MASTER INTERUNIVERSITARIO EN MEJORA GENÉTICA ANIMAL Y BIOTECNOLOGÍA DE LA REPRODUCCIÓN

Genomic evaluations using sequences of the 1000 bull genomes project

Tesis de Master
Valencia, Septiembre 2016

Kenza LAKHSSASSI

Director:
Dr. Oscar González Recio

Acknowledgments

This Master Thesis was carried out in the Department of Animal Breeding of the Spanish National Institute for Agricultural and Food Research and Technology (INIA) and was funded with a grant from International Center for Advanced Mediterranean Agronomic Studies (CIHEAM).

Thanks to CIHEAM-IAMZ for giving me the opportunity to do this Master. For having accompanied my first steps in Spain and keep doing so far.

I want to thank also my advisor Dr. Oscar González Recio. I appreciate all his contributions of time, ideas, and funding to make my Master Thesis experience productive and stimulating. The joy and enthusiasm he has for his research was contagious and motivational for me, even during tough times in the Thesis pursuit. I am also thankful for the excellent example he has provided as a successful researcher.

Thanks to CONAFE for allowing me to use their data for such an interesting study.

My sincere thanks go to all members of the Animal Breeding Department of INIA who provided me an opportunity to join their team and use their research facilities. In particular, I am grateful to Drs Ana Fernández, Maria Saura, Beatriz Villanueva, Raquel de Paz and Almudena Fernández for the preparation of sequence data, their help and their advices. They have contributed immensely to my personal and professional time at INIA. The group has been a source of friendships as well as good advice and collaboration.

I would like to thank also Prof. Agustín Blasco for organizing well the Master of Animal Breeding and Biotechnology of Reproduction. Thanks to all the professors who collaborated in this Master for their insightful comments and encouragement. Thanks for all.

For this dissertation I would like to thank my oral defense committee for their time, interest, and helpful comments and questions.

I am grateful to all the people I met during these two years. Thanks for all the love and the respect that they have shown to me. I have no words to express the immense joy of having met you.

Above all, I most specially thank my family. All believed on me sometimes more than I ever did. Without their unconditional love, support, guidance and encouragement, none of this would be possible.

Kenza

"Para mí no hay emoción comparable a la que produce la actividad creadora, tanto en ciencia como en arte, literatura u otras ocupaciones del intelecto humano. Mi mensaje dirigido sobre todo a la juventud es que si sienten inclinación por la ciencia, la sigan, pues no dejará de proporcionarles satisfacciones inigualables. Cierto es que abundan los momentos de desaliento y frustración, pero estos se olvidan pronto, mientras que las satisfacciones no se olvidan jamás."

Severo Ochoa [1905 - 1993]
Premio Nobel de Fisiología y Medicina
1959

Communications to conferences

The results obtained in this thesis have led to one oral communication in the National Meeting of Animal Breeding at the Polytechnic University of Valencia:

Lakhssassi K et González-Recio O., 2016. Evaluaciones genómicas utilizando secuencias del proyecto de los 1000 genomas bovinos. XVIII Reunión nacional sobre mejora genética animal. Valencia 2 y 3 de Junio 2016. España.

Contents

Resumen

Evaluaciones genómicas utilizando secuencias del proyecto de los 1000 genomas bovinos

Se presenta un enfoque de regresión sobre haplotipos para la evaluación genética usando una muestra de la población Holstein de 450 animales, con datos de secuencia completa procedente del proyecto 1000 genomas bovino. Este enfoque se basa en la hipótesis de que los haplotipos procedentes de datos de secuenciación están en un desequilibrio de ligamiento (LD) con los QTLs mucho mayor que los marcadores (SNP) procedentes de genotipados. Este estudio se centra en la extracción de los haplotipos en la población y su incorporación en el modelo de predicción de secuencia completa. En total, se incluyeron 38.319.258 (SNPs y indeles) procedentes de Next Generation Sequencing (NGS). Las variantes con menores frecuencias alélicas (MAF <0,025) fueron descartadas dejando un total de 13.912.326 SNPs disponible para los análisis.

Se usó el programa Findhap.f90 para la extracción de los haplotipos. El número de SNPs en el haplobloque varió de 799 en BTA 13 a 1285 en BTA 12, con una media de 924 SNP (166.552 pb). Los haplotipos con una frecuencia inferior al 1% fueron de alrededor del 97% en todos los cromosomas. Aquellos fueron ignorados dejando 153.428 haplotipos para los demás análisis. Cada haplotipo se identificó por cromosoma y el segmento en el que se encuentra, así como el número del haplotipo ordenado dentro del segmento. Los caracteres analizados fueron las pruebas MACE para proteína (Prot), Índice Global de Tipo (IGT), Recuento de Células Somáticas (SCS) y Días Abiertos (DO) proporcionadas por CONAFE. Los datos fenotípicos se fusionaron con el archivo de haplotipos y se usó un modelo bayesiano para predecir valores genómicos estimados (GEBV). Los haplotipos estimados mostraron una alta contribución a la varianza total de GEBV (entre 32 y 99.9%). Se observó que la mayoría de los haplotipos para Prot, IGT and DO están a frecuencia baja-intermedia, mientras que los haplotipos encontrados para SCS están mayoritariamente a bajas frecuencias. Por lo tanto, esperamos que nos aporten información adicional a los genotipados de SNP acerca de las variantes menos frecuentes para explorar su contribución en la variación genética, ya que los chips de SNPs están diseñados para marcadores a frecuencias intermedias-altas.

Con el fin de reducir el número de haplobloques necesarios para realizar la predicción genómica, se seleccionaron un subconjunto de haplobloques que contienen haplotipos con mayores efectos. Nuestro análisis estadístico detectó 1264, 1909, 851 y 1450 haplotipos distintos que tuvieron una estima del efecto superior a 3 desviaciones estándar

(sd) sobre la media para los kg de proteína, IGT, SCS y DO, respectivamente. En el segundo criterio, se seleccionaron los que superaron el umbral de 1 sd sobre la media y se detectaron un total de 44.319 haplotipos para Prot, 39.975 para IGT, 46.132 para SCS y 42.878 para DO. A continuación, los haplotipos seleccionados en cada criterio fueron sometidos a un nuevo análisis. La proporción de varianza de los valores genómicos estimados correspondiente al efecto de los haplotipos fue de 1.06, 5.24, 15.29 y 11.64% para Prot, IGT, SCS y DO, respectivamente, con aquellos haplotipos que superaron el primer criterio (3sd) y de 10.92, 101.62, 33.30 y 53.93% para el segundo criterio (1sd). Se esperaría que las predicciones genómicas utilizando solamente un conjunto de haplobloques adecuadamente seleccionados puede aportar información adicional a la predicción de GEBV, y deben ser considerados más en profundidad en los estudios.

Palabras clave: Evaluación genética, secuencia completa, Holstein, Findhap, haplotipos, modelo bayesiano.

Summary

Genomic evaluations using sequences of the 1000 bull genomes project

A haplotype regression approach for genetic evaluation using population sample of 450 Holstein animals, with full-sequence data from the 1000 bull genomes project is presented in this thesis. This approach is based on the assumption that haplotypes from sequencing data are in stronger linkage disequilibrium (LD) with Quantitative Trait Loci (QTL) than markers from SNP chips. This study focuses on the extraction of haplotypes in the population and their incorporation in the whole sequence prediction model. In total, 38,319,258 SNPs (and indels) from Next Generation Sequencing (NGS) were included. Variants with Minor Allele Frequency (MAF< 0.025) were discarded leading a total of 13,912,326 SNPs available for the analyses.

Haplotypes were obtained from version 3 of findfhap.f90 software. The number of SNPs in the haploblocks ranged from 799 in BTA 13 to 1285 in BTA 12, with a mean of 924 SNP (166,552 pb). The haplotypes with a frequency below 1% were around 97% in all chromosomes. These haplotypes were ignored leaving 153,428 haplotypes for subsequent analyses. Each haplotype was identified by chromosome and segment where it is located as well as the ordered number of the haplotype within the segment. The haploblocks were then used to predict four economically important traits: kg of protein (Prot), Global Type Index (IGT), Somatic Cell Score (SCS) and Days Open (DO). The phenotypic values were the MACE proofs provided by the Spanish Holstein Association CONAFE. The phenotypic data were merged with the haplotype file and a Bayesian model was implemented to predict Genomic Estimated Breeding Value (GEBV). Estimated haplotypes had a large contribution to the total variance of GEBV (between 32 and 99.9%). Most of the haplotypes for Prot, IGT and DO have low-intermediate frequencies while haplotypes found for SCS are mostly at low frequencies. We expect that these haplotypes will give us additional information to SNP genotypes on those less common variants, as SNP beadchips are designed to genotype intermediate-high MAF.

In order to reduce the number of haploblocks needed to perform genomic prediction, a subsets of haploblocks that contained haplotypes with large effects were selected. A total of 1264 haplotypes exceeded the genome wide threshold of 3 standard deviation (sd) above mean (in absolute value) for Prot, 1909 for IGT, 851 for SCS and 1450 for DO distributed along the genome. In the second criterion, those with effect estimate (in absolute value) larger than 1 sd above mean which led to a total of 44,319 haplotypes for

Prot, 39,975 for IGT, 46,132 for SCS and 42,878 for DO. Then, haplotypes selected in each criteria were subjected to a new analysis.

The proportion of the genetic variance estimated values corresponding to the haplotypes effect was 1.06, 5.24, 15.29 and 11.64% for Prot, IGT, SCS and DO, respectively, using the first criterion (3 sd) and 10.92, 33.30 and 53.93% for Prot, SCS and DO, respectively, using the second criterion (1 sd).

Genomic predictions using only a set of appropriately selected haploblocks can provide additional information to GEBV prediction, and should be considered in more in-depth studies.

Keywords: Genetic evaluation, full sequence, Holstein, Findhap, haplotypes, Bayesian model.

General Introduction

Introduction

The improvement of farm animals is a major concern for breeders looking for selecting the best candidates to obtain the best performing descendants and better adapted to current and next future farming conditions. Most of the economic characters in farm animals that are of interest for breeders commonly show continuous variation. There is a wide range of variability in these characters, which partly depends on the genes.

Traditional genetic improvement of livestock, using information on phenotypes and pedigrees to predict breeding values of the selection candidates based on Fisher's infinitesimal model, has been very successful. Nevertheless, we should be able to predict breeding value with higher accuracy using information from differences between animal DNA sequence (Goddard and Hayes, 2007).

Marker-assisted selection (MAS) has been proposed extensive, although in most cases it did not provide options for extra gains by increasing selection accuracy unless a sufficiently large number of markers were used (Meuwissen et al., 2001; Villanueva et al., 2005). However, the complexity of calculating breeding values including marker information was a further barrier to the application of MAS (Hayes et al., 2009). Its implementation has been limited and increments in genetic gain have been very limited (Dekkers, 2004).

New technological advances such as Single Nucleotide Polymorphism (SNP) discovery through deep sequencing and throughput SNP genotyping with SNP chips, have led to a new strategy of selection called genomic selection (GS) that has revolutionized breeding in some species such as dairy cattle, and at the same time posed new challenges (Hayes et al., 2009). This concept was introduced by Meuwissen et al., (2001), where genetic markers covering the whole genome were proposed to be in linkage disequilibrium (LD) with all quantitative trait loci (QTL). The SNP in close LD with QTLs enable us to divide the entire genome into thousands of relatively small chromosome segments. Then the effects of each chromosome segment are estimated simultaneously. Finally, the genomic breeding value equals to the sum of all estimated chromosome segment effects. However, the theory described by Meuwissen et al., (2001) was not applicable at this time because of the high cost of genotyping and the large number of markers required. Both limitations have been recently overcome by the dramatic development in sequencing technology,

which can sequence thousands of SNP, and by the development of genome-enhanced evaluations. GS offers many advantages at improving the rate of genetic gain in dairy cattle breeding programs. The most important factors that contribute to faster genetic gain include:

• Greater accuracy of predicted genetic merit for young animals.
• Shorter generation interval because of heavier use of young, genetically superior males and females.
• Larger selection intensity, because breeders can use genomic testing to screen a larger group of potentially elite animals.

The genetic gain (ΔG) in animal breeding programs can be calculated as:

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

where i: the intensity of selection; ρ: the accuracy of selection; σ_a: the additive genetic standard deviation; and L: the generation interval.

By increasing the accuracy and intensity of selection and shortening the generation interval, the rate of genetic progress for economically important traits can be approximately doubled (Van der Werf, 2013). Meuwissen et al., (2001), suggested by simulations that the breeding value could be predicted only from marker data with an accuracy of 0.85.

In practice, GS refers to selection decisions based on genomic estimated breeding values (GEBV) and other genome wide marker information. These GEBV are calculated by estimating SNP effects from prediction equations, which are derived from a subset of animals in the population (i.e., a reference population) that have SNP genotypes and phenotypes for traits of interest, and then used to predict the breeding values of new selection candidates (Hayes et al., 2009).

According to Goddard (2009) and Hayes et al., (2009), the accuracy of GEBV depends on 4 parameters. The first two of these are under the control of the experimenters while the last two are not:

1. The level of LD between the markers and the QTL;
2. The number of animals with phenotypes and genotypes in the reference population from which the SNP effects are estimated;
3. The heritability of the trait in question, or, if de-regressed breeding values are used, the reliability of these breeding values; and
4. The distribution of QTL effects.

There are some other parameters not mentioned by the authors such us:

1. The relationship between reference population animals and also between reference population and the candidates;
2. The statistical method used; and
3. The allelic frequencies of QTLs and markers.

In this section, we review a number of methodologies that have been proposed for estimating the single marker or haplotype effects across chromosome segment effects for GS and we will present a brief overview on our research context about genomic prediction using sequence data.

Approaches for genomic assisted predictions

Genomic selection depends on the possibility of predicting accurately the genetic merit of selection candidates based on their genotypes for SNP markers. The reasoning behind this process is that whenever marker density is high enough, most QTL will be in high LD with some markers, and marker effects estimation lead to accurate predictions of genetic merit for a trait.

Despite this, the amount of information to be analysed in this situation poses new challenges from statistical and computational point of view. The number of predictor variables (markers) is generally much higher than the number of observations (phenotypes), hence, there is lack of degrees of freedom to estimate all marker effects simultaneously, which is aggravated by the fact that models may suffer from multicollinearity, especially because markers in close positions are expected to be highly correlated.

Several approaches have been proposed to estimate the marker or haplotype effects across chromosome segments for GS. A key difference between these approaches is the

assumption they make about the variances of haplotype or single marker effects across chromosome segments.

Below, we briefly review some of the methods that have been proposed for genomic enhanced evaluations, based most of them on different regularization strategies.

Linear Least squares regression model

One of the simplest models in GS to predict the individual's breeding value by modeling the relationship between the individual's genotype and phenotype is:

$$y_i = \mu + \sum_{j=1}^{p} \mathbf{x}_{ij} g_j + e_i$$

Where $i = 1. . .n$ individual, $j = 1. . .p$ marker position/segment, y_i is the phenotypic value for individual i, μ is the overall mean, $\mathbf{x}_{ij}$ is an element of the incidence matrix corresponding to marker j, individual i, g_j is the effect associated with marker j, and e_i is a random residual term. Typically, e is chosen to have a normal distribution with mean of 0 and variance σ_e^2.

In order to estimate (μ, g), we can use least squares to minimize the sum of squared distance between the observed response and the estimated response.

We obtain the estimate of g obtained by solving the linear equations $\mathbf{X'X\ g = X'y}$, as $\widehat{g} = (\mathbf{X'X})^{-1}\mathbf{X'y}$. The elements of the design matrix $\mathbf{X}$ depend on the number of alternative alleles that the animal presents. For example, individuals having marker genotypes AA, Aa, aa, have elements coded as -1, 0, and 1 in x_{ij} respectively, although other codifications are also possible.

Usually the number of markers available is much greater than the number of individuals with phenotypic information, which means that p is much larger than n, and it is not possible to perform the estimation. Meuwissen et al., (2001) used a modification of least squares regression for GS using preselection. This approach makes no assumptions about the distribution of chromosome segment effects, because these effects are treated as fixed (Hayes, 2007). First, they performed least squares regression analysis on each segment separately using the model*:* $\mathbf{y} = \mu + \mathbf{x}_j g_j + \mathbf{e}$. Where $\mathbf{y}$ is the vector of the phenotypic information, μ is the overall mean vector, $\mathbf{x}_j$ is the j^{th} column of the design matrix corresponding to the j^{th} segment, g_j is the genetic effect associated with the j^{th} segment,

and **e** is the vector of the error terms. By plotting the log likelihood of this model, segments with significant effects were found.

Then the segments with the most significant effect were used for simultaneous estimation by the model:

$$y_i = \mu + \sum_{j=1}^{q} x_{ij} g_j + e_i$$

where q is the number of segments. This approach does not fully take advantage of the whole marker information because only markers with a significant effect are included in the final model. Further, this preselection procedure might bias the results. Other methods for GS have been introduced to overcome some of the drawbacks of the linear regression approach.

Ridge regression and BLUP

In ridge regression, estimates of the genetic effects are shrunk towards the mean. It is based on the assumptions that SNP marker effects are normally distributed, are uncorrelated, and have equal variances. The estimate of the regression coefficient is given by:

$$\hat{\mathbf{g}} = (\mathbf{X}'\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$$

The difficulty with ridge regression is that the choice of λ is arbitrary. Ridge regression BLUP uses the same estimator as ridge regression but estimates the penalty parameter by REML as $\lambda=\sigma_e^2/\sigma_\beta^2$, where σ_e^2 is the residual variance, σ_β^2 is the variance of the regression coefficients and var $(\beta) = \mathbf{I}\sigma_\beta^2$. These methods do not fit well for those cases where genes with large effect are involved (Xu, 2003).

G-BLUP

The genomic BLUP (G-BLUP) model proposed by (Meuwissen et al., 2001) is very close to pedigree BLUP (Henderson, 1975). In the G-BLUP, the markers effects are assumed to be randomly and normally distributed with uniform variance for all markers.

Goddard (2009) showed that the G-BLUP is equivalent to a traditional model BLUP replacing traditional pedigree relationship matrix A by a genomic relationship matrix **G** built from molecular information. 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.

Genomic relationship matrix **G** can be constructed in several ways (Gianola and Van Kaam, 2008), and various **G** matrices used in a genetic evaluation have resulted in different scaling and accuracies of GEBV (Aguilar et al., 2010; Forni et al., 2011).

Genomic relationship matrix **G** can be obtained by at least 3 methods (VanRaden, 2008): The first one uses the formula: $\mathbf{G} = \frac{\mathbf{ZZ'}}{2\sum p_i(1-p_i)}$ where p is the frequency of the second allele and i is the locus, **Z** is a matrix that results from the subtraction of **P** from **M**, being **P** = $2(p_i - 0.5)$, and **M** the matrix of genotypes codified as –1, 0, and 1 for the homozygote, heterozygote, and other homozygote, respectively.

The second method for obtaining **G** weights markers by reciprocals of their expected variance instead of summing expectations across loci and then dividing: **G = ZDZ′,** where **D** is diagonal with: $\mathbf{D}_{ii} = \frac{\mathbf{1}}{m[2p_i(1-p_i)]}$. That formula was proposed for human genetic studies (Leutenegger et al., 2003; Amin et al., 2007).

The third method for obtaining **G** does not require allele frequencies and instead adjusts for mean homozygosity by regressing **MM′** on **A** to obtain **G** using the formula:

$$\mathbf{G} = \frac{\mathbf{MM'} - g_0\,(\mathbf{11'})}{g_1}$$

where $\mathbf{MM'} = g_0\mathbf{11'} + g_1\mathbf{A} + \mathbf{E}$, g_0 is the intercept and g_1 is the slope. Matrix **E** includes differences of true from expected fractions of DNA in common plus measurement error.

The G-BLUP method does not suffer from large *p* small *n* problem since the amount of unknown effects is usually the same as in traditional BLUP (González-Recio et al., 2008).

Bayesian regression on markers

Bayes theorem is given as: P(x|y) ∝ P(y|x) P(x), where the symbol ∝ indicates 'is proportional to'. The probability P(x|y) is called the posterior probability. It is calculated from two terms. P(y|x) is a pseudo likelihood used by frequentists, and P(x) is called the prior probability. The different Bayesian methods used in GS are distinguished by the assumptions made concerning the distribution of SNP effects. According to Hayes (2007), we can make up our prior knowledge that there are some chromosome segments containing QTL of large effects, some segments with moderate to small effects, and some segments with no QTL at all when we estimate the effects of markers within the chromosome segments.

- ***Bayes A***

In Bayes A, Meuwissen et al., (2001) stated that the effects of SNP come from a normal distribution with a specific variance associated to each marker. The variances are modelled as an inverted χ^2 law. However, the specification of this model assumed the same a priori variance for all SNP effects, contrary to what was initially claimed in the original paper.

The prior distribution of SNP effect variances is: $P(\sigma_{gj}^2) \sim \chi^{-2}(v, S)$ where S is the scale parameter and v is the number of degrees of freedom. Gibbs sampling can be used to estimate the SNP effects and variances (Meuwissen et al., 2001).

- ***Bayes B***

The Bayes B has the advantage of selecting only markers with large effects. Meuwissen et al., (2001), proposed this model in which a proportion, π (arbitrarily set to 0.95), of markers has zero variance. The prior distribution is then:

$$\sigma_{gj}^2 = 0 \; with \; probability \; \pi$$
$$\sigma_{gj}^2 \sim \chi^{-2}(v, S) \; with \; probability \; (1 - \pi)$$

The Gibbs sampler described in method Bayes A cannot be used in method Bayes B, as it will not move through entire sampling space. This problem was resolved by sampling σ_{gj}^2 and g_j simultaneously using a Metropolis-Hastings algorithm.

The only difference between Bayes A and B is the prior for the variance components. Bayes B assumes that not all markers contribute to the genetic variation.

According to the results of Meuwissen et al., (2001) and the others studies, the Bayes B is often considered as the reference in terms of efficiency of genomic prediction, but it is extremely time consuming calculation. However, as Gianola et al., (2009) pointed out, this model is ill-posed because it originally specified the mixture distribution on the SNP effect variance. Assuming that 95% of the SNP effects have variances equal to zero implies that the effect is known without uncertainty. Further the choice of the degrees of freedom and the scale parameters of the scaled inverse chi-square distribution can influence the outcome.

- ***Bayes C***

Bayes C was proposed to overcome the statistical problems associated with the Bayes B, as the estimation of the probability π or the mixture distribution. The Bayes C model (Kizilkaya et al., 2010) differs from Bayes B by using a common variance for SNP with a non-zero effect, instead of a locus-specific variance. This variance is estimated, in contrast to G-BLUP, where it is supposed as known. The model is similar to the Bayes B model but for an uniform variance effect on all the loci:

$$\sigma_g^2 = 0 \; with \; probability \; 1 - \pi$$
$$P(\sigma_g^2) \sim \chi\text{-}2\,(v, S) \; with \; probability \;\; \pi$$

Using simulated data in a comparison, 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 Cπ & Dπ***

Habier et al., (2010) extended the panel of Bayesian methods with Bayes Cπ and Bayes Dπ. Bayes Cπ method assumes a common variance to non-zero effect markers with

probability 1-π and null effect with probability π. Additionally, the proportion π of markers is treated as unknown and is estimated from the data. Bayes Dπ denotes that each SNP has its own variance, but similar to Bayes Cπ, the π value is unknown (Habier et al., 2010). Both methods proved similar to the original methods regarding the accuracies.

- ***Bayesian LASSO***

De Los Campos et al., (2009b), González-Recio et al., (2009b) and Usai et al., (2009) proposed the Bayesian least absolute shrinkage and selection operator (LASSO) method for GS, where a double exponential prior distribution is assumed for the marker-effects with parameter λ (Park and Casella, 2008). This method performs a larger shrinkage on the marker-effects than other methods in a way that a large number of markers are estimated with a very small effect, and only a few markers are allowed to have larger effects.

The degree of shrinkage is determined by the parameter λ, which needs to be estimated. Park and Casella (2008) proposed the use of Empirical Bayes by Marginal Maximum Likelihood using an appropriate hyper prior for the estimation of λ. Legarra et al., (2011) proposed a modification of this method (BL2Var) which considers two different variances for the distribution of marker-effects and the residuals. Using Bayesian analysis, there is no need to pre-estimate the parameter λ as it is estimated from the data simultaneously with the marker effects and a gamma distribution can be assigned a priori.

The Bayesian LASSO appears to be an interesting alternative to the Bayes A method for performing linear regressions on markers. Legarra et al., (2011) and Ostersen et al., (2011) showed that Bayesian LASSO and G-BLUP gave comparable results for most traits, on real data sets of Montbéliarde and Holstein bulls, and on Danish Duroc pigs, respectively.

Up until now Bayesian LASSO has been widely applied for genomic evaluations as it provides accurate predictions for low density genotyping (Usai et al., 2009) and for traits that are regulated by many genes with a small effect (Cleveland et al., 2010).

- ***Bayes Stochastic search variable selection (SSVS)***

The technique was introduced by George and McCulloch (1993). It provides a method to maintain a constant dimensionality across all models but allows the SNPs in the predictive set to change. It allows this by instead of removing all non-significant parameters (those that would be excluded from the predictive set using the reversible jump algorithm) from the model, their effects are limited to values very close to zero (Verbyla et al., 2009). This method has a major advantage, which is that the posterior distribution of all parameters can be sampled directly using the Gibbs sampler, instead of using more computationally demanding algorithms such as the reversible jump algorithm (Verbyla et al., 2009)

The SSVS method has seen extensive use for applications to gene mapping (Yi et al., 2003; Meuwissen and Goddard, 2004). When it was used to predict genomic breeding values for real dairy data over a range of traits it produced accuracies higher or equivalent to other GS methods with significantly decreased computational and time demands than Bayes B. The faster speed of SSVS makes it more attractive.

However, one potential criticism of both Bayes B and Bayes SSVS is that the proportion of SNP in each distribution was not sampled appropriately, such that the means of the posterior distributions of the proportion of SNP with a zero or non-zero effect closely reflected the prior values of these proportions (Habier et al., 2011).

- ***Bayes R***

To overcome the drawback of Bayes B and Bayes SSVS, and for computational efficiency, Erbe et al., (2012) proposed a new method that assumes that the true SNP effects are derived from a series of normal distributions, the first with zero variance, up to one with a variance of approximately 1% of the genetic variance. The prior of the proportions of SNP in each distribution was the Dirichlet distribution.

The superior performance of Bayes R over other methods found by Erbe et al., (2012) probably results from using prior empirical knowledge about r^2, the assumed reliability. In Bayes R, $\sigma_g^2 = r^2\sigma^2$ is the assumed genetic variance, r^2 is the assumed reliability, and σ^2 is the variance of the target trait. Presumably, the assumption about r^2 is either model derived or based on prior cross-validation information, which is good Bayesian behavior,

normatively. Makowsky et al., (2011) gave evidence that what one assumes about genetic variance from inference in training data is not recovered in cross-validation.

- ***Elastic Net (EN)***

Croiseau et al., (2011) proposed the implementation of the EN algorithm for GS. This is a combination of G-BLUP and Bayesian LASSO weighted by a parameter α which takes values from 0 to 1. When α=0, a BLUP model is defined whereas α=1, a LASSO model is chosen.

$$\hat{\beta}_{EN=} \arg min \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x_i}\boldsymbol{\beta})^2 + \lambda((1-\alpha)\sum_j \beta_j^2 + \alpha \sum_j |\beta_j|) \right\}$$

where $\boldsymbol{\beta}= \{\beta_j\}$ is the vector of SNP effects, y_i is the phenotype of animal *i* and $\mathbf{x_i}$ is its vector of genotypes. The λ parameter corresponds to the intensity of the penalty.
Additionally, a pre-selection of markers can be applied prior to the analyses. EN, shares some variable selection properties with other methods (Bayes B, C$\boldsymbol{\pi}$…) and limits the number of SNPs with non-null estimated effects in the model. The purpose of this method is to provide a more flexible tool to deal with the large *p* small *n* problem. Limiting the number of SNP effects to estimate becomes important for an accurate prediction equation. The study of Croiseau et al., (2011) shows that EN provides better results than G-BLUP for most traits in the three breeds studied (Montbéliarde, Normande and Holstein). Furthermore, it resulted in encouraging results especially for small populations (Sánchez et al., 2010).

Machine learning algorithms

Machine learning methods have been used in genetic studies to explore the underlying genetic profile of disease and build models capable of detecting gene-gene interactions, predicting disease susceptibility, predicting cancer recurrence and predicting missing values of a marker (Szymczak et al., 2009). Machine learning methods are an interesting alternative for dealing with a higher predictive accuracy for routine genome-enhanced evaluations in a given population (Long et al., 2007). Several studies using machine learning approaches have been used for genome-enabled prediction in livestock and plants (González-Recio and Forni, 2011; Long et al., 2011b; a; Ober et al., 2011; Vazquez et al., 2012; Gonzalez-Recio et al., 2013; Crossa et al., 2014).

These methods aim at improving the predictive performance by learning from observations. They are model specification free, and may capture hidden information from large databases. This is appealing in a genomic information context in which multiple and complex relationships between genes exist (González-Recio and Forni, 2011). Some methods that have been proposed are:

1) Reproducing Kernel Hilbert Spaces Regression (RKHS) (Gianola et al., 2006) which resulted in accuracies similar or even higher than the ones obtained by the Bayesian methods (González-Recio et al., 2009b). It has been the most used one due to its similarity with BLUP, as shown by (De Los Campos et al., 2009a). The performance of RKHS has been shown to depend greatly on the choice of the space designed (González-Recio et al., 2008, 2009a; Konstantinov and Hayes, 2009; Ober et al., 2011).

2) Random Forest (RF), that considers all markers and gives the possibility of capturing interactions between genes and between genes and environment, which constitutes a major advantage in the study of complex diseases (Sun, 2010). Also, it presents a predictive ability equal or better than other parametric methods (González-Recio and Forni, 2010; González-Recio et al., 2011).

3) Neural Networks (NN), which proved to be useful for predicting complex traits as it can capture non-linear relations (Gianola et al., 2011). NN have been applied to genome-wide prediction in several studies (Long et al., 2011a; b). The comparison of RKHS and two different neural networks with some linear regression models (ridge regression, Bayesian LASSO, G-BLUP), showed an equal or better predictive ability for the machine learning methods (Tusell et al., 2013).

4) Support Vector Machines (SVMs) has been widely used in machine learning primarily for classification and it is also a particular case of RKHS (Moser et al., 2009; Pearce and Wand, 2006). Also, it performs robustified regression for quantitative responses by exploiting the relationships between observations by arraying predictors in observation space using a set of inner products (González-Recio et al., 2014).

5) Boosting, this is an ensemble method, which means that several models are somehow combined to improve the predictive ability just as RF. However, Boosting combines different predictors in a sequential manner with some shrinkage effect on each (Friedman,

2001). Thereby, it can handle interactions, automatically select variables, missing data and numerous correlated and irrelevant variables. It can construct variable importance in exactly the same way as RF (Ogutu et al., 2011), and is robust to outliers. The manner in which models are combined, labels the ensemble method and several criteria have been proposed (Friedman, 2001; Hastie et al., 2009; González-Recio et al., 2010). One of the most interesting modifications is the L2-Boosting algorithm for regression in high dimensional problems, which also has advantages when a non-null covariance structure between explanatory covariates exists, e.g. SNPs in high LD (González-Recio et al., 2010). Boosting has shown similar or better predictive ability than Bayes A or G-BLUP when it has been applied to genome wide prediction in chicken, swine and dairy cattle (González-Recio et al., 2010; González-Recio and Forni, 2011; Jiménez-Montero et al., 2013). Based on the experience from other studies, González-Recio et al., (2014) suggested the use of SVM and RF for classification problems, whereas RKHS and boosting may suit better regression problems.

Single-Step Genomic BLUP (ssGBLUP)

Obtaining genomic predictions via SNP arrays involves a multistep approach. A typical genomic evaluation requires a traditional evaluation with an animal model, extraction of pseudo-observations such as deregressed evaluations or daughter deviations, estimation of genomic effects for genotyped animals, and their combination with traditional parent averages and breeding values (Hayes et al., 2009; VanRaden et al., 2009). Because of its complexity, the multistep approach is prone to errors, which have been observed in many commercial releases in dairy cows. Considering that the genomic information can be included in a genomic relationship matrix, Misztal et al., (2009) proposed a single-step methodology where pedigree (matrix A) and genomic relationships (matrix G) are combined into matrix H, which is subsequently used in BLUP. Legarra et al., (2009) and Christensen and Lund, (2010) developed such a matrix, and Aguilar et al., (2010) demonstrated that a single-step methodology can be simple, fast, and accurate. This procedure is expected to improve the evaluation of not genotyped animals. Thus, the correct relationship matrix can be obtained by starting with the genotyped animals and then using the pedigree to calculate relationships involving ungenotyped descendants of these genotyped animals, i.e., going down the pedigree and accounting for the marker-based relationships of the ancestors of the pedigree (Meuwissen et al., 2016). The idea is

to replace pedigree with genomic relationships where available and retain the pedigree relationships where we do not have genomic relationships. The same idea can also be used up the pedigree, i.e., when ancestors are non-genotyped although it is not optimal in this case (Meuwissen et al., 2011).

The ssGBLUP has been used for several large-scale analyses including dairy (Tsuruta et al., 2011; Aguilar et al., 2011; VanRaden, 2012), pigs (Forni et al., 2011; Christensen et al., 2012), and chickens (Chen et al., 2011). In dairy cattle, ssGBLUP yields 0–2% more accuracy than multistep methods (Legarra et al., 2014), but for other species, which are less dominated by large sire families (i.e., where daughter averages are less able to summarize family information), the difference in accuracy between ssGBLUP and the multistep methods may be larger. Experiences indicate the following interesting properties for using ssGBLUP as it was quoted by Legarra et al., (2014):

- Automatic accounting of all relatives of genotyped individuals and their performances.
- Simultaneous fit of genomic information and estimates of other effects (e.g., contemporary groups). Therefore non loss of information.
- Feedback: the extra accuracy in genotyped individuals is transmitted to all their relatives (e.g. Christensen et al., 2012).
- Simple extensions. Because this is a linear BLUP-like estimator, the extension to more complicated models (multiple trait, threshold traits, and test day records) is immediate. Any model fit using relationship matrices can be fit using combined relationship matrices.
- Analytical framework. The Single Step provides an analytical framework for further developments. This is notoriously difficult with pseudo-data.

A more important feature of single-step models may be that they can account for pre-selection of young genotyped bulls, which might otherwise cause bias in the GEBV (Vitezica et al., 2011). According to Meuwissen et al., (2016), there is a clear need for a single-step method for the future that uses a “nonlinear” statistical method on sequence level data. Until recently, the size of the dataset to which ssBLUP could be applied is limited by the requirement that the **G** matrix must be inverted directly.

The Ancestor, Proven, Young Bull algorithm (APY) uses recursion to build a large component of the $\mathbf{G^{-1}}$ matrix directly, overcoming this limitation and expanding the application of ssBLUP to millions of genotyped animals (Fragomeni et al., 2015) but at the expense of some approximation in $\mathbf{G^{-1}}$.

The Reference population

Genomic breeding values are estimated from a reference population (RP) that contains animals with phenotypes and genotypes. The "RP" is used to train a statistical model on phenotypes and estimate the effects of each SNP or genomic combinations thereof. Accuracy of GS directly depends on the relationship between the RP and selection candidates. Most RP consist of proven bulls in national or international dairy cattle genomic selection programs (VanRaden et al., 2009).

The design of the RP is one of the key challenges for successful application of GS. It has a crucial impact on the accuracy of genomic prediction (Schöpke, 2014). Goddard and Hayes, (2009), Calus et al., (2013), Pszczola et al., (2012) and others, have emphasized the critical factors for assembling a RP which are the number and the composition of animals within this population and the phenotype accuracy. Usually, the size of the RP is limited by the number of genotyped animal or the availability of phenotypes if those are hard or expensive to record. The lower the heritability (h2), the larger the RP needs to be (Goddard, 2009). Moreover, the genomic structure of the population and the genetic trait architecture must be considered jointly at assembling a RP (Schöpke, 2014).

There are several factors to consider:

Size of the reference population

The economic aspect is the main limiting factor for the RP of a large population; either the trait is difficult or expensive to measure and/or the genotyping costs are large, and thus restrictive. The larger the size of the RP, the more accurately breeding values can be predicted. For numerically small breeds, assembling such a large reference population is challenging. Therefore, different approaches have been proposed to overcome these obstacles and enlarge the RP:

i) Combining populations from different countries or from different breeds: Consortia such as EuroGenomics or the US-Canadian collaboration show the large benefit resulting from the affiliation of several closely related populations into a single RP (Lund et al., 2011). Several studies have reported the accuracy of genomic prediction after combining phenotypes either difficult or expensive to measure using a multi-country RP (De Haas et al., 2012; Pryce et al., 2012a; b; Zhou et al., 2013), resulting in substantially increased reliabilities of GEBV. However, using one country for the RP and the other country for the validation set did not perform well (Pryce et al., 2014).

Nevertheless, the use of multiple breeds to calculate prediction equations in GS can be an attractive way to increase the size of the RP.

The prediction equation derived from one breed is only accurate for another breed when the marker-QTL LD persist across breeds. Therefore, a sufficient marker density is required. The more divergent the population is, the larger the density needs to be (De Roos et al., 2009). The extent of LD within a single populations and the consistency of LD between the populations are important factors and need to be considered when using combined RP.

ii) Adding genotyped cow to the RP: Thomasen et al., (2014) in their simulation study, showed that the inclusion of genotyped cows in the RP was an efficient way to increase the GEBV, and it would be a profitable investment for breeding schemes of small breeds. Wiggans et al., (2011) conducted the first empirical study of the inclusion of cows in the RP. They found an average gain in reliabilities of 3.5 and 0.9 percentage points respectively in Holstein and Jersey populations. Furthermore, Pryce et al., (2012b) demonstrated an improvement of 8 percentage points in the GEBV reliabilities by adding 10,000 genotyped cows to an RP consisting of approximately 3,000 bulls. However, Dassonneville et al., (2012) showed that the involvement of cow records in genomic evaluations can provoke over-estimation due to preferential treatment. In contrast, Lourenco et al., (2014) showed that including genotypes of elite females in genomic prediction using a ssGBLUP approach has no negative effect on evaluation accuracy. Besides that, Jiménez-Montero et al., (2012) have evaluated several female-selective genotyping strategies to increase the accuracy of GEBV. Depending on the population size, these authors either recommended a two-tailed selection (small populations), including females that exhibit upper and lower extreme values within the yield deviation

distribution, or they proposed a random selection for larger populations. According to Gao et al., (2015), the inclusion of genotyped females in the RP improves the reliability of genomic prediction by 1.9 to 4.5 percentage points. The benefit was larger for production traits than for conformation traits. These authors also showed that the addition of unselected females into the RP tends to reduce the prediction bias compared to adding selectively genotyped females.

iii) Imputing genotypes for related animals (complete un-genotyped animals): This strategy is particularly advantageous in cases where the phenotype of an animal has any added value but the genotype does not exist. Such imputations of un-genotyped individuals require a set of closely related genotyped animals. Bouwman and Veerkamp (2014) analyzed imputation accuracies for different scenarios of relatives with available genotypes and found it to be a helpful solution for including valuable phenotypes in genomic predictions, especially if genotyped offspring exists. Pimentel et al., (2013) have developed an algorithm to impute un-genotyped dams using known genotypes from the sire of each dam, one offspring, and the offspring's sire. The inclusion of these dams in the RP increased the accuracy of genomic predictions up to 37.14%. This approach was particularly beneficial for populations with lower levels of LD, for traits with low heritability, and for species with a limited RP. VanRaden et al., (2013), in their study of the accuracy of imputing HD from 500K and lower density genotypes reported that imputation to HD gave 99.3% correct genotypes from 50K, 96.1% from 6K, and 93.7% from 3K. Furthermore, a cost-effective strategy could be to sequence a small proportion of the population, and impute sequence data to the rest of the reference population. Druet et al., (2014) described strategies for selecting individuals for sequencing, based on either pedigree relationships or haplotype diversity. They demonstrated that the advantage of using imputed sequence data compared with dense SNP array genotypes was highly dependent on the allele frequency spectrum of the causative mutations affecting the trait. When this followed a neutral distribution, the advantage of the imputed sequence data was small. However, when all the causal mutations had low MAF, using the sequence data improved the accuracy of genomic prediction by up to 30%.

Genomic structure of population and genetic trait architecture

The RP composition and its relationship with selection candidates is also important. Pszczola et al., (2012) showed that the maximum reliability when using a cow RP can be reached if the relationship between animals in the RP is minimized, and if at the same time the relationship between validation set and RP is maximized. In some analyses, a randomly composed reference set appeared to be beneficial. Furthermore, RP need to be continuously updated, otherwise the relationship between reference and validation population decays and the accuracy of estimated SNP effects and therefore also those of the GEBV erodes (Habier et al., 2007; Pryce et al., 2012a; b). In addition, the accuracy of prediction may be affected by the properties of the QTL that control a trait, i.e. number of QTL, joint distribution of QTL allele frequencies across breeds, and distribution of QTL effects (Daetwyler et al., 2008; Goddard, 2009; Wientjes et al., 2015). Besides that, the genomic structure of population and genetic trait architecture are inextricably linked with each other. This must be taken into account when assembling a RP, designing a chip, or choosing appropriate statistical methods for genomic prediction (Schöpke, 2014).

Genomic prediction using sequence data

Motivation

Genomic predictions are now used routinely in selection of dairy cattle. The genetic gain that can be achieved is proportional to the accuracy of predictions. Thus, the challenge is to improve the accuracy of these predictions. The accuracy of genomic predictions based on SNP arrays depends on the proportion of the genetic variance captured by the array, determined by the LD between the SNP and the causative mutations affecting the trait (Druet et al., 2014). In contrast, GS from whole genome sequence data are expected to include the causal mutations responsible for trait variation (Meuwissen and Goddard, 2010). So, predictions should no longer depend on LD between SNPs and QTL, as the causal mutations are expected to be in the data set. According to MacLeod et al. (2013), inclusion of the causal mutations allows the effect of the QTL on a given trait to be estimated directly, which should increase the reliability of genomic predictions compared to using SNP genotypes, as well as the persistency of the reliability of predictions across generations.

Furthermore, with whole genome sequencing, at least the causative mutations, which do segregate across breeds, could be captured and this information can be used in multi-breed genomic predictions. According to Raven et al. (2014), a multi-breed reference would also benefit from the fact that LD across breeds is lower than that within breeds, so that causative mutations could be mapped more precisely.

Druet et al., (2014) showed that the accuracy of genomic breeding value may improve in the range of 2-30% (depending on trait). If the variation from rare alleles could be captured from the whole genome sequence data and exploited in genomic predictions. However, obtaining a higher persistency of reliabilities of genomic predictions over generations requires a large training set of thousands of sequenced individuals, QTL effects might be estimated with too much error and thus, there will be little advantage of using sequence data (Druet et al., 2014).

Furthermore, sequencing many individuals is still too expensive. Therefore, imputation to sequence data using SNP genotypes is an attractive and cost-effective approach to obtain a large training set of sequenced individuals. In this case, the lower density genotypes of the remaining individuals will be imputed to whole genome sequence genotypes using the sequenced individuals as reference (Van Binsbergen et al., 2014).

Challenges of sequence data in genomic prediction

The main challenges at dealing with whole genome sequence data for genomic prediction are: the huge number of variants, imputation accuracy of sequence variant, and statistical methods (Hayes et al., 2014). We will briefly describe next these challenges:

- ***The number of variants***

From the 1000 Bull Genomes Project, 31.8 million variants were detected in 2013. These variants were either SNP, short insertion deletions or copy number variation CNV. The use of these variants for genomic prediction presents a significant challenge. Hayes et al., (2014) recommended to use biological information to prioritise or filter variants.

This biological information comes in two forms, sites in the genome where variants are more likely to have an effect on any trait, for example coding regions or regulatory regions, and gene sets in which mutations are more likely to affect specific traits.

- ***Imputation of sequence data***

Imputation can be used to deduce the missing genotypes and could be helpful at increasing the accuracy of genome-enhanced breeding value. Imputation also allows for the use of low-density chips that may be more cost-effective, facilitating the widespread implementation of whole-genome selection (Weigel et al., 2010; Zhang and Druet, 2010).

Several imputation methods have been proposed and are implemented in programs like fastPHASE (Scheet and Stephens, 2006), Beagle (Browning and Browning, 2008), and findhap (Vanraden et al., 2010). These methods impute the missing genotypes based on reconstructed haplotypes informed by LD between SNPs.

Imputation of missing marker genotypes is based on available marker data from a given population. The population structure and frequencies of marker genotypes in the given population have an influence on the imputation accuracy (Druet et al., 2010; Dassonneville et al., 2011; Hickey et al., 2012). Because of differences in algorithms and different uses of information sources, the superiority of various imputation methods may differ in different imputation scenarios. In fact, FastPHASE and Beagle run slower as Bayesian method are applied for haplotype reconstruction, which may limit their practical use in large data sets. Findhap runs faster and is comparable to fastPHASE and Beagle in accuracy (Weigel et al., 2010).

Imputation accuracy in SNP chips was studied in cattle with 50,000 SNPs (Druet et al., 2010) and 777,000 SNPs (VanRaden et al., 2013). The general tendency in those studies was that the accuracy of imputation increased with an increasing number of SNPs, a shorter distance between the imputed SNP and the nearest SNP on the lower density marker panel, a larger MAF of the SNPs, a larger level of LD, and a larger number of close relatives between imputed and reference individuals (Van Binsbergen et al., 2014).

When using whole-genome sequence data, differences in extent of LD and population structure may affect imputation accuracies more in crop or livestock analyses than in human analyses (Van Binsbergen et al., 2014). A reliability of 0.83 was obtained at imputating from 777k SNP panels to sequence data with a reference set of 91 Holstein Friesian animals with whole-genome sequence data (Van Binsbergen et al., 2014).

Daetwyler et al., (2014) imputed genotypes for all sequence variants on chromosome 29 on Holsteins. The accuracy of imputation was reasonably high and varied along chromosomes, especially in regions where there were few SNPs on the BovineHD array or errors existed in the genome assembly. These authors observed that the accuracy of imputation decreased rapidly when MAF was below 0.1, suggesting that more sequenced individuals are required to accurately impute rare variants.

Sequence data in genomic prediction may not lead to higher accuracy if the accuracy of imputation to sequence data is too low. Hayes et al., (2014) used the 1000 bull genomes data to assess the accuracy of imputation to sequence in cattle using cross validation using Beagle4.0 (Browning and Browning, 2013). The accuracy was reasonable for variants with MAF greater than 5%. Accuracy of imputation rapidly declined for variants with MAF<5%. In their study, predictions were 2% more accurate than using 800k data set.

In the other hand, van Binsbergen et al., (2014) reported that it is necessary to aim for a large training set with a small average relationship between the animals, and possibly to pre-select SNPs based on functional information. Also, adding individuals of other breeds in a relatively large reference set will further increase imputation accuracy. In particular, it was reported that low MAF variants that segregate in other breeds can benefit from combining different breeds together (Bouwman and Veerkamp, 2014; Brondum et al., 2014).

- ***Methods for genomic prediction with full sequence data***

Many methods are available for genomic prediction but the choice of the best statistical method to derive the genomic predictions is still a challenge. Best linear unbiased prediction methods (BLUP) as described by Meuwissen et al., (2001), or GBLUP, (Habier et al., 2007) does not take full advantage of sequence because the priors used in these methods assume that all variants have an effect. Another problem with BLUP methodologies was identified by Verbyla et al., (2009) which is the severe shrinkage imposed on the marker effects, which means that the effect of a causative mutation is rarely captured by a single variant, rather the effect is split across several or many SNP.

van Binsbergen et al., (2015) reported GEBV reliabilities ranging from 0.37 to 0.52, with BSSVS performing better than GBLUP in all cases. Additionally, Meuwissen and

Goddard (2010) and MacLeod et al., (2013) showed the advantage of Bayesian methods over G-BLUP in simulation studies.

However, Ober et al. (2012) concluded that predictions from Bayes B were not better than predictions from G-BLUP when using real sequence data on *Drosophila melanogaster*. Furthermore, the results reported by van Binsbergen et al., (2015) did not show an advantage of using imputed sequence data compared to BovineHD genotype data for genomic prediction. These authors suggested using a large set of animals with small average relationships, along with other properties of the training set.

The 1000 bull genomes project

The 1000 Bulls Genome Project is an international collaboration between scientists in Europe, USA, and Australia. The project began in 2010, when scientists were looking for a way to share the huge cost of sequencing many entire genomes. The result was the 1000 Bulls Genome Project, which spreads the costs and shares the resources to help geneticists applying their knowledge collectively to improve genome-enhanced breeding value.

This project aims to assemble whole genome sequences of cattle from different institutions world-wide, and provide an extended data base for imputation of genetic variants for genomic prediction and genome wide association studies (GWAS) in all cattle breeds. It allows project partners to impute full genome sequences in bulls and cows that have been genotyped with SNP arrays which can be used for GS and more efficient discovery of causal mutations.

The project chose key ancestors animals for sequencing, because they are expected to have contributed substantially to nowadays population. Sequence data from these animals allows imputing the SNP chip genotypes of their descendants to whole sequence, allowing more accurate GWAS and genomic predictions. Spain, through INIA, joined into the 1000 bull genomes project consortium since 2015.

It is known that animal breeding programs are being transformed by the use of genomic data, which are becoming widely and cost-effective available to predict genetic merit. Most of the benefits of GS arise from the possibility of obtaining accurate predictions

early in the breeding cycle. A large number of genomic prediction studies have been published using both simulated and real data.

Recent developments in molecular and genotyping technology, combined with advances in statistical methodology in prediction of breeding values, have led to development and successful implementation of whole-genome selection methods in dairy cattle. The accuracy of GEBV prediction is important for a successful application of GS.

Additionally, genomic prediction with whole genome sequence data is now possible for cattle. The 1000 Bull Genomes Project provides a database from key ancestor bulls that can be imputed into RP genotyped with SNP arrays. Besides that, the genomic prediction methods to deal with such large data sets are under development.

Objectives

Main objective

The objective of this investigation was to develop strategies that incorporate sequence information in genetic evaluations. For that, the sequences of the 1000 bull genomes project will be used.

Specific Objectives

1. Evaluate the performance of the Findhap software to construct haplotypes from sequence data;
2. Detect sequence regions that are associated to traits of economic interest and can be incorporated in genomic evaluations in the Spanish dairy cattle;
3. Evaluate the proportion of genetic variance that can be explained by these regions.

Material and Methods

Data

In this study, the population sample consisted of 450 Holsteins animals with full-sequence data from run 5 of the 1000 Bull Genome Project and their pedigree information. In total, 38,319,258 SNPs (and indels) from NGS were included. However, a large percentage (50%) of these variants with low MAF are expected to be sequencing errors (Gonzalez-Recio et al., 2015). Hence, variants with MAF< 0.025 were discarded. The number of SNPs remained on each chromosome is given in Table 1.

Four economically important traits were used in this study: kg of protein (Prot), Global Type Index (IGT), Somatic Cell Score (SCS) and Days Open (DO). The phenotypic values were the MACE proofs provided by the Spanish Holstein Association CONAFE. Only animals with sequence and phenotype were kept for further analyses. In total, 361 sires were used.

Table1: Total and filtered SNP (MAF< 0.025) on each chromosome

Chromosome number	# Total SNP	# SNP remained
BTA1	2415624	859904
BTA 2	2062223	692662
BTA 3	1747334	620585
BTA4	1840583	672654
BTA5	1790983	663521
BTA6	1773469	679959
BTA7	1610969	571225
BTA8	1622084	573042
BTA9	1555596	566141
BTA10	1532614	575132
BTA11	1546812	559737
BTA12	1667137	711845
BTA13	1236102	409838
BTA14	1234979	428478
BTA15	1415166	522191
BTA16	1291009	427038
BTA17	1157679	447303
BTA18	964483	361249
BTA19	929690	334276
BTA20	1121685	394257
BTA21	1088553	379736
BTA22	892683	305698
BTA23	1016377	387518
BTA24	994429	346706
BTA25	670204	240877
BTA26	779371	286624
BTA27	698131	286641
BTA28	772863	276837
BTA29	890426	330652

Estimation of haplotypes in the population

Haplotype blocks may improve genomic predictions compared to individual SNPs, since haplotypes are in stronger LD with the QTL than individual SNPs are. It has also been hypothesized that an appropriate selection of a subset of haplotype blocks can result in similar or better predictive ability than using the whole set of haplotype blocks (Cuyabano et al., 2015). In this study, haplotypes were obtained from version 3 of Findfhap.f90 software (VanRaden et al., 2011). This program was designed to extract haplotypes in the population for future imputation.

Haplotyping algorithm in findhap

The algorithm begins creating a list of haplotypes from the genotypes, and the process is iterated for a fine haplotype construction. The steps in the algorithm are as follows:

1. Each chromosome is divided into segments with three progressively shorter lengths, long lengths to lock in identity by descent, and short lengths to fill in missing calls.
2. The first genotype is entered into the haplotype list as if it was a haplotype.
3. Any subsequent genotype that shared a haplotype is then used to fill the previous genotypes into haplotypes.
4. As each genotype is compared to the list, a match is declared if no homozygous loci conflicted with the stored haplotype.
5. Any remaining unknown alleles in that haplotype are imputed from homozygous alleles.
6. The individual's second haplotype is obtained by subtracting its first haplotype from its genotype, and the second haplotype is checked against remaining haplotypes in the list. If no match is found, the new genotype (or haplotype) is added to the end of the list. Unknown alleles in the genotype are stored as unknown alleles in the haplotype.
7. The list of currently known haplotype is stored from most to least frequent as haplotypes are found for efficiency so that more haplotypes are preferred.

In subsequent iterations, earlier created genotypes are matched again using haplotypes that occurred later. The first two iterations mainly focus on determination of haplotypes in the population. Only the highest-density genotypes are used in the first iteration, and then all genotypes are used in the second iteration.

After haplotyping, haplotypes are matched by using both pedigree and population in the following two iterations. Known haplotypes of genotyped parents were checked first, and either of the individual's haplotypes was not found with this quick check, then checking restarted from the top of the sorted list.

For example, the algorithm in Figure 1 could check haplotypes 5 and 8 first if parent genotypes are known to contain these haplotypes. The last two iterations did not search sequentially through the haplotype list and instead used only pedigrees to impute haplotypes of non-genotyped ancestors from their genotyped descendants, locate crossovers that created new haplotypes, and resolve conflicts between parent and progeny haplotypes.

If parent and progeny haplotypes differed at just one marker, the difference was assumed to be genotyping error, and the more frequent haplotype was substituted for the less frequent. FORTRAN program findhap.f90 requires little time and is available at:

http://aipl.arsusda.gov/software/index.cfm for download.

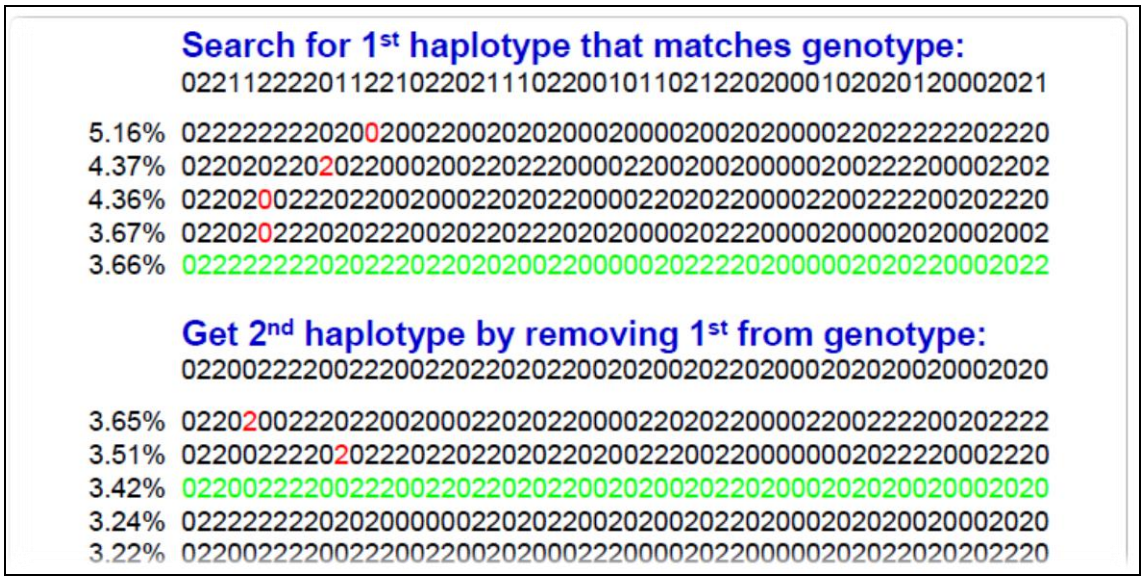

Figure 1: Demonstration of algorithm to find first and second haplotypes (VanRaden et al., 2011)

Implementing Findhap.f90

Genotypes were coded numerically as 0 if homozygous for the first allele (AA), 2 if homozygous for the second allele (aa), and 1 if heterozygous (Aa). To execute the findhap.f90 program, four input files are necessary:

- Genotypes.txt which contains: animal number, chip number and SNPs genotypes sorted by animal number.
- Chromosome.data which contains the SNP map with the list of all SNPs in the chromosome. Sorted by position within chromosome.
- Pedigree.file which contains the sex of animal, animal number, sire number, dam number, birthdate, animal ID and animal name sorted in ascending birth date order.
- Findhap.options which is a parameter file with user-defined options. These options include:

a) Error rate parameter

The error rate parameter is defined as the expected percentage of variants that are incorrectly called at sequencing. Indeed, with very large numbers of variants sequenced, the number of sequencing errors are likely to be considerable. Findhap program suggest 0.002 as error rate but in a recent study for distinguishing rare variants from sequencing errors, the authors observed that at MAF<0.01 up to 50% of variants are sequencing errors (Gonzalez-Reció et al., 2015). This creates haplotypes that appear only in one animal (singletons) and thus are not informative. These authors also estimate the sequencing error of 1% in variant calling. Hence, we have performed the findhap.f90 program for the 29 autosomes with an error rate = 0.01.

b) Haplotypes length

The haplotype length is defined as the number of SNP contained in the block (haploblock), and is provided by the user. This is one of the main parameters that are to be determined at implementing the algorithm. A previous study on haplotyping in German Holstein cattle (Qanbari et al., 2010), reported a mean block length of 164 kb. A proper definition of the haplotype length will minimize the probability of recombination within the block, and maximize the probability of transmitting the whole block to the progeny. Hence, the number of haplotype blocks and the haplotype length per chromosome were defined as follows:

$$\text{Number of blocks} = \frac{\text{Chromosome length (kb)}}{\text{Block length (kb)}}$$

$$\text{Number of SNP per block} = \frac{\text{Number of SNP remained}}{\text{Number of Blocks}}$$

where the average block length was considered 164 kb as proposed by (Qanbari et al., 2010). Haplotype blocks were built separately for each chromosome. According to the results obtained from the formulas, the options in Findhap were set to a minimum length of 800 SNPs, a maximum length of 100,000 SNPs and processing with 5 iterations per step. Singletons and low frequency haplotype alleles were ignored by excluding those with frequency <1%. After this filtering, **153,428** haplotypes were kept for subsequent analyses.

Haplotypes coding

Each haplotype was identified by the chromosome and segment where it is located as well as the ordered number of the haplotype within the segment. A program was developed in R software to define the number of alleles for each haplotypes that the animal carries (0, 1 or 2). Then, the phenotypic data were merged with the haplotype file for the subsequent analyses.

Incorporating sequence haplotypes in the whole sequence prediction model

The following linear equation represents the relationship between the phenotype of interest and the genetic effects (NGS variants and polygenic effect):

$$\mathbf{y} = \mathbf{1}'\mu + \mathbf{Wh} + \mathbf{Zg} + \mathbf{e}$$

where **y** is a vector with the phenotypic observations, μ is a population mean, **1**' is a vector of ones, **h** is the vector of haplotype effects assumed to be distributed as a double exponential (Laplace distribution) $\mathbf{h} \sim DE(\mu_h,\lambda)$, **g** is the vector of polygenic effects distributed as $\mathbf{g} \sim N(\mathbf{0},\mathbf{G}\sigma_g^2)$, **W** and **Z** are the corresponding incidence matrices**,** and **e** is the vector of random residual terms of the model, weighted by the MACE proof accuracy as proposed by De Los Campos et al., (2013), as $\mathbf{e} \sim N(0,\mathbf{D}\sigma_e^2)$.

The λ parameter is a smoothing parameter controlling the shrinkage of the double exponential distribution; λ^2 is distributed a priori as a gamma distribution with a shape and scale hyperparameters. **G** is the genomic relationship matrix built from Illumina Bovine 50K genotypes. 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:

$$G = \frac{ZZ'}{2\sum p_i(1-p_i)}$$

where p_i is the frequency of the minor allele at locus *i*, **Z** = (**M - P**) is a matrix that results from the subtraction of **P** from **M**, being **P** = 2(p_i − 0.5), and **M** the matrix of genotypes codified as –1, 0, and 1 for the homozygote, heterozygote, and other homozygote, respectively, following VanRaden (2008), where a more detailed description of this model is provided.

Then, **D** is a diagonal matrix with elements $\{\frac{1-r_i}{r_i}\}$, where r_i is the reliability of individual *i*. Finally, σ_g^2 and σ_e^2 are the additive polygenic and residual variances, respectively. The Bayesian model was solved for each chromosome separately using Gibbs Sampling, with a chain length of 10,000 and a burn-in period of 1000.

It should be noted that the total GEBV obtained from the prediction models consisted of the sum of the estimated haplotype effects and the polygenic effect estimate as:

$$\boldsymbol{GEBV} = \sum \textit{haplotype effects} + \textit{polygenic effects}$$

Haplotype Selection

Selection of a limited number of haplotypes, i.e. those with the largest prediction ability, is expected to be useful in routine genomic evaluations. Hence, haplotypes whose effect was larger / lower than the mean plus /minus 3 times the standard deviation and one standard deviation above the mean (μ_h) of the haplotypes effect distribution, were selected for each trait.

$$|\widehat{h}| > \mu_h + 3sd_h$$

$$|\widehat{h}| > \mu_h + sd_h$$

We attempted to estimate the effects of haplotype that exceeded these threshold with the goal to identify the most influential haploblocks for each trait. The analysis was repeated incorporating only haplotypes that exceeded each threshold using the model described above. Genetic variance explained by sequence data was calculated for each trait by analysing all chromosomes simultaneously.

Results and Discussion

Haplotype construction

In this study haplotype segments on each chromosome were extracted. The length and the number of these segments varied depending on the extent of LD present and on the chromosome length. Table 2 shows a descriptive summary of chromosomes and the number of haplotype blocks that were found on the 29 *Bos taurus* autosomal chromosomes (BTA) using Findhap algorithm. The total autosomal genome length was 2512.06 Mb with the shortest BTA 25 being 42.90 Mb and the longest BTA 1 being 158.33 Mb. The number of SNPs in the haploblock ranged from 799 in BTA 13 to 1285 in BTA 12, with a mean of 924 SNP (166,552 pb). The BTA 1 showed the highest number of haplotype blocks (961) and remaining haplotypes (9363) while the BTA 25 presented the smallest number of blocks (261) blocks and haplotypes (2788). Unique haplotypes were around 90% and haplotypes with a frequency below 1% were around 97% in all chromosomes. These haplotypes were not used in this analysis due to the difficulty of finding statistical effects when the haplotype is present in only a couple of individuals in our sample. Then, low frequency haplotype alleles (<1%) were ignored leaving 153,428 haplotypes for the other analysis.

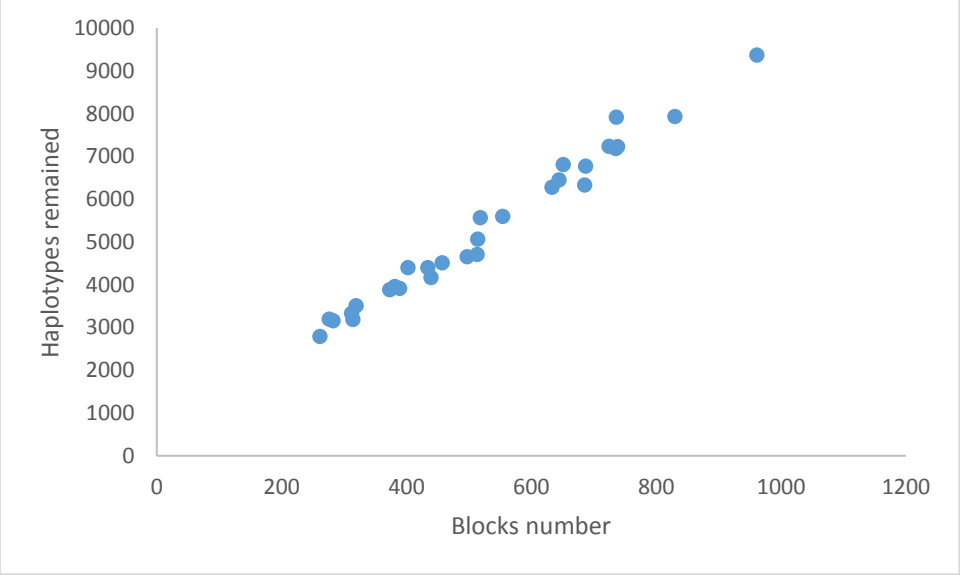

Figure 2: Distribution of haplotype block per chromosome

The number of genome-wide haplotype blocks is shown in Figure 2 against the number of haplotypes remaining after filtering. The distribution of haplotype block is proportional to the number of haplotypes. The larger the number of blocks the larger the number of

haplotypes. This suggests that the genetic variability within chromosome was proportional to the chromosome length, and we could not detect any chromosome with larger or lower variability than expected. The large haplotypes verified the existence of high LD in the BovineHD chip and validated the merits of haplotype analysis. LD in Holstein is very high due to the low effective size associated with the high intensity of selection and the bottleneck history of this breed that occurred around the middle 20th century.

Table 2: Genome-wide summary of haplotype blocks in the Holstein cattle of this study

Autosome	Autosome length (pb)	Haplotype Number	Blocks Number	Blocks length(SNP)	Unique haplotypes (%)	Haplotype freq<1% (%)	Haplotype remained
BTA1	158334731	397685	961	895	90.26	97.65	9363
BTA2	137060366	335343	830	835	89.69	97.64	7930
BTA3	121430266	302697	735	844	90.27	97.63	7179
BTA4	120825133	303697	736	914	89.78	97.39	7914
BTA5	121190985	314912	738	899	90.68	97.71	7224
BTA6	119458581	300751	724	939	90.15	97.59	7237
BTA7	112638649	285068	685	834	90.84	97.78	6328
BTA8	113383722	285048	687	834	90.31	97.62	6772
BTA9	105708161	272927	644	879	90.40	97.64	6446
BTA10	104304932	259351	633	909	89.25	97.58	6277
BTA11	107310498	272252	651	860	90.11	97.50	6807
BTA12	91163122	252401	554	1285	86.26	97.78	5597
BTA13	84240314	212863	513	799	90.48	97.79	4707
BTA14	84648338	206927	514	834	89.48	97.55	5060
BTA15	85295694	218218	518	1008	90.13	97.45	5565
BTA16	81724537	205580	497	859	90.40	97.74	4655
BTA17	75158596	197683	457	979	90.54	97.72	4507
BTA18	66003508	175914	402	899	90.66	97.50	4399
BTA19	64057258	166221	389	859	90.56	97.65	3910
BTA20	72041471	180605	439	898	89.87	97.69	4164
BTA21	71599084	183741	434	875	90.66	97.61	4398
BTA22	61435160	152167	373	820	89.49	97.45	3879
BTA23	52529233	137483	319	1215	89.30	97.45	3507
BTA24	62714571	155380	381	910	89.52	97.46	3953
BTA25	42904110	108716	261	923	89.26	97.44	2788
BTA26	51680365	128747	314	913	89.17	97.53	3182
BTA27	45407501	116888	276	1039	89.02	97.27	3195
BTA28	46312540	118038	282	982	89.37	97.33	3157
BTA29	51505224	134472	312	1060	90.57	97.53	3328

Alternative block lengths were also analyzed, considering the values recommended by VanRaden et al., (2011). These values were 100,000 and 2,000 SNP for the max and the

min length respectively. After filtering the singletons, 76,512 haplotypes were retained for the other analysis. The percentage of the variance of the GEBV that was explained by the haplotypes for each character was very low because of the reduced number of haplotypes included in the analysis. Given these non-satisfactory results, it was decided to increase the number of haplotypes by reducing their lengths based on the results of a recent study of Qanbari et al., (2010) as described in the part of Materials and Methods.

Haplotypes have been extensively explored in human genetics research (Curtis et al., 2001; Gabriel et al., 2002; Chapman et al., 2003; Curtis, 2007). More recent studies in animal breeding explore the use of haplotypes for genomic prediction of breeding values, but using low to medium density marker data (Calus et al., 2008; Villumsen et al., 2009; Boichard et al., 2012; Calus et al., 2009; Schrooten et al., 2013).

It is expected that there was an optimal haplotype length, which depends on the distance between the markers and extend of LD in the population. Reliabilities for GEBV were investigated by simulation to test the hypothesis that there is an optimal haplotype size for genomic predictions. Studies based on real data in dairy cattle are limited. Villumsen et al., (2009) in their study with 30K SNP chip, to test the hypothesis that there is an optimal haplotype size for genomic predictions and that genomic predictions are accurate for moderate and low heritability traits in a dairy cattle setting, showed a clear relationship between the size of haplotypes used in the prediction model and the reliabilities obtained. For their tested haplotype lengths, the optimal size of haplotypes was 10 SNP for heritabilities of 0.3 and 0.02. They observed a relationship between heritability and reliabilities; as heritability decreased so did the reliability.

The optimal haplotype size is very dependent on marker spacing and marker frequencies. If marker distance is low the nearest marker may not be the best predictor of the QTL effect, and a better predictor may be found at a larger distance (Zondervan and Cardon, 2004). On the other side, a recent study showed that better predictions in dairy cattle can be obtained by using a set of haploblocks with a fixed size (number of SNPs) (Boichard et al., 2012).

There are many published studies on haplotype block properties for cattle, which vary in many aspects (breed of interest, marker types, marker density, and chromosome regions), yielding average haplotype block sizes from a few kb in length: 5.7 kb considering 2 or

more SNPs (Villa-Angulo et al., 2009), 26.2 kb considering 4 or more SNPs (Kim and Kirkpatrick, 2009) to hundreds of kb in length: 700 kb (Gautier et al., 2007). However, these studies used smaller marker densities, with an average distance of 62 kb between adjacent markers (Qanbari et al., 2010).

Several methods have been used to construct haplotypes for genomic evaluation (Calus et al., 2008, 2009; Boichard et al., 2012; Cuyabano et al., 2014). Allele effect predictability can be defined as the expected prediction accuracy of the effect of haplotype alleles, and it is expected to have a significant effect on the performance of genomic prediction. However, none of the previously mentioned methods take into account any information on this predictability. The construction of haplotypes at a particular SNP position by merging this SNP with the flanking markers is straightforward. However, because of the short distance between the markers, the resulting haplotypes most frequently include a small number of over-represented alleles together with a large number of alleles with low frequencies within the population (Jónás et al., 2016).

The choice of using haplotypes to perform genomic prediction is a reasonable approach, under the hypothesis that haplotypes are expected to be in stronger LD to the causative mutations (or QTLs) than any single marker. Furthermore, when it comes to sequence data, haplotypes offer the possibility to reduce the number of explanatory variables in genomic prediction models compared with the individual SNP approach, depending on the chosen techniques to build haplotype blocks (haploblocks).

Haplotype effect estimates

Manhattan plots with estimated haplotypes effects were adopted to show the results from Prot, IGT, SCS and DO, respectively (Figs 3, 4, 5 and 6). Chromosomes 1–29 are shown separated by colors, and haplotypes effects are plotted as dots. In each plot, the genome wide threshold of 3 standard deviation is shown as a horizontal reference line. The figures show that it is possible to detect some regions on the genome that explain relevant effects for the studied traits. Many regions with large effects were detected. A total of 1264 haplotype exceeded the genome wide threshold for Prot, 1909 for IGT, 851 for SCS and 1450 for DO distributed along the genome.

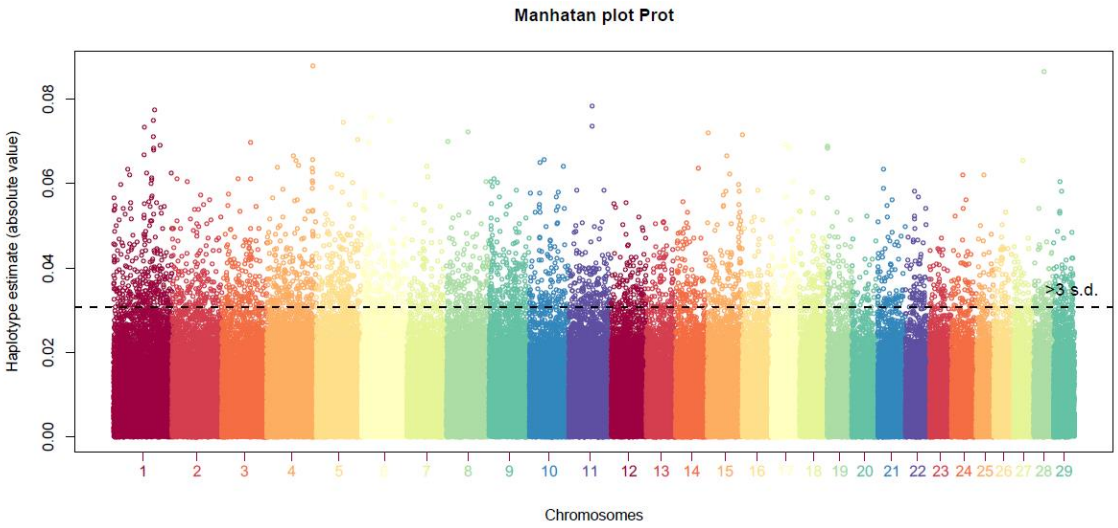

Figure 3: Manhattan plot with estimated haplotypes effects for kg of protein Prot

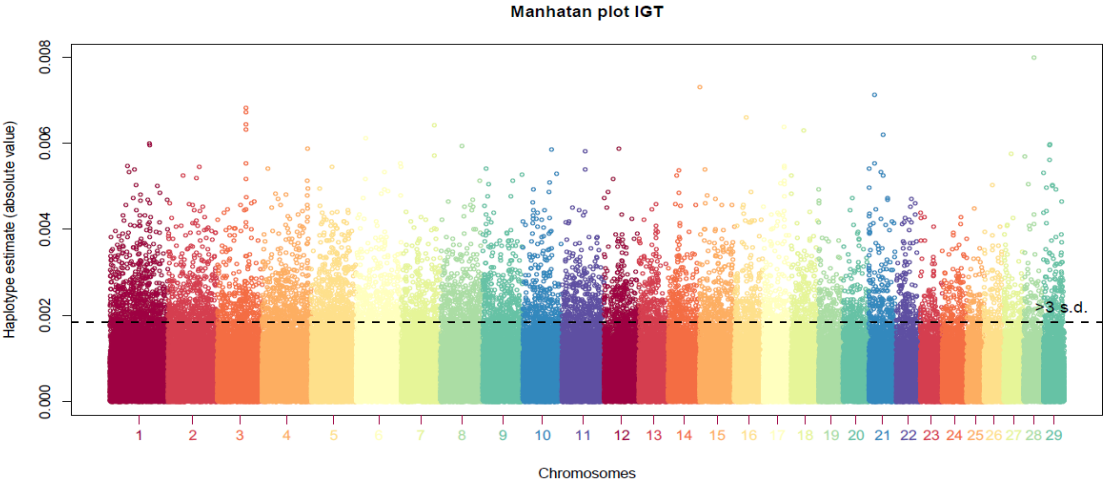

Figure 4: Manhattan plot with estimated haplotypes effects for IGT

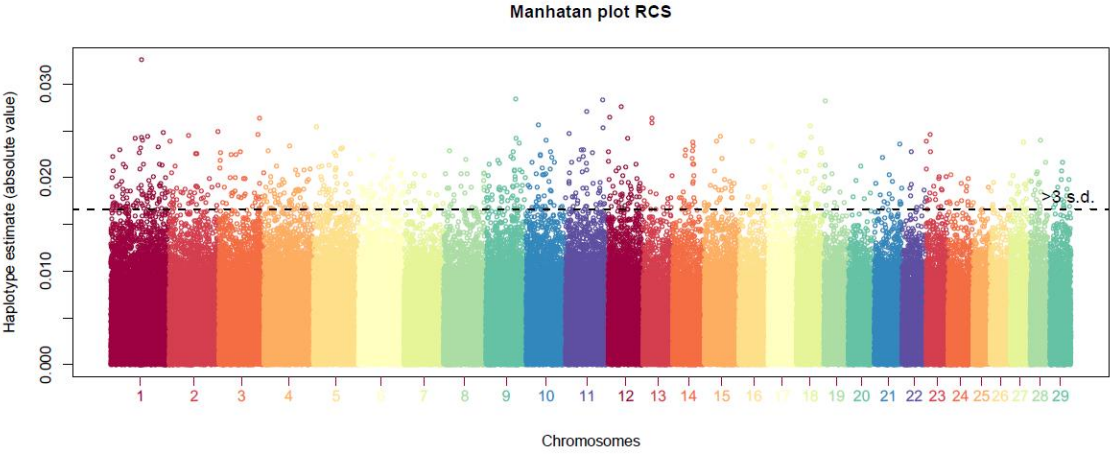

Figure 5: Manhattan plot with estimated haplotypes effects for SCS

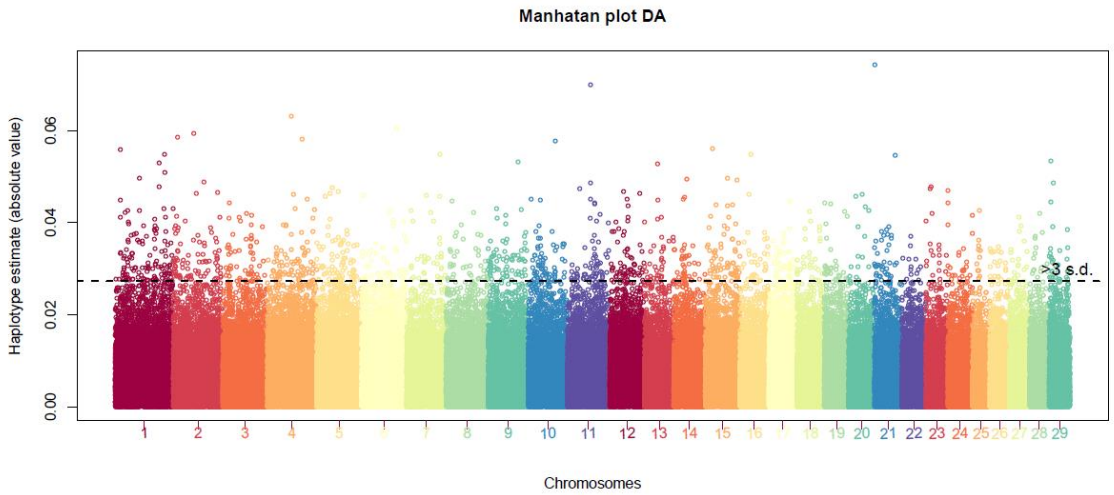

Figure 6: Manhattan plot with estimated haplotypes effects for DO

Our hypothesis is that the use of sequence data in genomic prediction would result in a larger predictive accuracy in genomic evaluations. Some studies showed increased predictive accuracy with sequence data using simulations (Meuwissen and Goddard, 2010; MacLeod et al., 2013). According to Wimmer et al., (2013), increasing the number of individuals in the training dataset or pre-selecting SNPs based on other sources of information might be necessary to increase prediction reliability based on sequence data, also reported in Hayes et al., (2014). These authors obtained a very small increase of 2 % in prediction reliability using imputed sequence data compared to BovineHD. However, they applied strict a-priori filtering steps for the SNPs and ended up with around 1.7 million variants. They claimed that advantage of sequence data compared to SNP Chip genotypes might be larger with large training set, and pre-selection of SNPs based on functional information. An efficient haplotype selection procedure from the haplotypes that exceed the threshold is required to identify the haplotypes most suitable for genomic evaluation purposes to achieve a high predictive accuracy.

Distribution of the allele frequency of haplotypes

Figure 7 shows the distribution of haplotypes allelic frequencies that have exceeded the threshold for each character. Most of the haplotypes for Prot, IGT and DO had low-intermediate frequencies while haplotypes found for SCS are at low frequencies, which may be of interest. Therefore, we expect that these haplotypes will give us additional information to SNP genotypes on those less common variants. It is necessary to explore their contribution to genetic variation.

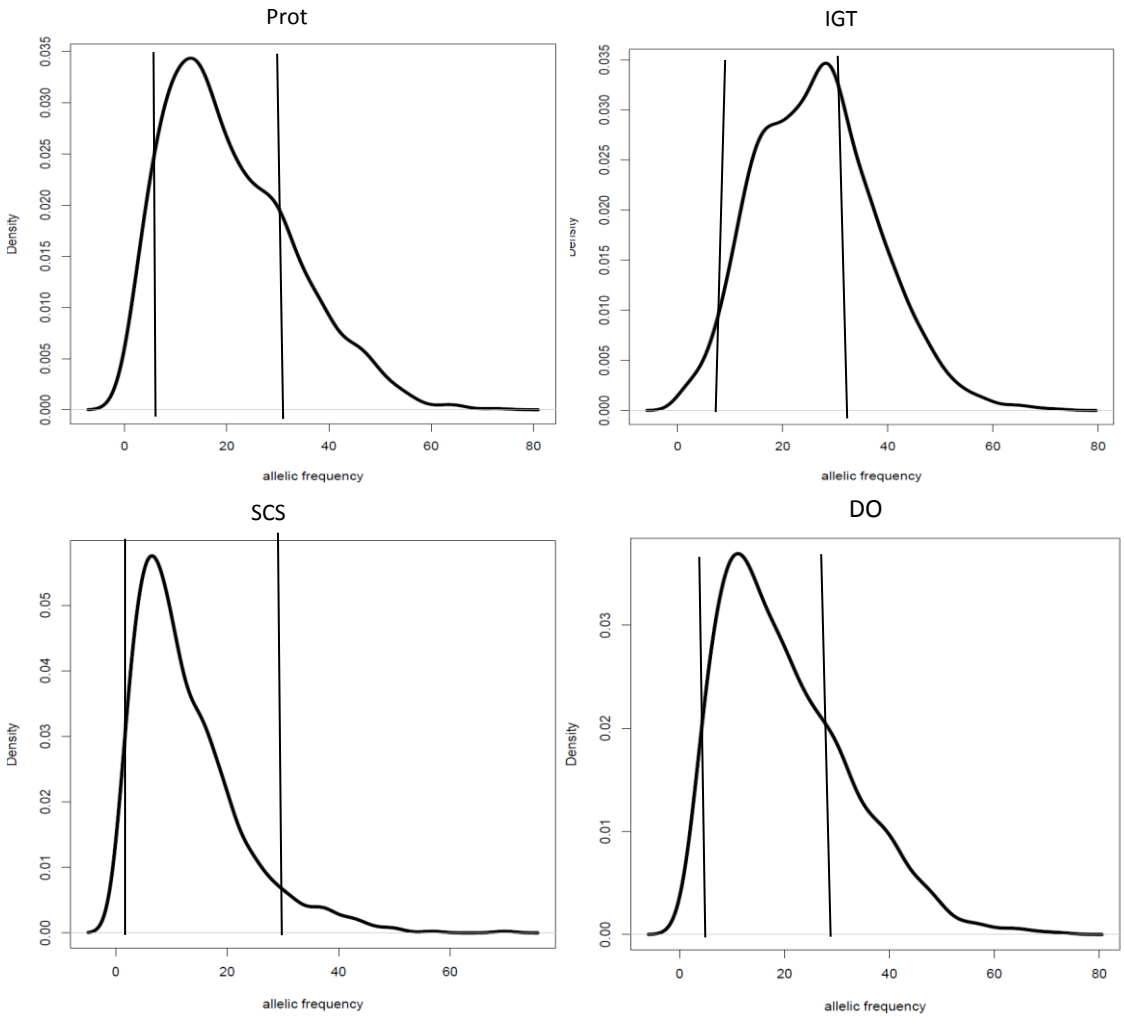

Figure 7. Allele frequencies distribution of haplotypes that have exceeded the threshold for each character

One motivation for using whole genome sequence data in genomic prediction and GWAS is that whole genome sequence data will include rare variants which may explain some extra variation in the targeted complex traits. SNP arrays have limited power to capture this variation, as the SNP on these arrays are selected to have high MAF, and are therefore unlikely to be in high LD with the rare variants (Hayes et al., 2015). Although, it is necessary to differentiate them from sequencing errors (Gonzalez-Recio et al., 2015).

An important advantage of haplotypes over single SNP markers is their higher ability to identify mutations. In animal breeding studies, SNPs are commonly bi-allelic and even when mutations have occurred it is possible that the allele frequencies remain (almost) unaltered. However, when haplotypes were analyzed, mutations in different loci tended to cause major changes in the haplotype frequencies (Curtis et al., 2001). Thus, a QTL

that is not in complete LD with any individual bi-allelic SNP marker may be in complete LD with a multi-marker haplotype.

Genetic variance explained by sequence data

The GEBV were estimated as the sum of the haplotypes effects plus the polygenic effect. Muir (2007) suggested that a polygenic component should be included in the GEBV, to capture any genetic variance not associated with the markers, for instance low-frequency QTL that may not be captured by the markers. This strategy has already been adopted by Australia, US, and New Zealand in their official genomic evaluations (Hayes et al., 2009).

The following Table 3 shows the proportion of GEBV variance that corresponds to the haplotype effects. The haplotypic genetic variance was estimated as the ratio of variance explained by haplotype over the total GEBV variance. Haplotypes estimated using the Bayesian LASSO model had a large contribution to the total variance of GEBV (between 32 and 99.9%). Haplotypes for SCS contributed with larger percentage (99.9%) compared to the other traits, although this seems likely to be an artifact caused by data structure, and the large p small n problem, and the lower heritability of the trait.

Table 3: Percentage of the genomic estimated values variance that was explained by the haplotypes for each character

	Kg Prot (%)	**IGT (%)**	**DO (%)**	**SCS (%)**
All	32.75	71.93	73.76	99.90
>1sd BL	10.92	N.C.[1]	53.93	33.30
>3sd BL	1.06	5.24	11.64	15.29

[1]No convergence obtained.

In order to reduce the number of haploblocks needed to perform genomic prediction, a subsets of haploblocks which contain haplotypes with large effects was selected. Two selection criteria were tested. The first one was 3sd (in absolute value above mean) which led to a total of 1264 haplotype for Prot, 1909 for IGT, 851 for SCS and 1450 for DO. The second criterion (1 sd in absolute value above mean) led to a total of 44,319 haplotypes for Prot, 39,975 for IGT, 46,132 for SCS and 42,878 for DO, distributed along the genome. Then, haplotypes that exceeded each threshold were subjected to a new analysis with Bayesian LASSO.

In this study, the larger number of haplotypes, the larger genetic variance that was captured for all traits under study. In this sense, using haplotypes that exceed the threshold of 1 sd captured larger proportion of variance than those exceeding the threshold of 3 sd.

Filtering for 3sd decreased the proportion of the genetic variance explained by the sequence data (haplotypes) for all the traits compared to filtering by 1 sd. This decline was 90% for kg Prot, 78% for DO and 54% for SCS. The model did not converge at the threshold of 1sd for IGT. This is probably due to a larger proportion of missing data for this trait, which accentuates the large *p* small *n* problem, as there was 39,965 haplotypes and only 348 phenotypes. We observed that the decline was more pronounced for kg of protein. This can be explained by a too strict criterion when filtering by 3sd. In this case selected haplotypes might be pointing to few genomic regions strongly associated to the traits, and with a large number of haplotypes each, but not representative of the whole genetic architecture (failing to identify/select other regions).

One limitation of this study is the reduced number of individuals (361) with phenotypic data were used to estimate the effects of over 46,000 haplotypes when filtering on the 3sd criterion. Thus, the number of haplotypes (*p*) was much larger than the number of observation (*n*). In this scenario, the QTL effect might be estimated with large error, which reduces the advantage of using sequence data compared to SNP genotypes for genomic prediction (Druet et al., 2014).

In addition, the choice of the prior distribution for λ^2 could potentially influence the results. Consistent results and convergence were observed when using scale hyperparameters of 0.0001 for 1sd and 0.00001 for 3sd. These hyperparameters affected the convergence of the Monte Carlo Markov Chain and should be chosen carefully, for example with a grid search, as done in this study: the hyperparameters for the lambda^2 prior distribution were set by a grid search with values ranging from 0.0000001 to 1.

Haplotypes provide valuable information on genetic variance and may lead to the development of more efficient strategies to identify genetic variants associated with traits of economic interest.

Genomic predictions using a set of appropriately selected haploblocks are expected to achieve higher prediction accuracy while reducing the amount of predictor variables in prediction models. Using preselected haploblocks for genomic prediction is an important area of research. Reliability of genomic prediction for a trait as well as persistency across generations are expected to improve by identifying the most influential haploblocks and include them in the prediction model. In addition, genomic predictive models including a selected group of haploblocks will reduce computing time considerably, compared to models using all haploblocks, and is more important when using whole-genome sequence data.

This study allowed us to observe the possibilities that exist at incorporating sequenced data from the 1000 bull genomes project in routine genomic evaluations.

Conclusions

This study utilized sequence data on 31.8 millions variants from 450 sires. We constructed haplotypes throughout the genome that were subsequently used as explanatory variables of progeny MACE proofs for Prot, IGT, DO and SCS. We concluded that:

The algorithm implemented in Findhap can extract haplotypes in the population studied, although it is highly dependent on the parameters set by the user for its implementation, and it is necessary to apply biological knowledge a priori to approximate an appropriate length of haplotypes based on data LD. In this study, the number of SNPs in the haploblock ranged from 799 in BTA 13 to 1285 in BTA 12, with a mean of 924 SNP (166.552 pb). The chromosomes 1 and 25 had the highest and lowest number of blocks and haplotypes respectively. The haplotype blocks were expected to be large because of the high LD in Holstein and confirmed the existence of high LD in the BovineHD chip.

Unique haplotypes and low frequency haplotype alleles appeared in a large proportion (97%) in all chromosomes, and must be utilized carefully or filtered out. Nonetheless, a large number of haplotypes (153,428) were still useful for genomic prediction.

Sequenced regions that are associated to traits of economic interest were detected and could be used in a Bayesian regression model incorporating a polygenic effect for genomic evaluations in the Spanish dairy cattle. The haplotypes with larger effect on the traits were those at low frequency, mainly for SCS. This reveals that NGS data will provide additional information to SNP genotyping on the less common variants and their contribution to genetic variation.

Haplotypes contributed highly to the variance of GEBV (ranging between 32 and 99.9%). It has been hypothesized that an appropriate selection of a subset of haplotype blocks can result in satisfactory or better predictive ability than SNP genotypes. The proportion of variance captured by sequence data is related not only to the nature of trait, but also to the number of haplotypes incorporated in the analysis model, and the available phenotypes, facing difficulties due to dimensionality problems (large p small n problem). Given the availability of data from the 1000 bull genome project, filtering by 3 sd would not be enough to capture a large proportion of genetic variance, whereas filtering by 1sd could be useful but caution must be taken in terms of model convergence and the choice of hyperparameters in the prior distribution of haplotype effects.

Genomic predictions using only a set of appropriately selected haploblocks can provide additional information to GEBV prediction, but increase in predictive accuracy must be checked in future studies. There is a need for statistical models that better capture genetic variance when under high dimensionality problems.

For future studies we recommend:

- The increase of the number of individuals with phenotypes and genotypes in the analysis model to capture a large proportion of genetic variance explained by sequence data;
- The use of less stringent thresholds to filter relevant haplotypes;
- Imputation of these haplotypes in the population and make cross-validation to determine the increase on predictive ability of these variants over SNP genotypes;
- Also, further research is needed to improve strategies to select optimal haplotype lengths.

References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 2010. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93:743–52. doi:10.3168/jds.2009-2730.

Aguilar, I., I. Misztal, S. Tsuruta, G. Wiggans, and T. Lawlor. 2011. Multiple trait genomic evaluation of conception rate in Holsteins. *J. Dairy Sci.* 94:2621–2624. doi:10.3168/jds.2010-3893.

Amin, N., C.M. van Duijn, and Y.S. Aulchenko. 2007. A genomic background based method for association analysis in related individuals. *PLoS One*. 2. doi:10.1371/journal.pone.0001274.

Boichard, D., F. Guillaume, A. Baur, P. Croiseau, M.N. Rossignol, M.Y. Boscher, T. Druet, L. Genestout, J.J. Colleau, L. Journaux, V. Ducrocq, and S. Fritz. 2012. Genomic selection in French dairy cattle. *Anim. Prod. Sci.* 52:115–120. doi:10.1071/AN11119.

Bouwman, A.C., and R.F. Veerkamp. 2014. Consequences of splitting whole-genome sequencing effort over multiple breeds on imputation accuracy. *BMC Genet.* 15:105. doi:10.1186/s12863-014-0105-8.

Brondum, R.F., B. Guldbrandtsen, G. Sahana, M.S. Lund, and G. Su. 2014. Strategies for imputation to whole genome sequence using a single or multi-breed reference population in cattle. *BMC Genomics*. 15:728. doi:10.1186/1471-2164-15-728.

Browning, B.L., and S.R. Browning. 2008. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.* 84:210–223. doi:10.1016/j.ajhg.2009.01.005.

Browning, B.L., and S.R. Browning. 2013. Improving the accuracy and efficiency of identity-by-descent detection in population data. *Genetics*. 194:459–471. doi:10.1534/genetics.113.150029.

Calus, M.P.L., Y. De Haas, M. Pszczola, and R.F. Veerkamp. 2013. Predicted accuracy of and response to genomic selection for new traits in dairy cattle. *Anim. Anim. Consort.* 7:183–191. doi:10.1017/S1751731112001450.

Calus, M.P.L., T.H.E. Meuwissen, A.P.W. De Roos, and R.F. Veerkamp. 2008. Accuracy of genomic selection using different methods to define haplotypes. *Genetics*. 178:553–561. doi:10.1534/genetics.107.080838.

Calus, M.P.L., T.H.E. Meuwissen, J.J. Windig, E.F. Knol, C. Schrooten, A.L.J. Vereijken, and R.F. Veerkamp. 2009. Effects of the number of markers per haplotype and clustering of haplotypes on the accuracy of QTL mapping and prediction of genomic breeding values. *Genet. Sel. Evol.* 41:11. doi:10.1186/1297-9686-41-11.

Chapman, J.M., J.D. Cooper, J.A. Todd, and D.G. Clayton. 2003. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* 56:18–31. doi:73729.

Chen, C.Y., I. Misztal, I. Aguilar, S. Tsuruta, T.H.E. Meuwissen, S.E. Aggrey, T. Wing, and W.M. Muir. 2011. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. *J. Anim. Sci.* 89:23–28. doi:10.2527/jas.2010-3071.

Christensen, O.F., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol.* 42:2. doi:10.1186/1297-9686-42-2.

Christensen, O.F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. *Animal*. 6:1565–1571. doi:10.1017/S1751731112000742.

Cleveland, M.A., S. Forni, N. Deeb, and C. Maltecca. 2010. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. *BMC Proc.* 4:S6. doi:10.1186/1753-6561-4-S1-S6.

Croiseau, P., A. Legarra, F. Guillaume, S. Fritz, A. Baur, C. Colombani, C. Robert-Granié, D. Boichard, and V. Ducrocq. 2011. Fine tuning genomic evaluations in dairy cattle through SNP pre-selection with the Elastic-Net algorithm. *Genet. Res. (Camb).* 93:409–417. doi:10.1017/S0016672311000358.

Crossa, J., P. Pérez, J. Hickey, J. Burgueño, L. Ornella, J. Cerón-Rojas, X. Zhang, S. Dreisigacker, R. Babu, Y. Li, D. Bonnett, and K. Mathews. 2014. Genomic prediction in CIMMYT maize and wheat breeding programs. *Heredity (Edinb).* 112:48–60. doi:10.1038/hdy.2013.16.

Curtis, D. 2007. Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association. *BMC Genet.* 8:49. doi:10.1186/1471-2156-8-49.

Curtis, D., B. V North, and P.C. Sham. 2001. Use of an artificial neural network to detect association between a disease and multiple marker genotypes. *Ann. Hum. Genet.* 65:95–107.

Cuyabano, B.C., G. Su, and M.S. Lund. 2014. Genomic prediction of genetic merit using LD-based haplotypes in the Nordic Holstein population. *BMC Genomics*. 15:1171. doi:10.1186/1471-2164-15-1171.

Cuyabano, B.C., G. Su, and M.S. Lund. 2015. Selection of haplotype variables from a high-density marker map for genomic prediction. *Genet. Sel. Evol.* 47:61. doi:10.1186/s12711-015-0143-3.

Daetwyler, H.D., A. Capitan, H. Pausch, P. Stothard, R. van Binsbergen, R.F. Brøndum, X. Liao, A. Djari, S.C. Rodriguez, C. Grohs, D. Esquerré, O. Bouchez, M.-N. Rossignol, C. Klopp, D. Rocha, S. Fritz, A. Eggen, P.J. Bowman, D. Coote, A.J. Chamberlain, C. Anderson, C.P. VanTassell, I. Hulsegge, M.E. Goddard, B. Guldbrandtsen, M.S. Lund, R.F. Veerkamp, D.A. Boichard, R. Fries, and B.J. Hayes. 2014. Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat. Genet.* 46:858–865. doi:10.1038/ng.3034.

Daetwyler, H.D., B. Villanueva, and J.A. Woolliams. 2008. Accuracy of predicting the genetic risk of disease using a genome-wide approach. *PLoS One*. 3. doi:10.1371/journal.pone.0003395.

Dassonneville, R., A. Baur, S. Fritz, D. Boichard, and V. Ducrocq. 2012. Inclusion of cow records in genomic evaluations and impact on bias due to preferential treatment. *Genet. Sel. Evol.* 44:40. doi:10.1186/1297-9686-44-40.

Dassonneville, R., R.F. Brøndum, T. Druet, S. Fritz, F. Guillaume, B. Guldbrandtsen, M.S. Lund, V. Ducrocq, and G. Su. 2011. Effect of imputing markers from a low-density chip on the reliability of genomic breeding values in Holstein populations. *J. Dairy Sci.* 94:3679–3686. doi:10.3168/jds.2011-4299.

De Haas, Y., M.P.L. Calus, R.F. Veerkamp, E. Wall, M.P. Coffey, H.D. Daetwyler, B.J. Hayes, and J.E. Pryce. 2012. Improved accuracy of genomic prediction for dry matter intake of dairy cattle from combined European and Australian data sets. *J. Dairy Sci.* 95:6103–12. doi:10.3168/jds.2011-5280.

Dekkers, J.C.M. 2004. Commercial application of marker-and gene-assisted selection in livestock: Strategies and lessons 1,2. *J. Anim. Sci*. 82:313–328.

De Los Campos, G., D. Gianola, and G.J.M. Rosa. 2009a. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. *J. Anim. Sci.* 87:1883–7. doi:10.2527/jas.2008-1259.

De Los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009b. Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*. 182:375–385. doi:10.1534/genetics.109.101501.

De Los Campos, G., A.I. Vazquez, R. Fernando, Y.C. Klimentidis, and D. Sorensen. 2013. Prediction of Complex Human Traits Using the Genomic Best Linear Unbiased Predictor. *PLoS Genet.* 9. doi:10.1371/journal.pgen.1003608.

De Roos, A.P.W., B.J. Hayes, and M.E. Goddard. 2009. Reliability of genomic predictions across multiple populations. *Genetics*. 183:1545–1553. doi:10.1534/genetics.109.104935.

Druet, T., I.M. Macleod, and B.J. Hayes. 2014. Toward genomic prediction from whole-genome sequence data: impact of sequencing design on genotype imputation and accuracy of predictions. *Heredity (Edinb).* 112:39–47. doi:10.1038/hdy.2013.13.

Druet, T., C. Schrooten, and a P.W. de Roos. 2010. Imputation of genotypes from different single nucleotide polymorphism panels in dairy cattle. *J. Dairy Sci.* 93:5443–54. doi:10.3168/jds.2010-3255.

Erbe, M., B.J. Hayes, L.K. Matukumalli, S. Goswami, P.J. Bowman, C.M. Reich, B.A. Mason, and M.E. Goddard. 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. *J. Dairy Sci.* 95:4114–29. doi:10.3168/jds.2011-5019.

Forni, S., I. Aguilar, and I. Misztal. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol.* 43:1–7. doi:10.1186/1297-9686-43-1.

Fragomeni, B.O., D.A.L. Lourenco, S. Tsuruta, Y. Masuda, I. Aguilar, A. Legarra, T.J. Lawlor, and I. Misztal. 2015. Hot topic: Use of genomic recursions in single-step genomic best linear unbiased predictor (BLUP) with a large number of genotypes. *J. Dairy Sci.* 98:4090–4094. doi:http://dx.doi.org/10.3168/jds.2014-9125.

Friedman, J.H. 2001. GREEDY FUNCTION APPROXIMATION: A GRADIENT BOOSTING MACHINE. *Ann. Stat.* 29:1189–1232.

Gabriel, S.B., S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S.N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E.S. Lander, M.J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. *Science*. 296:2225–9. doi:10.1126/science.1069424.

Gao, H., P. Madsen, U.S. Nielsen, G.P. Aamand, G. Su, K. Byskov, and J. Jensen. 2015. Including different groups of genotyped females for genomic prediction in a Nordic Jersey population. *J. Dairy Sci.* 98:9051–9. doi:10.3168/jds.2015-9947.

Gautier, M., T. Faraut, K. Moazami-Goudarzi, V. Navratil, M. Foglio, C. Grohs, A. Boland, J.G. Garnier, D. Boichard, G.M. Lathrop, I.G. Gut, and A. Eggen. 2007. Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics*. 177:1059–1070. doi:10.1534/genetics.107.075804.

George, E.L., and McCulloch. 1993. Variable selection via {Gibbs} sampling. *J Am Stat Assoc*. 91:883–904. doi:10.1080/01621459.1993.10476353.

Gianola, D., R.L. Fernando, and A. Stella. 2006. Genomic-Assisted Prediction of Genetic Value with Semiparametric Procedures. *Genetics*. 173:1761–1776. doi:10.1534/genetics.105.049510.

Gianola, D., and J.B.C.H.M. Van Kaam. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*. 178:2289–2303. doi:10.1534/genetics.107.084285.

Gianola, D., G. De Los Campos, W.G. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. *Genetics*. 183:347–363. doi:10.1534/genetics.109.103952.

Gianola, D., H. Okut, K.A. Weigel, and G.J. Rosa. 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet.* 12:87. doi:10.1186/1471-2156-12-87.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica*. 136:245–257. doi:10.1007/s10709-008-9308-0.

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. *J. Anim. Breed. Genet.* 124:323–30. doi:10.1111/j.1439-0388.2007.00702.x.

Goddard, M.E., and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programmes. *Nat Rev Genet*. 10:381–391. doi:nrg2575 [pii]\r10.1038/nrg2575.

González-Recio, O., J. Bazán, and S. Forni. 2011. An opensource library to implement random forests in genome-wide prediction. *EAAP. 62nd Annu. Meet. Stavanger*.

Gonzalez-Recio, O., H.D. Daetwyler, I.M. MacLeod, J.E. Pryce, P.J. Bowman, B.J. Hayes, and M.E. Goddard. 2015. Rare Variants in Transcript and Potential Regulatory Regions Explain a Small Percentage of the Missing Heritability of Complex Traits in Cattle. *PLoS One*. 10:e0143945. doi:10.1371/journal.pone.0143945.

González-Recio, O., and S. Forni. 2010. Random Forest algorithm with RanFoG. *XV Reun. Nac. Mejor. Genética Anim. Vigo*. 1–6.

González-Recio, O., and S. Forni. 2011. Genome-wide prediction of discrete traits using bayesian regressions and machine learning. *Genet. Sel. Evol.* 43:7. doi:10.1186/1297-9686-43-7.

González-Recio, O., D. Gianola, N. Long, K.A. Weigel, G.J.M. Rosa, and S. Avendaño. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. *Genetics*. 178:2305–2313. doi:10.1534/genetics.107.084293.

González-Recio, O., D. Gianola, G.J. Rosa, K.A. Weigel, and A. Kranis. 2009a. Genome-assisted prediction of a quantitative trait measured in parents and progeny: application to food conversion rate in chickens. *Genet. Sel. Evol.* 41:3. doi:10.1186/1297-9686-41-3.

Gonzalez-Recio, O., J.A. Jimenez-Montero, and R. Alenda. 2013. The gradient boosting algorithm and random boosting for genome-assisted evaluation in large data sets. *J. Dairy Sci.* 96:614–624. doi:10.3168/jds.2012-5630.

González-Recio, O., E.L. de Maturana, A.T. Vega, C.D. Engelman, and K.W. Broman. 2009b. Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model. *BMC Proc.* 3 Suppl 7:S63. doi:10.1186/1753-6561-3-s7-s63.

González-Recio, O., G.J.M. Rosa, and D. Gianola. 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. *Livest. Sci.* 166:217–231. doi:10.1016/j.livsci.2014.05.036.

González-Recio, O., K. a Weigel, D. Gianola, H. Naya, and G.J.M. Rosa. 2010. L2-Boosting algorithm applied to high-dimensional problems in genomic selection. *Genet. Res. (Camb).* 92:227–37. doi:10.1017/S0016672310000261.

Habier, D., R.L. Fernando, and J.C.M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*. 177:2389–2397. doi:10.1534/genetics.107.081190.

Habier, D., R.L. Fernando, K. Kizilkaya, and D.J. Garrick. 2011. Extension of the bayesian alphabet for genomic selection. *BMC Bioinformatics*. 12:186. doi:10.1186/1471-2105-12-186.

Habier, D., J. Tetens, F.-R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. *Genet. Sel. Evol.* 42:5. doi:10.1186/1297-9686-42-5.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning Data Mining, Inference, and Prediction. *Math. Intell.* 27:83–85. doi:10.1007/b94608.

Hayes, B. 2007. QTL Mapping, MAS, and Genomic Selection. *Program*. 118.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Genomic selection in dairy cattle: progress and challenges. *J. Dairy Sci.* 92:433–443. doi:92/2/433 [pii];10.3168/jds.2008-1646 [doi].

Hayes, B.J., P.J. Bowman, H.D. Daetwyler, and M.E. Goddard. 2015. WHY CAN WE IMPUTE SOME RARE SEQUENCE VARIANTS AND NOT OTHERS? *Proc. Assoc. Advmt. Breed. Genet*.

Hayes, B.J., I.M. MacLeod, H.D. Daetwyler, P.J. Bowman, a J. Chamberlain, C.J. vander Jagt, A. Capitan, H. Pausch, P. Stothard, X. Liao, C. Schrooten, E. Mullaart, R. Fries, B. Guldbrandtsen, M.S. Lund, D. Boichard, R.F. Veerkamp, C. Van Tassell, B. Gredler, T. Druet, A. Bagnato, J. Vilkki, D.-J. de Koning, E. Santus, and M.E. Goddard. 2014. Genomic Prediction from Whole Genome Sequence in Livestock: the 1000 Bull Genomes Project. *Proceedings, 10th World Congr. Genet. Appl. to Livest. Prod. Genomic*. 1–3.

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

Hickey, J.M., J. Crossa, R. Babu, and G. de los Campos. 2012. Factors affecting the accuracy of genotype imputation in populations from several maize breeding programs. *Crop Sci.* 52:654–663. doi:10.2135/cropsci2011.07.0358.

Jiménez-Montero, J.A., O. González-Recio, and R. Alenda. 2012. Genotyping strategies for genomic selection in small dairy cattle populations. *Animal*. 6:1216–1224. doi:10.1017/S1751731112000341.

Jiménez-Montero, J.A., O. González-Recio, and R. Alenda. 2013. Comparison of methods for the implementation of genome-assisted evaluation of Spanish dairy cattle. *J. Dairy Sci.* 96:625–34. doi:10.3168/jds.2012-5631.

Jónás, D., V. Ducrocq, M.-N. Fouilloux, and P. Croiseau. 2016. Alternative haplotype construction methods for genomic evaluation. *J. Dairy Sci.* 99:1–10. doi:10.3168/jds.2015-10433.

Kim, E.-S., and B.W. Kirkpatrick. 2009. Linkage disequilibrium in the North American Holstein population. *Anim. Genet.* 40:279–88. doi:10.1111/j.1365-2052.2008.01831.x.

Kizilkaya, K., R.L. Fernando, and D.J. Garrick. 2010. Genomic prediction of simulated multibreed and purebred performance using observed fifty thousand single nucleotide polymorphism genotypes. *J. Anim. Sci.* 88:544–551. doi:10.2527/jas.2009-2064.

Konstantinov, K. V, and B.J. Hayes. 2009. Comparison of BLUP and Reproducing kernel Hilbert spaces methods for genomic prediction of breeding values in Australian Holstein Friesian cattle. *InProceedings 9th World Congr. Genet. Appl. to Livest. Prod. Leipzig, Ger.*

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci.* 92:4656–4663. doi:10.3168/jds.2009-2061.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single Step, a general approach for genomic selection. *Livest. Sci.* 166:54–65. doi:10.1016/j.livsci.2014.04.029.

Legarra, A., C. Robert-Granié, P. Croiseau, F. Guillaume, and S. Fritz. 2011. Improved Lasso for genomic selection. *Genet. Res. (Camb).* 93:77–87. doi:10.1017/S0016672310000534.

Leutenegger, A.-L., B. Prum, E. Génin, C. Verny, A. Lemainque, F. Clerget-Darpoux, and E.A. Thompson. 2003. Estimation of the inbreeding coefficient through use of genomic data. *Am. J. Hum. Genet.* 73:516–523. doi:10.1086/378207.

Long, N., D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2011a. Marker-assisted prediction of non-additive genetic values. *Genetica*. 139:843–854. doi:10.1007/s10709-011-9588-7.

Long, N., D. Gianola, G.J.M. Rosa, and K.A. Weigel. 2011b. Application of support vector regression to genome-assisted prediction of quantitative traits. *Theor. Appl. Genet.* 123:1065–74. doi:10.1007/s00122-011-1648-y.

Long, N., D. Gianola, G.J.M. Rosa, K.A. Weigel, and S. Avendaño. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: Application to early mortality in broilers. *J. Anim. Breed. Genet.* 124:377–389. doi:10.1111/j.1439-0388.2007.00694.x.

Lourenco, D.A.L., I. Misztal, S. Tsuruta, I. Aguilar, E. Ezra, M. Ron, A. Shirak, and J.I. Weller. 2014. Methods for genomic evaluation of a relatively small genotyped dairy population and effect of genotyped cow information in multiparity analyses. *J. Dairy Sci.* 97:1742–52. doi:10.3168/jds.2013-6916.

Lund, M.S., S.P. de Ross, A.G. de Vries, T. Druet, V. Ducrocq, S. Fritz, F. Guillaume, B. Guldbrandtsen, Z. Liu, R. Reents, C. Schrooten, F. Seefried, and G. Su. 2011. A common reference population from four European Holstein populations increases reliability of genomic predictions. *Genet. Sel. Evol.* 43:43. doi:10.1186/1297-9686-43-43.

MacLeod, I.., B.. Hayes, and M.. Goddard. 2013. WILL SEQUENCE SNP DATA IMPROVE THE ACCURACY OF GENOMIC PREDICTION IN THE PRESENCE OF LONG TERM SELECTION? *Proc. Assoc. Advmt. Anim. Breed. Genet.*

Makowsky, R., N.M. Pajewski, Y.C. Klimentidis, A.I. Vazquez, C.W. Duarte, D.B. Allison, and G. de los Campos. 2011. Beyond missing heritability: Prediction of complex traits. *PLoS Genet.* 7. doi:10.1371/journal.pgen.1002051.

Meuwissen, T., and M. Goddard. 2010. Accurate prediction of genetic values for complex traits by whole-genome resequencing. *Genetics*. 185:623–631. doi:10.1534/genetics.110.116590.

Meuwissen, T., B. Hayes, and M. Goddard. 2016. Genomic selection: A paradigm shift in animal breeding. *Anim. Front.* 6:6. doi:10.2527/af.2016-0002.

Meuwissen, T.H., and M.E. Goddard. 2004. Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet. Sel. Evol.* 36:261–279. doi:10.1186/1297-9686-36-3-261.

Meuwissen, T.H.E., B.J. Hayes, and M.E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*. 157:1819–1829. doi:11290733.

Meuwissen, T.H.E., T. Luan, and J.A. Woolliams. 2011. The unified approach to the use of genomic and pedigree information in genomic evaluations revisited. *J. Anim. Breed. Genet.* 128:429–439. doi:10.1111/j.1439-0388.2011.00966.x.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. *J. Dairy Sci.* 92:4648–4655. doi:10.3168/jds.2009-2064.

Moser, G., B. Tier, R. Crump, M. Khatkar, and H. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. *Genet. Sel. Evol.* 41:56. doi:10.1186/1297-9686-41-56.

Muir, W.M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters. *J. Anim. Breed. Genet.* 124:342–355. doi:10.1111/j.1439-0388.2007.00700.x.

Ober, U., J.F. Ayroles, E.A. Stone, S. Richards, D. Zhu, R.A. Gibbs, C. Stricker, D. Gianola, M. Schlather, T.F.C. Mackay, and H. Simianer. 2012. Using whole-genome sequence data to predict quantitative trait phenotypes in Drosophila melanogaster. *PLoS Genet.* 8. doi:10.1371/journal.pgen.1002685.

Ober, U., M. Erbe, N. Long, E. Porcu, M. Schlather, and H. Simianer. 2011. Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics*. 188:695–708. doi:10.1534/genetics.111.128694.

Ogutu, J.O., H.-P. Piepho, and T. Schulz-Streeck. 2011. A comparison of random forests, boosting and support vector machines for genomic selection. *BMC Proc.* 5 Suppl 3:S11. doi:10.1186/1753-6561-5-S3-S11.

Ostersen, T., O.F. Christensen, M. Henryon, B. Nielsen, G. Su, and P. Madsen. 2011. Deregressed EBV as the response variable yield more reliable genomic predictions than traditional EBV in pure-bred pigs. *Genet. Sel. Evol.* 43:38. doi:10.1186/1297-9686-43-38.

Park, T., and G. Casella. 2008. The Bayesian Lasso. *J. Am. Stat. Assoc.* 103:681–686. doi:10.1198/016214508000000337.

Pearce, N.D., and M.P. Wand. 2006. Penalized Splines and Reproducing Kernel Methods. *Am. Stat.* 60:233–240. doi:10.1198/000313006X124541.

Pimentel, E.C.G., M. Wensch-Dorendorf, S. König, and H.H. Swalve. 2013. Enlarging a training set for genomic selection by imputation of un-genotyped animals in populations of varying genetic architecture. *Genet. Sel. Evol. GSE*. 45:12. doi:10.1186/1297-9686-45-12.

Pryce, J., J. Arias, P. Bowman, S. Davis, K. Macdonald, G. Waghorn, W. Wales, Y. Williams, R. Spelman, and B. Hayes. 2012a. Accuracy of genomic predictions of residual feed intake and 250-day body weight in growing heifers using 625,000 single nucleotide polymorphism markers. *J. Dairy Sci.* 95:2108–2119. doi:10.3168/jds.2011-4628.

Pryce, J., B. Hayes, and M. Goddard. 2012b. Genotyping dairy females can improve the reliability of genomic selection for young bulls and heifers and provide farmers with new management tools. *Proc. 38th ICAR Sess.*

Pryce, J.E., O. Gonzalez-Recio, J.B. Thornhill, L.C. Marett, W.J. Wales, M.P. Coffey, Y. de Haas, R.F. Veerkamp, and B.J. Hayes. 2014. Validation of genomic breeding value predictions for feed intake and feed efficiency traits. *J. Dairy Sci.* 97:537–42. doi:10.3168/jds.2013-7376.

Pszczola, M., T. Strabel, H. a Mulder, and M.P.L. Calus. 2012. Reliability of direct genomic values for animals with different relationships within and to the reference population. *J. Dairy Sci.* 95:389–400. doi:10.3168/jds.2011-4338.

Qanbari, S., E.C.G. Pimentel, J. Tetens, G. Thaller, P. Lichtner, A.R. Sharifi, and H. Simianer. 2010. The pattern of linkage disequilibrium in German Holstein cattle. *Anim. Genet.* 41:346–356. doi:10.1111/j.1365-2052.2009.02011.x.

Raven, L.-A., B.G. Cocks, and B.J. Hayes. 2014. Multibreed genome wide association can improve precision of mapping causative variants underlying milk production in dairy cattle. *BMC Genomics*. 15:62. doi:10.1186/1471-2164-15-62.

Sánchez, J.P., García-Gámez, Gutiérrez-Gil, and Y. Arranz. 2010. Evaluación preliminar de procedimientos de selección genómica en una población de ganado ovino lechero de raza churra. 2-4 pp.

Scheet, P., and M. Stephens. 2006. A Fast and Flexible Statistical Model for Large-Scale Population Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. *Am. J. Hum. Genet. Am. J. Hum. Genet.* 7878:629–644.

Schöpke, K. 2014. Assembling a Reference Population – from Genetic Architecture to New Phenotypes. *Proceedings*.

Schrooten, C., G. Schopen, A. Parker, A. Medley, and P. Beatson. 2013. Across-breed genomic evaluation based on bovine high density genotypes, and phenotypes of bulls and cows. *In* Proc. Assoc. Advmt. Anim. Breed. Genet. 138–141.

Sun, Y. V. 2010. Multigenic Modeling of Complex Disease by Random Forests. 72. 73-99 pp.

Szymczak, S., J.M. Biernacka, H.J. Cordell, O. Gonzlez-Recio, I.R. Konig, H. Zhang, and Y. V Sun. 2009. Machine learning in genome-wide association studies. *In* Genetic Epidemiology. 51–57.

Thomasen, J.R., C. Egger-Danner, A. Willam, B. Guldbrandtsen, M.S. Lund, and A.C. Sørensen. 2014. Genomic selection strategies in a small dairy cattle population evaluated for genetic gain and profit. *J. Dairy Sci.* 97:458–70. doi:10.3168/jds.2013-6599.

Tsuruta, S., I. Misztal, I. Aguilar, and T.J. Lawlor. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci.* 94:4198–4204. doi:10.3168/jds.2011-4256.

Tusell, L., P. Pérez-Rodríguez, S. Forni, X. Wu, and D. Gianola. 2013. Genome-enabled methods for predicting litter size in pigs: a comparison. *Animal*. 7:1739–49. doi:10.1017/S1751731113001389.

Usai, M.G., M.E. Goddard, and B.J. Hayes. 2009. LASSO with cross-validation for genomic selection. *Genet. Res. (Camb).* 91:427–436. doi:10.1017/S0016672309990334.

Van Binsbergen, R., M.C. Bink, M.P. Calus, F.A. van Eeuwijk, B.J. Hayes, I. Hulsegge, and R.F. Veerkamp. 2014. Accuracy of imputation to whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 46:41. doi:10.1186/1297-9686-46-41.

Van Binsbergen, R., M.P.L. Calus, M.C.A.M. Bink, F.A. van Eeuwijk, C. Schrooten, and R.F. Veerkamp. 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. *Genet. Sel. Evol.* 47:71. doi:10.1186/s12711-015-0149-x.

Van der Werf, J. 2013. Genomic selection in animal breeding programs. *Methods Mol. Biol.* 1019:543–61. doi:10.1007/978-1-62703-447-0_26.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91:4414–23. doi:10.3168/jds.2007-0980.

VanRaden, P.M. 2012. Avoiding Bias From Genomic Pre-Selection in Converting Daughter Information Across Countries. *Interbull Bull.* 0.

Vanraden, P.M., J.R.O. Connell, G.R. Wiggans, K.A. Weigel, and A. Improvement. 2010. Combining different marker densities in genomic evaluation. *Interbull Meet.* 1–4.

VanRaden, P.M., D.J. Null, M. Sargolzaei, G.R. Wiggans, M.E. Tooker, J.B. Cole, T.S. Sonstegard, E.E. Connor, M. Winters, J.B.C.H.M. van Kaam, A. Valentini, B.J. Van Doormaal, M.A. Faust, and G.A. Doak. 2013. Genomic imputation and evaluation using high-density Holstein genotypes. *J. Dairy Sci.* 96:668–78. doi:10.3168/jds.2012-5702.

VanRaden, P.M., J.R. O'Connell, G.R. Wiggans, and K. a Weigel. 2011. Genomic evaluations with many more genotypes. *Genet. Sel. Evol.* 43:10. doi:10.1186/1297-9686-43-10.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F.S. Schenkel. 2009. Invited review: reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci.* 92:16–24. doi:10.3168/jds.2008-1514.

Vazquez, A.I., G. de los Campos, Y.C. Klimentidis, G.J.M. Rosa, D. Gianola, N. Yi, and D.B. Allison. 2012. A comprehensive genetic approach for improving prediction of skin cancer risk in humans. *Genetics*. 192:1493–1502. doi:10.1534/genetics.112.141705.

Verbyla, K.L., P.J. Bowman, B.J. Hayes, and M.E. Goddard. 2010. Sensitivity of genomic selection to using different prior distributions. *BMC Proc.* 4:S5. doi:10.1186/1753-6561-4-S1-S5.

Verbyla, K.L., B.J. Hayes, P.J. Bowman, and M.E. Goddard. 2009. Accuracy of genomic selection using stochastic search variable selection in Australian Holstein Friesian dairy cattle. *Genet. Res. (Camb).* 91:307–311. doi:10.1017/S0016672309990243.

Villa-Angulo, R., L.K. Matukumalli, C.A. Gill, J. Choi, C.P. Van Tassell, and J.J. Grefenstette. 2009. High-resolution haplotype block structure in the cattle genome. *BMC Genet.* 10. doi:10.1186/1471-2156-10-19.

Villanueva, B., R. Pong-Wong, J. Fernández, and M.A. Toro. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. *J. Anim. Sci.* 83:1747–1752. doi:/2005.8381747x.

Villumsen, T.M., L. Janss, and M.S. Lund. 2009. The importance of haplotype length and heritability using genomic selection in dairy cattle. *J. Anim. Breed. Genet.* 126:3–13. doi:10.1111/j.1439-0388.2008.00747.x.

Vitezica, Z.G., I. Aguilar, I. Misztal, and A. Legarra. 2011. Bias in genomic predictions for populations under selection. *Genet. Res. (Camb).* 93:357–66. doi:10.1017/S001667231100022X.

Weigel, K.A., C.P. Van Tassell, J.R. O'Connell, P.M. VanRaden, and G.R. Wiggans. 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *J. Dairy Sci.* 93:2229–2238. doi:10.3168/jds.2009-2849.

Wientjes, Y.C., M.P. Calus, M.E. Goddard, and B.J. Hayes. 2015. Impact of QTL properties on the accuracy of multi-breed genomic prediction. *Genet. Sel. Evol.* 47:42. doi:10.1186/s12711-015-0124-6.

Wiggans, G.R., T. a Cooper, P.M. Vanraden, and J.B. Cole. 2011. Technical note: Adjustment of traditional cow evaluations to improve accuracy of genomic predictions. *J. Dairy Sci.* 94:6188–93. doi:10.3168/jds.2011-4481.

Wimmer, V., C. Lehermeier, T. Albrecht, H.-J. Auinger, Y. Wang, and C.C. Schön. 2013. Genome-wide prediction of traits with different genetic architecture through efficient variable selection. *Genetics*. 195:573–587. doi:10.1534/genetics.113.150078.

Xu, S. 2003. Estimating polygenic effects using markers of the entire genome. *Genetics*. 163:789–801.

Yi, N., V. George, and D.B. Allison. 2003. Stochastic search variable selection for identifying multiple quantitative trait loci. *Genetics*. 164:1129–1138.

Zhang, Z., and T. Druet. 2010. Marker imputation with low-density marker panels in Dutch Holstein cattle. *J. Dairy Sci.* 93:5487–5494. doi:10.3168/jds.2010-3501.

Zhou, L., X. Ding, Q. Zhang, Y. Wang, M.S. Lund, and G. Su. 2013. Consistency of linkage disequilibrium between Chinese and Nordic Holsteins and genomic prediction for Chinese Holsteins using a joint reference population. *Genet. Sel. Evol.* 45:7. doi:10.1186/1297-9686-45-7.

Zondervan, K.T., and L.R. Cardon. 2004. The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.* 5:89–100. doi:10.1038/nrg1314.