \mathbf{X})$, let the prediction of phenotypes be $\hat{F}_0 = \hat{\mu}$.

Then, for *m* in {1 to *M*}, with *M* being large, proceed as:

Step 1. Draw *mtry* out of *p* covariates from the original training set to construct a reduced training covariate matrix $\Psi^{(b)} = (\mathbf{y}, \mathbf{X}_{mtry})$ to train the algorithm in iteration *m*.

Step 2. Calculate the loss function $L\left(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, mtry_m)\right)$ for all *mtry* SNPs and select that minimizing $\sum_{i=1}^{n} L(y_i, F_{m-1}(\mathbf{x}_i) + h(y_i; \mathbf{x}_i, mtry_m))$ in the tuning set at iteration *m*, with $h(y_i; \mathbf{x}_i, mtry_m)$ being the prediction of the observation *i* in the tuning set using the learned parameters or coefficients of $h(\cdot)$ on the SNP $mtry_m$. These parameters or coefficients are learned using the training set as in the original algorithm. Note that if the resulting tuning set is not large enough, it may be recommended to select the SNP minimizing the loss function in the training set, without leaving a set aside set.

Step 3. Updated predictions at iteration *m* in the form $F_m(\mathbf{x}_i) = F_{m-1}(\mathbf{x}_i) + v \cdot h(y_i; \mathbf{x}_i, mtry_m)$ with *v* being some shrinkage factor, e.g. *v*=0.10.

Step 4. Update the residuals to be used in the next iteration as $y_i = y_i - F_m(\mathbf{x}_i)$.

Repeat steps 1 to 4 a large number of times (*M*).

This modification causes that the order in which SNPs are selected in the algorithm change regarding the original boosting, as not all SNPs will be tested at each iteration. However, the boosting algorithm is considered as a small step gradient descent technique (Bühlmann, 2006), therefore, for a sufficient small *v*, it is expected that the impact of the order in which the covariates are used to reduce the residual estimates has small or null effect on the final predictions. Nonetheless, note that small data set might yield different results for smaller mtry and less number of iterations.

CASE STUDY

Data

The algorithm and the proposed modification were implemented in a real data set composed by 1859 genotyped bulls. Full details on genotypes and the edition procedure can be found in Jiménez-Montero et al. (companion paper). After quality control 39,714 SNPs were kept in the analyses. Sires born before 2005 were used as a training sample (1601 and 1574individuals for production and type traits, respectively), whereas younger sires were used as a testing sample to evaluate predictive ability of the algorithm on yet to be observed phenotypes. De-regressed proofs (DRP) of four productive traits (milk yield (**MY**), fat yield (**FY**), and protein yield (**PY**) and fat percentage (**FP**)) and one type trait (udder depth (**UD**)) from January 2009 routine evaluations were used as dependent variables. The DRPs were

obtained following Jairath et al. (1998). Note that bulls in the testing set did not have progeny test proofs at that time. For convenience, the ordinary least square regression and the MSE were chosen as weak learner and loss function, respectively, as set up in the illustration example number one above. A 10-fold cross validation scenario was implemented in the training set. In each fold, 9/10 of the training data were used to calculate the regression coefficient estimates ($\hat{a}_m$ and $\hat{b}_m$), and the remaining 1/10 records were used as a tuning set to choose the SNP minimizing the MSE.

The respective DRP from the December 2011 routine evaluations were used to calculate the predictive ability of the predictions for sires in the testing set. Only sires with more than 20 effective daughter contribution were kept in the testing set (258 and 235 for production and type, respectively).The predictive accuracy was evaluated using Pearson correlation between predicted and observed (December 2011 DRP) response. The predicted bias was also calculated as $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)$, with *n* being the number of validation bulls.

The random boosting was applied to this data using a grid of values for *mtry* (1%, 5%, 10% and 50%), and was compared to the original boosting (*mtry*=100%). Further, different values for the smoothing parameter were tested (v=0.01, 0.10, and 0.20).

Results

Tables 4.1 and 4.2 show the Pearson correlation and bias, respectively, between predicted and observed phenotype in the testing set, regarding the smoothing parameter v and *mtry* for each trait. In general, the predictive ability of the algorithm was very similar regardless *mtry*, with differences of 1-2 points in Pearson correlation. Fat percentage showed better predictive

ability at larger *mtry* values. The known major genes (e.g. DGAT1), controlling this trait may partly explain this behavior, as sampling a small proportion of SNPs at each iteration may miss markers in these hot spots, hampering the predictive ability of the algorithm. Pearson correlation for *v*=0.10 and 0.20 were very similar, although *v*=0.10 showed equal or higher Pearson correlation than *v*=0.20 in all the analyses, excepting for UD with *mtry* equal to 5 and 10%.In terms of bias, the value of *mtry* did not show a clear trend, and differences were negligible. Convergence was slower for smaller values of *v*, because higher shrinkage is done on each coefficient estimate and a larger number of covariates are needed to explain the variance of the observed phenotypes. Nonetheless, the best combination of *v* and *mtry* was trait dependent. As a general recommendation, the random boosting algorithm may be used to speed up the calculus of genome- assisted evaluation without a relevant impact on the predicted ability, and in some cases with higher Pearson correlation between predicted and observed phenotypes in the testing set than using the original algorithm. Smaller values of *mtry* may be used without decreasing the predictive ability and with a significant reduction in the computation time. Nonetheless, *mtry* is genetic architecture dependent, and a large value is recommended to analyze traits with known major genes, as in the case of fat percentage. The choice of *mtry* and *v* is under discussion, and cross validation is currently the standard procedure. A more formal strategy with statistical properties could be studied in the future.

Table 4.1. Pearson correlation[1] between predicted and observed responses in the testing set using the original gradient boosting algorithm (mtry=100%) or its modified version "Random Boosting", for different values of percentage of SNPs sampled at each iteration (mtry) and smoothing parameter (v)

		mtry(%)				
	v	1	5	10	50	100
kg MILK	0.01	0.495	0.502	**0.508**	0.507	0.507
	0.10	0.487	0.500	0.503	**0.508**	0.503
	0.20	0.483	0.503	0.503	0.501	0.504
Kg FAT	0.01	0.552	0.561	0.559	0.559	0.559
	0.10	0.567	0.565	**0.569**	0.556	0.556
	0.20	0.551	0.554	0.562	0.550	0.551
Kg PROT	0.01	0.454	0.443	0.440	0.443	0.443
	0.10	**0.466**	0.441	0.445	0.444	0.444
	0.20	0.465	0.437	0.429	0.434	0.428
%FAT	0.01	0.746	0.753	0.748	0.763	**0.768**
	0.10	0.741	0.746	0.748	0.761	0.765
	0.20	0.728	0.737	0.740	0.753	0.767
UDDER DEPTH	0.01	0.496	0.504	0.502	0.509	0.503
	0.10	0.496	0.502	0.507	0.505	0.505
	0.20	0.490	0.505	**0.510**	0.502	0.507

[1] Highest value for each trait is in bold.

Table 4.2. Estimated bias[1] (measured as average difference between predicted and observed responses in standard deviation units) in the testing set using the original gradient boosting algorithm (mtry=100%) or its modified version "Random Boosting", for different values of percentage of SNPs sampled at each iteration (mtry) and smoothing parameter (v)

		mtry(%)				
	v	1	5	10	50	100
kg MILK	0.01	-0.040	-0.047	-0.037	-0.039	-0.039
	0.10	-0.044	-0.042	-0.041	-0.035	-0.038
	0.20	-0.032	-0.014	**-0.008**	-0.029	-0.026
Kg FAT	0.01	-0.113	-0.107	-0.104	-0.104	-0.104
	0.10	-0.121	-0.107	**-0.090**	-0.100	-0.099
	0.20	-0.095	-0.114	-0.103	-0.092	-0.095
Kg PROT	0.01	-0.049	-0.061	-0.071	-0.067	-0.070
	0.10	-0.029	0.062	-0.046	-0.047	-0.058
	0.20	**-0.025**	-0.034	-0.056	-0.068	-0.075
%FAT	0.01	0.039	0.051	0.053	0.046	0.045
	0.10	0.030	0.053	0.053	0.042	0.040
	0.20	0.032	0.048	0.055	**0.010**	0.041
UDDER DEPTH	0.01	-0.234	-0.233	-0.232	-0.219	-0.238
	0.10	**-0.217**	-0.226	-0.232	-0.233	-0.231
	0.20	-0.219	-0.220	-0.229	-0.241	-0.234

[1] Lowest value for each trait is in bold.

The original gradient boosting algorithm performed the complete genome-assisted evaluation (10-folds) in 171.67 hours with *v*=0.01, 69.17 hours with *v*=0.10 and 50 hours with *v*=0.20 (Table 4.3). The computation time was substantially reduced using the modification of the algorithm with *mtry* =0.01. The smaller times were 1.5, 0.83 and 0.67 hours for *mtry*=0.01 and *v*=0.01, *v*=0.10 and *v*=0.20, respectively. These computing times make Random boosting feasible for running frequent routine genome-assisted evaluations with large data sets without impairing the predictive accuracy.

Note that the parallelization of the code can be implemented at step 2 described above, when searching for the SNP minimizing the loss function. The parallelization would drastically decrease the computation time of the algorithm (not implemented in this study).

Table 4.3. Computation time[1] (in hours) to run 10-fold cross validations (a complete genomic assisted evaluation cycle) regarding the value of the smoothing parameter (*v*) and the proportion of SNPs sampled at each iteration (*mtry*)

Smoothing parameter (*v*)	*mtry*				
	1%	5%	10%	50%	100%[2]
v=0.01	1.50	8.33	16.33	86.33	171.67
v=0.10	0.83	3.34	6.67	35.00	69.17
v=0.20	0.67	2.83	5.33	25.00	50.00

[1]In an Intel Xeon QuadCore E5430 (4x2.66Ghz) proccesor with 8Gb RAM memory under Linux operating system.
[2]This value of *mtry* is equivalent to the original gradient boosting.

Concluding remarks

Incorporating high-density markers into models for prediction of genetic values poses important statistical and computational challenges. Machine learning algorithms can be used to deal with the curse of dimensionality and computational limitations when a large number of individuals have genotypic information. In particular, the boosting algorithm provides an efficient strategy to calculate additive genomic breeding values using high density SNP information. We have provided here a comprehensive description of the mechanisms of the algorithm and showed that it can be viewed as an additive gradient descent method that may be implemented as a SNP regression model. A modification of the algorithm has been also proposed to speed up computation of genomic breeding values, with a minimum impact in the predictive ability. The companion study by Jiménez-Montero et al. (*Companion paper*) provides comparison of boosting and random boosting with other methods commonly used in the genome-assisted evaluations.

Acknowledgments

The authors acknowledge funds from the project CDTI-P080250866 UPM and the agreement INIA-CC10-046, and to CONAFE, ASCOL, ABEREKIN, XENETICA FONTAO and GENETICAL for providing genotypes and phenotypes used in this study.

Referentes

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., Lawlor, T.J., 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science 93, 743–752.

Bühlmann, P., 2006. Boosting for High-Dimensional Linear Models. The Annals of Statistics 34, 559–583.

De los Campos, G. de los, Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting Quantitative Traits With Regression Models for Dense Molecular Markers and Pedigree. Genetics 182, 375–385.

Freund, Y., Schapire, R., 1996. Experiments with a New Boosting Algorithm. Presented at the International Conference on Machine Learning, pp. 148–156.

Friedman, J.H., 2000. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189–1232.

Gianola, D., Campos, G. de los, Hill, W.G., Manfredi, E., Fernando, R., 2009. Additive Genetic Variability and the Bayesian Alphabet. Genetics 183, 347–363.

Gianola, D., Fernando, R.L., Stella, A., 2006. Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics 173, 1761–1776.

Gianola, D., Okut, H., Weigel, K., Rosa, G., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genetics 12, 87.

González-Recio, O., Forni, S., 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genetics Selection Evolution 43, 1–12.

González-Recio, O., Gianola, D., Long, N., Weigel, K.A., Rosa, G.J.M., Avendaño, S., 2008. Nonparametric Methods for Incorporating Genomic Information Into Genetic Evaluations: An Application to Mortality in Broilers. Genetics 178, 2305–2313.

González-Recio, O., Gianola, D., Rosa, G., Weigel, K., Kranis, A., 2009. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. Genetics Selection Evolution 41, 3.

González-Recio, O., Weigel, K.A., Gianola, D., Naya, H., Rosa, G.J.M., 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. Genetics Research 92, 227–237.

Hastie, T., Tibshirani, R., Friedman, J., Franklin, J., 2005. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer 27, 83–85.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B., Kolstad, B., 1998. Genetic Evaluation for Herd Life in Canada. Journal of Dairy Science 81, 550–562.

Jiménez-Montero, J.A., González-Recio, O., Alenda, R., 2012. Genotyping strategies for genomic selection in small dairy cattle populations. animal 6, 1216–1224.

Jiménez-Montero, J.A., González-Recio, O., Alenda, R., n.d. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. Journal of Dairy Science 96, 625–634.

Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline functions. Journal of Mathematical Analysis and Applications 33, 82–95.

Long, N., Gianola, D., Rosa, G.J.M., Weigel, K.A., Kranis, A., GonzáLez-Recio, O., 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. Genetics Research 92, 209–225.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Nadaraya, E.A., 1964. On Estimating Regression. Theory of Probability & Its Applications 9, 141–142.

Schaid, D.J., 2010. Genomic Similarity and Kernel Methods II: Methods for Genomic Information. Human Heredity 70, 132–140.

Seni, G., Elder, J., 2010. Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions. Morgan and Claypool Publishers.

VanRaden, P., O'Connell, J., Wiggans, G., Weigel, K., 2011. Genomic evaluations with many more genotypes. Genetics Selection Evolution 43, 1–11.

Watson, G.S., 1964. Smooth Regression Analysis. Sankhyā: The Indian Journal of Statistics, Series A (1961-2002) 26, 359–372.

5

Predictive ability of dairy cattle genotypes imputed from different density platforms

J. A. Jiménez-Montero

D. Gianola

K. Weigel

R. Alenda

O. González-Recio

Submitted to Journal of Dairy Science

The imputation of genotypes from lower to higher density platforms is an essential tool to optimize genomic selection programs. Imputation from 3K and 6K assays to 50K density and later to HD is performed with the aim to compare predictive ability and selection efficiency of imputed genotypes. This study considers the performance of these sets of data in terms of their efficiency when used for selection of top animals in a dairy cattle population.

Abstract

The aim of this study was to evaluate different density genotyping platforms for genotype imputation and genomic prediction. Genotypes from customized Golden Gate Bovine3K BeadChip (**LD3K**) and BovineLD BeadChip (**LD6K**) platforms were imputed to BovineSNP50v2 BeadChip (**50K density**). In addition, LD3K, LD6K and 50K genotypes were imputed to BovineHD BeadChip (**HD**) 800K density, and subsequently evaluated and compared. Comparisons of prediction accuracy were carried out using Random Boosting (**R-Boost**) and Genomic BLUP (**G-BLUP**) algorithms. Four traits under selection in the Spanish Holstein population were used: milk yield (**MY**), fat percentage (**FP**), somatic cell count (**SCC**), and days open (**DO**). Training sets at 50K density for imputation and prediction included 1632 genotypes. Testing sets for imputation from LD to 50K contained 834 genotypes while testing sets for genomic evaluation included 383 bulls. The reference population genotyped at HD included 192 bulls. Imputation using Beagle software was efficient for the reconstruction of dense 50K and HD genotypes, even when a small reference population was used.

R-Boost out-performed G-BLUP in terms of prediction accuracy, mean squared error and selection efficiency of top animals when FP was considered; for other traits otherwise there were no clear differences between methods.

Overall prediction accuracy was 2% greater for LD6K than LD3K after imputation to 50K density. No differences were found between imputed LD and 50K genotypes, while evaluation of HD genotypes was on average 4% more accurate than 50K prediction. Similar bias in regression coefficients were found across data sets, but notably coefficients were 0.32 units closer to unity for DO when genotypes were imputed to HD density. The HD

predictions also resulted in smaller MSE when compared to 50K predictions in three out of four traits.

Regarding selection efficiency of top animals, more (2%) top bulls were classified correctly with imputed LD6K genotypes as compared with LD3K case. When original 50K genotypes were used, correct classification of top bulls increased by 1%, and when those genotypes were imputed to HD 3% more top bulls were detected (3%).

Differences in performance between 3KLD and 6KLD genotypes were more noticeable for imputation accuracy than for prediction ability. In general, genotypes imputed from LD performed similarly to those obtained for the animals originally genotyped at 50K. However, selection efficiency could be slightly enhanced for certain traits like FP, SCC or DO when genotypes are imputed to HD.

Introduction

Genomic selection (GS) in dairy cattle started in December 2006 (De Roos et al., 2009), when high-density single nucleotide polymorphism (SNP) panels became affordable for non-human applications (Tassell et al., 2008). The first official direct genomic values (DGV) were provided to dairy farmers in January 2009 (Wiggans et al., 2009b). Despite the improvement in reliability of young selection candidates achieved after genotyping and genomic evaluations (Wiggans et al., 2011),the commercial price of high density SNP chips may limit their use to males and elite females in many populations. The next key objective in GS programs is to optimize of the use of genomic information (Pryce and Daetwyler, 2012). Low density SNP panels and posterior imputation is a promising way to reduce genotyping costs; because these low density genotyping platforms could greatly increase the number of genotyped animals in commercial dairy herds. The optimal size of such platforms depends on population characteristics, such as the extend of linkage disequilibrium, genetic architecture of traits under selection, number and proportion of animals with high-density SNP genotypes, and relationships between these animals and the future candidates (Weigel et al., 2010a). Imputation methods work by combining data from a reference panel of individuals genotyped at a dense set of polymorphic sites (usually SNPs) with data from a study sample collected from a genetically similar population and genotyped at a subset of these sites (Howie et al., 2009). There is a need to integrate different density SNP panels in the genomic breeding programs for genomic evaluation. For accurate imputation of missing SNPs, the reference population must include a sufficient number of individuals with good representation of the whole population's SNP frequencies (Hao et al., 2009). Imputation accuracy is also related to the degree of relationship between the reference population and the animals to be imputed (Meuwissen and Goddard, 2010). Theoretically, heavily

represented bulls in the population, or animals from the most common matings (sire x maternal grandsire) are optimum animals to be genotyped as the reference population. High density genotypes (50K) can be imputed with accuracy above 90% from a low density (2-4K) genotypes (Weigel et al., 2010b).

Currently, the most commonly used chip for cattle is the BovineSNP50.v2 Beadchip (Illumina Inc., San Diego, CA) and imputation strategies are focused on imputation from 3K and 6K to 50K. Recently, the availability of the 800K SNP BovineHD BeadChip (Illumina Inc.) offers the option for imputation from 50K to this high density platform. Genotyping a large reference population at extra large high density could be cost prohibitive. However, genotyping a subset of this reference population and then imputing the rest of the genotypes may be an efficient strategy if the predictive ability of subsequent genomic evaluations exceed that obtained before imputation. In addition, imputed SNPs from low density 3K and 6K platforms to high density must be assessed in terms of predictive ability.

Several methods have been developed for imputation, and software is publicly available for these methods, (Howie et al., 2009; Kong et al., 2008; Scheet and Stephens, 2006). Beagle (Browning and Browning, 2009) has become one of the preferred options for imputation of large data sets (Boichard et al., 2012; Erbe et al., 2012). This software uses a hidden Markov model to infer haplotype phase with both typed and untyped SNPs. Its competitive imputation accuracy and computational advantages when compared with other methods have been widely reported (Calus et al., 2011; Nothnagel et al., 2009; Segelke et al., 2012; Sun et al., 2012).

After imputation, is possible to estimate DGV with similar accuracy as that obtained from high density genotyping (Berry and Kearney, 2011), Accuracy of DGV for selection candidates can be increased by imputation compared

with estimation based on low density SNPs, and recent studies have shown that low density animals with enough phenotypic information, can be added to the reference population after imputation to increase the overall accuracy of estimation (Weigel et al., 2010a).

The objective of this study was to compare imputation accuracy, predictive ability, and selection efficiency for selection candidates genotyped at different densities using the Random Boosting (R-Boost) and G-BLUP algorithms.

Material and methods

Genotypes and Phenotypes

A total of 2658 genotyped bulls were used in this study, using the BovineSNP50.v2 Beadchip for 2226 bulls and the BovineSNP50.v1 Beadchip (Illumina Inc,) for 240 bulls. These 2658 bulls make up the 50K Holstein Spanish population that will subsequently be referred to as 50K. An additional 192 were genotyped using the BovineHD BeadChip; these bulls make up the Spanish HD population and will subsequently be referred as **HD**.

Before carrying out genomic evaluations, SNPs with greater than 5% incidence of missing genotypes across individuals and SNPs with minor allele frequency (MAF) less than 5% were discarded, leaving 39,714 and 540501 SNPs for the 50K and HD evaluations, respectively. Animals with call rates lower than 90% were also excluded of the genomic evaluation process.

Four complex traits were examined, including milk yield (**MY**), fat percentage (**FP**), somatic cell count (**SCC**), and days open (**DO**). These

traits were selected to show differences regarding heritability of the trait and amount of phenotypic information available.

Deregressed MACE progeny proofs (**DRP**) from January 2009 Interbull evaluation (Uppsala, Sweden) calculated as described by (Jairath et al., 1998), and the genotypes from progeny tested bulls in the training sets were used to estimate marker coefficients. Bulls in the testing sets had DGV in December 2011 that was based on 20 or more effective daughter contributions (**EDC**).

Imputation from LD to 50K

Design of reference and testing sets for imputation process and genomic evaluation is detailed in Figure 5.1. Design of the training and testing sets followed the recommendations of (Mäntysaari et al., 2010), although the recommended four year gap between training and testing sets was reduced to three years because of the small size of the reference population, thereby leaving more bulls in the training set to maximize the accuracy of estimated DGV.

For the validation set, low density genotypes were created *in silico*. During that process, those SNPs included in the 50K assay that were not included in the GoldenGate Bovine 3K (**LD3K**) (Illumina Inc.) and the Bovine LD (**LD6K**) (Illumina Inc.) assays, respectively were masked, and the original 50K SNPs were also included in the study for comparison purposes.

Thereafter, phased haplotypes from the reference population (**TRAIN50K**) filled in by BEAGLE were used as reference set for imputation in the LD3K and LD6K validation sets as well as for the imputation of missing SNPs in the original 50K set. The outcomes were referred as **3K50K**, **6K50K** and **TEST50K** data sets.

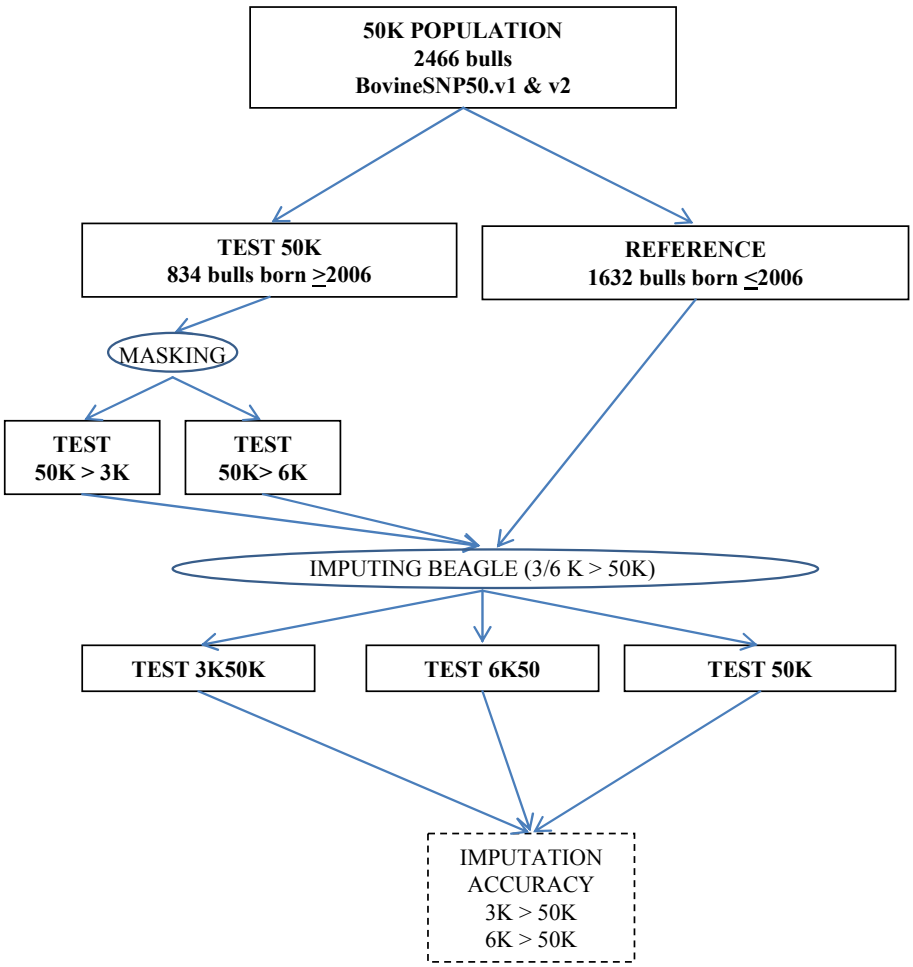

Figure 5.1. Diagram of the design of reference and validation sets and process of imputation accuracy evaluation from 3K and 6K to 50K.

From the 50K genotypes, reference and test data sets were generated based on year of birth. A total of 1632 bulls born before 2006 with progeny test results in January 2009 were used as the reference for imputation and genomic evaluation of production traits (MY and FP), Whereas 1629 and 1412 bulls were used as the reference populations for SCC and DO genomic evaluation, respectively. Bulls that were born between 2006 and 2010 were used as the validation sets, resulting in 834 genotypes to be imputed, of which 382 were progeny tested bulls in December 2011 that were used as the

testing set for MY and FP genomic predictions, as compared with 380 bulls for SCC and 216 bulls for DO.

Imputation from 50K to HD

The final process involved the imputation of 3K50K, 6K50K, TEST50K and TRAIN50K to HD (**3KHD, 6KHD, 50KHD** and **TRAIN50KHD**), using the original HD population as reference. Imputation to HD was consequently performed in two steps for the 3KLD and 6KLD sets, with the aim to optimize the accuracy of imputation (VanRaden et al., 2013).

Accuracy of imputation

The accuracy of the imputation process was measured using the allele error rate (AER). The number of errors was counted as 0 when the imputed and actual marker types were identical, 1 if the actual marker type was homozygous and the imputed genotype was heterozygous (or vice versa), and 2 if the actual and imputed marker types were opposite homozygotes. Error counting only considered markers/animals where observed marker types were not missing in the original non-imputed data set. The AER was calculated as 100 times the total number of errors divided by the twice the number of imputed loci. This gave the number of falsely predicted alleles, which is an appropriate measure when using an additive prediction model, as in this study.

Genomic Evaluation Models

It should be noted that the original HD population was used only for imputation purposes, and genomic evaluation comparisons were carried out using the original 50K data set. Two different genomic evaluation models were used:

R-Boost. The Random Boosting algorithm (González-Recio et al., 2013) is a machine learning technique that combines different predictors and a shrinkage factor (Friedman, 2000), Boosting methods iteratively adds basis functions, such that each addition further reduces the selected loss function (Hastie et al., 2005). Ordinary least square estimation was chosen as the basis function and was successively applied to the residuals of the previous iteration in a sequential manner. The MSE of prediction was used as the loss function to be minimized. The marker effect shrinkage parameter *v* of the algorithm was fixed as 0.10, while the parameter *mtry* that sets the proportion of markers sampled in each iteration was fixed to 0.05.

G-BLUP. The G-BLUP is similar to traditional BLUP evaluations where, the pedigree relationship matrix is replaced with the genomic relationship matrix (G) built from molecular information. Pairs of individuals sharing the same genotype for a large number of markers will be more similar genomically, and will have higher values in the corresponding off diagonal cells of the matrix, as is the case for pairs of related animals in a pedigree-based relationship matrix. The genomic relationship matrix was computed as $G = \frac{ZZ'}{2\sum p_i(1-p_i)}$, following (VanRaden, 2008), where p is the frequency of the second allele and i is the locus, Z is a matrix that results of the subtraction of P from M, with P = 2(pi -0.5) and M the matrix of genotypes

codified as -1, 0 and 1 for the homozygote, heterozygote and other homozygote respectively.

Criterion for Comparisons

Accuracy, Bias in Regression Coefficients and MSE. The prediction accuracy of genomic evaluations was computed as the Pearson correlation between the predicted DGVs and the December 2011 DRPs. The regression coefficients of the realized December 2011 DRPs on the estimated DGVs were also calculated, because this parameter is commonly used as a measure of prediction bias in genomic evaluations (Mäntysaari et al., 2010). Finally, the MSE of predictions was also estimated. Means and confidence intervals were estimated using bootstrapping for each evaluation output (Efron, 1986). Pairs were defined as the predicted phenotypes in the testing set and its corresponding observed (known) phenotype. Then, 1000 samples were drawn with replacement from the whole testing set, and predictive correlation estimates, regression coefficients and MSE correlation were computed for each of the bootstrap samples. The MSE should be the criterion of choice when animals with different amount of information are compared, (Vitezica et al., 2011). Finally the confidence interval was considered as the narrowest gap containing 95% of the replicates.

Selection Efficiency. Selection efficiency was measured as $\alpha_{top} / \alpha_{sel,}$ where α_{sel} represent a given percentage of bulls ranked by their predicted DGVs and α_{top} represents the percentage of bulls selected by the model that were in the same percentile according to their realized DPR. Selection efficiency can

be interpreted as the minimum number of top young bulls evaluated based on their genomic information that actually include at least one actual top bull, or similarly, the number of actual top bulls included in a given set of those top genotyped bulls.

Selection efficiency could be also measured using confusion matrices. These matrices are a common way to show results in classification problems, it is performed by comparing predictions with the observations in a validation dataset. In this study predictions and observations across traits and methods were split in to five classes according to their observed DRP and predicted DGV rankings. Therefore each class included 20% of bulls in the testing set. Observations were classified in rows and predictions in columns. Correct predictions fall on the diagonals (a_{ii}), and misclassifications on the off-diagonal $(a_{ij}, i \neq j)$ of the confusion matrix. Elements above the diagonal represent bulls that were under-classified by the genomic method while outcomes below the diagonal represent bulls that were over classified. Confusion matrices allow computation of overall selection efficiency as the proportion of correctly classified observations for a given C matrix: $efficiency = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}, (i=j)}{C}$, where c_{ii} are the elements on the diagonal. From the point of view of dairy cattle breeding programs, there may be more interest in the efficiency of selection for top bulls; this can be computed from the confusion tables as the proportion of correctly classified bulls in the first class: $efficiency_{top} = \frac{\sum c_{11}}{\sum_{i=1}^{n}\sum c_{i1}}$. Similarly, efficiency of selection was also computed for the top 60% to provide a measure of

efficiency for low selection intensity scenarios: $efficiency_{60\%} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3} c_{ij}}{\sum_{i=1}^{n}\sum_{j=1}^{3} c_{ij}}$.

Results and discussion

Imputation Performance

To measure imputation accuracy for missing SNPs in the original data sets a pilot study was carried out. Known SNP's were masked mimicking missing marker rate of the population, and the imputation AER(x100) was 0.08.

Imputation performances from customized LD3K and LD6K to 50K density in terms of AER resulted 3.1 and 1.3, respectively. Those results are in accordance to previous studies using similar population sizes (Berry and Kearney, 2011; Dassonneville et al., 2012; Zhang and Druet, 2010). Based on these results, the use of the LD6K array should be an important improvement, especially in cases where the genotyped population is not very large.

It must be noted that results from this study may be slightly optimistic, because LD genotypes are masked instead of directly genotyped, especially in the case of 3KLD due to the different chemistries used (Dassonneville et al., 2012). Regarding imputation from 50K to HD, a small number of HD genotypes could be enough for accurate imputation in some populations (Schrooten et al., personal communication), despite the fact that this accuracy can be enhanced when more HD genotypes are included within the range of genotypes used in this study. In a previous pilot study, AER(x100) after imputation from 50K to HD was 0.9 when 192 HD bulls were used as reference.

Validation of genomic evaluations

Accuracy. Prediction accuracy results obtained by the two methods considered are shown in Table 5.1. Both methods resulted in similar accuracy, R-Boost was the preferred method for FP and G-BLUP for MY and DO, while for SCC results were case dependant.

Table 5.1. Accuracy for the genomic estimation of two evaluation methods indexed for four traits of economic interest in dairy cattle after the imputation from 3K, 6K and 50K to 50K and HD. Mean of the 1000 replicates after Bootstraping and confidence intervals considered as the narrowest gap containing 95% of the replicates

Trait	Method	3K50K	6K50K	TEST50K	3KHD	6KHD	50KHD
Milk Yield (MY)	G-BLUP	**0.59**	**0.59**	**0.59**	0.54	0.54	0.55
	C. I.	0.53 0.66	0.51 0.65	0.52 0.66	0.47 0.61	0.47 0.61	0.47 0.62
	R-Boost	0.53	0.55	0.57	0.52	0.54	0.54
	C. I.	0.45 0.60	0.47 0.62	0.50 0.64	0.44 0.59	0.46 0.61	0.47 0.62
Fat Percentage (FP)	G-BLUP	0.59	0.60	0.60	0.54	0.55	0.55
	C. I.	0.53 0.65	0.53 0.67	0.54 0.67	0.46 0.61	0.47 0.62	0.48 0.63
	R-Boost	0.73	0.78	0.78	0.74	0.79	**0.80**
	C. I.	0.67 0.78	0.74 0.82	0.74 0.82	0.67 0.79	0.75 0.83	0.75 0.84
Somatic Cell Count (SCC)	G-BLUP	0.49	0.49	0.48	0.44	**0.50**	0.47
	C. I.	0.40 0.58	0.40 0.58	0.39 0.57	0.34 0.54	0.41 0.59	0.38 0.56
	R-Boost	0.46	0.45	0.46	0.46	**0.50**	0.49
	C. I.	0.37 0.55	0.36 0.55	0.36 0.54	0.36 0.55	0.41 0.58	0.40 0.58
Days Open (DO)	G-BLUP	0.25	0.29	0.19	0.29	**0.32**	0.31
	C. I.	0.10 0.40	0.14 0.43	0.02 0.33	0.14 0.44	0.18 0.46	0.18 0.46
	R-Boost	0.20	0.19	0.22	0.25	0.20	0.28
	C. I.	0.04 0.36	0.02 0.33	0.08 0.38	0.10 0.39	0.05 0.37	0.12 0.43

In bold: The preferred method within trait and set criteria
[1]Methods: R-Boost (Random Boosting) and G-BLUP

Regarding SNP density, the imputation of low density genotypes to a 50K platform resulted in accuracies similar to have observed for the original 50K genotypes, especially for the LD6K case, agreeing with results reported previously by (Segelke et al., 2012). Differences found in imputation performance between LD cases showed prediction accuracy across traits 2% higher for LD6K than the LD3K. In practice, this difference is expected to

be larger since the LD3K assay was developed using a different technology than the 6K and 50K assays (Dassonneville et al., 2012). For traits related to fertility, imputed genotypes were competitive compared with the 50K data set. Previously, higher accuracies had been reported for imputed genotypes when compared with the high density assay for fertility traits (Erbe et al., 2012b).

Prediction accuracy results for the genotypes imputed from 50K to HD depended on the trait considered. No improvement was found for MY, but modest improvement occurred for FP, SCC and DO. Data sets imputed from LD and 50K and then to HD performed similarly. This result was in accordance with results previously reported for other Holstein populations, where HD estimates were only slightly better than 50K (Erbe et al., 2012b; VanRaden et al., 2013).

Confidence intervals estimated by bootstrapping showed that distributions regarding prediction accuracy were widely overlapped across methods and sets for MY and SCC, Suggesting that for the considered data, there were no differences. However, R-Boost estimates were more accurate than G-BLUP for FP. As expected, large CI were found for DO, probably due to reference population size and the amount of information for each bull.

Bias in Regression coefficients The estimated "b" values of the regression of realized DRP on estimated DGV for the considered traits, methods, and genotyped sets are shown in Table 5.2. Estimated regression coefficients were close to unity for MY, SCC and DO when the prediction was carried out using R-Boost, while G-BLUP showed closer values to unity for FP. Values were within the range of previously reported values for other populations (Olson et al., 2011).

Table 5.2. Regression coefficients for the genomic estimation of two evaluation methods indexed for four traits of economic interest in dairy cattle after the imputation from 3K, 6K and 50K to 50K and HD. Mean of the 1000 replicates after Bootstrapping and confidence intervals considered as the narrowest gap containing 95% of the replicates

Trait	Method	3K50K	6K50K	TEST50K	3KHD	6KHD	50KHD
Milk Yield (MY)	G-BLUP	0.74	0.72	0.72	0.68	0.67	0.67
	C. I.	0.63 0.84	0.61 0.82	0.62 0.83	0.57 0.79	0.56 0.76	0.58 0.79
	R-Boost	0.86	0.88	**0.90**	0.84	0.85	0.85
	C. I.	0.72 1.01	0.75 1.03	0.77 1.05	0.71 0.99	0.69 0.97	0.70 0.98
Fat Percentage (FP)	G-BLUP	1.08	1.02	**1.01**	0.93	0.94	0.93
	C. I.	0.94 1.23	0.88 1.15	0.87 1.14	0.78 1.08	0.80 1.11	0.78 1.08
	R-Boost	1.17	1.29	1.30	1.13	1.24	1.25
	C. I.	1.05 1.28	1.19 1.40	1.20 1.41	1.01 1.22	1.15 1.34	1.15 1.36
Somatic Cell Count (SCC)	G-BLUP	0.59	0.60	0.58	0.62	0.61	0.67
	C. I.	0.48 0.73	0.47 0.72	0.45 0.69	0.47 0.78	0.49 0.74	0.51 0.81
	R-Boost	1.03	0.98	**1.00**	1.03	1.06	1.03
	C. I.	0.79 1.27	0.76 1.23	0.77 1.22	0.79 1.29	0.83 1.26	0.82 1.25
Days Open (DO)	G-BLUP	0.39	0.45	0.25	0.47	0.51	0.51
	C. I.	0.14 0.64	0.21 0.70	0.03 0.45	0.21 0.70	0.28 0.78	0.26 0.75
	R-Boost	0.74	0.64	**0.99**	0.85	0.58	0.91
	C. I.	0.15 1.43	0.10 1.14	0.31 1.70	0.30 1.33	0.14 1.06	0.35 1.44

In bold: The preferred method within trait and set criteria
1Methods: R-Boost (Random Boosting) and G-BLUP

No relevant differences were found when 3KLD and 6KLD were compared in terms of the "b" values. As reported for prediction accuracy, the use of imputed genotypes resulted in similar performance when compared with the 50K data set. For most of the cases considered, data sets including HD resulted in similar departures from unity as the evaluations using 50K. However regarding DO the imputation to HD lead to a "b" value notably closer to unity when G-BLUP was used. (Su et al., 2012a) also reported slightly better performance at HD density for fertility traits regarding bias in the regression coefficient. Still, smaller values of the coefficient for this case should increase when new animals are included in the reference population

and their DRP are based on larger EDC. R-Boost confidence intervals included unity for MY in the 50K data set, and all data sets for SCC and DO, the latter as a result of a large uncertainty. However, for G-BLUP only FP estimates included unity.

MSE. The MSE of prediction showed notable differences between evaluation methods (Table 5.3). R-Boost showed smaller MSE in all four traits. Compared with G-BLUP, the MSEs of R-Boost predictions were 12%, 54%, 12%, and 5% smaller for MY, FP, SCC, and DO, respectively. Those results, and the aforementioned accuracies and regression coefficients were in agreement with (Jiménez-Montero et al., 2013) when they compared different methods using a similar population but different traits.

When 3KLD and 6KLD were imputed to 50K and therefore, genomically evaluated, the 6KLD set was preferred for MY and FP but no clear differences were found for SCC and DO; for those traits 6KLD out-performed 3KLD when G-BLUP was the estimation method but showed larger MSE for the R-Boost algorithm. Smaller MSE was found when the 50K set was imputed to HD, except for MY, and the improvement averaged across traits was just 1 %. The 50KHD set also out-performed LD sets after imputation to HD. The MSE averaged across traits and methods was 1% and 3% smaller for the 5OKHD set when compared with 6KHD and 3KHD, respectively.

Regarding MSE, CI overlapped for MY, SCC and DO, showing similar levels of uncertainty when methods were compared, but R-Boost was preferred; only for FP R-Boost resulted in smaller MSE with little or no overlap with the G-BLUP CI. Estimation with 50K sets was preferred over HD for MY, as with accuracy and regression slope. Nevertheless, MSE for HD predictions was smaller for FP, SCC and DO, but differences were not clear and CI widely overlapped.

Table 5.3. Mean Squared Errors for the genomic estimation of two evaluation methods indexed for four traits of economic interest in dairy cattle after the imputation from 3K, 6K and 50K to 50K and HD. Mean of the 1000 replicates after Bootstrapping and confidence intervals considered as the narrowest gap containing 95% of the replicates

Trait	Method	3K50K		6K50K		TEST50K		3KHD		6KHD		50KHD	
Milk	G-BLUP	256		255		258		277		276		278	
Yield	C. I.	217	291	220	296	222	299	237	317	240	317	238	315
(MY)	R-Boost	244		236		**229**		247		241		240	
	C. I.	206	280	201	274	198	266	212	284	206	278	205	281
Fat	G-BLUP	0.044		0.044		0.044		0.048		0.048		0.047	
Percentage	C. I.	0.038	0.051	0.038	0.050	0.039	0.051	0.042	0.055	0.041	0.054	0.041	0.054
(FP)	R-Boost	0.034		0.030		0.030		0.032		0.028		**0.027**	
	C. I.	0.028	0.039	0.026	0.034	0.026	0.035	0.026	0.038	0.024	0.032	0.023	0.032
Somatic	G-BLUP	157.8		155.1		154.5		149.2		152.0		143.3	
Cell	C. I.	134.5	178.9	130.2	175.7	132.3	177.4	126.7	171.3	129.1	172.2	122.9	164.5
Count	R-Boost	136.6		138.3		137.4		137.1		**131.7**		133.4	
(SCC)	C. I.	117.5	157.7	118.8	160.5	115.4	158.0	117.8	156.4	115.0	153.6	114.2	154.5
Days	G-BLUP	562.6		548.6		636.3		537.5		530.2		535.1	
Open	C. I.	460.1	673.3	446.4	656.1	518.8	758.7	433.5	642.0	437.5	638.8	440.1	644.1
(DO)	R-Boost	531.7		541.9		523.4		**519.4**		546.4		519.6	
	C. I.	429.6	638.2	441.1	641.5	429.5	634.7	405.7	618.7	426.5	650.9	415.7	619.2

In bold: The preferred method within trait and set criteria
1Methods: R-Boost (Random Boosting) and G-BLUP
2Values x 1000

Selection Efficiency. When TEST50K and 50KHD were compared in terms of selection accuracy for top animals, both methods used in this study selected top ranked bulls similarly with respect to their observed DRP. Only for FP was R-Boost clearly more efficient (Figure 5.2). Selection accuracy of top bulls was slightly more efficient with HD genotypes. A large percentage of actual top bulls for MY, FP and DO were ranked on top when R-Boost was the evaluation method and HD genotypes were used. Similar performance was showed for DO and G-BLUP. For other cases, 50K and HD resulted in similar patterns. R-Boost seems more sensitive to the number

of markers than G-BLUP, probably because it is based on the estimation of individual SNP effects instead of an averaged genomic similarity between pairs of individuals.

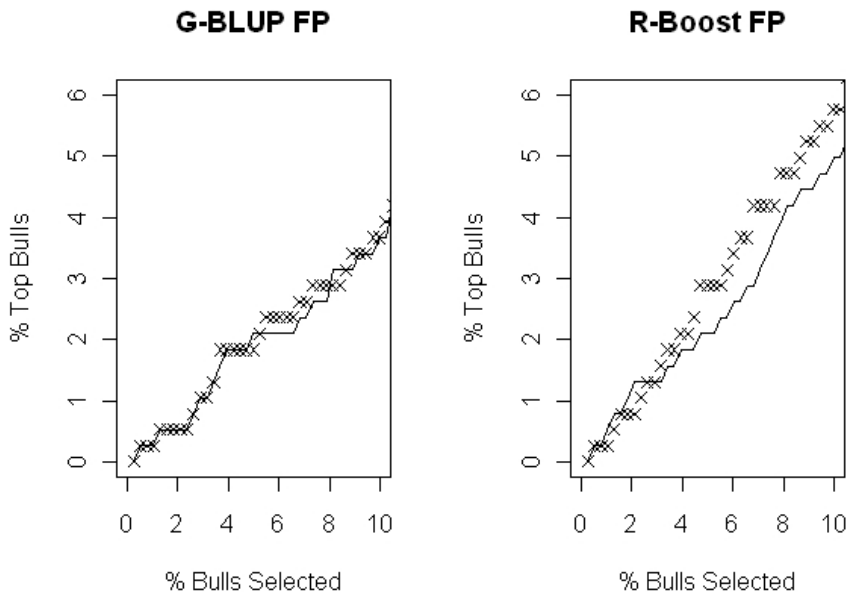

Figure 5.2. Percentage of common bulls in the observed and predicted rankings when less or equals than top 10% of genomically evaluated bulls are selected regarding fat percentage. Comparison between 50K (—) and HD (×) genotypes.

Another way to measure selection efficiency is through confusion tables. Bulls in the testing sets were classified according to observed DRPs and predicted DGV rankings. Classifications were then compared using confusion matrices. As example 6K50K and 50KHD cases are shown in table 5.4 and the full table is available in the online appendix II. Results of confusion matrices are also shown in Table 5.5, including overall rate of correctly classified animals, rate of correctly classified in the first class (Top 20%), and rate of correctly classified within the top three classes (Top 60%).

Table 5.4. Confusion matrices for the classification of animals in five classes according to their ranking regarding observed DRPs of four traits of economic interest in dairy cattle using two evaluation methods after the imputation from 6K to 50K and from 50K to HD. Observed and predicted classes in rows and columns respectively

	6K50K					50KHD				
Milk Yield (MY)										
G-BLUP	34	23	14	3	3	31	24	15	3	4
	17	26	14	9	10	17	23	14	12	10
	15	15	19	21	6	16	16	17	18	9
	10	8	17	21	20	12	9	15	25	15
	1	4	12	22	38	1	4	15	18	39
R-Boost[1]	33	23	10	8	3	33	23	9	9	3
	20	19	20	9	8	17	24	16	10	9
	10	20	20	14	12	15	15	21	17	8
	9	10	17	18	22	7	9	20	18	22
	5	4	9	27	32	5	5	10	22	35
Fat Percentage (FP)										
G-BLUP	37	19	12	7	2	38	17	10	9	3
	18	24	15	15	4	16	21	21	12	6
	11	16	20	14	15	13	14	23	12	14
	9	11	17	17	22	6	18	11	23	18
	2	6	12	23	34	4	6	11	20	36
R-Boost[1]	44	21	9	3	0	**46**	18	13	0	0
	23	22	20	10	1	17	27	22	7	3
	9	21	21	16	9	11	21	18	22	4
	1	10	19	29	17	3	8	16	31	18
	0	2	7	18	50	0	2	7	16	52
Somatic Cell Conunt (SCC)										
G-BLUP	31	24	5	12	4	**42**	7	10	10	7
	19	17	12	16	12	13	21	15	17	10
	14	21	19	14	8	10	20	16	16	14
	9	9	20	18	20	7	20	14	17	18
	3	5	20	16	32	4	8	21	16	27
R-Boost[1]	31	18	13	11	3	32	21	11	8	4
	15	23	16	11	11	15	22	17	14	8
	12	13	22	20	9	13	13	23	12	15
	14	12	13	17	20	13	9	17	16	21
	4	10	12	17	33	3	11	8	26	28

Days Open (DO)

G-BLUP	12	15	7	5	4	12	16	5	5	5
	11	5	9	11	7	9	6	10	11	7
	7	14	9	3	10	11	9	9	7	7
	7	4	10	9	13	7	6	9	9	12
	6	5	8	15	10	4	6	11	10	13
R-Boost[1]	13	7	12	9	2	13	13	7	4	6
	6	13	7	9	8	7	13	7	7	9
	9	10	7	6	11	9	9	8	10	7
	8	5	6	12	12	8	4	8	13	10
	7	8	11	7	11	6	4	13	9	12

In bold: The preferred method within trait and set criteria for the selection of bulls in the first class (Ranked in the top 20% according their DRPs)
1Method: R-Boost (Random Boosting)

Small differences were found between data sets and methods. In three out of four cases, G-BLUP correctly classified more animals in the first class (Top 20%). Also, for three out of four traits, HD estimates correctly classified more animals as belonging to the top class than predictions based on 50K genotypes. The rate of animals correctly classified ranged between 0.21 for DO to 0.46 for FP, and only for MY, did classification based on HD correctly classify fewer animals.

For the classification of top animals 6K50K increased averaged efficiency by 2% over 3K50K and TEST50K out-performed imputed genotypes by 1%. Also 3% greater efficiency was found when genotypes were imputed to high density. On average across traits, methods, and data sets, 42% of the top 20% bulls were correctly classified, while the overall rate of correctly classified bulls was 32%.

Table 5.5 Rate of animals correctly classified according to their ranking in five classes each one containing 20% of the values (Overall), correctly classified in the first class (Top 20%), or within the three highest classes (Top 60 %). Results showed for four traits of economic interest in dairy cattle using two evaluation methods after the imputation from 3K, 6K and 50K to 50K and HD

	3K50K	6K50K	50K50K	3KHD	6KHD	50KHD
Milk Yied (MY)						
G-BLUP						
Overall	0.35	0.36	**0.37**	0.34	0.34	0.35
Top 20%	0.45	0.44	**0.48**	0.40	0.40	0.40
Top 60%	0.77	0.77	**0.78**	0.75	0.76	0.76
R-Boost[1]						
Overall	0.33	0.32	0.34	0.34	0.32	0.34
Top 20%	0.43	0.43	0.45	0.44	0.43	0.43
Top 60%	0.75	0.76	**0.78**	0.73	0.74	0.76
Fat Percentage (FP)						
G-BLUP						
Overall	0.35	0.35	0.36	0.36	0.36	0.37
Top 20%	0.48	0.48	0.49	0.48	0.51	0.49
Top 60%	0.76	0.75	0.76	0.75	0.76	0.76
R-Boost[1]						
Overall	0.43	0.43	0.45	0.44	0.43	**0.46**
Top 20%	0.52	0.57	0.56	0.57	0.56	**0.60**
Top 60%	0.82	0.83	**0.84**	0.82	0.83	**0.84**
Somatic Cell Count (SCC)						
G-BLUP						
Overall	0.31	0.31	0.32	0.30	0.31	0.32
Top 20%	0.42	0.41	0.42	0.50	0.39	**0.55**
Top 60%	0.71	0.71	0.72	0.68	0.72	0.68
R-Boost[1]						
Overall	0.32	0.33	0.34	0.32	**0.35**	0.32
Top 20%	0.39	0.41	0.39	0.39	0.42	0.42
Top 60%	0.71	0.71	0.72	0.71	**0.73**	**0.73**

			3K50K	6K50K	50K50K	3KHD
Days Open (DO)						
G-BLUP						
Overall	0.23	0.21	0.22	0.25	0.24	0.23
Top 20%	0.28	0.28	0.28	**0.33**	**0.33**	0.28
Top 60%	0.67	**0.69**	0.65	0.67	0.67	0.67
R-Boost[1]						
Overall	0.22	0.26	0.23	0.22	0.25	**0.27**
Top 20%	0.23	0.30	0.30	0.28	0.23	0.30
Top 60%	0.67	0.65	0.64	0.65	0.67	0.67

In bold: The preferred method within trait and set criteria for the correct classification of bulls (overall), correct classification in the first class (Top 20%) or within three top classes (Top 60%)
[1]Method: R-Boost (Random Boosting)

Conclusions

Imputation using Beagle software was efficient for the reconstruction of 50K genotypes from low density chips. Also, BEAGLE performed well for imputation of HD genotypes from 50K, even when a small HD reference population was available.

Genomic evaluation methods (R-Boost and G-BLUP) resulted in similar prediction ability for the traits and genotypes included in this study. R-Boost showed clearly better performance for FP, while for the other traits no clear differences were found.

Differences between LD3K and LD6K were more noticeable for imputation accuracy than for prediction ability or selection efficiency. In general, genotypes imputed from LD performed similarly to those obtained for the animals originally genotyped at 50K in terms of prediction ability. LD genotyping and imputation could be an interesting approach in order to reduce genotyping costs, as no remarkable lack on selection efficiency is produced by the imputation process. Imputation could be useful for pre-selection of progeny testing candidates, genomic mating programs, or to increases the reliability of low heritability traits through the inclusion of

some of these animals in the reference population. In addition, LD chips could allow genomic selection programs to be implemented in other species or breeds where it is not affordable at current costs.

Imputation to HD showed similar overall predictive performance to 50K evaluations in terms of Pearson correlation, MSE, and regression coefficients. However, selection efficiency could be slightly enhanced for certain traits like FP, SCC or DO, especially when the aim of the evaluation is detect top animals in the population. Imputation to HD may be justified due to the larger number of actual top bulls identified as selection candidates.

Acknowledgments

The authors acknowledge funds from the project CDTI-P080250866 UPM and the agreement INIA-CC10-046, to CONAFE, EUROGENOMICS consortium, ASCOL, ABEREKIN, XENETICA FONTAO and GENETICAL for providing biological samples and phenotypes used in this study and to "Dirección General de Producciones y Mercados Agrarios" ,"Laboratorio Central de Veterinaria del Ministerio de Agricultura, Alimentación y Medio Ambiente" for support of genotyping process and specially Dr. Dassonneville for helpful suggestions and comments.

References

Berry, D.P., Kearney, J.F., 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5, 1162–1169.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., Van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A., Wiggans, G.R., For the Bovine LD Consortium, 2012. Design of a Bovine

Low-Density SNP Array Optimized for Imputation. PLoS ONE 7, e34130.

Browning, B.L., Browning, S.R., 2009. A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. The American Journal of Human Genetics 84, 210–223.

Calus, M.P.L., Veerkamp, R.F., Mulder, H.A., 2011. Imputation of missing single nucleotide polymorphism genotypes using a multivariate mixed model framework. J ANIM SCI 89, 2042–2049.

Dassonneville, R., Fritz, S., Ducrocq, V., Boichard, D., 2012. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of Dairy Science 95, 4136–4140.

De Roos, A.P.W., Schrooten, C., Mullaart, E., Van der Beek, S., De Jong, G., Voskamp, W., 2009. Genomic selection at CRV. Interbull Bulletin 39, 47.

Efron, B., 1986. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statist. Sci. 1, 54–75.

Erbe, M., Hayes, B.J., Matukumalli, L.K., Goswami, S., Bowman, P.J., Reich, C.M., Mason, B.A., Goddard, M.E., 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. Journal of Dairy Science 95, 4114–4129.

Friedman, J.H., 2000. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics 29, 1189–1232.

González-Recio, O., Jiménez-Montero, J.A., Alenda, R., 2013. The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. Journal of Dairy Science 96, 614–624.

Hao, K., Chudin, E., McElwee, J., Schadt, E., 2009. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. BMC Genetics 10, 27.

Howie, B.N., Donnelly, P., Marchini, J., 2009. A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies. PLoS Genet 5, e1000529.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B., Kolstad, B., 1998. Genetic Evaluation for Herd Life in Canada. Journal of Dairy Science 81, 550–562.

Jiménez-Montero, J.A., González-Recio, O., Alenda, R., 2013. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. Journal of Dairy Science 96, 625–634.

Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T., Sulem, P., Mouy, M., Jonsson, F., Thorsteinsdottir, U., Gudbjartsson, D.F., Stefansson, H., Stefansson, K., 2008. Detection of sharing by descent, long-range phasing and haplotype imputation. Nature Genetics 40, 1068–1075.

Mäntysaari, E., Liu, Z., VanRaden, P., 2010. Validation Test for Genomic Evaluations. InterbullBull 41, 17–22.

Meuwissen, T., Goddard, M., 2010. The Use of Family Relationships and Linkage Disequilibrium to Impute Phase and Missing Genotypes in Up to Whole-Genome Sequence Density Genotypic Data. Genetics 185, 1441–1449.

Nothnagel, M., Ellinghaus, D., Schreiber, S., Krawczak, M., Franke, A., 2009. A comprehensive evaluation of SNP genotype imputation. Human Genetics 125, 163–171.

Olson, K.M., VanRaden, P.M., Tooker, M.E., Cooper, T.A., 2011. Differences among methods to validate genomic evaluations for dairy cattle. Journal of Dairy Science 94, 2613–2620.

Pryce, J.E., Daetwyler, H.D., 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim. Prod. Sci. 52, 107–114.

Scheet, P., Stephens, M., 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. The American Journal of Human Genetics 78, 629–644.

Segelke, D., Chen, J., Liu, Z., Reinhardt, F., Thaller, G., Reents, R., 2012. Reliability of genomic prediction for German Holsteins using imputed genotypes from low-density chips. Journal of Dairy Science 95, 5403–5411.

Su, G., Brøndum, R.F., Ma, P., Guldbrandtsen, B., Aamand, G.P., Lund, M.S., 2012. Comparison of genomic predictions using medium-density (□54,000) and high-density (□777,000) single nucleotide polymorphism marker panels in Nordic Holstein and Red Dairy Cattle populations. Journal of Dairy Science 95, 4657–4665.

Sun, C., Wu, X.-L., Weigel, K.A., Rosa, G.J.M., Bauck, S., Woodward, B.W., Schnabel, R.D., Taylor, J.F., Gianola, D., 2012. An ensemble-based approach to imputation of moderate-density genotypes for genomic selection with application to Angus cattle. Genetics Research 94, 133–150.

Tassell, C.P.V., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5, 247–252.

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91, 4414–4423.

VanRaden, P.M., Null, D.J., Sargolzaei, M., Wiggans, G.R., Tooker, M.E., Cole, J.B., Sonstegard, T.S., Connor, E.E., Winters, M., Van Kaam,

J.B.C.H.M., Valentini, A., Van Doormaal, B.J., Faust, M.A., Doak, G.A., 2013. Genomic imputation and evaluation using high-density Holstein genotypes. Journal of Dairy Science 96, 668–678.

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genet Res (Camb) 93, 357–366.

Weigel, K.A., De los Campos, G., Vazquez, A.I., Rosa, G.J.M., Gianola, D., Van Tassell, C.P., 2010a. Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. Journal of Dairy Science 93, 5423–5435.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M., Wiggans, G.R., 2010b. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. J. Dairy Sci. 93, 2229–2238.

Wiggans, G.R., VanRaden, P.M., Bacheller, L.R., Ross Jr, F.A., Sonstegard, T.S., Te Meerman, G., Van Tassell, 2009. Transition of genomic evaluation from a research project to a production system. J. Dairy Sci. 87 (E-Suppl. 2), 313–314.

Wiggans, G.R., VanRaden, P.M., Cooper, T.A., 2011. The genomic evaluation system in the United States: Past, present, future. Journal of Dairy Science 94, 3202–3211.

Zhang, Z., Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science 93, 5487–5494.

6

General Discussion

Dairy cattle market is changing since high reliable breeding values can be obtained early in the animal life using GS, with no need of own or close relatives phenotype. Therefore, genetic gains of properly designed genomic programs largely overcome traditional approaches.

The field implementation of genomic selection will first affect to AI centers at selection of candidates for progeny testing. GS is expected to increase reliability of predictions regarding traditional PI. Further, marketing genomic candidates yet to be progeny proved, would contribute as a new product in the portfolio offered by AI companies.

Then, GS became also beneficial for commercial farms. More reliable genomic young bulls replace those still to be proven based on PI. Genomic young bulls usually have outstanding breeding values and due to their reliabilities should be marketed at lower prices than top proven bulls. The use of sets of 4-6 young bulls should be a recommended strategy to avoid risks due to lower reliability.

Most of the programs have some restrictions about the rights for obtain genomic breeding values of males, but this situation could be modified in the near future, then some farmers can keep some extra benefits from bulls born on their farms.

Replacement and culling decisions could be done more accurately if females are genotyped. In addition, selection of bull dams could be done in a fairer situation than previously.

In summary, this new technology could be greatly beneficial for those actors using it properly.

The results in this thesis suggest that female genotypes are valuables as RP. If females are genotyped, predictive ability depends largely on the genotyping strategy. To genotype just the top ranking cows as reference result in poor results. However, predictive ability is notably enhanced if cows in the opposite tail of the distribution are also included.

The Spanish genomic population had more than 1600 highly reliable proven bulls. This number seemed insufficient to obtain accurate genomic predictions that allow exploiting the potential of GS. Therefore, it was necessary to increment the size of the RP with new animals. Before joining the Eurogenomic consortium that solved the problem for the Spanish population, different options had to be evaluated to obtain information about interesting animals. Different exchange conventions and strategic alliances were planned with the aim to share this valuable data. The exchange of genomic information is needed, but also genetic values or phenotypic information associated with the genotypes must be shared. Both sources of information are required to obtain future predictions.

The first objective of the thesis was to evaluate different genotyping strategies based on simulated data in a context of limited proven bull population. To increase the size of the RP the suggested alternative to that was the use of females as a complementary option independent of the exchange of genotypes with other counties. It must be noted that phenotypic records from females are more affected by environmental factors, despite the fact they can be corrected by statistical methods, probably estimates will be less reliable. On the other hand we must not underestimate the potential of phenotypes because they are the best source of direct information from the genotypes. Therefore females should be taken into account, not only as complement to the information provided by bulls, but probably in the future as RP. Results from this thesis showed that predictions from a RP built with females from the two tails of the distribution over-performed accuracies

from a male RP. That information allows the option of genotyping females as RP. It's possible to increase accuracy of genomic evaluation for those populations with a limited number of highly reliable progeny tested sires. However the implementation of genomic evaluations from a two tailed RP raises certain difficulties of application:

A good farming recording scheme is a mandatory pre-requisite, as well as the definition of selection criteria for the two tails. Our simulation dealt with two different heritability traits but in a single trait design. Today's dairy cattle selection is based on multi trait evaluation. Genotyping a different RP for each trait does not seem possible in the current scenario of chip prices. In contrast it could be created a divergent RP respect to a combined index or the most interesting traits.

Females from the two tails of the distribution may be highly influenced by environmental conditions that distort their breeding values. For that reason, a prior filtering and a deep understanding of the industry and their production systems are required. Preferential treatments or particular stress situations affecting these females should be avoided. There may be disagreement on the genotyping of animals located in the low percentiles of the distributions. However, based on the results of this work is necessary to avoid the genotyping of only the best animals as RF. Similar results have been published in another simulation study (Ehsani et al., 2010). These works showed the importance of consider animals with a poor phenotypic result and in the opposite extreme of the distribution than the selected individuals. The inclusion of these animals improved the results of a random genotyping regarding predictive ability. Data recording of those animals and genotyping costs may be justified by the significant improvement in the accuracy of the evaluations and the benefits for the rest of the population.

In a recent study, Boligon et al. (2012) carried out a comparable simulation to that detailed in Chapter 2. They used similar genotyping strategies including a 0.5 heritability trait, different selection intensities and a scenario were genotyping strategies were applied to an indicator trait. Results were evaluated for a target trait, with the genetic correlation between the two traits set to 0.50. Their result agreed with the results of this thesis, genotyping strategies based on one tail of the distribution resulted in low prediction ability. Also, extreme animals in both tails of the distribution, were the most informative when training GS models. This was the best strategy to obtain the highest correlation between genomic predictions and simulated true breeding values for a correlated trait, but it was not the preferable strategy in terms of mean squared error. Their simulations and the included in this thesis provide a wide range of scenarios among trait heritability, dependent variable, generations of selection, intensity of selection and selection on correlated traits. However, the ranking of performance of the selective genotyping strategies was consistent across studies and should be maintained across a wide range of scenarios.

Studies, based on simulations offer the advantage of modeling different scenarios, study behaviors in fully controlled situations and take decisions with a greater degree of security in cases where there is not availability of real data. We must take into account the limitations of the simulations at drawing conclusions.

Results from this thesis are of interest in other small populations or those with a limited number of highly reliable individuals. For example, in beef cattle, it is almost unfeasible to achieve reasonable reliabilities for young animals for carcass traits. Such traits are generally expressed late in life, require slaughtering the animals, and incur a high cost of measurement. In such cases, it is possible to use marker information from a set of animals in previous generations to predict performance in the next generation. The

results of this study show that the predictive ability of breeding values will depend, among other factors, on which animals are genotyped in the RP.

After genotyping the RP, development of methods that were able to deal with high-density markers was required. Therefore, the second objective of this thesis was the development of a competitive and reliable genomic evaluation in terms of prediction accuracy, computationally efficient and flexible for further future developments.

Previously, the national RP was evaluated. Descriptors of the genomic structure showed that the Spanish population is similar to other Holstein dairy cattle populations in terms of MAF, Linkage disequilibrium and heterozygosity, as expected (Wiggans et al., 2009a; Banos and Coffey, 2010; Habier et al., 2010). Based on this similarity, genomic evaluations of genotyped animals for recorded traits included in the milk recording scheme should be feasible.

Machine learning algorithms can be used to deal with the curse of dimensionality, and computational limitations when a large number of individuals have genotypic information. This thesis describes the R-Boost algorithm that is compared with B-LASSO, Bayes-A and G-BLUP in terms of accuracy, bias and MSE.

B-LASSO provided the highest Pearson correlations averaged across traits. However, differences in accuracy between methods were small with the exception of FP where R-Boost achieved greatest accuracy. There were no relevant differences between R-Boost and the additive models based on marker regression, except for FP. Although machine learning techniques are expected to accommodate cryptic relationships in the data, the use of dependent variables that represent previously computed (additive, linear and smoothed), sire EBVs could mask such differences. R-Boost seems to

provide some advantages over Bayesian regression when a small number of QTL regulate the trait under purely additive regulation (González-Recio and Forni, 2011).

When the genomic predictions of young bulls are compared with highly reliable, progeny-tested bulls, biases from genomic predictions must be taken into account. The DGV of bulls in the testing set showed an average deviation over the realized DRP of 0.08 genetic SD across methods and traits. Standardized bias showed greater differences between methods than Pearson correlations. R-Boost resulted in nearly unbiased predictions for MY and FP and also produced the least bias for PY, whereas B-LASSO, produced the least bias in predictions for FY, FP and UD. Bayes-A showed a similar bias to R-Boost for PY. G-BLUP tended to provide, more biased predictions for all traits, with the exception of UD.

The coefficients of regressing realized DRP on estimated DGV are commonly used as a measure of bias in genomic evaluations. The expected value for this slope coefficient is unity if evaluations predict the actual magnitude of differences between bulls, if the genotyped young bulls are a representative sample of the bulls in the population. However, the genotyped young bulls are typically pre-selected by the AI centers based on their EBV or Sire-PI (Mäntysaari et al., 2010).In our study, regression coefficients ranged between 0.58 for Bayes-A (MY) and 1.19 for the R-Boost (FP). R-Boost provided slope coefficient closest to unity for four of the five traits (0.87 for MY, 0.99 for FY, 0.80 for PY and 0.82 for UD). These Regression coefficients were within the range reported in other studies in similar dairy cattle populations (Olson et al., 2011; Tsuruta et al., 2011).

MSE may be a more appropriate comparison criterion than the Pearson correlation, as it combines accuracy and bias. R-Boost was the preferred

method across traits in terms of MSE providing the smallest MSE on average, followed by B-LASSO, Bayes-A, and G-BLUP respectively.

Based on those results, R-Boost was considered as an efficient method to calculate additive genomic breeding values using high-density marker information and large data sets. R-Boost predictions resulted especially competitive in terms of mean prediction error and coefficient of the regression of realized DRP on estimated DGV. In addition, this methodology also produces lower MSE estimates. MSE is considered a measurement of overall fit of the model to the data, accounting for both accuracy and bias. It is recommended when animals with different amount of information are compared (Vitezica et al., 2011) as is the case of Dairy Cattle.

Currently, large amounts of dairy cattle females are being genotyped (Faust and Olson, 2012) and they become more representative of the overall population than males (Boichard et al., 2012). Under this situation the use of proper phenotypes and female genotypes as main source of genomic information will be a reasonable scenario. Flexible prediction methods able to deal with complex genetic and environment interactions should be valuable. R-Boost is expected to deal with these scenarios properly.

When the Spanish genomic program joint the Eurogenomics consortium, the RP size was increased largely. Therefore, the evaluation methodology should be adequate to the new requirements. A modification of the original Boosting algorithm was proposed to speed up computation of genomic breeding values, with a minimum impact in the predictive ability. This modifications included sampling a percentage (*mtry*) of markers on each iteration instead of the whole set and the inclusion of the shrinkage factor (v) over the predictions. The original gradient boosting algorithm performed the complete genome-assisted evaluation (10-folds) in 171.67 hours with v=0.01, 69.17 hours with v=0.10 and 50 hours with v=0.20. The computation

time was substantially reduced using the modification of the algorithm with mtry =0.01. The smaller times were 1.5, 0.83 and 0.67 hours for mtry=0.01 and v=0.01, v=0.10 and v=0.20, respectively. These computing times make Random boosting feasible for running frequent routine genome-assisted evaluations with large data sets without impairing the predictive accuracy. The choice of mtry and v is under discussion, and cross validation is currently the standard procedure. A more formal strategy with statistical properties could be studied in the future.

Cost efficiency is a key point in genomic selection programs. The use of inexpensive low density chips and posterior imputation is an efficient strategy for increasing the number of genotypes and therefore, multiply the benefits of genomics. Accordingly, the third objective of the thesis was to implement a flexible and efficient imputation design for different density genotypes.

Imputation performances using BEAGLE from customized LD3K and LD6K to 50K density in terms of AER resulted 3.1 and 1.3, respectively. Those results are in accordance to previous studies using similar population sizes (Berry and Kearney, 2011; Dassonneville et al., 2012; Zhang and Druet, 2010). Regarding imputation from 50K to HD, a small number of HD genotypes could be enough for accurate imputation in some populations (Schrooten et al., personal communication), despite the fact that this accuracy can be enhanced when more HD genotypes are included within the range of genotypes used in this study. In a previous pilot study, AER(x100) after imputation from 50K to HD was 0.9 when 192 HD bulls were used as reference.

Differences between LD3K and LD6K were more noticeable for imputation accuracy than for prediction ability or selection efficiency. In general, genotypes imputed from LD performed similarly to those obtained for the

animals originally genotyped at 50K in terms of prediction ability. LD genotyping and imputation could be an interesting approach in order to reduce genotyping costs, as no remarkable lack on selection efficiency is produced by the imputation process. Imputation could be useful for pre-selection of progeny testing candidates, genomic mating programs, or to increases the reliability of low heritability traits through the inclusion of some of these animals in the RP. In addition, LD chips could allow genomic selection programs to be implemented in other species or breeds where it is not affordable at current costs.

Imputation to HD showed similar overall predictive performance to 50K evaluations in terms of Pearson correlation, MSE, and regression coefficients. However, selection efficiency could be slightly enhanced for certain traits like FP, SCC or DO, especially when the aim of the evaluation is detect top animals in the population. Imputation to HD may be justified due to the larger number of actual top bulls identified as selection candidates.

The results in this thesis suggest that genotyping at 6K density doesn't affect future predictive ability and selection decisions, if genotypes were previously imputed to 50K. In addition, accurate imputations can be performed from 50K to 700K density using small numbers of ultra high density genotypes as reference. Those HD imputed genotypes are expected to enhance genomic selection in some scenarios using adequate evaluation methods.

Implementation of Spanish genomic program

This thesis has been developed in parallel with the implementation of the Spanish genomic program adapting the objectives of research to the industry requirements.

The Spanish dairy cattle population is mostly Holstein breed (99%). More than 60% of the Holstein cows are already registered in CONAFE (data from 2011), the national breeder association that includes all regional associations. CONAFE is in charge of the basic recording scheme that includes 26 traits included in the national genetic index (ICO). Also CONAFE runs traditional genetic evaluations and is part of Interbull.

There are four main testing scheme programs ABEREKIN, ASCOL, GENETICAL and XENETICA FONTAO reaching 140 progeny tested bulls per year. Those bulls are tested all around the country over more than 400,000 cows. As a result more than 900 bulls have been currently tested in Spain. Some of those bulls are nationally and internationally marketed after testing. The amount of progeny tested sires was clearly not enough to build a RP able to provide reliable predictions. Among other alternatives as genotype share, inclusion of females as reference population should be evaluated. Within this scenario the thesis was titled "Genomic selection in small dairy cattle populations" and the first objective was to study different genotyping strategies including females as RP detailed in the Chapter 2 of this thesis.

Within the scope of genomic selection in dairy cattle, some moments were of great importance for the future implementation of the national program. The first scientific manuscript dealing with methodology published in 2001(Meuwissen et al., 2001). The first SNPs assay for dairy cattle marketed in December 2007 (Tassell et al., 2008). The first genomic evaluation carried out in North America in August 2008, and the first official evaluation in the early 2009 (Wiggans et al., 2009b). During this period CONAFE and the national breeding programs realized the significance of GS and the necessity of implement it.

The Spanish GS program started in 2011 with an agreement between the breeders association and the progeny test programs with the inclusion of a scientific partner INIA and the support of the Spanish government for the genotyping of 2,000 reference bulls. In addition agreement was signed with Xenetica Fontao to be one of the labs performing genotyping. In the same year Spain joint to the Eurogenomics consortium that shares over 22,000 genotypes of progeny tested bulls.

The progeny testing programs provide bull genotypes while CONAFE was in charge of phenotyping and control the process, the scientific partner was in charge of the develop of genomic evaluations methodology including imputation, and reliability estimation of genomic values. The studies carried out in this thesis were considered for the implementation of the genomic program. The R-Boosting method was developed and implemented. Currently, first official genomic evaluation has been carrying out using this algorithm and they have been validated by Interbull. The strong reduction in computing requirements achieved by the proposed modification of the algorithm allowed the use the Eurogenomics population as RP efficiently. Computation times per trait averaged 18 hours.

Beagle was the software selected for routine imputations based on the results of this thesis and other pilot studies carried out by the author. Since October 2012, LD genotypes are monthly imputed to 50K density in a previous step to genomic evaluation. LD chips have been used for preselection of young bulls by the AI centers but also for the selection of cows and heifers as bull dam candidates.

Since the availability of a large RP of progeny tested bulls, genotyping strategies involving females as RP was no longer a priority for primary traits, but it could be valuable for the inclusion of new traits in the breeding program. The key of a breeding program for an A.I. center is at the selection

of future bull dams. Genomics can provide an increase in the number of candidate dams and the reliability of their genetic merit.

First genomic evaluations were carried out for those traits included in the Chapter 4 of this thesis and were used for AI centers in September 2011. The Eurogenomic population was included in November and first complete genomic evaluation for those traits included in the Spanish index (ICO) was carried out in February 2012. In May 2012 Spanish genomic evaluation for protein yield was validated by Interbull. Finally, in November the 30th of 2012 first official genomic evaluation were published on-line by CONAFE (http://www.conafe.com/noticias/20121130a.htm). Official genomic evaluations were published as GEBV blending DGV estimated using R-Boosting and traditional EBV from traditional evaluations. Weights of both sources of information take into account the reliability of the breeding value over the reliability of the GEBV. Reliability of the GEBV was calculated as the original reliability of the EBV plus reliability gain due to genomic predictions.

Effects of genomics in dairy selection

Use of young bulls evaluated based on their genomic information

Genomic selection has modified dairy cattle market, some progeny programs have reduced the number of bulls to approximately one half of the number previously tested per year (Spelman et al., 2012). This reduction could lead to lower accuracies in the future if the number of recent proved sires in the RP decreases (Lillehammer et al., 2011). Bulls entering in the progeny testing are previously genomically selected from large groups of genotyped candidates.

Those pre-candidates are calves born from elite sires and top pedigree, or genotyped cows. In some cases, the number of young bulls sampled

(genotyped) has already increased dramatically, and there is a strong trend within breeding companies of purchasing bull dams to ensure exclusivity (Dürr and Philipsson, 2012). Other programs have focused on genomic selection, and sons of outstanding top genomic young bulls are retained. Simulation studies suggest greater genetic response following the second strategy (Lillehammer et al., 2010; Pryce and Daetwyler, 2012). However, lower relationship between those young bulls and the RP implies lower accuracy of DGV. In addition, it must be taken into account that initial results have shown that some genomic estimates were over-estimated (Spelman et al., 2012). Especially in the case of outstanding genomic values of young bulls some shrinkage over the average is expected when new information is added. Schefers and Weigel (2012) suggested the use of teams of genomic young bulls to avoid the risk of low individual accuracy.

In some countries as New Zealand, over 40% of the inseminations are made from genomically selected bulls. This proportion is consistent with the rate of use of young bulls evaluated based on their genomic information in a number of other countries, including Australia and Ireland (Cromie et al., 2012). In France, a formal progeny test is no longer undertaken. It is expected that, through time, other dairy breeding schemes will move to this scenario once a greater degree of confidence is reached with genomic technology. However genomic selection required strong recording schemes to fulfill expectations. Modifying current testing process does not imply to avoid data recording but just the opposite.

Female genotyping

Some programs decide pre-screening more young bulls instead of elite dam (Spelman et al., 2012). However, both paths are used for some programs. For instances, Ireland started an initiative in 2012 encouraging farmers, through a slightly reduced cost, to genotype maiden heifers with the view of

including them in the training population from 2013. Heifers were targeted to avoid any possible selection bias since only high producing fertile cows remain to later lactations. Some Spanish testing programs routinely search top national females as bull dam candidates. To include them in the genomic program could be appealing in the near future. As the number of proven bulls may be limited in the near future, the potential use of females, with their own performance records, to estimate marker effects becomes increasingly important, especially in countries with small populations. Simulated studies show that the inclusion of female phenotypic and genomic information increase the rate of genetic gain, compared with a traditional BLUP breeding program. The generation interval of the males also decreases (Mc Hugh et al., 2011).

Genotyped cows provide less biased prediction that can enhance the selection of bull dams (Bouquet and Juga). Combined with biotechnological techniques as multiple ovulation and embryo transfer (**MOET**), this improvement could be maximized. However, inbreeding rate also increases (Pryce et al., 2010).

It must be noted, dairy cattle genetic market is highly globalize. However, only males are routinely compared through MACE by Interbull (Uppsala Sweden). Preselection of bull dams are typically based on national evaluation, own performances, deeply knowledge of their pedigrees and intuition. Currently GS, provides the opportunity to compare objectively cows worldwide. This is a great advance due GS.

Genomic selection in farms

Selection decisions

Based on simulation studies, Weigel et al. (2012) concluded that on commercial dairy farms selection and culling decision could be more

successful using genomic testing especially for selection of heifers. However the expected gain depends on selection intensity applied in each farm. Genomic selection will be more advantageous for animals with missing or incomplete pedigree and also for those farms with lower replacement requirements. Up to now, it is not the case of the Spanish dairy cattle population. Expected gains for lactating cows that had phenotypes don't clearly reimburse costs of genotyping. Cost effective genotyping strategies include pedigree index presorting by traditional parent average and genotype the set of heifers were selection should be made.

To make genomic selection feasible at commercial farm level, genotyping costs should be affordable, low density genotypes and imputations as shown in Chapter 4 of the thesis should be considered. Reliability of genomic values after imputation allows confident selection decisions.

At farm level, the incomes are based on milk sales. The breeders look for a bull that compensates the weakness of the cow to produce a replacement heifer. GS could be interesting when the replacement ratio is high and breeding decision must be accurate.

Inbreeding

Simulation studies showed that genomic breeding programs have the potential of decreasing the rate of inbreeding compared with conventional selection methods (Daetwyler et al., 2007). However, first results from realized performance show that inbreeding is increasing since genomic selection is available in Canada (Schenkel, 2012).

Increases in inbreeding could be minimized through genomic optimal contribution selection (De Cara et al., 2011) or mating programs based on optimal selection (gain & inbreeding) and minimum coancestry (Toro and Varona, 2010). Female's genotyping facilitates the identification of the least

related animals more accurately than the traditional relationship matrix, with lower rates of inbreeding associated with the genotyping of a large number of females (Mc Hugh et al., 2011).

Genomic mating programs

Since genomic evaluations are routinely carried out, dairy cattle industry should optimize the use of this new information. Next challenge in a genomic program is to optimize mating programs to maximize genetic gain and international competitiveness.

Traditional matings programs have been an important breeding tool for dairy cattle farmers in Spain. For instance ABEREKIN mating program is run for more than 100,000 females per year. Afterward, ASCOL, XENETICA FONTAO and CONAFE have developed their own mating programs. Genomics mating programs favors finding the ideal sire to mate a given cow (Cole and VanRaden, 2010). To cope with this goal, SNPs effects for a given trait and their respective positions need to be estimated or known, providing insight of the genomic areas of greater interest. This knowledge would allow designing matings with the aim to obtain the interesting combinations from the most complementary parents (Weigel and Cowan, 2009).

It is possible to conduct genomic evaluations at chromosome level rather than whole genome, even by regions within a chromosome. In low prolific species, multiple ovulation and embryo transfer programs acquire a greater importance to increase the probability of obtaining the desired combination for each mating. Currently, “velo-genetics” or “whizzo-genetics” are not implanted but they may become a reality soon. This programs consist on genotyping embryos or oocites to make the selection on them and increase the genetic response by further reduction of the generation interval (Meuwissen, 2003). Other applications of genomic mating program include

the potential use of dominant and epistatic effects on the commercial animals, or the more convenient combination in crossbreeding.

References

Banos, G., Coffey, M.P., 2010. Short communication: Characterization of the genome-wide linkage disequilibrium in 2 divergent selection lines of dairy cows. Journal of Dairy Science 93, 2775–2778.

Berry, D.P., Kearney, J.F., 2011. Imputation of genotypes from low- to high-density genotyping platforms and implications for genomic selection. Animal 5, 1162–1169.

Boichard, D., Chung, H., Dassonneville, R., David, X., Eggen, A., Fritz, S., Gietzen, K.J., Hayes, B.J., Lawley, C.T., Sonstegard, T.S., Van Tassell, C.P., VanRaden, P.M., Viaud-Martinez, K.A., Wiggans, G.R., For the Bovine LD Consortium, 2012. Design of a Bovine Low-Density SNP Array Optimized for Imputation. PLoS ONE 7, e34130.

Boligon, A.A., Long, N., Albuquerque, L.G., Weigel, K.A., Gianola, D., Rosa, G.J.M., 2012. Comparison of selective genotyping strategies for prediction of breeding values in a population undergoing selection. J ANIM SCI.

Bouquet, A., Juga, J.,Integrating genomic selection into dairy cattle breeding programmes: a review. animal FirstView, 1–9.

Cole, J.B., VanRaden, P.M., 2010. Visualization of results from genomic evaluations. Journal of Dairy Science 93, 2727–2740.

Cromie, A.R., Berry, D.P., Wickham, J.F., Kearney, J.F., Pena, J., vanKaam, J.B., Gengler, N., 2012. International Genomic Co-operation; Who, what, when, where, why and how?. Interbull Bulletin 42, 72.

Daetwyler, H. d., Villanueva, B., Bijma, P., Woolliams, J. a., 2007. Inbreeding in genome-wide selection. Journal of Animal Breeding and Genetics 124, 369–376.

Dassonneville, R., Fritz, S., Ducrocq, V., Boichard, D., 2012. Short communication: Imputation performances of 3 low-density marker panels in beef and dairy cattle. Journal of Dairy Science 95, 4136–4140.

De Cara, M. a. r., Fernández, J., Toro, M. a., Villanueva, B., 2011. Using genome-wide information to minimize the loss of diversity in conservation programmes. Journal of Animal Breeding and Genetics 128, 456–464.

Dürr, J., Philipsson, J., 2012. International cooperation: The pathway for cattle genomics. Animal Frontiers 2, 16–21.

Ehsani, A., Janss, L., Christensen, O.F., 2010. Effects of Selective Genotyping on Genomic Prediction, in: Proc. 9th World Cong. Genet. Appl. Livest. Prod. Presented at the 9th World Congr. Genet. Appl. Livest. Prod., Leipzig (Germany).

Faust, M.A., Olson, K., 2012. Selection and application-Making more profitable Holstein.

González-Recio, O., Forni, S., 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. Genetics Selection Evolution 43, 1–12.

Habier, D., Tetens, J., Seefried, F.-R., Lichtner, P., Thaller, G., 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genetics Selection Evolution 42, 5.

Lillehammer, M., Meuwissen, T.H.E., Sonesson, A.K., 2010. Effects of alternative genomic selection breeding schemes on genetic gain in dairy cattle., in: Proc. 9th World Cong. Genet. Appl. Livest. Prod.

Presented at the 9th World Cong. Genet. Appl. Livest. Prod., Leipzig (Germany).

Lillehammer, M., Meuwissen, T.H.E., Sonesson, A.K., 2011. A comparison of dairy cattle breeding designs that use genomic selection. Journal of Dairy Science 94, 493–500.

Mc Hugh, N., Meuwissen, T.H.E., Cromie, A.R., Sonesson, A.K., 2011. Use of female information in dairy cattle genomic breeding programs. Journal of Dairy Science 94, 4109–4118.

Meuwissen, T.H.E., 2003. Genomic selection: The future of marker assisted selection and animal breeding. Presented at the Marker Assisted Selection: a fast track to increase genetic gain in plants and animal breedeng?, Turín (Italia) FAO.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. Genetics 157, 1819–1829.

Pryce, J.E., Daetwyler, H.D., 2012. Designing dairy cattle breeding schemes under genomic selection: a review of international research. Anim. Prod. Sci. 52, 107–114.

Pryce, J.E., Goddard, M.E., Raadsma, H.W., Hayes, B.J., 2010. Deterministic models of breeding scheme designs that incorporate genomic selection. Journal of Dairy Science 93, 5455–5466.

Schefers, J.M., Weigel, K.A., 2012. Genomic selection in dairy cattle: Integration of DNA testing into breeding programs. Animal Frontiers 2, 4–9.

Schenkel, F.S., 2012. Inbreeding using genomics and how it can help.

Spelman, R.J., Hayes, B.J., Berry, D.P., 2012. Use of molecular technologies for the advancement of animal breeding: Genomic selection in dairy cattle populations in Australia, Ireland and New Zealand, in: AUSTRALASIAN DAIRY SC I ENCE SYMPOSIUM 2012

Proceedings. Presented at the AUSTRALASIAN DAIRY SC I ENCE SYMPOSIUM 2012, Melbourne, Australia.

Tassell, C.P.V., Smith, T.P.L., Matukumalli, L.K., Taylor, J.F., Schnabel, R.D., Lawley, C.T., Haudenschild, C.D., Moore, S.S., Warren, W.C., Sonstegard, T.S., 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nature Methods 5, 247–252.

Toro, M.A., Varona, L., 2010. A note on mate allocation for dominance handling in genomic selection. Genet Sel Evol 42, 33.

Vitezica, Z.G., Aguilar, I., Misztal, I., Legarra, A., 2011. Bias in genomic predictions for populations under selection. Genet Res (Camb) 93, 357–366.

Weigel, K., Cowan, M., 2009. Genomic selection of dairy cattle: Opportunities and challenges.

Weigel, K.A., Hoffman, P.C., Herring, W., Lawlor Jr., T.J., 2012. Potential gains in lifetime net merit from genomic testing of cows, heifers, and calves on commercial dairy farms. Journal of Dairy Science 95, 2215–2225.

Wiggans, G.R., Sonstegard, T.S., VanRaden, P.M., Matukumalli, L.K., Schnabel, R.D., Taylor, J.F., Schenkel, F.S., Van Tassell, C.P., 2009a. Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. Journal of Dairy Science 92, 3431–3436.

Wiggans, G.R., VanRaden, P.M., Bacheller, L.R., Ross Jr, F.A., Sonstegard, T.S., Te Meerman, G., Van Tassell, 2009b. Transition of genomic evaluation from a research project to a production system. J. Dairy Sci. 87 (E-Suppl. 2), 313–314.

Zhang, Z., Druet, T., 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. Journal of Dairy Science 93, 5487–5494.

Final Conclusions

In this thesis the fundamental steps for the implementation of a genomic selection program in dairy cattle have been studied and Spanish genomic reference population has been evaluated. Final conclusions of the thesis are:

1) Female genotypes are valuables as reference population.
2) If females are genotyped as reference population, predictive ability depends largely on the genotyping strategy.
3) To genotype just the top ranking cows as reference population produce poor predictive ability.
4) Predictive ability is notably enhanced if cows in the opposite tail of the distribution are also included.
5) Machine learning algorithms can be used to deal with the curse of dimensionality, and computational limitations when a large number of individuals have genotypic information.
6) Random-Boosting algorithm is an efficient method to calculate additive genomic breeding values using high-density SNP information and large data sets.
7) Random-Boosting predictions resulted especially competitive in terms of mean prediction error and coefficient of the regression of realized DRP on estimated DGV. In addition, this methodology also produces low mean squared error estimates.
8) The use of less expensive low density chips and posterior imputation is an efficient strategy for increase the number of genotypes and therefore, multiply the benefits of genomics.
9) The effect of imputation from 6K to 50K is minimal in terms of future predictive ability and selection decisions.
10) Accurate imputations can be performed from 50K to 700K density using small numbers of ultra high density genotypes as reference.
11) Genotypes imputed to HD enhance genomic selection in some scenarios when adequate methods for genomic evaluation are used.

In summary, to genotype the most informative animals as reference population, predict genomic values using an appropriate methodology in terms of prediction ability and exploit the advantages of imputation methods are prerequisites to maximize the profitability of a genomic selection program in dairy cattle.