UNIVERSITA' CATTOLICA DEL SACRO CUORE
PIACENZA

Scuola di Dottorato per il Sistema Agro-alimentare

Doctoral School on the Agro-Food System

cycle XXIII

S.S.D: AGR/17

New trends in dairy cattle genetic evaluation

Candidate: Ezequiel Luis Nicolazzi
Matr. n.: 3611481

Coordinator: Ch.mo Prof. Gianfranco PIVA

Tutor: Prof. Paolo Ajmone Marsan

Academic Year 2009/2010

To Valentina

É isso aí, como a gente (não) achou que ia ser
a vida tão simples é boa, quase sempre
É isso aí, os passos vão pelas ruas
ninguém reparou na lua, a vida sempre continua
Eu não sei parar de te olhar
Não vou parar de te olhar
Eu não me canso de olhar
Não sei parar
De te olhar

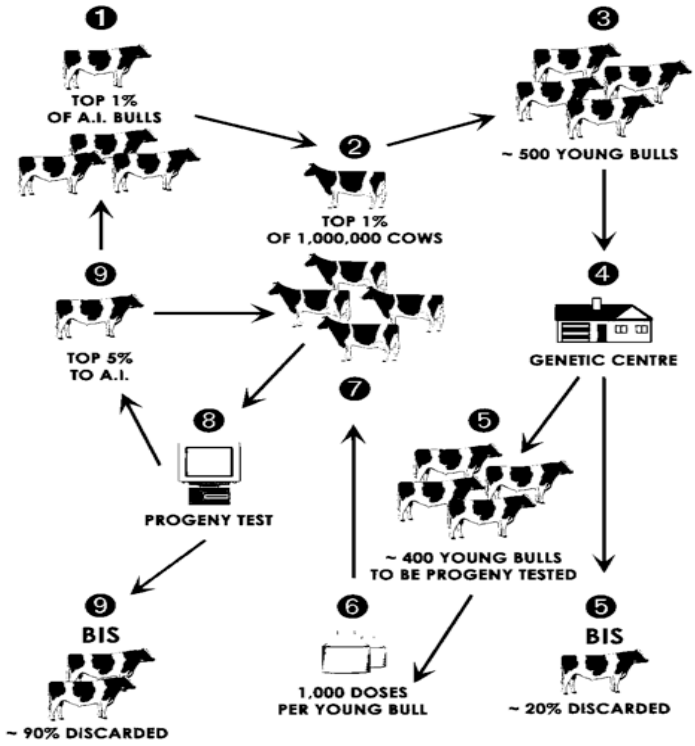
INDEX

INTRODUCTION

In species in which artificial insemination is widely used, such as dairy cattle, males contribute more than females to the genetic make up of the next generations. Accurate estimation of the genetic value of sires is therefore essential to maximize genetic progress. Currently, dairy cattle breeding relies on modelling of phenotypic and pedigree information to estimate the genetic value of animals, and so to obtain the Best Linear Unbiased Prediction (BLUP) of a bull genetic merit. The BLUP method enables the disentangling of the genetic and non-genetic components of the phenotypic values (Henderson, 1975). Since in dairy cattle many traits included in the selection index are sex-limited (i.e. milk yield and quality traits) and can only be recorded on cows, phenotypes are recorded on the daughters of bulls under evaluation. This selection scheme, called Progeny Test (PT), is designed to increase genetic progress by optimizing the accuracy of selection and the generation interval (Figure 1) and is implemented in dairy cattle breeding in most countries. Its main limitation is the long time required (~5/7 years) to collect sufficient information to obtain a sufficiently accurate (first) estimation of bulls' genetic merit. Long generation intervals imply both reduced potential genetic progress and high costs related to PT. Unable to intervene on generation intervals without heavily penalising accuracies, dairy cattle geneticists worldwide focused on obtaining more accurate predictions by improving the model used for the estimation of BV. The "sire" model, that included only the genetic effects of fathers to explain the phenotypes of daughters, was followed by the "sire maternal grandsire" model, that considers also the genetic effects of the mothers through the maternal grandsires, hence correcting sires genetic value estimates for preferential crossing. Later on a model was developed for the joint evaluation of both male and female animals in a population (animal model; Henderson, 1984; Mrode, 2005). Progress in computational technologies and

advances in animal breeding theory also allowed a change in the type of phenotypic data used in the genetic evaluations for longitudinal data. The introduction of the fixed and random regression test day model enabled the use of multiple records during the lactation (test days) as phenotypic records, instead of a single lactation record. This model represents a change of paradigm, because single measurements are considered as correlated traits. This implies that different genes may control the traits during different phases of the lactation and across different lactations.

Figure 1. Example of the selection scheme adopted in Italy for the Holstein-Friesian breed, based on a progeny test (ANAFI, 2010).



In the last 20 years, the exchange of genetics has become an international business and has fostered the exchange of information across countries to improve the accuracy of (national) genetic evaluations (Powell et al., 2000). This international network of estimated breeding values (EBVs) is managed by an international

organization (Interbull centre), which provides an across-country measure of the genetic value of the animals. International EBVs are useful for the improvement of the accuracy of national evaluations, and valuable for countries interested in the importation of foreign bulls (i.e. bull' semen). These accurate predictions of bulls' genetic merit (as confirmed by Powell et al., 2004 and Brochard et al., 2006 in the USA and France, respectively) allow a greater exchange of bull semen across countries, boosting, as a consequence, genetic progress worldwide.

The success obtained by this "traditional" selection in the last decades is reflected in the remarkable genetic progress achieved over the years, mainly in traits with medium-high heritability (i.e. yield traits). However, traditional selection is less efficient in improving traits with low heritability (i.e. fertility), traits with phenotypes which are difficult or expensive to collect.

In traditional selection, genetic merit of animals is estimated considering the genome as a "black box", hence ignoring the number, location and relative effect of genes controlling the traits under selection. The collection of molecular information is a first fundamental step towards the understanding of this black box. For example, high density genetic marker panels offer the opportunity of identifying genomic regions having relevant effects on complex traits. In this scenario, the integration of molecular data into breeding can help to reduce (or to solve) present inefficiencies and, as a consequence, to increase genetic gain

The idea of using molecular information in breeding is not new. Neiman-Sorensen and Robertson (1961) showed an association between blood groups and production traits in three Danish dairy cattle breeds, and proposed a large-scale use of this information. In the following years, technological and scientific progress permitted the use of more informative genetic markers like: Amplified Fragment Length

Polymorphism (AFLP), Restricted Fragment Length Polymorphism (RFLP), micro- and mini-satellite and Single Nucleotide Polymorphisms (SNP). Using these technologies, many efforts were addressed to detect quantitative trait loci (QTL) controlling the expression of traits under selection in most important breeds and different species (e.g. Wimmers et al., 2002; Khatkar et al., 2004; Sharma et al., 2006; Milanesi et al., 2008).

However, even if many QTLs (Animal genome web-site, 2010) and few causative mutations have been identified (Grisart et al.,2002, Cohen-Zinder et al., 2005, Varvio et al., 2008), the impact of the use of this type of molecular information in breeding (e.g. Marked and Gene Assisted Selection, MAS and GAS) has been almost negligible until recently. Several reasons explain why this highly valuable information has found only limited application in breeding. Low density marker maps and the investigation of segregating populations fosters the identification of markers that often are rather distant from the causative mutation(s) causing the QTL effects. This means that the association phase between marker and QTL alleles may vary in different families and is not consistent across the population. In addition, the (high) physical distance between molecular markers results in phases that can be reversed in very few generations (i.e. low linkage disequilibrium) even in the same family. In these conditions marker-QTL phase is to be verified across generations, to avoid the risk of selecting in favour of an undesired QTL. This increases the cost of MAS, and reduces its potential impact on genetic progress of the entire population. Furthermore, the low power of QTL detection with sparse marker maps permits the identification of only a few QTL with major effects, whereas several studies have demonstrated that most QTL have moderate to small effects. In this scenario, QTL effects tend to be overestimated, possibly introducing bias in genetic evaluations

(Beavis, 1998). To validate QTL effects, these are to be investigated in independent samples of the same population or in independent populations, resulting in further increase of costs and possible loss of information (as some of the QTL can fail validation). Finally, costs involved in QTL detection and routine genotyping of animals were to be added to the already expensive PT. In MAS, molecular information was intended as additional information to augment accuracy of EBV, but not as a substitute of traditional selection.

Because of all these reasons, only France and Germany introduced the systematic use of this type of molecular information in their breeding programs (Druet et al., 2005, Liu et al., 2004).

In the last few years, next-generation technologies and highly parallel SNP genotyping offered new options for the use of molecular data into breeding. These are changing livestock selection as we presently know it. Meuwissen et al. (2001) theorized on the possibility of using dense genome-wide genetic markers to select animals based only on genotypic information. This method was named "genomic selection" (GS). Briefly, GS uses dense marker panels to estimate direct genomic breeding values (DGV) in a training population in which both phenotypes[1] and genotypes are known. The DGV of young animals can the be predicted before their (daughter) phenotypic information is available, avoiding PT. This theory was initially developed and tested on a simulated dataset, sincea technology able to genotype many thousand markers on a few thousand animals at low cost was not available in 2001. In their study, Meuwissen et al. (2001) stated that:"*[…] the advent of DNA chip technology may make genotyping of many animals for many of these markers feasible (and perhaps even cost effective) […]* ".

[1] The term "phenotypes" in GS is generally used to indicate the dependent variable. As for now, the dependent variables used in GS are EBV, DYD (daughter yield deviations) or DRP (deregressed proofs), and not actual phenotypes.

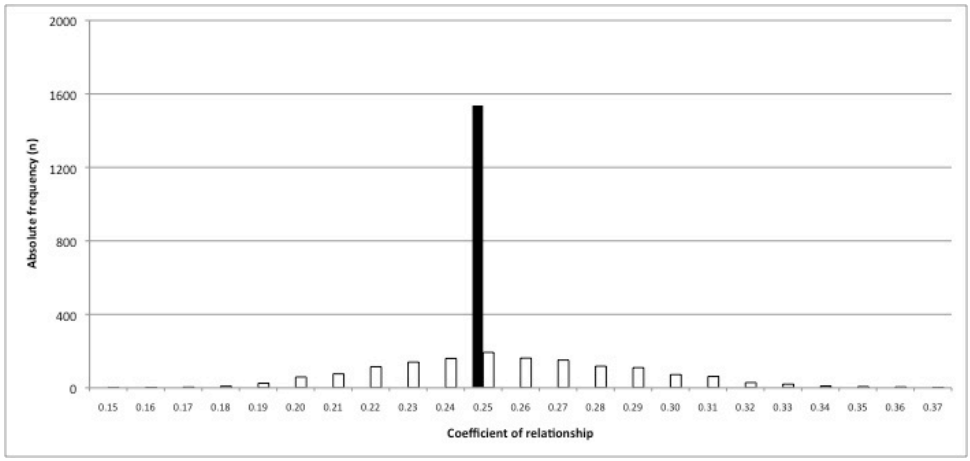
Only recently, the availability of cost-effective high throughput molecular information theorized by Meuwissen et al. (2001) became reality. Compared to MAS, dense maker maps allow carrying out population-wide studies, exploiting the genetic structure existing within breeds that determines a high level of linkage disequilibrium (LD) between neighbouring markers. In fact, GS fits the *markers* in the genetic model, not explicitly accounting for the QTLs controlling the trait. In this condition, marker-marker and marker-QTL phases (the latter considered only indirectly) are likely to be conserved across families within a breed.

Currently, thousands of animals are being genotyped worldwide with the (54K) bovine SNP beadchip (VanRaden and Sullivan, 2009; Goddard and Hayes, 2009). The importance of GS is highlighted by the potential advantages of this methodology over current “traditional” selection. They range from an increase of genetic gain caused by a strong reduction of the generation interval (young animals receive their genomic evaluations nearly at birth), to the increase of accuracies of breeding value estimates in the female population (Shaeffer, 2006), including a more accurate genome-wide based estimation of the relationships between animals (VanRaden,2007; VanRaden, 2008;Figure 2) and a substantial reduction of the costs for estimating the genetic value of breeding animals (Konig et al., 2009).

However, the statistical and computational issues raised by the large amount of information in GS are yet to be solved. In fact, no golden standard statistical method is agreed and many alternative statistical procedures for the estimation of DGV are under evaluation. Currently, the performances of different BLUP, non-linear or Bayesian models are being tested in different traits and breeds (VanRaden et al., 2009; Hayes et al., 2009; Gredler et al., 2009; van der Werf, 2009;Legarra et al., 2008; Gonzalez-Recio et al., 2009). In addition, nowadays it seems that completely

cancelling PT is not a viable option as previously stated, since accurate breeding values are needed of generations immediately before the current generation to maintain sufficiently accurate DGVs.

Figure 2. Comparison between average and genomic relationship for half-sib Italian Holstein bulls. Only bulls with $0.25 \leq a_{average} < 0.26$ were considered.



The black bar indicates the average relationship (i.e. currently used in 'traditional' genetic evaluations), whereas the white bars indicate the genomic relationship for the same bulls. Unpublished data.

Genome-wide marker data are not only exploited for GS. Genomic information can also be used to study the genetic population structure (McKay et al., 2008), perform fine-mapping QTL (Armin et al., 2007; Lionikas et al., 2010), discover causative mutations (Charlier et al., 2008), genome-wide association studies (GWAS; Sherman et al., 2008; Goddard and Hayes, 2009), and tracing signatures of selection in the genome (Barendse et al., 2009; MacEachern et al., 2009; Stella et al., 2010), amongst others. The aim of this thesis is to study different aspects of both traditional and genomic selection, testing and suggesting new methods in both simulated and real datasets.

REFERENCES

Animal Genome web-site. 2010. Cow. QTL Maps. [www.animalgenome.org].

Barendse, W., B.E. Harrison, R.J. Bunch, M.B. Thomas, and L.B. Turner. 2009. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. BMC genomics, 10:178.

Beavis, W.D. 1998. QTL analyses: power, precision, and accuracy. Molecular Dissection of Complex Traits, 1998:145–161.

Charlier, C., W. Coppieters, F. Rollin, D. Desmech, J.S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J.C. Frennet, R. Hanset, X. Hubin, C. Jorgensen, L. Karim, M. Kent, K. Harvey, B.R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R.J. Harvey, and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nature Genetics, 40(4): 449-454.

Cohen-Zinder, M., E. Seroussi, D.M. Larkin, J.J. Loor, A. Everts-van der Wind, J.H. Lee, J.K. Drackley, M.R. Band, A.G. Hernandez, M. Shani, H.A. Lewin, J.I. Weller, and M. Ron. 2005. Identification of a missense mutation in the bovine ABCG2 gene with a major effect on the QTL on chromosome 6 affecting milk yield and composition in Holstein cattle. Genome Res., 15:936-944.

Druet, T., S. Fritz, J.J. Colleau, M. Gautier, A. Eggen, M.N. Rossignol, M.Y. Boscher, A. Malafosse, and D. Boichard. 2005. Genetic markers in breeding programs. Pages 1-8 in Proceedings of the 26th European Holstein and Red Holstein Conference: May 2005; Prague, Czeck Republic.

Goddard, M.E., and B.J. Hayes. 2009. Mapping genes for complex traits in domestic animals and their use in breeding programs. Nature reviews, 10:381-391.

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

Gredler, B., K.G. Nirea, T.R. Solberg, C. Egger-Danner, T.H.E. Meuwissen, and J. Sölkner. 2009. A comparison of methods for genomic selection in Austrian dual purpose Simmental cattle. Pages 569-572 in Proceeding of the Association for the Advancement of Animal Breeding and Genetics: 28 September-1 October 2009; Barossa Valley, South Australia.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and Snell R. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res, 12:222-231

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

Henderson, C.R. 1984. Applications of linear models in animal breeding. Guelph, Canada: University of Guelph Press.

Khatkar, M.S, P.C. Thomson, I. Tammen, and H.W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Gen. Sel. Evol., 36:163-190.

Konig, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci., 92:382-391.

Legarra, A., C. Robert-Granie, E. Manfredi, and J.M. Elsen. 2008. Performance of genomic selection in mice. Genetics, 180:611-618.

Lionikas, A., R. Cheng, J.E. Lim, A.A. Palmer, and D.A. Blizard. 2010. Fine-mapping of muscle weight QTL in LG/J and SM/J intercrosses. Physiol Genomics. *In press.*

Liu, Z., F. Reinhardt, J. Szyda, H. Thomsen, and R. Reents. 2004. A marker assisted genetic evaluation system for dairy cattle using a random QTL model. Pages 170-174 in Proc. of the Interbull mtg: 29-31 May 2004; Sousse, Tunisia. Interbull No. 32, Interbull, Uppsala, Sweden.

MacEachern, S., B.J. Hayes, J. McEwan, and M.E. Goddard. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle populations (Bos Taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genetic diversity in Domestic cattle. BMC genomics, 10:181.

McKay, S.D., R.D. Schnabel, B.M. Murdoch, L.K. Matukumali, J. Aerts, W. Coppieters, D. Crews, E. Dias Neto, C.A. Gill, C. Gao, H. Mannen, Z. Wang, C.P. van Tassel, J.L. Williams, F.J Taylor, and S.S. Moore. 2008. An assessment of population structure in eight breeds of cattle using a whole genome SNP panel. BMC genetics, 9:37.

Milanesi, E., R. Negrini, F. Schiavini, L. Nicoloso, R. Mazza, S. Biffani, F. Canavesi, F. Miglior, A. Valentini, A. Bagnato, P. Ajmone-Marsan. 2008. Identification of QTLs influencing milk protein percentage in Italian Friesian cattle integrating selective genotyping, DNA pooling, AFLP and micro satellite markers. J. Dairy Res., 75:430-438.

Mrode, R.A. 2005. Linear models for the prediction of animal breeding values (2nd Edition). Wallingford, UK: Cabi Publishing.

Neiman-Sorensen, A., and Robertson. 1961. The association between blood groups

and several production characters in three Danish cattle breeds. Acta Agric Scand. 11:163–196.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet. 123:218-223.

Sharma, B.S., G.B. Jansen, N.A. Karrow, D. Kelton, and Z. Jiang. 2006. Detection and characterization of amplified fragment length polymorphism markers for clinical mastitis in Canadian Holsteins. J. Dairy Sci., 89:3653-3663.

Sherman, E.L., J.D. Nkrumah, B.M. Murdoch, and S.S. Moore. 2008. Identification of polymorphisms influencing feed intake and efficiency in beef cattle. Anim. Genet., 39(3): 225-231.

Schmitt, A.O., H. Al-Hasani, J.M. Cheverud, D. Pomp, L. Bünger, and G.A. Brockmann. 2007. Fine Mapping of Mouse QTLs for Fatness Using SNP Data. OMICS: A Journal of Integrative Biology, 11(4): 341-350.

Stella, A., P. Ajmone-Marsan, B. Lazzari, and P. Boettcher. 2010. Identification of selection signatures in cattle breeds selected for dairy production. Genetics, 185:1451-1461.

van der Werf, J.H.J. 2009. Potential benefit of genomic selection in sheep. Pages 38-41 in Proceeding of the Association for the Advancement of Animal Breeding and Genetics: 28 September-1 October 2009; Barossa Valley, South Australia.

VanRaden, P.M. 2007. How relatives are related. Pages 33-36 in Proc. of the Interbull mtg: 23-26 August 2007; Dublin, Ireland. Interbull No. 37, Interbull, Uppsala, Sweden.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91:4414-4423.

VanRanden, P.M., and P. Sullivan. 2009. National and International genomic evaluations for dairy cattle. J. Anim. Sci., 87:E-suppl.2:175

Varvio, S.L., T. Iso-Touru, J. Kantanen, S. Viitala, I. Tapio, A. Maki-Tanila, M Zerabruk, and J. Vilkki. 2008. Molecular anatomy of the cytoplasmic domain of bovine growth hormone receptor, a quantitative trait locus. Proc. Biol. Sci, 275:1525-1534.

Wimmers, K., E. Murani, S. Ponsuksii, M. Yerle, and K. Scellander. 2002. Detection of quantitative trait loci for carcass traits in the pig by using AFLP. Mamm. Genome, 13:206-10.

CHAPTER I

BACKGROUND

From 'traditional' genetic evaluations

to genomic selection

1) NATIONAL AND INTERNATIONAL GENETIC EVALUATIONS

1.A. NATIONAL EVALUATIONS AND MODELS APPLIED

The use of direct (DGV) and genomic (G-EBVs) breeding values, the latter obtained by combining DGV with pedigree based estimates, needs further validation in many breeds and countries, comprised Italy. The use of genomic data is therefore often still "experimental" and awaiting to enter in the routine of genetic evaluations. In the meantime, both national and international evaluations are still based upon what is know as "traditional" evaluations (i.e. PT based). In simple words, this means that phenotypes and known degrees of relationship (i.e. from pedigree information) are combined to obtain an estimation of the genetic value of an animal. The idea behind this is that genetic differences between animals can be estimated from phenotypes, once considered the relationships among them and excluded all non-genetic effects in a genetic model. This is the (simplistic) explanation of the concept driving all statistical methods currently applied in genetic evaluations.

First approaches to selection of livestock were simply based on the raw comparison of phenotypes between animals. Since Best Linear Unbiased Predictions (BLUP) became available, these comparisons were performed on the estimated breeding value (EBV) of animals (Henderson, 1950). The evolution of more advanced features of this method generally corresponded to advances in computational and theoretical techniques. For example, early applications of BLUP methods evaluated in the model only the male population (sire, sire and maternal grandsire models). An evaluation of all animals in the pedigree (BLUP animal model) became feasible only in the late '80s, when computer technology allowed performing the (much) higher amount of

calculations required. Since then, many developments have been proposed and adopted in different countries.

The success of this now considered “traditional” selection is demonstrated by the great genetic progress achieved. The greatest gains were mainly achieved in medium-high heritability traits, such as production related traits, on which selection has focused for decades. However, as a consequence, reduced (or even negative) genetic progress has been obtained in functional, low heritability traits or traits negatively correlated with production (Figure 1).

Figure 1. Italian Holstein (progeny tested) bulls’ genetic trend for main productive and functional traits, by bulls’ year of birth.

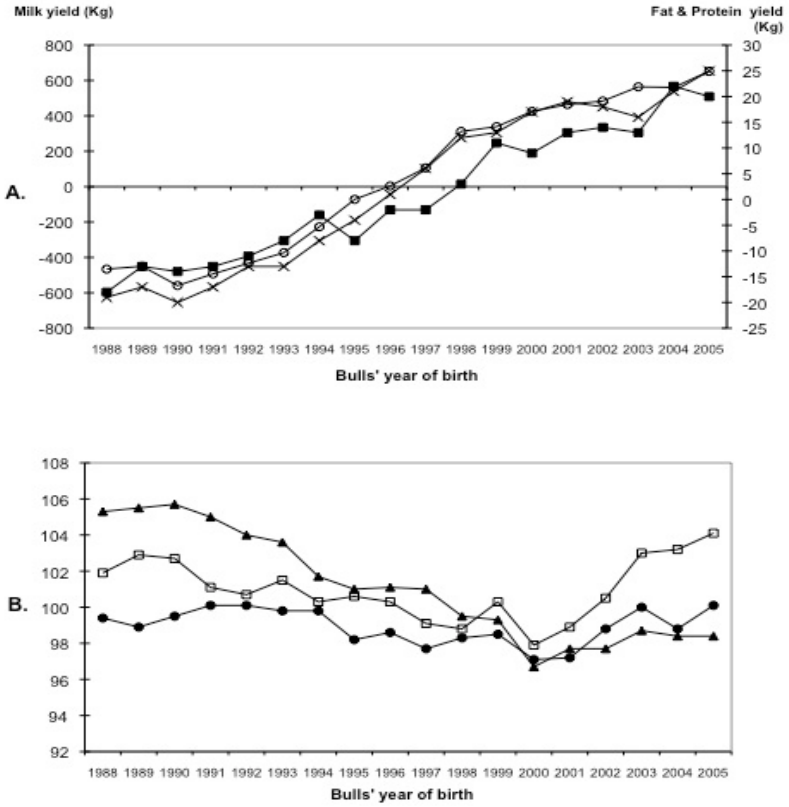


Figure “A” shows the genetic trend for milk (white circles), fat (black squares) and protein yield (black crosses). Figure “B” shows main functional traits trends for functional longevity (white squares), somatic cell count (black circles) and female fertility (black triangles). Note that there is a clear positive genetic trend for all productive traits, a nearly stable trend for functional longevity and somatic cell count and a negative trend for female fertility (data provided by the Italian Holstein Association, ANAFI; January 2010).

Below is presented a brief description of the basis of past and present models used in "traditional" genetic evaluations, to introduce some of the methods currently applied in genetic evaluation systems worldwide.

Sire model (SM) and Sire and Maternal Grandsire model (MGS)

Early applications of BLUP methods evaluated only bulls based on their progeny performances. The sire model in matrix form is defined by:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zs} + \mathbf{e} \qquad [1.1]$$

where **y** is a vector of phenotypes; **X** is an incidence matrix linking phenotypes to the fixed effects; **b** is a vector of solutions of the fixed effects; **Z** is an incidence matrix linking phenotypes to the random sire effect; **s** is a vector of sire random effects with $s \sim N(0, A\sigma^2_s)$, where **A** is the additive relationship matrix and σ^2_s is the sire variance; and ***e*** is a vector of random residuals with $e \sim (0, I\sigma^2_e)$, where **I** is an identity matrix and σ^2_e is the residual variance. Note that since only sires are evaluated, the genetic effect considered is *half* of the additive genetic effect $\hat{A}$ of the daughter (the other half, corresponding to the dam, is not accounted for). Thus, $\sigma^2_s = var(½\,\hat{A}) = (½)^2 \times (\hat{A})^2 = ¼\sigma^2_a$, where σ^2_a is the (animal) additive genetic variance.

The (factorized) mixed model equations (MME) of this model result:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{s} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \qquad [1.2]$$

,where $\lambda = \sigma^2_e/\sigma^2_s = (4–h^2)/h^2$, and h^2 is the heritability of the trait.

Nearly all countries applied this methodology in their genetic evaluation systems until the late '80s. However, this model has some shortcomings. For example, in the case of assortative mating, not accounting for the genetic merit of dams biases sires' BV estimates (Schaeffer, 1983). In a dairy cattle breed, a farmer may mate top bulls with

top dams to produce high-quality offspring (positive assortative mating) or to dams of much lower genetic level, to improve the genetic level of the breed (negative assortative mating). Since SM considers only sire effects, the effect of the mating is not accounted for: sires are assumed to mate dams of equal genetic merit and not related to each other (usually not the case in dairy cattle). This model leads to an over- or under-estimation of sires EBV.

Alternative models have been proposed to reduce this bias, as the model that considers the *maternal* grandsires in the evaluation of a sire genetic merit (MGS; Everett, 1979; Schaeffer, 1983). However, the best solution to this issue is to evaluate together all animals within a breed (i.e. bulls *and* cows), in an "Animal model".

Animal model (AM)

In an AM, all animals are considered in the evaluation model. The number of MME is much higher than in the SM, as in dairy cattle the female is generally much larger than the male population.

In matrix notation, an AM is similar to [1.1]:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e} \qquad [1.3]$$

however, in an AM, the term "**a"** is a vector of *animal* random effects with $\mathbf{a} \sim N(0, A\sigma^2_a)$.The (factorized) MME in the AM result:

$$\begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + A^{-1}\lambda \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X'y \\ Z'y \end{bmatrix} \qquad [1.4]$$

,with $\lambda = \sigma^2_e/\sigma^2_a = (1 - h^2)/h^2$.

The AM started to be adopted in dairy breeds genetic evaluation systems in the late 80's, when computer technology allowed enough calculation power to store and solve (i.e. invert) large matrices, which can easily reach the order of millions.

Modelling longitudinal phenotypes

Longitudinal data are repeated measurements of a variable (i.e. milk yield) along a trajectory (i.e. a lactation). The statistical model that analyses this kind of phenotypic data should consider that measurements along the trajectory are somehow correlated.

Among the statistical solutions available, the repeatability animal model (repAM) was the model used for productive traits by almost all countries until 1999 (Interbull, 2010). The repAM allows to account for random environmental effects (including non-additive genetic effects) important to define the (co)variance structure between measurements of the same animal over time. However, the repAM assumes an equal variance for all measurements and a correlation of one between pairs of measurements, hence considering successive records as repeated measurements of the same trait (Mrode, 2005). These unrealistic assumptions make this a simplistic model for the description of the complex behaviour of longitudinal data.

In terms of model complexity and computing time, the easiest way to analyse longitudinal phenotypes is to combine single lactation records (test-day records) into a 305-day lactation record. This 305-combined record is usually estimated from a mathematical function that takes into account the shape of the lactation curve. Some examples of such mathematical functions are the "test interval method" (Everett and Carter, 1968) or the "multiple trait prediction" (Schaeffer and Jamrozik, 1996). Using these methods, covariables are used to describe the shape of the lactation curve of

cows within fixed sub-classes (i.e., cows of the same productive region, age, parity and season of calving), and all factors that may influence single lactation records are averaged together (i.e. specific conditions of the test day). As a standard lactation curve is assumed for all cows in the same fixed sub-class, the EBV evidences the differences in the height of the lactation curves. Ptak and Schaeffer (1993) first approached a solution to this problem by directly using single test day records as phenotypic records, to consider that temporary environmental effects may differ across single test day records. The model they described bypasses the need to estimate the 305-day lactation record, although the amount of information to be stored for each lactation and the number of equations to be solved by the model was much higher (i.e. around 10 single test day records/lactation, instead of a single 305-day lactation record). However, this solution still assumed a correlation of one between lactation records in a same lactation, and cow lactation curves weren't allowed to assume different shapes within the same sub-class. These issues were solved with the introduction of random regression test day models (RRTDM; Schaeffer and Dekkers, 1994). In this model, the shape of the lactation curves for single cows are considered to be different, thus allowing persistency of cows[1] to be evaluated and environmental effects affecting the lactation of a cow to be better estimated (Schaeffer et al., 2000). In a RRTDM the lactation curve of a cow is defined by both fixed and random regressions on days in milk (DIM). The general shape of the lactation curve is defined by fixed regressions for cows in the same (fixed) sub-class (as in the approach by Ptak and Schaeffer, 1993), whereas random regressions allow describing the deviation of a specific cow from that (general) fixed regression (Jamrozik and Schaeffer, 1997). The functions used to describe the shape

[1] Lactation persistency of a cow refers to the degree of milk production that is maintained form the beginning to the end of the lactation. In Italian Holstein, for example, persistency is evaluated as rate of milk production between the 280th and the 60th day of lactation.

of the lactation curve span from Wilmink functions (Wilmink, 1987) to Legendre polynomials (Kirkpatrick et al., 1990) or cubic splines (White and Brotherstone, 1997), among others.

The main drawback of this methodology is the much higher computational requirements in terms of storage of information and equation solving. Other shortcomings are the sensitivity of this model with respect to the accuracy of collection of phenotypes and the higher complexity of the model itself, that difficult the dissemination of information to breeders.

Test day model in Italy

In Holstein-Friesian dairy cattle population the RRTDM is the model currently adopted for the estimation of BV for milk, fat, protein and SCS in Italy, Canada, Germany and the Netherlands, among others (Interbull, 2010). The multiple-trait multiple-lactation RRTDM was adopted in Italy in November 2004 (Canavesi et al., 2004). The term "multiple-trait" means that the four aforementioned traits are analysed simultaneously; and "multiple-lactation" refers to the fact that each of the first three lactations is considered as a different (correlated) trait. Thus, the model analyses 12 traits simultaneously: milk, fat, protein yield and SCS, for the first 3 lactations of a cow. Test day records, in this model used as dependent variables, are pre-adjusted for both heterogeneity of variance (Schaeffer et al., 2000) and number of days of pregnancy at the test date (Canavesi et al., 2009).

The model is as follows (after Muir et al., 2007):

$$y_{ijkptd} = HTDP_{ipt} + \sum_{m=0}^{4} \phi_{dm}\beta_{kmpt} + \sum_{m=0}^{4} \phi_{dm}u_{jmpt} + \sum_{m=0}^{4} \phi_{dm}pe_{jmpt} + e_{ijkptd} \qquad [1.5]$$

where $\boldsymbol{y}_{ijkptd}$ is the (corrected) test-day record on trait *t* of cow *j* in lactation *p* on DIM *d* within herd-test-day-parity *i* and in the *k* fixed regression subclass for "TRAPS": Time

(considered as year effect of production), Region (four regions are currently considered: northern Italy, centre of Italy, southern Italy and Parmigiano-Reggiano region), Age by Parity, and Season; HTDP are the herd-test-day-parity effects; β are the fixed regression coefficients; u and pe are the random genetic and random permanent environmental effects of cow *j*; ø are the *m* fourth-order Legendre polynomials; and e is the random residual.

In matrix notation:

y = **Xb** + **Za** + **Wpe** + **e** [1.6]

where **y** is a vector of test-day records; **X** is the incidence matrix for HTDP and fixed regressions; **b** is a vector of solutions for HTDP and fixed regressions, **Z** and **W** are incidence matrices for the animal and the permanent environmental effects, respectively; **a** is the random regression for animal genetic effects, with **a** ~ *N*(0, A⊗G), where **A** is the additive relationship matrix, and **G** is the (co)variance matrix of the additive genetic random regression coefficients for the four traits and three lactations (thus, of order 60 because each curve is defined by five Legendre polynomials parameters); **pe** is the random regression vector for permanent environmental effects, with **pe** ~ *N*(0,I⊗P), where **I** is an identity matrix and **P** is the (co)variance matrix of the permanent environmental random regression coefficients (of order 60); and **e** is a vector of random residuals, with **e** ~ *N*(0,R), where **R** is a (co)variance matrix of random residuals, where covariances among traits can assume values different from zero. The MME for the RRTDM are:

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z & X'R^{-1}W \\ Z'R^{-1}X & Z'R^{-1}Z + A^{-1} \otimes G & Z'R^{-1}W \\ W'R^{-1}X & W'R^{-1}Z & W'R^{-1}W + P \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \\ p\hat{e} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ X'R^{-1}y \\ W'R^{-1}y \end{bmatrix} \quad [1.7]$$

1.B. INTERNATIONAL EVALUATIONS

In the early '70s, the development of reproductive technologies (i.e. frozen semen) and the resulting international trade of livestock genetic material, evidenced the need to standardize the expression of bulls' EBV to make them comparable across countries. In 1975 the European Association for Animal Production (EAAP) and the International Dairy Association (IDF) formed a working group to investigate possible solutions to this issue. In the following years, studies by Hinkovski et al. (1979) and Stolzman et al. (1981) evidenced differences in the genetic level of bulls coming from different countries. In particular, Stolzman et al. (1981) compared the productive performances of Polish Holstein cows, sired by bulls coming from 10 countries. Their results revealed large differences in milk yield, consistent with sires' country of origin. Thus, international evaluations suddenly became a necessity not only for the international trade of bull' semen, but also to enhance the genetic progress of the breed worldwide.

In 1983, the EAAP, the IDF and the International Committee of Animal Recording (ICAR) founded Interbull, in order to organize, harmonize and structure the development of the international evaluations (Philipsson, 2005). In 1991, the Interbull centre was officially founded, and placed in the Swedish University of Agricultural Sciences (SLU) in Uppsala, Sweden. Five years later, the EU Commission designated Interbull as the official reference centre for international evaluations.

One of the first tasks of the newborn Interbull group, was to gather all the information regarding the different breeding programs, traits analysed, models, methods and criteria for publication of EBV used in the different countries. Such reviews further evidenced that research and harmonization of methods were necessary. The very first attempt to compare bulls across countries was the use of conversion formulae of

bulls' EBV (Goddard, 1985; Wilmink, 1986). Those formulae were simple empirical methods to predict the EBV of a bull in a country *i*, given the EBV of the same bull in another country *j*:

$$\mathbf{EBV_i = a + b \times EBV_j} \qquad [1.8]$$

where a is the intercept; b is the regression coefficient, obtained as in Goddard (1985) or Wilmink (1986). However, a set of *common* bulls with high EBV reliability[2] was needed to derive the parameters of the conversion formulae (i.e. not the entire population, but only a highly selected group). Powell (1988) showed that different thresholds of EBV reliabilities (i.e. different datasets of bulls) resulted in very different estimates.

Research for a statistical method to perform a comprehensive across country evaluation in an international framework led, in 1994, to the application of a multiple across country evaluation (MACE; Schaeffer, 1994). MACE is a multiple-trait, sire and maternal grandsire model that is currently the official method for international evaluations. This method includes all known pedigree relationships between animals both within and across populations, accounting for differences in production systems across countries[3]. Using genetic correlations among countries, MACE allows to predict breeding values for all bulls in all countries participating to the international genetic evaluation, even those with no daughter information available locally (Figure 2).

In matrix notation this model is (after Schaeffer, 1994):

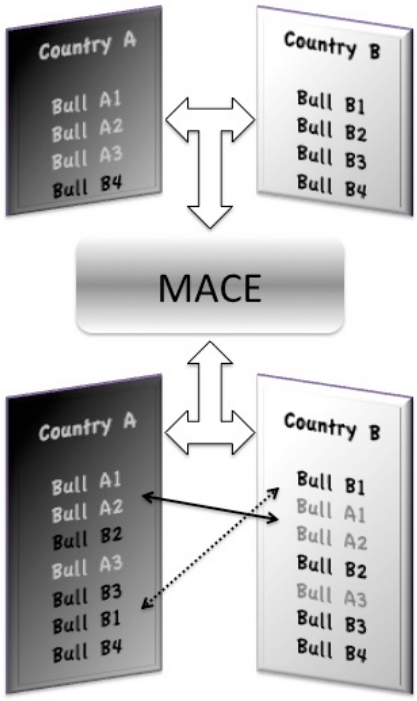
$$\mathbf{y_i} = \mathbf{1}\mu_i + \mathbf{Z_i Q p_i} + \mathbf{Z_i s_i} + \mathbf{e_i} \qquad [1.9]$$

[2] Interbull officially adviced to include only bulls with EBV accuracy higher than 75% to estimate the parameters of the conversion formula (Powell, 1988).

[3] Different genotypes may respond differently to environmental conditions, a phenomenon known as genotype by environment interaction. The importance of taking into account this interaction in the model becomes clear when multiple countries and productive systems (i.e, grazing and stable, seasonal calving, etc) are analysed simultaneously.

where $\mathbf{y}_i$ is a vector of deregressed proofs[4] (DRP) of the country *i* for one trait (i.e. milk yield); 1 is a vector of ones; $\boldsymbol{\mu}_i$ is the mean of country *i*; $\mathbf{Z}_i$ is a design matrix relating phenotypes to sires in country *i*; **Q** is a design matrix relating sires to genetic (or phantom) groups (i.e. unknown parents grouped, for example, by breed, year of birth, country of origin and path of selection); $\mathbf{p}_i$ is a vector of genetic effects for phantom groups; $\mathbf{s}_i$ is a vector of genetic effects for sires for country *i*; $\mathbf{e}_i$ is a vector of random residuals.

Figure 2. Role of MACE on bulls' international EBV comparison.



MACE analyses domestic EBV (based on domestic daughter information) in two countries participating to the International genetic evaluation. After MACE evaluation, the bull ranking in a country may differ from that in another country and also from the ranking in the country of origin.
In this example, the relative ranking of bulls prior to MACE evaluation in country A is maintained after MACE evaluation in country B (Bull A1 > A2 > A3 > B4), but not vice-versa (dashed black line indicates that the top bull ranked in country B prior to MACE (bull B1) is ranked almost as the worst bull in country A). This happens because some animals perform better in some environments than in others.

[4] Deregressed proofs are obtained by de-regressing national EBV in order to make them independent of country group effects and relationships among all bulls included in the international evaluation (Mrode, 2005).

This model also takes into account the precision of the calculation of the dependent variables (DRP) by using effective daughter contribution[5] (EDC; Fickse and Banos, 2001) as weighting factors (see (co)variance structure [1.10]).

The (co)variance structure for two countries is (after Mrode, 2005):

$$\text{var}\begin{bmatrix} p_1 \\ p_2 \\ s_1 \\ s_2 \\ e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} A_{pp}g_{11} & A_{pp}g_{12} & A_{pn}g_{11} & A_{pn}g_{12} & 0 & 0 \\ A_{pp}g_{21} & A_{pp}g_{22} & A_{pn}g_{21} & A_{pn}g_{22} & 0 & 0 \\ A_{np}g_{11} & A_{np}g_{12} & A_{nn}g_{11} & A_{nn}g_{12} & 0 & 0 \\ A_{np}g_{21} & A_{np}g_{22} & A_{nn}g_{21} & A_{nn}g_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & D_1\sigma^2_{e1} & 0 \\ 0 & 0 & 0 & 0 & 0 & D_2\sigma^2_{e2} \end{bmatrix} \qquad [1.10]$$

where n and p are the number of bulls and groups, respectively; **A** is the additive genetic relationship matrix of *n* bulls and *p* phantom groups, based on MGS relationships; $\mathbf{g}_{ij}$ is the sire genetic (co)variance between countries *i* and *j*; $\mathbf{D}_i$ is a diagonal matrix containing the reciprocal of EDC of bulls in country *i*; and $\boldsymbol{\sigma}^2_{ei}$ is the residual variance matrix for the i^{th} country.

Thus, MACE MME equations come from a multiple-trait MGS with genetic (phantom) grouping:

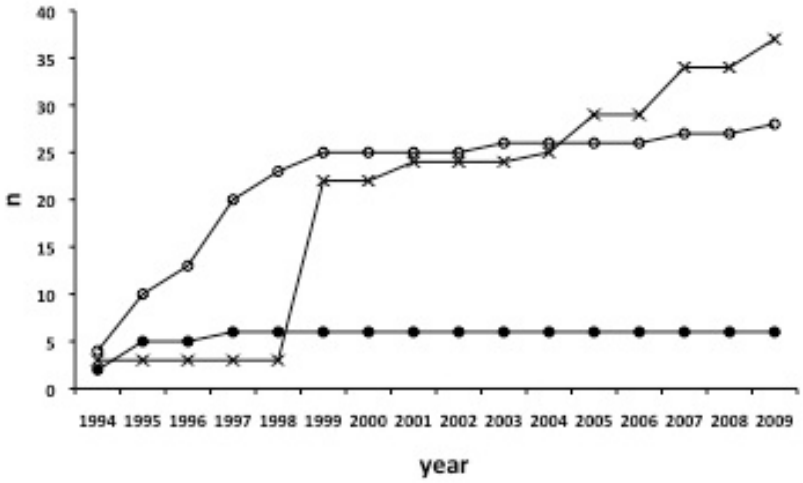
$$\begin{bmatrix} X'R^{-1}X & Z'R^{-1}X & 0 \\ Z'R^{-1}X & Z'R^{-1}Z + A^{-1} \otimes G^{-1} & -A^{-1}Q \otimes G^{-1} \\ 0 & -Q'A^{-1} \otimes G^{-1} & Q'A^{-1}Q \otimes G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ Q\hat{w} + \hat{s} \\ \hat{w} \end{bmatrix} = \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \\ 0 \end{bmatrix} \qquad [1.11]$$

As stated previously, the Interbull centre still adopts MACE for the estimation of international breeding values (ITB-EBV). However, some updates were introduced in the past (Interbull, 2010) and new developments are currently under study. Some examples of future developments are: i) the use of an sire-dam pedigree (van der

[5] Effective daughter contribution is a measure of the precision of daughter information used to calculate DRP (Fikse and Banos, 2001).

Linde et al., 2005; Jakobsen and Fickse, 2009); and ii) the introduction of a multiple-trait multiple-country MACE (MT-MACE) for female fertility (Nilforooshan et al., 2009). The Interbull centre currently distributes MACE-based EBVs three times a year (January, April and August) to all participating countries. The first official publication of international EBV was performed over production traits in Nordic countries. Since then, the number of countries, traits and breed analysed increased constantly (Figure 3).

Figure 3. Number of countries, traits and breeds analysed by Interbull center from first official publication of international evaluations to present



White circles indicate the number of countries, black crosses indicate the number of traits and black circles indicate the number of breeds analysed (*source: Interbull, 2010).*

Currently, a total of 33 countries are involved in the international evaluations for at least one of the 7 major trait groups (production, conformation, udder health, direct longevity, calving traits, female fertility and workability) and breeds (Brown Swiss, Guersey, Jersey, Holstein, European red dairy breeds and Simmental).

2) THE NEW FRONTIER: THE USE OF GENOMIC INFORMATION

The recent availability of the bovine genome sequence and the related Hapmap project, led to the discovery of millions single nucleotide polymorphisms (SNP)

spread across the genome (Bovine genome sequencing and analysis consortium, 2009; Bovine Genome web-page, 2010; Bovine HAPMAP consortium, 2009). In the last few years, the development of new technologies allowed a dramatic reduction in the cost of genotyping, increasing in turn the number of species in which this technology can be applied. Genome-wide SNP chips are currently available for human, bovine, ovine, porcine, canine and equine populations (Illumina, 2010). Research is ongoing to augment the density of SNP chips for the species already available, and to increase the number of species that can be analysed with this technology. For example, in cattle 54K SNP beadchips are currently available; however, next generation bovine SNPchip (recently available) are providing genotypic information at much higher density (nearly 800K).

The availability of dense marker panels throughout the genome has marked a paradigm shift in the way bovine livestock populations are investigated and analysed, either for studying the population genetic structure (McKay et al., 2008), searching for patterns of recent and past selection (MacEachern et al., 2009), searching for QTL controlling complex traits (Kolbehdari et al., 2008), or to perform a genome-wide marker-enhanced selection on young animals (Harris et al., 2008; VanRaden et al., 2009; Hayes et al., 2009), among others.

Until recently, population geneticists based their inferences assessing only a few (i.e. around 30) multi-allelic microsatellite or bi-allelic dominant AFLP markers (MacNeil et al., 2007; Negrini et al., 2007). Deriving population structure from whole-genome SNP data, provides a more precise picture of the true extent of genomic diversity and structure within and across cattle breeds. The importance of correctly assessing genetic population structure (i.e. populations stratification, linkage disequilibrium) and precisely inferring population demographic trajectories (i.e. bottlenecks, founder

effects) has a direct impact on the robustness of population genetic studies as well as the optimization of the strategies for genetic diversity conservation (i.e. control of inbreeding, optimal mating management). For example, the HapMap project analysed the population structure of 19 breeds (which included *Bos Taurus*, *Bos Indicus* and admixed breeds) genotyped for 37,470 SNP, and showed that "[…] *difference in diversity* [among breeds] *is mainly due to progenitor population diversity and bottleneck effects at, and before, breed formation rather than differences in the intensity of natural or artificial selection post-domestication*. […]" (The Bovine HapMap Consortium,et al., 2009). Moreover, the interest in detecting the hidden genetic structure relies on the fact that population stratification may heavily influence results in all genomic studies. In fact, one of the sources of bias in these studies is admixture, either considered as the presence of different breeds in the sample, or as the existence of relationship among the animals (Goddard and Hayes, 2009). In most of the techniques used for the aforementioned genomic studies, both sources of admixture can generally be accounted for in the statistical models used to analyse the data.

The study and comparison of the genetic structure of breeds permits to identify genomic regions that are under selection. Such regions can be detected by comparing the distribution of allele frequencies at marker loci within or between populations (or groups of populations), in search for markers significantly departing from neutral behaviour. The comparison of the distribution of allele frequencies can be either direct or through different statistics, function of allelic or genotypic frequencies, as the fixation index F_{st} (The Bovine HapMap Consortium et al., 2009)

or the linkage disequilibrium[6] (LD; Ennis, 2007). In addition, specific tests for detecting significant effects have been developed, as the integrated standardized extended haploptype homozygosity, or integrated haplotype score (iHS; Voight et al., 2006) that compares the LD of markers flanking a selected allele.

The F_{st} index is a widely used, robust and simple methodology (Cavalli-Sforza, 1966; Weir et al., 2006; Barendse et al., 2009). It is obtained from the formula: $F_{st}=1-H_s/H_t$, where H_s is the expected within sub-population heterozygosity following Hardy-Weinberg equilibrium, and H_t is the total expected heterozygosity, assuming no differentiation between sub-populations. This method uses differences in allele frequencies to determine the differentiation between sub-populations. However, such allele frequency differences might be determined by causes other than selection, as genetic drift caused by finite population size or inbreeding. On the other hand, deviations of F_{st} values caused by inbreeding or genetic drift can be identified, as inbreeding affects the entire genome in a similar way and genetic drift affects all loci randomly, not showing any pattern of LD between successive loci (MacEachern et al., 2009). A deviation of F_{st} in a small region of the genome will be only observed in case of selection, as it only *affects* small blocks of the genome (that is, the selected locus and the genetic markers linked to it). The length of the genomic region affected by selection depend on a series of variables, as number of generations of selection, the recombination rate in the region, etc.

There are other methods available to identify genomic regions under selection, as Tajima's D (Tajima, 1989) or Fay and Wu's H tests (Fay and Wu, 2000) that use

[6] Linkage disequilibrium is defined as the correlation of alleles at two loci, and is dependent on the distance between the loci. The greater the distance, the higher is the probability that a recombination event might occur between the two loci. On the other hand, if any two marker loci are very close to each other, then the probability of recombination of their alleles is very low, determining that the same alleles will (almost always) segregate together in the population. Assuming there is a QTL in between these marker loci, the same allele of the QTL will be transmitted jointly with the same alleles of the two marker loci. Consequently, following the segregation of the markers, we can also track the QTL.

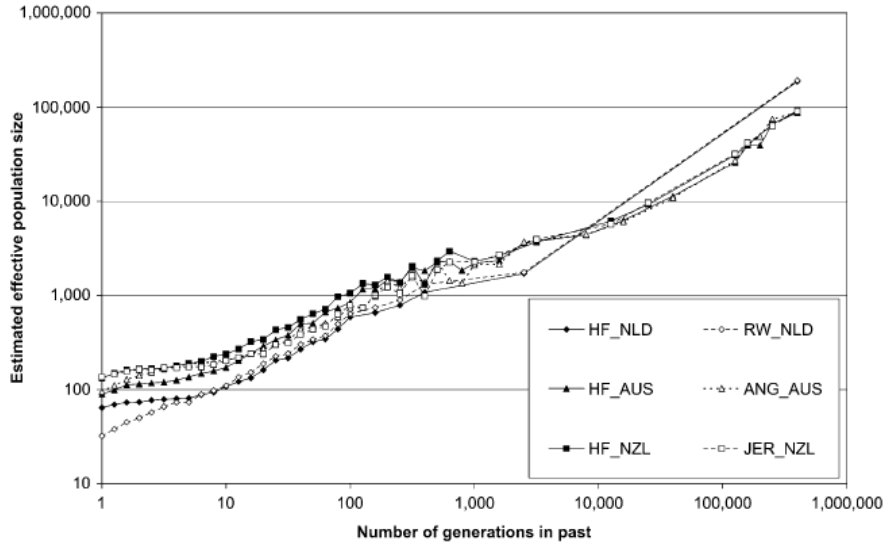
frequency distribution summaries. These methods, however, can be influenced by the demographic history of the breeds included in the analysis. For example, a reduction in effective population size (N_e), can mimic selection if a marker has a rare allele, as increased inbreeding (derived by the reduction of N_e) can reduce the frequency of the rare allele.

The use of LD to identify selection signatures in the genome is an appealing method; however, considering LD between intervals gives only small overall information in the study of selection history, as LD can also arise from different causes not related to selection (some of which will be discussed next) (Hill, 1981). On the other hand, comparing the LD of flanking markers to a specific (selected) core allele (known as "iHS method") has been proved to be an effective method to detect positive selection both in humans (Voight et al., 2006) and in cattle (Hayes et al., 2007). The main problem with this methodology is that the information on ancestral alleles, which are inferred or assumed as the monomorphic alleles in wild relatives of *Bos Taurus* (as, for example, Bison or Yak), is usually low in cattle so that iHS can only be applied to a reduced subset of SNP. This issue will probably be bypassed when full-sequence on wild relatives will be available.

The LD level and the (genetic) history of cattle populations are closely interconnected, as the demographic and evolutionary history of the population (in particular the effective population size N_e) are the main drivers of the genome-wide level of LD. Four are the main drivers of LD in a population: selection, mutation, migration and genetic drift. As discussed before, selection (and mutation) are influencing LD only in a small part of the genome, and migration is a cause of high genome-wide LD when pure-lines are crossbred (not usually the case in cattle breeding), that is greatly reduced after only a small number of generations. On the

contrary, a high level of drift induced by severe bottlenecks or non random mating decreasing N_e means that few common ancestors, only few generations ago, gave origin to almost all the alleles that are segregating in the population (Goddard and Hayes, 2009). This means that the study of the LD level maintained against distance in a population can be used to study the population history itself (Tenesa et al., 2007; MacLeod et al., 2009). Using the LD information in some dairy cattle breeds, de Roos et al. (2008) showed that the N_e of these breeds was higher than 50.000 before domestication, declining to a few thousands after domestication and is currently around 100 in most modern breeds (Figure 4). This means that LD in cattle will be maintained at distances higher that 1 cM and, thus, currently available bovine SNPchips can potentially identify markers in LD with all the QTL involved in the expression of the traits currently under selection.

Figure 4 Effective population size for main cattle breeds, inferred from LD (de Roos et al., 2008; *Copyright © 2008 of Genetics Society of America, with kind permission of the Editor-in-Chief*).



Breeds analysed were: Dutch Holstein-Friesian (HF_NLD), Australian Holstein-Friesian (HF_NLD), New Zealand Holstein-Friesian (HF_NZL), Dutch red-and-white Holstein-Friesian (RW_NLD), Australian Angus (ANG_AUS) and New Zealand Jersey (JER_NZL).

In the past, QTL discovery studies were performed using sparse multi-allelic microsatellites, which traced QTL by linkage within large half-sib families. This experimental design was needed to cope with the large distance between markers (that could reach 20cM, or even more), which made the underlying causative mutation difficult to identify and the QTL impossible to track at the population level.

The higher density of markers currently available, however, can exploit population-wise LD to identify more QTL and more precisely, in a genome-wide association study (GWAS). The appealing feature of applying GWAS with current SNP technology, is that this approach potentially identifies SNP that are in strong LD with all QTL that control a trait. GWAS analyses have been carried out with good success in different cattle breeds (Pryce et al., 2010), either for traits controlled by single genes (Charlier et al., 2008), and for traits controlled by many genes of small effect (Kolbehdari et al., 2008). These studies not only allowed to confirm QTL previously reported, but also identified new regions affecting either quantitative or qualitative traits. For example, Pryce et al. (2010) identified a new putative QTL on BTA 18 for female fertility, analysing Australian Holstein and Jersey breeds. Furthermore, using a 60K SNP chip, Charlier et al. (2008) identified the five regions causing five different inherited defects in three different breeds (Belgian Blue, Italian Chianina and Danish Red Holstein).

Indeed, GWAS experience in humans showed that markers with significant (and validated) effects explain only a small proportion of the additive genetic variance of complex traits, even of those with high heritability (i.e. human height) and even when searched with several hundred thousand markers on tens of thousand samples (Manolio et al., 2009). This means that many of the underlying mutations controlling such traits are still uncounted for reasons still to be understood. Technical inability

may be an explanation, since standard SNP typing technologies are not highly effective in detecting some type of polymorphisms as copy number variations (CNV). Epigenetic effects may play a role, or simply complex traits are really controlled by a very large number of genes with very small effect, as modelled by Fisher's infinitesimal model. In the latter case, signals remain below the significance threshold unless a huge population is analysed. In cattle, where the number of individuals analysed in GWAS is much lower than in humans, the power to identify QTL with low effects is reduced, although the level of LD at any distance is higher than in the human population.

Furthermore, the generally high levels of false discovery rate[7] found in GWAS performed in cattle, suggests that a (medium-large) proportion of the significant markers are expected to be significant just by chance. Further research, with higher power to detect QTL with small effects (i.e. larger datasets) or different models that consider gene-gene interactions, or genotype-environment interactions is ongoing.

As specified before, genome-wide associations studies were boosted by the availability of dense markers spread throughout the genome, based on the fact that all QTL are (or should be) in medium-high LD with at least one SNP. Based on the same assumptions, Meuwissen et al. (2001) suggested the possibility of using this technology in breeding. They called this methodology genomic selection (GS), and tested it on simulated data using different statistical methods. GS uses dense marker panels to estimate direct genomic breeding values (DGV) in a training population in which both phenotypes and genotypes are known, and predicts DGV on young animals before their phenotypic information is available (Figure 5). Briefly, single

[7] False discovery rate (FDR) is the expected percentage of (significant) markers that are false positives. It is obtained from the formula: [n x *P*(k)]/k, where n is the number of markers tested; *P*(k) is the largest P-value of the marker that exceeded the significance threshold; and k is the number of markers that exceeded the significance threshold.

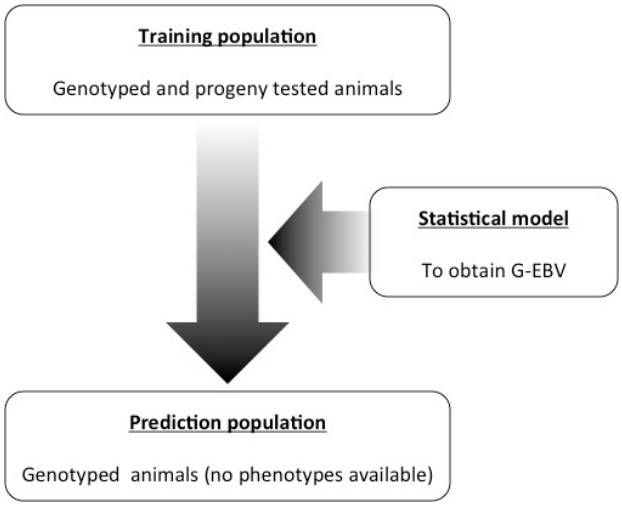
SNP effects are estimated in the training population using a prediction equation (i.e., a statistical model) and are then used to estimate the DGV on young animals using molecular information only[8].

This means that, theoretically, progeny test could be avoided, highly reducing the cost of the selection process. Furthermore, remarkable gains in genetic progress are expected, as the generation interval could be markedly reduced and the genome-enhanced EBV (DGV or G-EBV) accuracies of both (young) bulls and cows could be increased (Shaeffer, 2006). Other potential advantages of GS over current "traditional" selection include a more accurate control of inbreeding in the population (Daetwyler et al., 2007), a more accurate estimation of relationship matrices (VanRaden, 2007; VanRaden, 2008) and a marked reduction of costs (Konig et al., 2009).

Meuwissen et al. (2001) tested statistical methods that ranged from ordinary least squares (OLS) and BLUP to Bayesian models (BayesA and BayesB). OLS required a two step procedure to reduce the dimensionality of the dataset, as this methodology does not allow the number of independent variables (i.e. marker effects) to be higher than the number of dependent variables (i.e. phenotypes). OLS-derived DGV accuracies (obtained as correlation between DGV and simulated true breeding values) were the lowest of all methods tested, and marker estimates obtained with this method resulted highly biased.

[8] As it will be specified later on, there are several methods available to integrate genomic information into breeding. Most methods currently applied in national breeding programs worldwide integrate "traditional" and molecular information together. Originally, the term G-EBV was used to indicate a genetic evaluation based on markers only. Only very recently, a distinction of the terms "DGV" and "G-EBV" was adopted internationally. Currently, the term DGV means "Direct Genomic Value", that considers molecular information only, whereas G-EBV (Genome-Wide Estimated Breeding Values) blends traditional and molecular information together. *Please note that, because of this recent change in terminology, the term "G-EBV" actually means "DGV" in Chapters III and IV of this thesis.*

Figure 5. Simple scheme of genome-wide selection.



Training population is the population of animals genotyped with phenotypes available. All this information is combined in the prediction statistical model to obtain the DGV. The DGV obtained from the training population will be used to predict genomic values for the (young) animals in the "prediction" population (i.e. without phenotypes available).

The BLUP method assumed an equal contribution of each locus to the genetic variance. However, when considering molecular data in the genetic model, BLUP infinitesimal assumptions should be reconsidered, as only a small proportion of the loci included in the model are assumed to be actually contributing to the genetic variance of a trait (i.e. those SNP that are in LD with the QTL). In terms of estimation of BV, BLUP can still obtain good results, as the effects of all QTL (signalled by the nearby SNP) are aggregated across many segments. In fact, in Meuwissen et al. (2001), the BLUP method obtained nearly twofold DGV accuracies compared to OLS.

The two Bayesian methods (BayesA and BayesB) proposed were tested to consider a specific variance for each marker. The prior SNP variance distribution was a scaled inverted chi-squared distribution, to consider that most of the markers should have a nearly (or exact) zero effect (i.e. markers not linked to any QTL) and only few should obtain large effects. This leads to a t-distribution for SNP effects. The t-distribution is

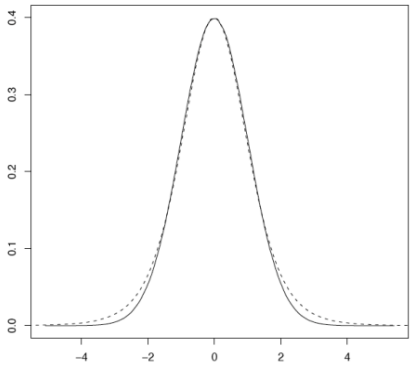
somewhat similar to the normal distribution, but has thicker tails, allowing a higher proportion of markers to assume high effects (Figure 6).

The main difference between the two Bayesian models is in the specifications for the prior distribution of SNP variances: BayesA allows SNP effects to have SNP variances that are ***close*** to 0, whereas BayesB allows a certain proportion of the markers to obtain a variance ***equal*** to zero (Figure 7).

This means that some of the variables analysed by BayesB are actually excluded from the analysis. In Meuwissen et al. (2001), both these models obtained accuracies of DGV higher than OLS and BLUP, with BayesB reaching a DGV accuracy of nearly 85% in a simulated trait with an heritability of 0.5.

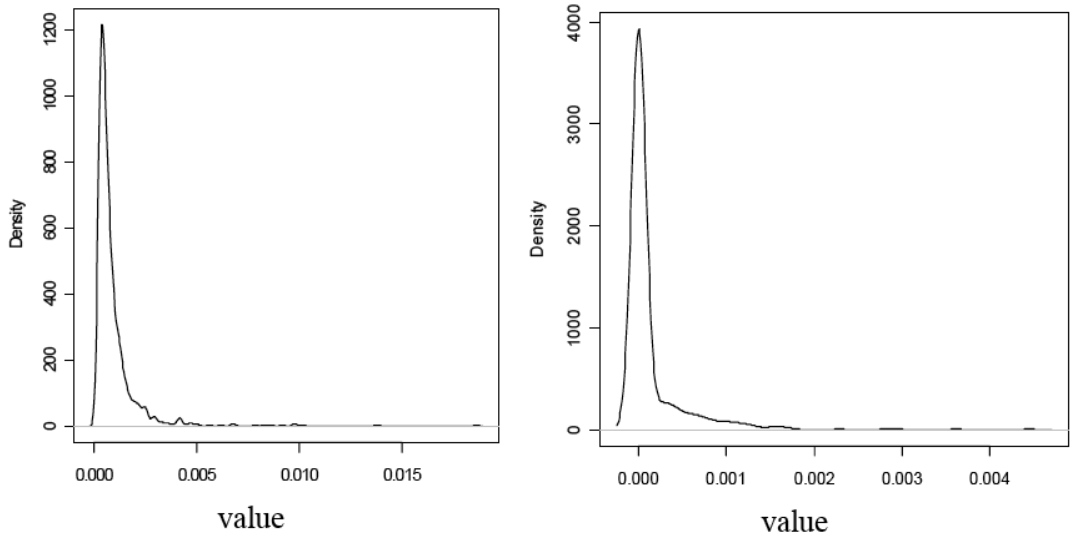
The differences between BayesA and BayesB methods, introduce an interesting concept about the variables that are taken into account in the predictive model. In general, the number of phenotypes (i.e. genotyped bulls) is much lower than the number of independent variables (i.e. SNP effects) that have to be estimated.

Figure 6. Comparison of standardized normal and t-student distributions used as prior distribution of SNP effects in BLUP and BayesA, respectively.



The continuous line indicates the normal distribution (BLUP), whereas the dashed line indicates a t-student distribution (BayesA).

Figure 7. Prior distribution for SNP variances in BayesA and BayesB.



Density of the scaled inverted chi-squared distribution for SNP variances used in BayesA (left) and BayesB (right). Note that BayesA allows SNP variances to obtain values close to 0, while BayesB allows SNP variances to assume a value equal to 0. From Ben Hayes course notes, September 2008, Salzbourg (Austria), with permission.

Such data asymmetry raises several statistical issues, such as co-linearity among predictors and multiple testing (Gianola and van Kaam, 2008). Until recently three main approaches were used to reduce the dimensionality of the set of regression variables:

i) to use a method that *directly* reduces the number of the "original" variables. Some examples are BayesB (Meuwissen et al., 2001), Least angle regression (LARS) or Least absolute shrinkage and selection operator (LASSO; Tibshirani,1996);

ii) to reduce dimensionality of "original" variables in an *indirect* way (in a two-step procedure, by preselecting the SNP *before* using the predictive model); or

iii) to find a (reduced) number of linear combinations of the variables (as, for example, partial least squares regression and principal component regression; Solberg et al., 2009).

In the last few years, many other methods have been tested to perform GS either reducing the dimensionality of the dataset or using all variables available[9]. For example, Meuwissen (2009) suggested an alternative BayesC method (a derivate from BayesB), that instead of considering a null SNP variance for a proportion of the total amount of markers, assigns a common (small) variance to these markers in an attempt to capture a part of the genetic variance given by QTL with small effects. Other developments resulted in more advanced methods, like BayesCpi, BayesCCsub, etc.

A common characteristic of most of the aforementioned methods is that hyper-parameters[10] of all the prior distributions are considered fixed, thus, are defined *a priori*. In particular, for BayesA and BayesB, where the prior influences heavily the resulting shrinkage of estimates of the model (Gianola et al., 2009), this means that the same fixed hyper-parameters might perform better in one simulated data than in the other, depending of the simulated genetic structure of the population. Consequently, when analysing real data, these methods are expected to perform differently depending on the QTL distribution of the traits considered. For example, the diacylglycerol-acyltransferase1 (DGAT1) is a mutation that explains more than 30% of the variation for fat percentage (Grisart et al., 2002). In such scenario, methods that allow non-normal marker effects will more accurately estimate the resulting DGV. In fact, VanRaden et al. (2009) compared linear and non-linear models over 21 traits, obtaining the greatest difference in performance of the two models in this trait (non-linear model obtained 8% higher DGV accuracy). However, for most of the traits currently under selection, there is no such large-effect QTL

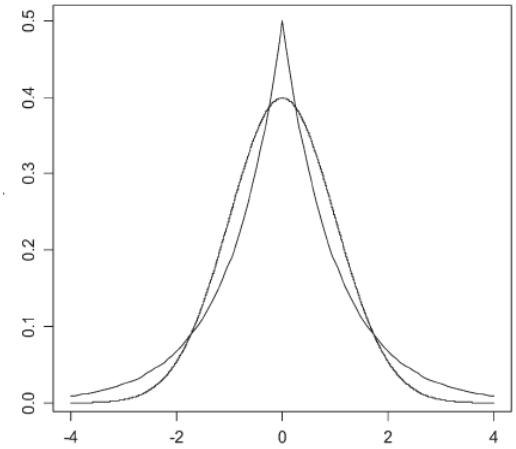
[9] From the many methods (and developments) that have been proposed to analyze genomic data, only those directly or indirectly related to this thesis are briefly considered in this introduction.
[10] Hyper-parameters are the parameters that define the prior distribution. For example, hyper-parameters for a Gaussian prior distribution are its mean and variance.

affecting the traits. Thus, more flexible (or different) assumptions are needed to account for the different QTL distributions (and heritability) of complex traits in real data. Some of the Bayesian methods that have been proposed take such issue into account, allowing to obtain hyper-parameters that are conditional on the data. This means that the resulting shrinkage of the estimates of the model are regulated by the data itself. Yi and Xu (2008) and de los Campos (2009) suggested a computationally efficient method to approach this, implementing a hierarchical approach called the Bayesian LASSO. In the Bayesian LASSO, SNP variances (i.e. this method uses all the variables available, differently from the "original" LASSO) are sampled from an exponential distribution. This leads to SNP effects sampled from a double-exponential distribution, which has greater mass towards zero than the normal or the t-distribution (Figure 8).

This is not much different from BayesA, with the only difference that SNP variances and effects are sampled from a more stringent distribution. The main difference between these two methods is that in BayesA the hyper-parameters of the prior distribution are fixed, whereas in the Bayesian LASSO the hyper-parameter of the exponential prior distribution is sampled from the data. This means that is the data itself that regulates the amount of shrinkage of the estimates.

Although research on new and more efficient methods to predict DGV are currently being investigated, most of the countries that are actually applying GS in their breeding programs use BLUP as predictive model. In Australia, Canada, Ireland, New Zealand and the USA, the BLUP method is used to perform genome-enhanced estimates which are blended with parent average or "traditional" EBV by selection index in order to obtain the "final" G-EBV (Nieuwhof et al., 2010; van Doormaal, 2009; Kearney et al., 2009; Harris et al., 2008; VanRaden et al., 2009).

Figure 8. Comparison of standardized normal and exponential distributions, used as prior distributions for SNP effects in BLUP and Bayesian LASSO, respectively (de los Campos et al., 2009; *Copyright © 2009 of Genetics Society of America, with kind permission of the Editor-in-Chief*).



The exponential distribution has greater mass towards zero than a normal or a t-distribution. However, the thicker tails indicate that a greater proportion of marker effects can obtain high values than in a normal distribution.

These and many other countries tested several methodologies that span from Bayesian to non-linear or non-parametric methods (Gianola and van Kaam, 2008). However, most countries reported that G-EBV accuracies obtained with BLUP were only slightly lower than those obtained with other methods in most traits, and the need in terms of computational time and resources were much lower (VanRaden et al., 2009; Gredler et al., 2009). These results suggest that for most traits, there is a large proportion of QTL of small effect and only few (or none) of large effect, thus, the infinitesimal assumption in BLUP is close to reality (Hayes et al., 2009).

The aforementioned countries are only a part of the list of countries that have implemented or are close to implement GS in their breeding programs. Other countries are: Austria (Gredler et al., 2009), Germany (Reinhardt et al., 2009), Italy

(van Kaam et al., 2009); France (Ducrocq et al., 2009), the Netherlands[11] (de Roos et al., 2009), the Nordic countries (Lund and Su, 2009) and Poland (Szyda et al., 2009), among others. Furthermore, international cooperation to obtain a multiple-country genomic evaluation, coordinated by the Interbull Centre, is currently taking place for some breeds as, for example, Holstein, Jersey (Interbull, 2010) and Brown Swiss (InterGenomics, Jordani et al., 2010),

REFERENCES

Barendse, W., B.E. Harrison, R.J. Bunch, M.B. Thomas, and L.B. Turner. 2009. Genome wide signatures of positive selection: The comparison of independent samples and the identification of regions associated to traits. BMC genomics.10:178.

Bovine HapMap Consortium et al. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science, 324:528-532.

Canavesi, F., S. Biffani, and F. Biscarini. 2004. Test day model for production traits and SCS for the Italian Holstein. J. Dairy Sci., 87(Suppl 1):40 (Abstr).

Canadesi, F., S. Biffani, G. Bramante, and R. Finocchiaro. 2009. Improving the stability of test day model evaluation for production traits in the Italian Holstein. Ital. J. Anim. Sci., 8(Suppl. 2):39-41.

Cavalli-Sforza, L.L. 1966. Population structure and human evolution. Proc. R. Soc. Lond. B. Biol. Sc., 164:362-379.

Charlier, C., W. Coppieters, F. Rollin, D. Desmecht, J.S. Agerholm, N. Cambisano, E. Carta, S. Dardano, M. Dive, C. Fasquelle, J.C. Frennet, R. Hanset, X. Hubin, C.

[11] The Netherlands reported early applications of GS on breeding programs at national scale (van der Beek, 2007). In this case, young bulls available for progeny testing were pre-selected based on (SNP) molecular information; however, SNP panels used in this case included only 3000 SNP

Jorgensen, L. Karim, M. Kent, K. Harvey, B.R. Pearce, P. Simon, N. Tama, H. Nie, S. Vandeputte, S. Lien, M. Longeri, M. Fredholm, R.J. Harvey, and M. Georges. 2008. Highly effective SNP-based association mapping and management of recessive defects in livestock. Nature Genetics 40:449-454.

Daetwyler, H.D., B. Villanueva, P. Bijma, and J.A. Wooliams. 2007. Inbreeding in genome-wide selection. J. Anim. Breed. Genet., 124: 369-376.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182:375-385.

de Roos, A.P.W., C. Schroten, E. Mullaart, S. van der Beek, G. de Jong, and G. Voskamp. Genomic selection at CRV. Pages 47-50 in Proc. of the Interbull international workshop, Uppsala, Sweden. Interbull No. 39. Interbull, Uppsala, Sweden.

de Roos, A.P.W., B.J. Hayes, R.J. Spelman, and M.E. Goddard. 2008. Linkage disequilibrium and persistence of phase in Holtesin-Friesian, Jersey and Angus Cattle. Genetics, 179:1503-1512.

Ducrocq, V., S. Fritz, F. Guillaume, and D. Boichard. 2009. French report on the use of genomic evaluation. Pages 17-22 in Proc. of the Interbull international workshop, Uppsala, Sweden. Interbull No. 39, Interbull, Uppsala, Sweden.

Ennis, S. 2007. Linkage disequilibrium as a tool for detecting signatures of natural selection. Meth. Mol. Biol., 376:59-70.

Everett, R.W., and H.W. Carter. 1968. Accuracy of test interval method of calculating Dairy herd improvement association records. J. Dairy Sci., 51:1937.

Everett, R.W., R.L. Quaas, and A.E. McClintock. 1979. Daughters' maternal grandsires in sire evaluation. J. Dairy Sci., 62:1304–1313

Fay, J.C., and C.I. Wu. 2000. Hitchhiking under positive Darwinian selection. Genetics, 155:1405-1413.

Fikse, W.F., and G. Banos. 2001. Weighting factors of sire daughter information in International genetic evaluations. J. Dairy Sci., 84:1759-1767.

Gianola, D., G. de los Campos, W. Hill, E. Manfredi, and R. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics, 183:347-363.

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.

Goddard, M.E. 1985. A method of comparing sires evaluated in different countries. Livest. Prod. Sci, 13:321-331.

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

Gredler, B., K.G. Nirea, T.R. Solberg, C. Egger-Danner, T.H.E. Meuwissen, and J. Solkner. 2009. Genomic selection in Fleckvieh/Simmental – First results. Pages 209-213 in Proc. of the Interbull mtg., Barcelona, Spain. Interbull No. 40, Interbull, Uppsala, Sweden.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res., 12:222-231

Harris, B.L., D.L. Johnson, and R.J. Spelman. 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 11-16 in Proc. of the 36th ICAR Biennial Session, Niagara Falls, USA.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci., 92:433-445.

Hayes, B.J., S. Lien, H. Nilsen, H.G. Olsen, P. Berg, S. MacEachern, S. Potter, and T.H.E. Meuwissen. 2007. The origin of selection signatures on bovine chromosome 6. Anim. Genet., 39:105-111.

Henderson, C.R. 1950. Estimation of genetic parameters (abstract). Ann. Math. Statist., 21:309-310.

Hill, W.G. 1981. Estimation of effective population size form data on linkage disequilibrium. Genet. Res., 38:209-216.

Hinkovski, T., A. Alexiev, B. Lindhé, and C.G. Hickman. 1979. The red and red-and-white cattle breed comparison in Bulgaria. World Anim. Rev., 29:8-12.

Interbull. 2010. Public area. Genetic evaluation. [http://www-interbull.slu.se/eval/framesida-genev.htm]

Jakobsen, J., and W.F. Fikse. 2009. Sire-Dam Pedigree in MACE – Results from a Full-Scale Pilot Study. Pages 91-98 in Proc. of the Interbull international workshop, Uppsala, Sweden. Interbull No. 39, Interbull, Uppsala, Sweden.

Jarmrozik, J., and L.R. Schaeffer. 1997. Estimates of genetic parameters for a test day model with random regressions for yield traits of first lactation Holsteins. J. Dairy Sci., 80:762-770.

Jorjani, H., B. Zumbach, J. Durr, and E. Santus. 2010. Joint genomic evaluation of BSW populations. Pages 1-6 in Proc. of the Interbull international workshop, Paris,France. Interbull No. 41, Interbull, Uppsala, Sweden.

Kearney, F., A. Cromie, and D.P. Berry. 2009. Implementation and uptake of genomic evaluations in Ireland. Pages 227-230 in Proc. of the Interbull mtg., Barcelona, Spain. Interbull No. 40, Interbull, Uppsala, Sweden.

Kirkpatrick, M., D. Lofsvold, and M. Bulmer.1990. Analysis of inheritance, selection and evolution of growth trajectories. Genetics, 124:979:993.

Kolbehdari, D., Z. Wang, J.R. Grant, B. Murdoch, A. Prasad, Z. Xiu, E. Marques, P. Stothard, and S.S. Moore. 2008. A whole-genome scan to map quantitative trait loci for conformation and functional traits in Canadian Holstein bulls. J. Dairy Sci., 91:2844-2856.

Konig, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci., 92:382-391.

Lund, M.S., and G. Su. 2009. Genomic selection in the Nordic countries. Pages 39-42 in Proc. of the Interbull international workshop, Uppsala, Sweden. Interbull No. 39, Interbull, Uppsala, Sweden.

MacEachern, S., B.J. Hayes, J. McEwan, and M.E. Goddard. 2009. An examination of positive selection and changing effective population size in Angus and Holstein cattle population (Bos Taurus) using a high density SNP genotyping platform and the contribution of ancient polymorphism to genetic diversity in Domestic cattle. BMC genomics, 10:181.

MacLeod, I.M., T.H.E. Meuwissen, B.J. Hayes, and M.E. Goddard. 2009. A novel predictor of multilocus haplotype homozygosity: comparison with existing predictors. Genet. Res., 91:413-426.

MacNeil, M.D., M.A. Cronin, H.D. Blackburn, C.M. Richards, D.R. Lockwood, and L.J. Alexander. 2007 Genetic relationships between feral cattle from Chirikof Island, Alaska and other breeds. Anim. Genet., 38:193-197.

Manolio, T.A., F.S. Collins, N.J. Cox, D.B. Goldstein, L.A. Hindorff, D.J. Hunter, M.I. McCarthy, E.M. Ramos, L.R. Cardon, A. Chakravarti, J.H. Cho, A.E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C.N. Rotimi, M. Slatkin, D. Valle, A.S. Whittemore, M. Boehnke, A.G. Clark, E.E. Eichler, G. Gibson, J.L. Haines, T.F.C. Mackay, S.A. McCarroll, and P.M. Visscher. 2009. Finding the missing heritability of complex diseases. Nature, 461:747-753.

McKay, S.D., R.D. Schnabel, B.M. Murdoch, L.K. Matukumalli, J. Aerts, W. Coppieters, D. Crews, E. Dias Neto, C.A. Gill, G. Chuan, H. Mannen, Z. Wang, C.P. van Tassel, J.L. Williams, J.F. Taylor, and S.S. Moore. 2008. An assessmet of population structure in eight breeds of cattle using a whole genome SNP panel. BMC genetics, 9:37.

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.

Meuwissen, T.H.E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet. Sel. Evol., 41:35.

Mrode, R.A. 2005 Linear models for the prediction of animal breeding values (2nd Edition). Wallingford, UK: Cabi Publishing.

Muir, B.L., G. Kistemaker, J. Jamrozik, and F. Canavesi. 2007. Genetic parameters for a multiple-trait multiple-lactation random regression Test-day model in Italian Holsteins. J. Dairy Sci., 90:1564-1574.

Negrini, R., I.J. Nijman, E. Milanesi, K. Moazami-Goudarzi, J.L. Williams, G. Erhardt, S. Dunner, C. Rodellar, A. Valentini, D.G. Bradley, I. Olsaker, K. Kantanen, P. Ajmone-Marsan, and J.A. Lenstra. 2007. Differentitation of European cattle by AFLP fingerprinting. Anim. Genet., 38:60-66

Nieuwhof, G.J., K.T. Beard, K.V. Konstantinov, P.J. Bowman, and B.J. Hayes. 2010. Implementation of Genomics in Australia. *In press* in Proc. of the Interbull mtg.: 31 May- 4 June 2010; Riga, Latvia. Interbull, Uppsala, Sweden.

Nilforooshan, M.A., J. Jakobsen, W.F. Fikse, B. Berglund, and H. Jorjani. 2009. MT-MACE for female fertility and milk yield. Pages 685-71 in Proc. of the Interbull mtg., Barcelona, Spain. Interbull No. 40, Interbull, Uppsala, Sweden.

Philipsson, J. 2005. Interbull – how it began and some achievements. Pages 131-135 in Proc. of the Interbull mtg., Uppsala, Sweden. Interbull No. 33, Interbull, Uppsala, Sweden.

Powell, R.L. 1988. Comparison of procedures to convert sire evaluations between countries. J. Dairy Sci., 71:1609-1615.

Pryce, J., S. Bolormaa, A. Chamberlain, P. Bowman, K. Savin, M.E. Goddard, and B.J. Hayes. 2010. A validated genome wide association study in two dairy cattle breeds for milk production and fertility traits using variable lengths haplotypes. J. Dairy Sci., 93:3331-3345.

Ptak, E., and L.R. Schaeffer. 1993. Use of test day yields for genetic evaluation of dairy sires and cows. Livest. Prod. Sci., 34:23.

Reinhart, F., Z. Liu, F. Seefried, and G. Thaller. 2009 Implementation of genomic evaluation in German Holsteins. Pages 219-226 in Proc. of the Interbull mtg., Barcelona, Spain. Interbull No. 40. Interbull, Uppsala, Sweden.

Schaeffer, L.R., and J. Dekkers. 1994. Random regression in animal models for test-day production in dairy cattle. Pages 443-446 in Proc. of the 5th World congress on genetics applied to livestock production: 7-12 August 1994; Guelph, Canada.

Schaeffer, L.R., and J. Jamrozik. 1996. Multiple-trait prediction of lactation yields for dairy cows. J. Dairy Sci., 79:2044-2055

Schaeffer, L.R., J. Jarmrozik, G.J. Kistemaker, B.J. van Doormaal. 2000. Experience with Test-Day Model. J. Dairy Sci., 83:1135-1144.

Schaeffer, L.R. 1983. Effectiveness of model for cow evaluation intraherd. J. Dairy Sci., 66:874–880

Schaeffer, L.R. 1994. Multiple-Country comparison of dairy sires. J. Dairy Sci., 77:2671-2678.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet., 123:218-223.

Solberg, T.R., A.K. Sonesson, J. Woolliams, and T.H.E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol., 41:29

Stolzman, M., H. Jasiorowski, Z. Reklewski, A. Zarnecki, and G. Kalinowska. 1981. Friesian cattle in Poland, preliminary results of testing different strains. World Anim. Rev., 38:9.

Szyda, J., A. Zarnecki, and S. Kaminski. 2009. The Polish genomic breeding value estimation project. Pages 43-46 in Proc. of the Interbull international workshop: 26-29 Jannuary 2009; Uppsala, Sweden. Interbull No. 39, Interbull, Uppsala, Sweden.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics, 123:585-595.

Tenesa, A., P. Navarro, B.J. Hayes, D.L. Duffy, G.M. Clarke, M.E. Goddard, and P. Visscher. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res., 17:520-526.

Tibshirani, R. 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc., 58:267-288.

van der Beek, S. 2007. Effect of genomic selection on national and international genetic evaluations. Pages 115-118 in Proc. of the Interbull mtg., Dublin, Ireland. Interbull No.37, Interbull, Uppsala, Sweden.

van der Linde, R., A.P.W. de Roos, A.G.F. Harbers, and G. de Jong. 2005. Mace with sire-mgs and animal pedigree. Pages 3-7 in Proc. of the Interbull mtg., Uppsala, Sweden. Interbull No. 33, Interbull, Uppsala, Sweden.

van Doormaal, B.J., G.J. Kistemaker, P.G. Sullivan, M. Sargolzaei, and F.S. Schenkel. 2009. Canadian implementation of genomic selection. Pages 214-218 in Proc. of the Interbull mtg., Barcelona, Spain. Interbull No. 40, Interbull, Uppsala, Sweden.

van Kaam, J.C.H.M., G.B. Jansen, E.L. Nicolazzi, and F. Canavesi. 2009. Work in progress on Italian Holstein genomic evaluation. Pages 231-234 in Proc. of the Interbull mtg., Barcelona, Spain. Interbull. 40, Interbull, Uppsala, Sweden.

VanRaden, P.M., C.P. van Tassell, G.R. Wiggans, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and F. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. J. Dairy Sci., 92:16-24.

VanRaden, P.M. 2007. How relatives are related. Pages 33-36 in Proc. of the Interbull mtg., Dublin, Ireland. Interbull No. 37, Interbull, Uppsala, Sweden.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci., 91:414-4423.

Voight, B.F., S. Kundaravalli, X. Wen, and J.K. Pritchard. 2006. A map of recent positive selection in the Human genome. PLos Biol., 4(4):e154.

Weir, B.S., L.R. Cardon, A.D. Anderson, D.M. Nielsen, and W.G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Res., 15:1468-1476.

White, I.M.S., and S. Brotherstone. 1997. Modeling lactation curves with cubic splines. Pages 80-82 in Proc. of the 1997 Interbull mtg., Vienna, Austria. Interbull No. 16, Interbull, Uppsala, Sweden.

Wilmink, J.B.M. 1986. Conversion of breeding values for milk from foreign populations. Livest. Prod. Sci., 14:223-229.

Wilmink, J.B.M. 1987. Adjustment of test-day milk, fat, and protein yields for age season and stage of lactation. Livest. Prod. Sci., 16:335-348.

Yi, N., and S. Xu. 2008. Bayesian LASSO for quantitative trait loci mapping. Genetics, 179:1045-1055.

CONTENTS OF THE THESIS

This thesis is structured in an introduction, five main chapters and a conclusion. The five main chapters (Chapters II to VI) contain papers published or submitted to national or international journals. Papers in Chapters II to V were published (or submitted) to peer-reviewed journals. The paper included in Chapter VI was published in the Interbull bulletin, which is not peer reviewed.

Chapter II assess the ability of international genetic evaluations in predicting domestic breeding values of Holstein bulls. This is an important issue, as international EBV are used in most countries either to improve bulls' national evaluations or to decide whether to import or not to import semen of a foreign bull. Previous studies have reported a greater accuracy of international genetic evaluations compared to other possible sources of information (i.e. pedigree relationships, or "country of origin" foreign evaluations). These have been carried out in single countries, considering different time periods, using different methods, analysing different traits and including a different amount of genetic evaluations. The study described in Chapter II uses a common approach on bulls progeny tested in six countries with different production systems and evaluation models, and is focused on yield traits (milk, fat and protein). This to assess the value of international evaluations across countries in a wide range of possible scenarios.

Chapter III investigates marker pre-selection methods in genome-wide simulated data. Marker pre-selection decreases the over-parameterization of models used in the estimation of breeding values with genome-wide information. Two indirect methods, chosen for their simple application and based on statistical significance of SNP association with the trait analysed, were tested: i) Bonferroni correction of the significance threshold and ii) Permutation test to obtain the reference distribution of

the null hypothesis. Associations were estimated by single SNP regression and DGV with a BLUP method.

In Chapter IV the aforementioned simulated dataset was used to test an alternative Principal Component Regression (PCR) approach to reduce the number of estimates in the model. The appealing feature of PCR is that it reduces the number of estimates without actually excluding SNP from the estimation by using (independent) linear combinations of all markers. The idea is not new, as interesting results on simulated data have already been reported (Solberg et al., 2009). The study included in this thesis tested the use of eigenvalues in the mixed model equations, to account for the fact that eigenvalues actually *quantify* the contribution of each PC to the original marker covariance structure (avoiding to consider a homogeneous variance for all PCs which would be incorrect).

Chapter V tested four different methods for genome-wide selection in both simulated and real datasets: Bayes-BLUP, BayesA, and two different Bayesian LASSO models. Simulated data were the same as in Chapters III and IV. The real dataset consisted in nearly a thousand progeny tested Australian Holstein bulls. The Bayesian LASSO had already been proposed for QTL mapping and for indirectly pre-select markers for genomic selection. However, there are no direct comparisons available of the performance of this method with other well-known methods as BLUP and BayesA.

The paper in Chapter VI searched for genome-wide selection signatures in Italian Brown comparing fixation index (F_{st}) profiles of 2682 bulls belonging to the dairy Italian Brown and Italian Holstein, the dual purpose Italian Simmental, and the beef Marchigiana and Piedmontese breeds. Thereafter, different combinations of markers with outlier F_{st} values were used in a genomic prediction of two production traits (milk

yield and protein percentage), a conformation trait (Udder Score) and an economic index (Total Economic Index), using BayesA as prediction method.

OBJECTIVES OF THE THESIS

The general objective of this thesis is to investigate new problems and opportunities presently faced by the dairy cattle industry. On the one side, the need for unbiased methods for estimating breeding values across countries, since semen and embryos international trading is continuously increasing in importance and value. On the other side, the new opportunities offered by novel developments in DNA technologies to include or even base the genetic evaluations on genomic data. Both are hot topics that promise to have a great impact in the sector in the near future. In particular, specific objectives are:

- To assess the accuracy of international genetic evaluation when predicting dairy cattle domestic breeding values of foreign Holstein bulls in a multi-country framework.
- To test different methods (BLUP with and without marker pre-selection, principal component regression, different Bayesian methods) in the prediction of breeding values using genome-wide marker data.
- To analyse the different performance of the methods in simulated and real scenarios.
- To search for selection signatures in five Italian breeds genotyped genome-wide by the Italian SELMOL project.
- To test genome-wide predictions in Italian Brown using markers carrying significant selection signatures in this breed.

CHAPTER II

Assessment of the value of international genetic evaluations for yield in predicting domestic breeding values for foreign Holstein bulls.

E.L. Nicolazzi*, F. Foraboscoٹ and W. F. Fikseٹ

**Istituto di Zootecnica, Facoltà di Agraria, Università Cattolica del Sacro Cuore, Piacenza, Italy.*

ٹInterbull Centre, Dept. of Animal Breeding and Genetics, SLU, Uppsala, Sweden

ٹDept. of Animal Breeding and Genetics, SLU, Uppsala, Sweden

Submitted to: Journal of Dairy Science

ABSTRACT

International genetic evaluations are a valuable source of information for decisions about the importation of (semen of) foreign bulls. This study analysed data from six countries (Australia, Canada, Italy, France, the Netherlands and the United States), and compared international evaluations for production traits of foreign bulls (i.e. when no national daughter information was available) with their national breeding values in August 2009, based only on domestic daughters' data. A total of 821 bulls with highly reliable estimated breeding values (EBV) for milk, fat and protein yield were included in the analysis. No evidence of systematic over or underestimation was found in most of the countries analysed. Observed correlations between national and international evaluations were close to 0.9 and for most of the countries generally close to their expected values (calculated from national and international EBV reliabilities). In Italy, however, higher differences between observed and expected correlations and significant mean differences between EBVs for more than one trait were observed for bulls progeny tested in the USA and in (other) European countries (with differences up to 33.1% of the genetic standard deviation). These results were probably induced by a relatively recent change in the model for national evaluation. The findings in this study reflect a conservative outcome of the real value of international evaluations, as changes in methodologies in either the national or the international evaluations reduced the ability of past international evaluations to predict current national evaluations. Nevertheless, our results indicate that international evaluations based on foreign information for Holstein bulls were reasonably accurate predictions of the future national breeding values based only on domestic daughters.

(**Key words**: Holstein, production trait, international genetic evaluation)

INTRODUCTION

International estimated breeding values (**EBV**) are expected to accurately predict the future performance of bull daughters in all countries participating in the international genetic evaluations. Since 1995, the method used to estimate international EBV is a multiple-trait sire model called Multiple Across Country Evaluation (**MACE**) by Schaeffer (1994). Interbull genetic evaluation service provides international EBV for 6 dairy breeds and 7 major trait groups. In the August 2009 evaluation, MACE predictions for the Holstein breed were distributed to 26 participant countries for production traits. International information on foreign bulls is widely used by all countries not only to improve national bull evaluations but also to decide whether or not to import a foreign bull (or its semen). When domestic daughters' data for a foreign bull are not available in the importing country, possible sources of information are pedigree relationships and information on daughters in another (foreign) country, either as "country of origin" foreign evaluations or as (MACE) international evaluations.

Over the past fifteen years, Interbull implemented several changes in the procedure for international evaluations to improve the ability of the method to convert breeding values across countries. For example, a time edit for (date of birth of) bulls was considered to ensure that the base population is similar for all countries (Weigel and Banos, 1997; de Jong, 2003). In 2000 effective daughter contributions (EDC; Fikse and Banos, 2001) replaced number of daughters as weighting factors to account for the precision of national bull evaluations, which reduced bias in sire variance estimates and resulted in improved approximations of reliabilities. The procedure for estimation of genetic correlations was reviewed in 2004 (Wilmink and Fikse, 2004), and changes were implemented during the same year. Furthermore, other

improvements are under development, such as the inclusion of a sire-dam pedigree, which is expected to greatly reduce the problem related to the influence of phantom grouping (de Jong, 2003; van der Linde et al., 2005; Jakobsen and Fikse, 2009), or the implementation of multiple-trait MACE (Nilforooshan et al., 2009). Interbull evaluations are under constant scrutiny and other concerns (Canavesi et al., 2005; Ducrocq et al. 2003) have not directly resulted in changes. Hence, monitoring of the quality of international evaluations is warranted.

Powell et al. (2000) showed that accuracy of the United States national dairy bull evaluations (i.e. based only on national data) was improved when including foreign daughter information. Later on, Powell et al. (2004) showed that parent average for production traits was not a good alternative to international EBV based on foreign daughters. Parent average underestimated bulls' EBV based on US daughters after semen importation. On the other hand, international evaluations were on average close to the US national evaluation based on US daughters. Furthermore, Van der Linde (2004) showed that correlations in conformation traits for Holstein bulls between "country of origin" foreign evaluations (i.e. on foreign country scale) and Dutch national evaluations were up to 32% lower than those obtained comparing international and national evaluations. In addition, Brochard et al. (2006) analysed production, udder health and conformation traits and confirmed the accuracy of international evaluations for predicting the future performance of Holstein foreign bulls in France.

While these previous studies were of great importance to document the accuracy of international EBV in predicting future performance of imported bulls in different countries, comparability of the results is somewhat limited because of differences in the approaches adopted in the analyses. In this study we compared data of six

countries from three continents using a common approach (i.e., time period, editing criteria and statistical analysis). This facilitates a multiple-country assessment of the value of international EBV for production traits (milk, fat and protein yield) as unbiased predictors of future performance of Holstein bulls' daughters in countries other than the test country.

MATERIALS AND METHODS

Six reference countries from three continents were studied: Australia (**AUS**), Canada (**CAN**), France (**FRA**), Italy (**ITA**), the Netherlands (**NLD**) and the United States of America (**USA**). Two sources of information were compared: i) August 2009 national evaluations in each reference country (i.e. EBVs based only on daughter information in the reference country, and hereinafter referred to as **DOM2009**) and ii) previous international genetic evaluations based only on foreign daughters' information (Interbull evaluations from 2001 and onwards, hereinafter referred to as **INTPRED**). For each reference country we retrieved a bull's last international EBV that did not include information on daughters from the reference country. Bulls missing one or both types of information were not of interest and deleted from the analysis.

International EBVs of bulls in MACE are based on pedigree relationships and progeny information. If a bull has no daughters in a particular country, MACE allows predicting breeding values for this bulls and country by using genetic correlations among countries and daughter information from other countries. In addition, information on ancestors, currently sire and maternal grandsire, is utilized. In matrix notation MACE is (after Schaeffer, 1994):

$$\mathbf{y}_i = \mathbf{1}\mu_i + \mathbf{Z}_i\mathbf{Q}\mathbf{g}_i + \mathbf{Z}_i\mathbf{s}_i + \mathbf{e}_i,$$

where $\mathbf{y}_i$ is a vector of deregressed proofs (DRP) of country *i* for a trait (e.g. milk yield); $\mathbf{1}$ is a vector of ones; μ_i is the mean of the country *i*; $\mathbf{Z}_i$ is a design matrix

relating phenotypes to sires in country *i*; **Q** is a design matrix relating sires to phantom groups; $\mathbf{g}_i$ is a vector of phantom parent group effects; $\mathbf{s}_i$ is a vector of genetic effects for sires in country *i*; and $\mathbf{e}_i$ is a vector of random residuals. Phantom groups of unknown parents are formed by year of birth, country of origin and path of selection (sire, maternal grand-sire, maternal grand-dam).

To express historical evaluations on the same base and scale, linear regression were performed of MACE EBVs from August 2009 on MACE EBVs from May 2001 onwards. These regressions were done separately for each country, and were based on all bulls with domestic daughters for each pair of MACE evaluations. Intercept and slopes were used to convert all international EBV to the most recent genetic base, in order to allow comparisons across time. In general, intercepts were negative (i.e. because of base changes), regression coefficients were close to one and correlations were high (i.e. higher than 94%), except when a change in the predictive model of a reference country was introduced (Figure 1).

To ensure comparisons between reliable EBV, animals with Interbull reliability lower than 70% or less than 100 daughters in either the foreign or the reference country were discarded. For the same reason, only bulls with a first international evaluation before January 2004 and uninterruptedly present in the Interbull distribution files up to August 2009 were retained.

All bulls that met the aforementioned requirements were analysed either all together (i.e. bull progeny tested in all the foreign countries included in our study) or grouped by the foreign country of test. The foreign country of test of a bull was defined as the country where the bull had the highest number of daughters in the last international genetic evaluation considered (August 2009). The assumption was that most bulls would have the highest number of daughters in the country of test. Although this

assumption may incorrectly assign the foreign country of some bulls, it was considered more accurate than using the country of first registration. For all reference countries, bulls progeny tested in Italy, France, Germany and the Netherlands were grouped together (and named **EUR**), because the number of bulls per country was less than 50.

Mean and SD of differences, average reliabilities, regressions and correlations between INTPRED and DOM2009 were calculated. Significances of differences between INTPRED and DOM2009 were tested for all three traits and in all foreign countries with a paired t-test. A Bonferroni correction of type I error threshold was applied to reduce false positives due to multiple testing.

All comparisons were performed on the national scale of the reference country. Mean and SD of differences between EBVs were expressed in % of (animal) genetic SD of the reference country (obtained from Interbull, 2010). Domestic reliabilities ($REL_{domestic}$) were obtained as a function of EDC and heritability (Liu et al., 2004):

$$REL_{domestic} = \frac{EDC}{EDC + k} \times 100$$

, where

$$k = \frac{\sigma_e^2}{\sigma_s^2} = \frac{4 - h^2}{h^2}$$

Since Interbull made reliabilities of international evaluations for fat and protein yield available only from March 2007 onwards, reliabilities for these traits prior to January 2007 were considered to be equal to milk reliabilities. As a result, reliabilities of international EBV in this study are not expected to have large variation across traits. Observed correlations were compared to their expected value, obtained following Brochard et al. (2006):

$$Corr_{\exp} = \sqrt{REL_{domestic} \times REL_{ITB}} \ ,$$

where $Corr_{exp}$ is the expected correlation, $REL_{domestic}$ is the mean reliability of bulls in August 2009 national evaluation of the reference country, and REL_{ITB} is the (previous) international mean reliability of bulls on their latest international evaluation without daughters in the reference country.

During the time period considered (from May 2001 to August 2009) models and procedures have changed for both national and international evaluations. Detailed information on these changes is available on the Interbull website (Interbull, 2010). Briefly, major changes on international evaluations were: the introduction of checks on national EBV and on pedigree data received by the countries and a modification of the procedure for estimation of genetic correlations.

With respect to national evaluations for production traits, Italy changed its model from a lactation model to a random regression multiple-trait multiple-lactation test day model (**TDM**) in November 2004. In May 2003, Canada introduced Legendre polynomials to model the curves of their TDM. The Netherlands introduced a random regression single-trait multiple-lactation TDM inNovember 2002. Finally, the United States introduced new adjustments to its model in February 2005 and May 2007.

RESULTS AND DISCUSSION

The total number of bulls considered was 821, with 518 foreign bulls present in only one reference country and the remaining 303 bulls present in more than one reference country (Table 1). A total of 1,337 cases (bulls receiving a domestic evaluation after semen import in one of the reference countries) were obtained after the editing in the six reference countries (Table 1). The year of birth of the bulls included in the analyses ranged from 1986 to 1999. The United States and Australia had the lowest number of observations (134 and 137, respectively), whereas the

number of foreign bulls in Italy was the highest (336). Only 29 European bulls were found when Canada was the reference country; thus results in this analysis must be read carefully because they might be influenced by the low number of observations. The majority of bulls had their last Interbull evaluation without daughters in the reference country before 2007. The number of observations in the last class, generally corresponding to younger bulls, was usually lower than the other two classes (Table 2).

Table 1. Number of bulls in the reference countriesand mean (international and domestic) reliabilities for milk yield[1].

Reference countries		Foreign countries[2]			
		ALL	CAN	USA	EUR
Australia	bulls (n)	137		54	83[5]
	reliability int.[3]	79.3		79.5	79.2
	reliability dom.[4]	95.9		95.5	96.1
Canada	bulls (n)	180		151	29[6]
	reliability int.[3]	90.6		90.3	92.4
	reliability dom.[4]	97.8		97.9	97.4
France	bulls (n)	288	61	155	72[7]
	reliability int.[3]	90.4	90.9	90.1	90.5
	reliability dom.[4]	97.9	98.4	97.7	98.1
Italy	bulls (n)	336	52	199	85[8]
	reliability int.[3]	88.1	88.9	89.1	85.2
	reliability dom.[4]	97.3	96.9	97.7	96.5
The Netherlands	bulls (n)	262		161	101[9]
	reliability int.[3]	89.1		89.4	88.7
	reliability dom.[4]	97.4		97.3	97.7
United States	bulls (n)	134	57		77[10]
	reliability int.[3]	86.7	87.6		86.0
	reliability dom.[4]	95.2	95.7		94.9

[1] Mean reliabilities in fat and protein yield do not differ more than 1.5% from those obtained in milk, thus, are omitted for clarity.
[2] Foreign countries: ALL=all countries in this study, analysed together; CAN=Canada; USA=the United States of America; EUR= European countries
[3] Mean of 'previous' international reliability.
[4] Mean of August 2009 domestic reliability.
[5] Italy(20), Germany (17), France(18) and the Netherlands(28) were grouped together.
[6] Italy(10), Germany(7) and the Netherlands(12) were grouped together.
[7] Italy(36), Germany(12) and the Netherlands(24) were grouped together.
[8] France(20), Germany(30) and the Netherlands(35) were grouped together.
[9] Italy(26), France(36) and Germany(39) were grouped together.
[10] Italy(22), Germany(12), France(11) and the Netherlands(32) were grouped together.

Table 2. Number of bulls in the reference countriesby classes of latest international genetic evaluation without domestic daughters.

Reference countries	Latest international evaluation	Foreign countries[1]			
		ALL	CAN	USA	EUR
Australia	[2001:2003][2]	68		32	36[5]
	[2004:2006][3]	45		15	30[5]
	[2007:2009][4]	24		7	17[5]
Canada	[2001:2003][2]	61		56	5[5]
	[2004:2006][3]	72		62	10[5]
	[2007:2009][4]	47		33	14[5]
France	[2001:2003][2]	169	34	91	44[6]
	[2004:2006][3]	76	19	43	14[6]
	[2007:2009][4]	43	8	21	14[6]
Italy	[2001:2003][2]	130	12	82	36[7]
	[2004:2006][3]	149	25	89	35[7]
	[2007:2009][4]	57	15	28	14[7]
The Netherlands	[2001:2003][2]	159		115	44[8]
	[2004:2006][3]	52		38	14[8]
	[2007:2009][4]	22		8	
United States	[2001:2003][2]	63	32		31[5]
	[2004:2006][3]	54	20		34[5]
	[2007:2009][4]	17	5		12[5]

[1] Foreign countries: ALL=all countries in this study, analysed together; CAN=Canada; USA=the United States of America; EUR= European countries
[2] Latest international evaluations without domestic daughter information from May 2001 to November 2003.
[3] Latest international evaluations without domestic daughter information from February 2004 to November 2006.
[4] Latest international evaluations without domestic daughter information from February 2007 to April 2009.
[5] Italy, Germany, France and the Netherlands were grouped together.
[6] Italy, Germany and the Netherlands were grouped together.
[7] France, Germany and the Netherlands were grouped together.
[8] Italy, France and Germany were grouped together.
.

Domestic reliabilities of bulls in August 2009 national evaluation were on average higher than 95% (Table 1). The reliabilities of the last Interbull evaluation without daughters in the reference country were somewhat lower (generally around 90%), especially for Australia (79%) due to the lower genetic correlation for production

between Australia and most of the other countries in the Interbull evaluations (Interbull, 2010).

Australia

When all 137 foreign bulls were analysed together mean differences between INTPRED and DOM2009 were 9.4, 0.1 and 15.1% of the genetic SD for milk, fat and protein yield, respectively (Table 3).

Table 3. Mean and standard deviations of differences [(1)] for milk, fat and protein yield traits, between LAST [(2)] and PREV, expressed in % of genetic standard deviation. Significance thresholds were corrected for multiple testing.

		Foreign country[(3)]							
		ALL		CAN		USA		EUR	
Reference country	Trait	mean[(1)]	SD[(1)]	mean[(1)]	SD[(1)]	mean[(1)]	SD[(1)]	mean[(1)]	SD[(1)]
Australia	milk yield	9.4	47.9			11.6	49.3	8.0	47.2[(4)]
	fat yield	0.1	47.3			-4.8	40.5	3.5	51.5[(4)]
	protein yield	15.1	52.4			7.7	49.7	20.5	54.1[(4)]
Canada	milk yield	8.1	40.4			8.9	41.9	4.2	31.5[(4)]
	fat yield	-9.7*	36.9			-11.8*	37.5	1.3	31.8[(4)]
	protein yield	11.0	42.7			10.8	44.4	12.2	32.7[(4)]
France	milk yield	-0.2	37.5	1.9	32.6	-0.1	37.8	-2.2	41.0[(5)]
	fat yield	-5.6	34.4	-0.2	25.2	-9.1	36.6	-2.3	35.8[(5)]
	protein yield	6.8	40.4	12.3	36.2	4.2	41.3	7.6	41.7[(5)]
Italy	milk yield	18.4***	46.7	2.6	44.2	21.0***	45.8	22.0**	48.6[(6)]
	fat yield	7.1	45.7	-2.3	45.3	4.6	44.4	18.4*	47.2[(6)]
	protein yield	22.0***	50.3	1.6	48.0	22.6***	48.1	33.1***	53.2[(6)]
The Netherlands	milk yield	-3.7	40.5			-3.8	39.1	-3.6	42.9[(7)]
	fat yield	-3.5	40.1			-3.6	39.7	-3.4	40.8[(7)]
	protein yield	-2.5	44.5			-3.5	42.9	-0.8	47.1[(7)]
United States	milk yield	4.9	21.5	1.6	20.1			7.2	22.4[(4)]
	fat yield	3.1	21.9	0.0	20.3			5.3	22.8[(4)]
	protein yield	6.7*	23.2	0.9	21.6			11.1**	23.5[(4)]

* $P < 0.05$ ** $P < 0.01$ *** $P < 0.001$

[(1)] Mean and standard deviations of differences, expressed as % of genetic standard deviation. EBV mean differences were calculated as: 100 × (latest national EBV– international EBV)/σ_g, where σ_g is the animal genetic standard deviation.

[(2)] Obtained using only domestic daughters information.

[(3)] Foreign countries: ALL=all countries in this study, analysed together; CAN=Canada; USA=the United States of America; EUR= European countries

[(4)] Italy, Germany, France and the Netherlands were grouped together.

[(5)] Italy, Germany and the Netherlands were grouped together.

[(6)] France, Germany and the Netherlands were grouped together.

[(7)] Italy, France and Germany were grouped together.

None of these differences were significant ($P > 0.05$). SD of the differences were generally the highest of all the reference countries analysed in this study. Such variability of differences was expected based on the lower reliability of international EBV (Table 1) resulting from the lower genetic correlations between Australia and the other reference countries analysed (Interbull, 2010). Coefficients of regression of DOM2009 on INTPRED ranged from 1.08 to 1.16, but were not significantly different from one. Observed correlations between INTPRED and DOM2009 were at most 1% lower than their expectation (Table 4).

When bulls were grouped by foreign countries of test, mean differences between INTPRED and DOM2009 for bulls tested in the USA were 11.6, -4.8 and 7.7 % of the genetic SD for milk, fat and protein yield, respectively. For bulls tested in EUR, mean differences between EBVs for the same traits were 8.0, 3.4 and 20.5% of the genetic SD. None of these differences were significant.

Coefficients of regression of DOM2009 on INTPRED ranged from 1.01 to 1.18, and none were significantly different from one. McClintock et al. (2004) reported a similar trend for foreign US bulls between 2000 and 2003, for a combination of all three production traits (Australian Selection Index). Observed correlations between INTPRED and DOM2009 for bulls tested in the USA were as expected for milk yield and 1% higher than expected for fat (Table 4). However, for protein yield the observed correlation (0.84) was somewhat lower than expected (0.87). This difference could be traced to three closely related bulls (two half-sib bulls, and a third bull having the same maternal grandsire with one of the half-sibs) with more than 2 SD of difference between national and international EBV.

Table 4. Expected[1] and observed[2] correlations for milk, fat and protein yield on each for each reference country.

Reference country	Trait	Foreign country[3]							
		ALL		CAN		USA		EUR	
		Exp.[1]	Obs.[2]	Exp.[1]	Obs.[2]	Exp.[1]	Obs.[2]	Exp.[1]	Obs.[2]
Australia	milk yield	0.87	0.88			0.87	0.87	0.87	0.88[4]
	fat yield	0.87	0.86			0.87	0.88	0.87	0.82[4]
	protein yield	0.87	0.86			0.87	0.84	0.87	0.86[4]
Canada	milk yield	0.94	0.90			0.94	0.89	0.95	0.94[4]
	fat yield	0.94	0.93			0.94	0.92	0.94	0.93[4]
	protein yield	0.94	0.91			0.94	0.91	0.94	0.94[4]
France	milk yield	0.94	0.92	0.95	0.92	0.94	0.93	0.94	0.90[5]
	fat yield	0.94	0.92	0.95	0.95	0.94	0.92	0.94	0.90[5]
	protein yield	0.94	0.93	0.94	0.91	0.94	0.92	0.94	0.92[5]
Italy	milk yield	0.93	0.88	0.93	0.87	0.93	0.87	0.91	0.88[6]
	fat yield	0.93	0.87	0.93	0.87	0.93	0.88	0.91	0.83[6]
	protein yield	0.92	0.85	0.93	0.82	0.93	0.86	0.90	0.83[6]
The Netherlands	milk yield	0.93	0.89			0.93	0.90	0.93	0.87[7]
	fat yield	0.93	0.90			0.93	0.91	0.93	0.87[7]
	protein yield	0.93	0.91			0.93	0.91	0.93	0.90[7]
United States	milk yield	0.91	0.89	0.92	0.88			0.90	0.89[4]
	fat yield	0.91	0.90	0.92	0.88			0.90	0.90[4]
	protein yield	0.91	0.91	0.92	0.86			0.90	0.91[4]

[1] Expected correlations were obtained from the square root of the product of mean international and domestic reliabilities.
[2] Observed (Pearson) correlations.
[3] Foreign countries: ALL=all countries in this study, analysed together; CAN=Canada; USA=the United States of America; EUR= European countries
[4] Bulls tested in Italy, Germany, France and the Netherlands were grouped together.
[5] Bulls tested in Italy, Germany and the Netherlands were grouped together.
[6] Bulls tested in France, Germany and the Netherlands were grouped together.
[7] Bulls tested in Italy, France and Germany were grouped together.

On the other hand, observed correlations for bulls progeny tested in EUR were almost as expected for milk and protein yield. However, the observed correlation for fat yield was 5% lower than its expected value. Again, a thorough analysis revealed two bulls with more than 2 SD of difference between national and international EBV. These five outlier bulls had in common that their reliability of international EBV was low (on average 74%), and had information on daughters and sons in several countries contributing to the latest international EBV.

Canada

Analysed foreign countries in Canada were the same as those considered in Australia as a reference country (Table 1). However, large differences were observed in either the total amount of bulls (180) or the number of bulls considered for each of the foreign countries considered. In fact, the number of bulls progeny tested in the USA was nearly threefold (151), and almost a third of the bulls tested in EUR (29). No bulls progeny tested in France were retained after editing. When bulls were analysed all together, mean differences between INTPRED and DOM2009 were 8.1, -9.7 and 11.0% of the genetic SD for milk, fat and protein yield, respectively (Table 3). A significant difference ($P < 0.05$) was found for fat yield, which could be attributed to bulls progeny tested in the USA. Coefficients of regression of DOM2009 on INTPRED ranged from 1.04 to 1.11 and were not significantly different from unity. Observed correlations between INTPRED and DOM2009 were always lower than the expected correlations (-4, -1 and -3% for milk, fat and protein yield, respectively). For bulls progeny tested in the USA, mean differences between INTPRED and DOM2009 for milk and protein yield were similar to those for US bulls in Australia as a reference country, but the SD of differences were 8 and 5% lower for both traits, respectively (Table 3). A significant mean difference between EBVs of -11.8% of the genetic SD was observed for fat yield ($P < 0.05$). This was somewhat unexpected, especially considering the large amount of common bulls in both countries' pedigrees. There were 11 bulls with differences between EBV from INTPRED and DOM2009 greater than 2 SD in at least one of the three production traits, and the latest international evaluation without daughters in Canada for all these bulls was prior to November 2005. Many major changes in methodologies and models were introduced in both countries before that evaluation. For example, in May 2003, Canada changed from

Wilmink to Legendre polynomials to model the curves used in their TDM and in February 2005 the USA introduced a package of adjustments to its national model (USDA, 2010). The impact of these changes on the results obtained in this study was tested retaining only bulls with latest international evaluation without Canadian daughters after November 2005 (50 bulls). When considering only these bulls, mean differences between INTPRED and DOM2009 were greatly reduced (-2.01, -4.67 and 0.53% of the genetic SD for milk, fat and protein yield, respectively) and non-significant, and the SD of differences between EBVs were nearly halved (18.90, 18.25 and 21.67% of the genetic SD). Thus, the significant difference for protein yield between INTPRED and DOM2009 was more likely caused by changes in the models and the procedures rather than by an actual bias of international EBV.

None of the coefficients of regression of DOM2009 on INTPRED were significantly different from unity, with values ranging from 0.99 for milk yield for bulls tested in EUR to 1.11 for fat yield for bulls tested in the USA. Observed correlations for bulls tested in EUR were close to their expected values (-1, -1 and 0% differences between observed and expected correlations for milk, fat and protein yield, respectively). Observed correlations for bulls tested in the USA were 0.89, 0.92 and 0.91 for milk, fat and protein yield, respectively, and somewhat lower than expected (Table 4). Considering only US bulls with the last international evaluation without domestic daughters after November 2005, observed correlations matched their expected value for milk yield, and were 1% higher than expected for fat and protein yield.

France

Foreign bulls progeny tested in CAN, the USA and EUR were considered (Table 1). When all bulls where analysed together, mean differences between INTPRED and

DOM2009 were generally small and non-significant for all traits. A significant coefficient of regression of DOM2009 on INTPRED of 0.92 was found for fat yield, whereas the other two not significant regression coefficients ranged from 0.90 to 0.91. Observed correlations between INTPRED and DOM2009 were always lower than expected, but only with a maximum of 2% difference (for milk and fat yield).

The highest mean differences between EBVs (expressed in % of the genetic standard deviation) for bulls grouped by foreign country of test were found for protein yield, for bulls progeny tested in CAN and in EUR, and for fat yield for foreign US bulls (12.3, 7.6, and -9.1%, respectively). However, none of the mean differences were significant.

Coefficients of regression of DOM2009 on INTPRED ranged from 0.83 for fat yield for bulls tested in EUR (the only regression coefficient found significant, $P < 0.05$) to 1.01 for fat yield for bulls progeny tested in CAN. Observed correlations between INTPRED and DOM2009 ranged from 0.90 for protein yield for bulls tested in EUR to 0.95 for fat yield for bulls tested in CAN (Table 4). These results were reasonably similar to those reported by Brochard et al. (2006). In fact, Brochard et al. (2006) obtained slightly higher observed correlations between national and international EBV for bulls tested in CAN and the USA, although lower observed correlations were reported for bulls tested in EUR. However, differences in the methods implemented in both studies make comparisons of results difficult. In particular, in Brochard et al. (2006) the grouping of bulls was by country of first registration, whereas in the present study it was the country with the highest number of daughters in August 2009. Another reason is the different time period and editing criteria adopted in this paper, resulting in a different number of bulls analysed. Brochard et al. (2006) considered a shorter time period (from 2000 to 2005) whereas our study includes

four more years of international evaluations. Finally, the editing criteria in Brochard et al. (2006) were less strict for international evaluations (only a reliability higher than 70% was required to include bulls in the analysis) and more stringent for national evaluations (at least 150 daughters, corresponding to a reliability higher than 90%) than the criteria adopted in this study.

Italy

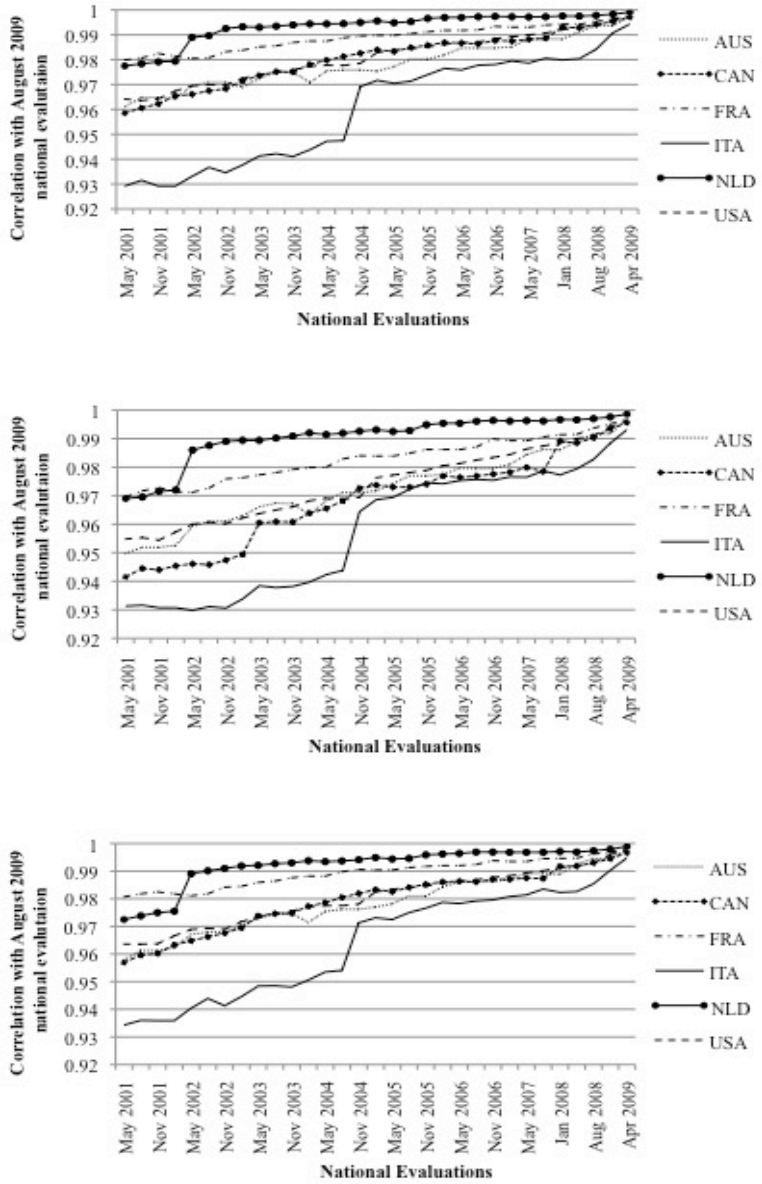
Bulls progeny tested in CAN, the USA and EUR were analysed. Large mean differences between INTPRED and DOM2009 were found for all traits (18.42, 7.06 and 22.0% of the genetic SD for milk, fat and protein yield, respectively) when all bulls were analysed together (Table 4). The differences for milk and protein yield were highly significant ($P < 0.001$). Significant regression coefficients of DOM2009 on INTPRED of 1.03 and 1.02 were found for the same traits. Observed correlations between INTPRED and DOM2009 were between 5% (milk yield) and 7% (protein yield) lower than expected correlations. Possible causes of these results are discussed next, when bull tested in the different foreign countries were analysed separately. In this case, lower mean differences between INTPRED and DOM2009 were found for all traits for bulls tested in CAN, although the SD of differences were similar to those obtained for the other foreign countries (Table 3). Lower mean differences between EBV of ITA and CAN were expected, considering that since November 2004 both countries use a similar TDM model for production traits (Canavesi et al.,2004). Highly significant ($P < 0.001$) mean differences between EBVs were found for milk and protein yield for US progeny tested bulls and for protein yield for bulls tested in France, Germany and the Netherlands. Moreover, for foreign bulls tested in EUR, significant mean differences of EBV were also found for milk ($P < 0.01$) and fat yield ($P < 0.05$). Surprisingly, the SD of differences between

INTPRED and DOM2009 for bulls tested in the USA and in EUR were similar to, and in two cases higher than, the ones obtained considering Australia as a reference country.

Coefficients of regression of DOM2009 on INTPRED for the three foreign countries were not significant and ranged from 0.95 for protein yield in EUR to 1.06 for milk yield in the USA, indicating that in the countries and traits with significant mean differences there is a systematic underestimation of international predictions. Observed correlations between INTPRED and DOM2009 were lower than expected in all foreign countries and traits (Table 4). Differences between observed and expected correlations ranged from -11% (protein yield for bulls tested in CAN) to -3% (milk yield for bulls tested in EUR).

The impact of the relatively recent introduction of TDM on national EBV correlations in all traits was more profound in Italy compared to the introduction of TDM in other countries like Canada or the Netherlands (Figure 1). With respect to international evaluations, the introduction of TDM affected genetic correlations between Italy and the other countries. In addition, in November 2004 Interbull introduced a new method for the estimation of genetic correlations that caused further impact on genetic correlations across countries (Interbull, 2010). Considering that these changes in methodologies may have a large impact on results, further analyses were performed. Foreign bulls in the USA and in EUR were divided into two sub-groups corresponding to the two periods with different predictive models in Italy. The first sub-group (**LAC**) included bulls with the latest international genetic evaluation without Italian daughters from May 2002 to May 2004 (i.e. corresponding to the time period during which Italy used a lactation model). The remaining bulls formed a second sub-group (TDM), corresponding to the time period after the introduction of TDM in Italy.

Figure 1. Within reference country correlations of national EBV in August 2009 with previous evaluations for milk (a), fat (b) and protein yield (c). Reference countries were coded as: AUS (Australia), CAN (Canada), FRA (France), ITA (Italy), NLD (the Netherlands) and USA (the United States of America).



International evaluations for bulls included in the LAC sub-group were compared with August 2004 Italian EBVs, and for bulls included in the TDM sub-group were compared with August 2009 national EBVs. Observed correlations for bulls tested in EUR decreased from 0.93, 0.93 and 0.91 for milk, fat and protein yield in the LAC sub-group, to 0.81, 0.80 and 0.79 for the same traits in the TDM sub-group.

Surprisingly, for bulls progeny tested in the USA, lower differences in observed correlations between the two sub-groups were observed. Observed correlations for all traits in the LAC sub-group were 0.87, and only 1 and 2% lower for milk and protein yield, and 3% higher for fat yield compared to the TDM sub-group. Note that bulls progeny tested in CAN were omitted because the number of observations in both sub-groups was low. Higher variability of correlations between national and international evaluations across time within TDM evaluations had already been reported by Van Kaam et al. (2008). In recent years many efforts have been made to successfully deal with this problem of stability of TDM indices in Italy (Canavesi et al., 2008a,b, 2009). These improvements, however, are too recent to evaluate their impact in the present study.

The Netherlands

Bulls progeny tested in the USA and EUR were considered in the analysis (Table 1). When all bulls were analysed together, mean differences between INTPRED and DOM2009 were always small and negative (-3.7, -3.5, and -2.5% of the genetic SD for milk, fat and protein yield). None of these differences were significant. Non-significant regression coefficients of DOM2009 on INTPRED ranged from 0.94 (milk yield) to 0.96 (fat and protein yield). Observed correlations between INTPRED and DOM2009 were always lower than their expectations (-4%, -3% and -2% for the aforementioned traits, respectively).

Mean differences between INTPRED and DOM2009 for bulls grouped by single country of test were negative for all traits, with at most -3.8% of the genetic SD difference for milk yield (for bulls progeny tested in the USA). None of the mean differences between EBVs were significant (Table 3). Van der Linde (2004) reported

similar mean differences between EBVs considering Canada, Germany and France as foreign countries, with a lower number of bulls included in the analysis.

Coefficients of regression of DOM2009 on INTPRED were lower than one in all traits with values ranging from 0.93 to 0.98, except for protein yield for bulls progeny tested in EUR (1.03). Again, none of these regression coefficients were significant. Observed correlations between INTPRED and DOM2009 for US progeny tested bulls were about 0.9 (from 2% to 3% lower than expected). Lower observed correlations were found for bulls tested in EUR, with values ranging from 0.87 for milk and fat yield to 0.90 for protein yield (Table 4). In this case, ten bulls were found with differences greater than 2 SD in at least one of the production traits. Eight of these bulls had been progeny tested in Italy and France, two countries with different predictive models from the Netherlands. Other than this, no clear pattern for the outlier bulls was found: international reliabilities were only 2% lower than the average of all other bulls; national reliabilities were on average with the non-outlier bulls; the time of the last international evaluation without Dutch daughters ranged from November 2002 to January 2007; birth years of the bulls ranged from 1994 to 1998; and except for three half-sibs, no clear pedigree structure was found.

The United States

Foreign bulls progeny tested in CAN and EUR were considered in the analysis (Table 1). When considering all bulls together, mean differences between INTPRED and DOM2009 were low, positive and not significant for milk and fat yield (4.9 and 3.1% of the genetic SD, respectively). A significant mean difference between EBVs of 6.7% of the genetic SD was obtained for protein yield. Non significant regression coefficients of DOM2009 on INTPRED ranging from 1.01 to 1.05 were observed for the three production traits. Observed and expected correlations between INTPRED

and DOM2009 were generally similar. Observed correlations were 2% and 1% lower than expected for milk and fat yield, respectively, whereas they matched for protein yield. As for the previous reference countries, specific analyses for each bulls' foreign country of test were performed. For CAN progeny tested bulls, low mean differences between INTPRED and DOM2009 were observed (1.6, 0 and 0.9% of genetic SD for milk, fat and protein, respectively). Larger mean differences between EBVs were observed for bulls tested in EUR, with a significant ($P < 0.01$) mean difference of 11.1% of the genetic SD for protein yield. SD of differences for all traits and both foreign countries were the lowest of all the reference countries analysed (values ranged from 20.1 to 23.5% of the genetic SD).

Coefficients of regression of DOM2009 on INTPRED ranged from 0.90 for protein yield for bulls tested in CAN to 1.05 for fat yield for bulls progeny tested in EUR. None of these values deviated significantly from unity. Observed correlations between INTPRED and DOM2009 for bulls tested in CAN (0.88, 0.88 and 0.86 for milk, fat and protein yield, respectively) were similar to those reported by Powell et al. (2004), although both studies differ not only by the period of time considered (from 1995 to 2004 in Powell's study) but also by the number of bulls included in the analysis. Powell et al. (2004) retained bulls with higher national and international EBV reliabilities (80%) than those used in this study but did not set a threshold on the number of daughters for both national and international evaluations. As a result, the editing adopted in the present study was more stringent, as indicated by the higher average reliabilities (+1.7 and +3.6 average reliability in national and international evaluations, respectively). Similar to the results obtained for the USA (foreign) bulls with Canada as a reference country, the six (foreign) CAN bulls progeny tested in the USA found with differences greater than 2 SD between national and international

EBV for at least one of the three production traits, had their latest international evaluation without US domestic daughters before November 2005. Furthermore, four of these bulls had common pedigree with the outliers when Canada was reference country (only sires and maternal grandsires were checked).

Observed correlations between INTPRED and DOM2009 for bulls tested in EUR were almost as expected, with a maximum of ± 1% differences between observed and expected correlations for all traits. For these bulls, comparisons of realized correlations with those obtained by Powell et al. (2004) were not possible.

Final considerations

Mean differences between INTPRED and DOM2009 based on two independent (national and foreign) groups of daughters were generally small and in most cases not significant. Nearly all regression coefficients of DOM2009 on INTPRED did not deviate significantly from unity. In fact, only one regression coefficient was significant (fat yield for bulls progeny tested in EUR when considering France as a reference country). Observed correlations between INTPRED and DOM2009 were generally similar to their expected values.

Although possibly influenced by the low number of bulls included in the analyses, these results indicate that, except for Italy, there is no evidence of systematic problems in the international evaluation of production traits or specific foreign countries.

Modifications in methodologies in either the national or the international evaluations might have influenced our findings (i.e. Italy, Canada). As a consequence, this study may show a conservative outcome of the real value of international foreign evaluations, as these changes could actually be responsible for some of the variation found.

The expected correlations were based on reliability of domestic and international evaluations, which in turn were based on EDCs. Both EDCs and the methods to compute reliabilities are approximate methods, and the expected correlations in this study are thus subject to approximation errors. Approximation methods typically lead to overestimation of reliability, the degree of which depends on genetic evaluation models and data structure. This means that expected correlations in the present study probably imply an upper bound.

The choice to retain only bulls with 70% EBV reliability and at least 100 daughters in both the foreign and the reference country had a direct impact on the amount of young bulls included in the analysis (Table 2). By requiring such highly reliable EBV, the variability of results obtained in this study was more likely to be affected by past changes in predictive models and procedures in both national and international evaluations. To verify the impact of the editing criteria, we performed all analyses using bulls with at least 60% EBV reliability in the domestic evaluation as the only editing criteria. Although observed and expected correlations were somewhat lower, as expected, we confirmed that the agreement between observed and expected correlations was as good as the results in Table 4. Thus it appeared that our observations also apply to younger bulls with lower reliabilities. In addition, all analyses were performed by weighting the information by bull's effective daughter contribution (**EDC**) for bulls retained after the editing, but results were very similar to those of the unweighted analyses.

The range of birth years among the bulls included was 13 years, but there were relatively few young bulls as a consequence of our editing criteria. The amount and sources of information will differ based on the age of the bull, where older bulls typically have higher reliabilities due to larger daughter group sizes and more

information on other close relatives like full/half-sibs brothers. To determine whether the agreement between international and domestic evaluations depended on the age, each dataset was divided in two halves based on year of birth ("old": born before 1994; and "young": born after 1995). No specific trend in regression coefficients was found for the two groups (data not shown). In general lower observed and expected correlations were observed for the group of "young" bull, corresponding to a lower reliability of their EBV.

The criterion used to assign the foreign bulls' country of origin might not have been optimal. In fact, this criterion might have wrongly assigned some bulls, especially in countries where the use of imported bulls is intensive (i.e. Italy), contributing to an increase in the variability of results.

This study analysed performances of bulls tested in different production systems and with different national evaluation models, genetic parameters and amount of genetic links between countries. However, even if countries were selected to cover a wide variability of scenarios, these results can only be considered as a general indication for countries not involved in the current analysis**.**

CONCLUSIONS

Results obtained in this study show that international genetic evaluation for yield traits based on foreign daughters' performance are reasonably accurate predictions of bulls' national EBV in most countries. In fact, except for Italy, no strong bias was observed and realized correlations between national and international EBV obtained were generally close to 0.90. Italian results were most probably influenced by a number of circumstances that might have affected the outcome in this analysis.

Nevertheless, results obtained in this study agreed well with previous studies that compared national and international evaluations.

ACKNOWLEDGEMENTS

Authors would like to acknowledge Daniel Abernethy and Gert Nieuwhof (ADHIS), Brian Van Doormaal (CDN), Sofie Mattalia (INRA), Fabiola Canavesi (ANAFI), Gerben de Jong (CRV) and Paul VanRaden (USDA) for their comments and suggestions, and for providing the necessary information for this study. Prof. Paolo Ajmone-Marsan and João Dürr are kindly acknowledged. Helpful comments and suggestions from the 3 anonymous reviewers are also acknowledged.

ELN was partly founded by AGRISYSTEM Ph.D. fellowship (ciclo XIII).

REFERENCES

Brochard, M., S. Minèry, and S. Mattalia. 2006. Accuracy of international evaluations in predicting French estimated breeding values of foreign Holstein bulls. Pages 67-71 in Proc. Interbull Annu. Mtg., Kuopio, Finland. Interbull No. 35, Interbull, Uppsala, Sweden.

Canavesi, F., S. Biffani, and F. Biscarini. 2004. Test day model for production traits and SCS for the Italian Holstein. J. Dairy Sci. 87(Suppl. 1):40. (Abstr.)

Canavesi, F., G.B. Jansen, E.L. Nicolazzi, S. Biffani, and R. Finocchiaro. 2008. Impact of data editing on fat and protein content on stability of test day model evaluation. Pages 32-35 in Proc. Interbull Annu. Mtg., Niagara Falls, USA. Interbull No. 38, Interbull, Uppsala, Sweden.

Canavesi, F., S. Biffani, E.L. Nicolazzi, and R. Finocchiaro. 2008. Improving stability

of test day model bull proofs. J. Dairy Sci. 91(Suppl. 1):617:619.

Canavesi, F., S. Biffani, G. Bramante, and R. Finocchiaro.2009.Improving the stability of test day model evaluation for production traits in the Italian Holstein. Ital.J.Anim.Sci., 8(Suppl. 2):39-41.

de Jong, G. 2003. MACE – Options for improvement. Pages 112-116 in Proc. of the Interbull technical workshop. Beltsville, MD, USA. Interbull No.30, Intrebull, Uppsala, Sweden.

Fikse, W. F., and G. Banos. 2001. Weighting factors of sire daughter information in international genetic evaluations. J. Dairy Sci., 84:1759-1767.

Interbull. 2010. Public area. Genetic evaluation. Production. [http://www-interbull.slu.se/eval/framesida-prod.htm. Accessed March 24, 2010].

Jakobsen, J., and F. Fikse. 2009. Sire-Dam Pedigree in MACE – Results from a Full-Scale Pilot Study. Pages 91-98 in Proc. Interbull International Workshop. Uppsala, Sweden. Interbull No. 39, Interbull, Uppsala, Sweden.

Liu, Z., F. Reinhardt, A. Bunger, and R. Reents. 2004. Derivation and Calculation of Approximate Reliabilities and Daughter Yield-Deviations of a Random Regression Test-Day Model for Genetic Evaluation of Dairy Cattle. J. Dairy Sci., 87:1896–1907

McClintock, S., K. Beard, and R. Poole. 2003. Interbull proofs are a reasonably unbiased prediction of future performance in Australia for imported bulls. Pages 169-170 in Proc. Interbull Annu. Mtg., Rome, Italy. Interbull No.31, Interbull, Uppsala, Sweden.

Nilforooshan, M.A., J. Jakobsen, W.F. Fikse, B. Berglund, and H. Jorjani. 2009. MT-MACE for female fertility and milk yield. Pages 68-71 in Proc. Interbull Annu. Mtg., Barcelona, Spain. Interbull No.40, Interbull, Uppsala, Sweden.

Powell, R.L., A.H. Sanders, and H.D. Norman. 2004. Accuracy of foreign dairy bull evaluations in predicting United States evaluations for yield. J. Dairy Sci., 87:2621-2626.

Powell, R.L., H.D. Normal, and G. Banos. 2000. Improving prediction of national evaluations by use of data from other countries. J. Dairy Sci., 83:368.

Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. J. Dairy Sci. 77:2671-2678.

USDA. 2010. Documents of evaluations. Changes across time. February 2005. Base change (general documentation).
http://www.aipl.arsusda.gov/reference/base2005.htm. Accessed March 24, 2010.

van der Linde, R., and M. Nooijen. 2004. De waarde van Interbull. Pages 12-14 in Veeteelt, Januari 1/2.

van der Linde, R., A.P.W. de Roos, A.G.F. Harbers, and G. de Jong. 2005. Mace with sire-mgs and animal pedigree. Pages 3-7 in in Proc. Interbull Annu. Mtg. Uppsala, Sweden. Interbull No. 33, Interbull, Uppsala, Sweden.

Van Kaam, J.B.C.H.M., R. Finocchiaro, F. Canavesi, and S. Biffani. 2008. Assessment of predictive ability of MACE for production traits of Italian and foreign bulls. Pages 158-162 in Proc. Interbull Annu. Mtg. Niagra Falls, USA. Interbull No. 38, Interbull, Uppsala, Sweden.

Weigel, K.A. and G. Banos. 1997. Effect of time period of data used in International dairy sire evaluations. J. Dairy Sci., 80:3425–3430.

CHAPTER III

Use of different marker pre-selection methods based on single SNP regression in the estimation of Genomic-EBVs

E. L. Nicolazzi[1], R. Negrini[1], C. Dimauro[2]

[1]Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, Italy;
[2]Dipartimento di Scienze Zootecniche, Università di Sassari, Italy.

*Published in:*Italian Journal of Animal Science 2009. Vol. 8: 2s. ISSN 1828-051X.

ABSTRACT

Two methods of SNPs pre-selection based on single marker regression for the estimation of genomic breeding values (G-EBVs) were compared using simulated data provided by the XII QTL-MAS workshop: i) Bonferroni correction of the significance threshold and ii) Permutation test to obtain the reference distribution of the null hypothesis and identify significant markers at P<0.01 and P<0.001 significance thresholds. From the set of markers significant at P<0.001, random subsets of 50% and 25% markers were extracted, to evaluate the effect of further reducing the number of significant SNPs on G-EBV predictions. The Bonferroni correction method allowed the identification of 595 significant SNPs that gave the best G-EBV accuracies in prediction generations (82.80%). The permutation methods gave slightly lower G-EBV accuracies even if a larger number of SNPs resulted significant (2,053 and 1,352 for 0.01 and 0.001 significance thresholds, respectively). Interestingly, halving or dividing by four the number of SNPs significant at P<0.001 resulted in an only slightly decrease of G-EBV accuracies. The genetic structure of the simulated population with few QTL carrying large effects, might have favoured the Bonferroni method.

*Key words:*Genomic Selection, SNP pre-selection, Bonferroni correction, Permutation test

INTRODUCTION

The recent availability of high-density SNP panels for the bovine genome boosted fine-mapping QTL studies, association studies with functional traits, and the search for causative mutations. However, the highest expectation is in Genomic Selection (GS), which uses dense marker panels for predicting genomic estimated breeding values (G-EBVs) on young animals before phenotypic information is available

(Meuwissenet al., 2001). A major statistical and computational limitation to be solved in GS is the estimation of tens of thousands of marker effects based only on few thousands of phenotypes. The size of available SNP panels (54K in cattle) largely affects the dimension of matrices in the mixed model equations and the required computational resources for data storage and algorithm solving (Legarra and Misztal, 2008). To face these problems, an important issue is whether or not to include all the available SNPs in the predictive model (Gonzalez-Recio et al., 2008). In spite of likely decreasing G-EBV accuracies, SNPs pre-selection will sensibly reduce the number of equations in the model. The choice of a suitable predictive model, able to combine adequateG-EBV accuracies with reasonable computing requirements, is another key issue. In simulated data, Meuwissen et al. (2001) using Bayesian MCMC methods obtained values of accuracies ranging from 6 to 11% higher than those obtained using BLUP. However, Bayesian methods require substantially longer computing time compared to BLUP. Moreover, early results on real data indicate that G-EBV accuracies obtained with BLUP are only 2-3% lower than those obtained with Bayesian methods (Harris et al., 2008). Therefore, BLUP predictions based on pre-selected SNPs seem a reasonable compromise between loss of accuracy and computational effort. In this paper we tested two single marker regression based methods to reduce the number of equations in the modelcomparing the variations in G-EBV accuracies.

MATERIAL AND METHODS

The simulated data set comprised 5,865 individuals structured in 7 generations. Pedigree relationships and genotypes at 6,000 SNPs evenly distributed across six chromosomes were available for all individuals, whereas phenotypic information was provided for the first 4 generations only. A total of 4,665 individuals from generation 0

to 3 were considered as training animals and 1,200 individuals from generations 4 to 6 as prediction young animals. True breeding values (TBV), calculated by summing QTL effects, were available for all animals. Although the SNP phases were known, all the analyses were performed by single markers because the level of linkage disequilibrium (LD) of the dataset (r^2=0.21 at 0.1 cM distance) greatly reduces the potential advantage of using haplotypes (Hayes et al., 2007).SNP pre-selection using Bonferroni correction was performed (Bolding, 2006) fixing an empirical threshold of $1.6E^{-6}$ (i.e., 0.01/6000) for the *P* values of the F test.

SNPs pre-selection by Permutation test was performed considering 1,000 iteration (a good compromise between statistical significance and computational time) and fixing two different significance thresholds: 0.01 and 0.001. Furthermore, two subsets comprising 50% and 25% of SNPs significant at 0.001 threshold were randomly assembled, to assess the effect of the number of SNPs on G-EBV accuracies. Random sampling procedure was iterated three times for each subset. Marker effects were estimated with the following mixed linear model:

$$y_{ijk} = \mu + SEX_i + GEN_j + \sum_{k=1}^{m} \mathbf{H}_k \mathbf{b}_k + e_{ijk}$$

where **y** is the trait value, μ is the overall mean, **SEX** is the fixed effect of sex (i=1, 2), **GEN** is the fixed effect of generation (j=0-6), **b** is a vector of genotype random effects for all *m* significant SNPs, **H** is the corresponding design matrix, and **e** is the random residual. An equal contribution of each locus to the genetic variance was considered (e.g.: σ^2_a*1/m), thus was calculated as $\sigma^2_e/(\sigma^2_a$ /number of m significant SNPs). Moreover, no interaction effect between SNPs was assumed. G-EBVs for training and prediction generations were obtained as:

$$G\text{-}EBV_i = m + \sum_{k=1}^{m} h'_k b_k$$

Variance components were calculated with the MTDFREML package, and accuracies were estimated by calculating the correlation between G-EBVs and TBVs.

RESULTS AND DISCUSSION

Theadditive variance (σ^2_a) of the trait was 1.324 and the residual variance (σ^2_e) was 3.142. The heritability was 0.30.The polygenic animal model for traditional EBV estimation produced accuracies of 71% for training and 33% for prediction generations.

Bonferroni correction method retained 595 out of 6,000 markers, whereas the permutation approach yielded 2,053 and 1,352 significant SNPs for 0.01 and 0.001 significance thresholds, respectively. All the Bonferroni-selected markers overlapped those selected with permutation test, with the exception of one marker at the 0.001 threshold.

High conservative Bonferroni correction showed its drawback failing to retain markers close to 8 small effect QTLs out of the 44 QTLs embedded in the dataset. However, Bonferroni-selected markers yielded higher accuracies in prediction generations (Table 1).

Conversely, permutation test was able to identify all QTLs but the cost for this sensitivity was a “background noise” - due to the higher number of false positives - that negatively affected G-EBV accuracies. Indeed, SNPs significant at 0.001 threshold performed better than those significant at 0.01 threshold, albeit only 2/3 of the markers were used in G-EBV estimation(81.11% *vs.* 79.37% accuracy in prediction generations).Given the hard computation effort needed to further decrease the significance threshold (e.g., to 1/10,000), a lower number of markers was tested just creating subsets of randomly selected SNPs among those passing the 0.001 threshold. Interestingly, randomly halving the number of SNPs used in the estimation,

G-EBV accuracies decreased only 2% on average. Indeed, many of the 1,352 SNPs were located nearby the 44 QTLs and the random selection of marker subsets still tagged all or most QTLs. When decreasing the number of markers fourfold, G-EBV accuracies decreased further (84.97 and 74.26 for training and prediction, respectively).

Table 1. Accuracies obtained with Bonferroni and Permutation methods.

			Bonferroni correction		Permutation test 0.01 threshold		Permutation test 0.001 threshold	
			Training	Prediction	Training	Prediction	Training	Prediction
4 training generations	All SNPs	"	89.00	82.80	89.20	79.37	89.19	81.11
	50% SNPs	%	-	-	-	-	87.79 (0.008)	78.40 (0.001)
	25% SNPs	"	-	-	-	-	84.97 (0.008)	74.26 (0.017)
1 training generation	All SNPs	"	84.85	71.15	83.12	64.55	83.93	68.46
	50% SNPs	%	-	-	-	-	83.25 (0.006)	65.79 (0.023)
	25% SNPs	"	-	-	-	-	81.20 (0.007)	61.67 (0.014)

The accuracy values obtained in this paper, combining pre-selection methods based on single marker regression and BLUP estimation of G-EBV, were lower than those reported in literature for Bayesian methods while higher than those obtained by the polygenic animal model. These results were also comparable with those reported in simulated data with similar marker density and models (Kolbedhari et al., 2007; Muir, 2007). In traits where few QTLs explain large proportions of genetic variance - as in this simulated data set - Bonferroni correction seems a better pre-selection method compared to Permutation test at 0.001 significance threshold.

ACKNOWLEDGMENTS

The authors wish to thank Prof. N. Macciotta, Prof. P. AjmoneMarsan, and Dr. G. Gaspa for their contribution.Research funded by the Italian Ministry of Agricultural Policies (research project SELMOL)

REFERENCES

Bolding, D.J. 2006. A tutorial on statistical methods for population association studies.Nature Reviews Genetics. 7:782-791.

Gonzalez-Recio, O., D. Gianola, N. Long, K.A. Weigel, G.J.M. Rosa, and S. Avendano. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. Genetics. 178:2305-2313.

Harris, B.L., D.L. Johnson, and R.J. Spelman. 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 11-16 in Proc. 36th ICAR Biennial Session, Niagara Falls, USA.

Hayes, B.J., A.J. Chamberlain, H. McPartlan, I. Macleod, L. Sethuraman, and M.E. Goddard. 2007.Accuracy of marker-assisted selection with single markers and marker haplotypes in cattle. Genet. Res. Camb. 89:215-220.

Kolbedhari, D., L.R. Schaeffer, and J.A.B. Robinson. 2007. Estimation of genome-wide haplotype effects in half-sibs design. J. Anim. Breed. Genet. 124:356-361.

Legarra, A., and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection. J. Dairy. Sci. 91:360-366.

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

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.

CHAPTER IV

Using eigenvalues as variance priors in the prediction of Genomic breeding values by principal component analysis

N. P. P. Macciotta,[*] G. Gaspa,[*] R. Steri,[*] E. L. Nicolazzi,[§] C. Dimauro,[*] C. Pieramati[†] and A. Cappio-Borlino[*]

[*]*Dipartimento di Scienze Zootecniche, Università di Sassari, Sassari, Italy 07100*
[§]*Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza Italy 29100*
[†]*Centro di Studio del Cavallo Sportivo, Università di Perugia, Perugia, Italy 06100*

Published in: Journal of Dairy Science 2010. Vol. 93: 2765-2778. doi:10.3168/jds.2009-3029

ABSTRACT

Genome-wide selection aims to predict genetic merit of individuals by estimating the effect of chromosome segments on phenotypes using dense single nucleotide polymorphism (SNP) marker maps. In the present paper, principal component analysis was used to reduce the number of predictors in the estimation of genomic breeding values for a simulated population. Principal component extraction was carried out either using all markers available or separately for each chromosome. Priors of predictor variance were based on their contribution to the total SNP correlation structure. The principal component approach yielded the same accuracy of predicted genomic breeding values obtained with the regression using SNP genotypes directly, with a reductionin the number of predictors of about 96% and computation time by 99%. Although these accuracies are lower than those currently achieved with Bayesian methods, at least for simulated data, theimproved calculation speed together with the possibility of extracting principal components directly on individual chromosomes may represent an interesting option for predicting genomic breeding values in real data with a large number of SNP. The use of phenotypes as dependent variable instead of conventional breeding values resulted in more reliable estimates, thus supporting the current strategies adopted in research programs of genomic selection in livestock.

Key words: single nucleotide polymorphism, genomic selection, principal component analysis, eigenvalue.

INTRODUCTION

Marker assisted selection programs had limited commercial applications until the early 2000s because of the fact that most of reported marker-QTL associations had been found within families but were in linkage equilibrium across the population (Hayes and Goddard, 2001; Dekkers, 2004; Khatkar et al., 2004). The availability of genome-wide dense marker maps for several animal species has recently allowed the prediction of genomic breeding values (GEBV) by estimating marker haplotype effects on phenotypes (Meuwissen et al., 2001; Goddard and Hayes, 2007). Genome-wide selection relies on highly dense markers whose effects on phenotypes are estimated on a training population and then used to calculate GEBV for both training individuals and animals with only marker genotypes available (for example, young animals without phenotypes or EBV). A reduction in generation interval, an increase of accuracy in the cow side of the pedigree, and a decrease of selection costs are the expected advantages of an efficient genome-wide selection over traditional selection (Schaeffer, 2006; Konig et al., 2009).

High density SNP maps fulfill the basic requirement of genome-wide selection (i.e., the analysis of genome bits having large and persisting population-wide linkage disequilibrium; Muir, 2007). However, the use of dense marker platforms results in a large number of effects to be estimated (many thousands) in comparison with the relatively small number of phenotypes available (often just a few thousand). Such data asymmetry raises several statistical issues, such as collinearity among predictors and multiple testing (Gianola and van Kaam, 2008). To cope with such a problem, several methods of reduction of the number of predictors without a large decrease in accuracy have been proposed.

Selection of relevant SNP by single marker regression on phenotypes may improve results in genome-wide association studies (Aulchenko et al., 2007; Long et al.,2007), but it leads to a decrease of GEBV accuracy (Meuwissen et al., 2001). Bayesian methods that select SNP by evaluating their individual contribution to the variance of the trait, such Bayes B method (Meuwissen et al., 2001; Fernando et al., 2007; VanRaden, 2008), usually give best GEBV accuracies when simulated data with few QTL are modeled. However, results on actual data indicate that BLUP estimation, which assumes an equal contribution of all marker intervals to the genetic variance, performs only slightly worse than Bayesian methods in GEBV prediction (Hayes et al.,2009; VanRaden et al., 2009). Moreover in all the abovementioned techniques, markers are selected according to their relevance on the variability of the phenotype analyzed. Consequently, specific sets of markers may be required for different traits (Habier et al., 2009).

Multivariate dimension reduction techniques may offer an alternative approach based on the evaluation of the contribution of each marker locus to the total SNP (co)variance structure. Principal component analysis (PCA) has been used for analyzing complex genetic patterns in human genetics (Cavalli Sforza and Feldman, 2003; Paschou et al., 2007) and for selecting markers in genome-wide association studies. Solberg et al. (2009) used PCA and partial least squares regression to reduce the dimensionality of predictors in genomic selection. Both PCA and partial least squares regression showed comparable accuracies with Bayes B when lower marker densities were fitted, whereas the gap between methods increased with the number of markers used. Solberg et al. (2009) concluded that reduction in computational complexity provided by multivariate methods did not counterbalance their lower accuracy compared with Bayes B. Such considerations are justified by the low cost of

calculation time and by the computational speed that can be provided by optimized techniques such as parallel computing. On the other hand, it is reasonable to expect that denser SNP platforms will be available very soon for livestock species and dimensionality will again represent a relevant problem.

In their proposal, Solberg et al. (2009) regressed phenotypes on principal component (PC) scores extracted from the SNP matrix using the single value decomposition approach with an assumption of equal variance of each PC score. The choice of priors of marker effects represents a crucial point for genomic models (de los Campos et al., 2009). On the other hand, the ordinary method for calculating PC relies on the eigenvalues of the correlation matrix of starting variables that measure the contribution of each PC to the original variance of predictors. Thus, eigenvalues can be used as priors of predictor effect for the calculation of GEBV. It is worth remembering that eigenvalues have already been incorporated in mixed model algorithms to optimize calculations for variance component estimation (Dempster et al., 1984; Taylor et al., 1985).

In the present paper, PCA is used to perform a BLUP prediction of GEBV in a simulated data set to test the ability of this technique to reduce the number of predictors without decreasing GEBV accuracy. Moreover, the feasibility of extracting PC from dense, commercially available SNP platforms is tested.

MATERIALS AND METHODS

Data

The data set was generated for the XII QTLs–MAS workshop (http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html).The base population consisted of 100 individuals (50 males, 50 females). The genome had 6

chromosomes (total length 6 M), with 6,000 biallelic SNP, equally spaced at a distance of 0.1 cM. A total of 48 biallelic QTL were generated, with positions sampled from the genetic map of the mouse genome. Quantitative trait loci effects were sampled from a gamma distribution with parameters estimated by Hayes and Goddard (2001). Initial allelic frequencies of both SNP and QTL were set to 0.5. Then 50 generations of random mating followed. Generations 51 to 57 were used to create the experimental population of 5,865 individuals. Generations 51 to 54 (4,665 individuals; **TRAIN** data set) had pedigree, phenotype, and marker information available. The last 3 generations (1,200 individuals; **PRED** data set) had only pedigree and marker information available. True breeding values (**TBV**) were considered as the sum of all QTL effects across the entire genome. Phenotypes were generated by adding environmental noise to the TBV. Further details on the simulation can be found in Lund et al. (2009). Polygenic breeding values, being among the most frequently used dependent variable in GEBV prediction with real data, were also predicted.

Polygenic breeding values and additive genetic (σ^2a) and residual (σ^2e)variance components were estimated with a single trait animal model that included the fixed effects of sex and generation and the random additive genetic effect of the animal. The pedigree relationship matrix included 5,939 animals.

PCA analysis

Principal component analysis aims to synthesize information contained in a set of n observed variables (M_1,…, M_n) by seeking a new set of k (k < n) orthogonal variables (PC_1,…, PC_k) named PC, which are calculated from the eigen decomposition of the covariance (or correlation) matrix **M**. The jth PC is a linear combination of the observed variables

$$PCj = \alpha_{1j}M_1 + \ldots + \alpha_{nj}M_n$$

where coefficients α_{ij} are the elements of the eigenvector corresponding to j^{th} eigenvalue. Principal components are usually extracted in a descending order of the corresponding eigenvalue that measures the quota of variance of original variables explained by each PC (Morrison, 1976; Krzanowsky, 2003).

An SNP data matrix **M** with m rows (m = 5,865, the number of individuals in the entire data set) and n columns (n = 5,925, the number of SNP markers that were found to be polymorphic) was created. Each element (i,j) corresponded to the genotype at the the j^{th}marker for the i^{th} individual. Genotypes were coded as −1, 0, or 1 according to the notation used by Solberg et al. (2009).

Data editing is usually recommended when handling dense marker maps (Wiggans et al., 2009), either to correct for data quality (i.e., genotyping not successfully performed) or to avoid possible estimation biases because of a severe imbalance of genotypes. However, considering that in the present simulated data only 288 markers had minor allele frequency <0.05, whereas 47 deviated significantly ($P < 0.01$) from the Hardy-Weinberg equilibrium, this deviation may be attributable to drift; only the 75 monomorphic SNP were discarded from the analysis. Such a choice is at least partially supported by results of Chan et al. (2009), who pointed out that SNP attributes commonly considered in SNP data editing, such as minor allele frequency or deviation from Hardy-Weinberg equilibrium, have actually a very small effect on overall false positive rate in genomewide association studies.

Principal component analysis was carried out on **M**, and the number of PC (k) retained for further analysis was based on both the sum of their eigenvalues and the obtained GEBV accuracy. Principal component extraction was performed either on all SNP simultaneously (**PC_SNP_ALL**) or separately for each chromosome (**PC_SNP_CHROM**). Scores of the k selected PC were calculated for all individuals.

Marker haplotypes may be more efficient than genotypes in capturing marker-QTL association, especially in outbred populations where it may differ between families (Calus et al., 2008). Thus, PCA was performed also on haplotypes constructed from pairs of adjacent marker loci, using either

all loci together (**PC_HAP_ALL**) or separately per chromosome (**PC_HAP_CHROM**).

Predictor effect estimation and GEBV calculations

Dependent variables used in the analysis were either phenotypes or polygenic breeding values. For the estimation of the effects of predictors, records of the 4,665 individuals of the TRAIN data set were analyzed with the following mixed linear model:

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

where **y** is the vector of either phenotypes or polygenicbreeding values, **X** is the design matrix of fixed effects(mean, sex = 1, 2, generation = 1, 2, 3, 4 for phenotypes;only mean for polygenic breeding values); **b** isthe vector of solutions for fixed effects; **Z** is the (m ×k) design matrix of random effects, where each elementcorresponds to the score of the k^{th} component for themth animal of the training generations; **g** is the vector ofsolution for random regression coefficients of PC scores;and **e** is the random residual. Covariance matrices ofrandom PC effects (**G**) and residuals (**R**) were modelled as diagonal $\mathbf{I}(\sigma^2_{ai})$ and $\mathbf{I}(\sigma^2_{e})$, respectively. The BLUPmethods used for estimating SNP effects usually assumean equal contribution of each SNP locus to the varianceof the trait, sampled from the same normal distribution(i.e., $\sigma^2_{aj} = \sigma^2_a/n$; Meuwissen et al., 2001; VanRaden etal., 2009). In the present work, 2 different options werecompared. The first is the above-mentioned equalityof variances. The second starts from the considerationthat PC scores were used as predictor variables andtheir contribution to the original SNP covariance

structureis quantified by the corresponding eigenvalue (λ).Thus, variances of PC effects were calculated as $\sigma^2_{aj}=(\sigma^2_a/k) \times \lambda_j$.

The **G** matrix diagonality, commonly implementedin BLUP methodologies for estimating SNP marker effects(Meuwissen et al., 2001; VanRaden, 2008), relieson the assumption that marker effects in a large populationare uncorrelated (VanRaden et al., 2009). Withthe use of PC scores, such an assumption is consistentwith the orthogonality between PC (Morrison, 1976).The BLUP solutions were estimated using Henderson'snormal equations (Henderson, 1985).

To have a comparison with the most straightforwardestimation method, SNP effects were estimated directlyby using the same mixed linear model but with **Z** indicatingthe design matrix of the 5,925 polymorphic SNPgenotypes [coded as 0, 1, and 2 (i.e., on the basis of thenumber of alleles)]. Covariance matrix **G** was assumedto be diagonal as $\mathbf{I}(\sigma^2_a/n)$. A Cholesky decompositionwas used to solve mixed model equations (Harville,1997).

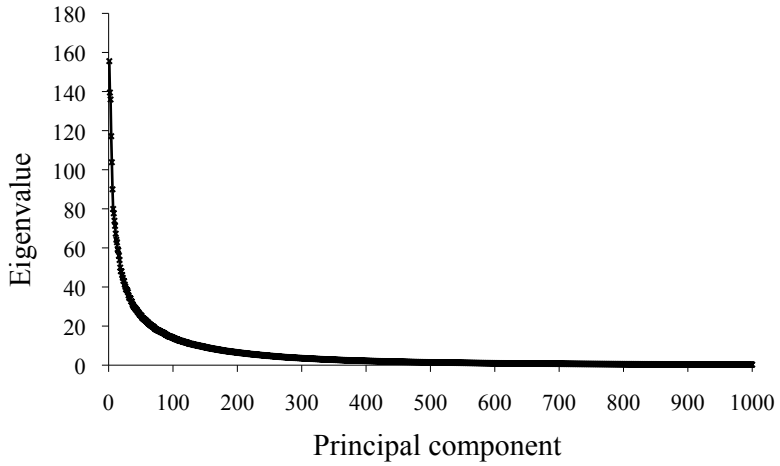
Overall mean and effects of PC scores or SNP genotypes($\hat{\mathbf{g}}$) estimated on the TRAIN data set were thenused to predict GEBV both in TRAIN and PREDindividuals as

$$\mathbf{GEBV} = \mu + \mathbf{Z}\hat{\mathbf{g}}$$

,where **GEBV** is the vector of predicted GEBV and **Z** is the matrix of the PC scoresor SNP genotypes of all individuals**.**

Accuracies of prediction where evaluated by calculating Pearson correlations between GEBV and TBV for the PRED generations. Bias of prediction was assessed by examining the regression coefficient of TBV on GEBV (Meuwissen et al., 2001).

Figure 1. Pattern of the eigenvalues of the correlation matrix of SNP markers.



Goodness of prediction wasevaluated also by the mean squared error of prediction (**MSEP**), calculated as

$$MSEP = \sum_{i=1}^{n} \frac{\left[TBV_i - GEBV_i\right]^2}{n}$$

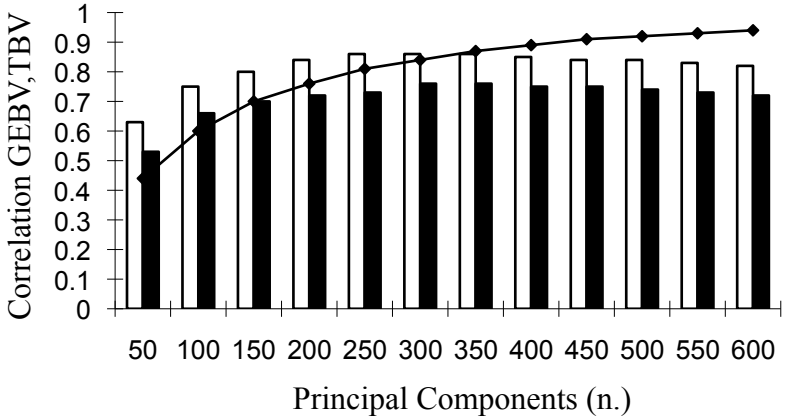
where n is the number of individuals in the PRED generations, and by its partition in different sources of variation related to systematic and random errors of prediction (Tedeschi, 2006).

RESULTS

The pattern of eigenvalues of the correlation matrix of SNP genotypes obtained with PCA of all markers simultaneously is reported in Figure 1 (only the first 1,000 eigenvalues are plotted for brevity). A smooth decrease in the amount of variance explained by each successive PC can be observed, with a plateau between 250 and 300 PC (about 84% of variance explained). Thus, between 200 and 300 PC could be considered adequate for describing the original variance of the system. The GEBV accuracies for different numbers of retained PC (50–600) using all SNP simultaneously and eigenvalues as variance priors are reported in Figure 2. Accuracy

for both training and prediction generations increases until a plateau, reached at about 250 to 300 PC. Further increasing the number of retained PC does not result in an increase of accuracy, probably because of the small amount of variance explained by each additional variable. Similar results were obtained by Solberg et al. (2009), who report best accuracies when 350 PC were extracted from 8,080 biallelic markers distributed on 10 chromosomes. However, Solberg et al. (2009) found a rather decreasing trend of the correlation between GEBV and TBV for larger numbers of PC. Based on the accuracy of GEBV prediction, 279 PC (83% of the original variance) were retained in the present work for PC_SNP_ALL and PC_HAP_ALL approaches. In the analysis carried out on individual chromosomes, to keep the same number of predictors of the previous approach, 46 and 47 PC for chromosomes 1 to 3 and 4 to 6, respectively, were retained.

Figure 2. Pattern of correlations between genomic breeding values (GEBV) and true breeding values (TBV) when principal components are extracted from all SNP genotypes simultaneously and eigenvalues are used as priors, for different number of retained PC (white bars = training individuals, black bars = prediction individuals).



The continuous line represents the amount of variance explained by the corresponding number of PC.

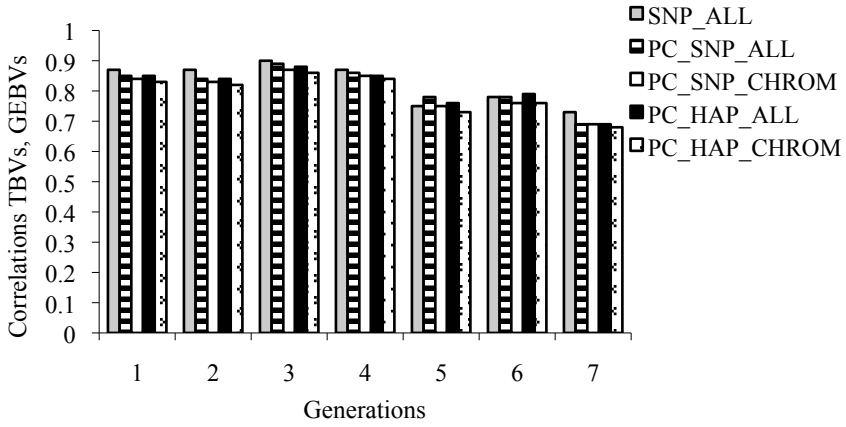
Average GEBV accuracies obtained using phenotypes are, for the 3 prediction generations, around 0.70 (Table 1) when an equal contribution of PC score on the variance of the trait is assumed, similar to those reported by Solberg et al. (2009).

Table 1. Pearson correlations between predicted genomic breeding values and true breeding values, for different estimation methods, using either phenotypes or polygenic breeding values for the prediction generations and assuming either equal variance contribution for each principal component or eigenvalues as variance priors.

Method[1]	Phenotype	Polygenic breeding values
SNP_ALL	0.76	0.41
Equal variance		
PC_SNP_ALL	0.69	0.53
PC_SNP_CHROM	0.70	0.55
PC_HAP_ALL	0.68	0.54
PC_HAP_CHROM	0.71	0.56
Eigenvalues		
PC_SNP_ALL	0.76	0.57
PC_SNP_CHROM	0.73	0.56
PC_HAP_ALL	0.75	0.56
PC_HAP_CHROM	0.73	0.55

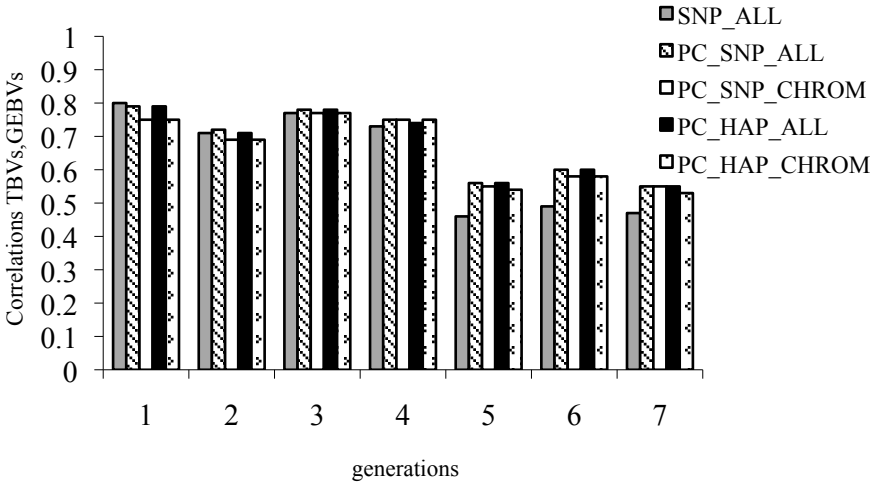
[1] SNP_ALL = all 5,925 SNPs; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PC_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PC_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PC_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome.

Figure 3. Correlations between genomic breeding values (GEBV) and true breeding values (TBV) in the different approaches when phenotypes were used as dependent variables.



SNP_ALL = all 5,925 SNP; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PCA_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PCA_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PCA_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome

Figure 4. Correlations between genomic breeding values (GEBV) and true breeding values (TBV) in the different approaches when polygenic breeding values were used as dependent variables.



SNP_ALL = all 5,925 SNP; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PCA_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PCA_HAP_ALL = principal components extracted from all SNPS haplotypes simultaneously; PCA_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome

Accuracies increase by about 10% (to an average of 0.75) when eigenvalues are used in the diagonal of the $\mathbf{G}^{-1}$ matrix of mixed model equations. In general, results are of the same order as in previous literature reports for BLUP estimation on simulated (Meuwissen et al., 2001, 2009; Fernando et al., 2007) and real (Hayes et al., 2009; VanRaden et al., 2009) data. Correlations obtained when all SNP were used as predictors are equal to those obtained with PC with eigenvalues as priors. On the other hand, a remarkable difference in calculation speed between the 2 methods has been observed: about 6 h for the SNP_ALL approach and 3 min for the PC, using a computer with a dual core processor (2.33 GHz and 3.26 MB of random access memory). Slight differences can be observed between estimates of PC carried on all chromosomes or separately for each of them. Moreover, the same results have been basically obtained when genotypes at single markers or haplotypes were used, in

agreement with previous reports for high density markers (Hayes et al., 2007; Calus et al., 2008).

The GEBV accuracies are larger when phenotypes instead of polygenic breeding values are used as dependent variables (Table 1). This is particularly evident when all SNP are used as predictors (on average 0.73 vs. 0.55 for phenotypes and polygenic breeding values, respectively). Also, the decrease in accuracy between TRAINING and PRED generations is more evident for polygenic breeding value-based predictions (Figures 3 and 4). These findings are confirmed by values of regression coefficients of TBV on GEBV (Table 2). Moreover, b-values for methods based on PC are similar to those reported by Solberg et al. (2009) when equal variances were assumed, whereas they are closer to 1 (about 0.85) when eigenvalues are used as variance priors.

Table 2. Regression coefficients ($b_{TBV,GEBV}$) of true breeding value (TBV) on predicted genomic breeding value (GEBV) for the different estimation methods using either phenotypes or polygenic breeding values for the prediction generations and assuming either equal variance contribution for each principal component or eigenvalues as variance priors.

	Phenotype		Polygenic breeding value	
Method[1]	$b_{TBV,GEBV}$	SE	$b_{TBV,GEBV}$	SE
SNP_ALL	1.08	0.027	1.15	0.073
Equal variance				
PC_SNP_ALL	0.63	0.019	1.08	0.049
PC_SNP_CHROM	0.67	0.019	1.13	0.048
PC_HAP_ALL	0.61	0.019	1.08	0.049
PC_HAP_CHROM	0.65	0.018	1.11	0.047
Eigenvalues				
PC_SNP_ALL	0.88	0.021	1.33	0.055
PC_SNP_CHROM	0.84	0.022	1.28	0.055
PC_HAP_ALL	0.88	0.022	1.32	0.056
PC_HAP_CHROM	0.83	0.023	1.26	0.056

[1]SNP_ALL = all 5,925 SNPs; PC_SNP_ALL = principal components extracted from all SNP genotypes simultaneously; PC_SNP_CHROM = principal components extracted from SNP genotypes separately for each chromosome; PC_HAP_ALL = principal components extracted from all SNP haplotypes simultaneously; PC_HAP_CHROM = principal components extracted from haplotypes separately for each chromosome
The decomposition of the MSEP for some of the considered scenarios is reported in Table 3.

Table 3. Mean squared error of prediction (MSEP) decomposition (%) and coefficient of determination (r^2) for the prediction generations in some scenarios using either phenotypes or polygenic breeding values[1,2].

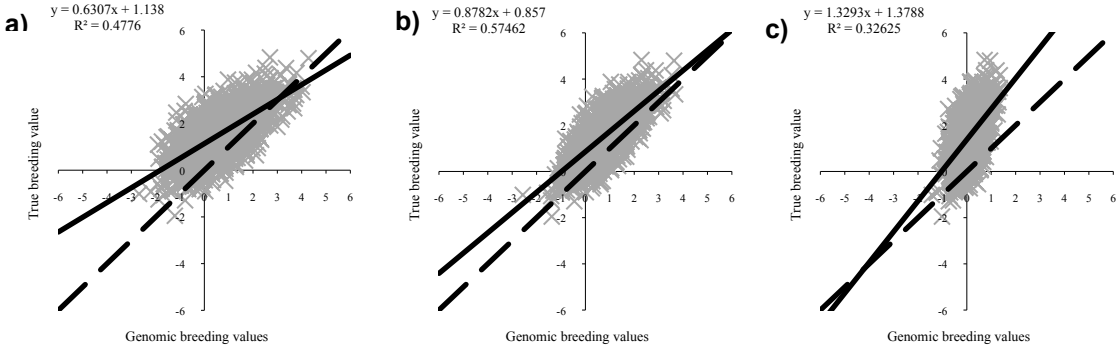
Item	SNP_ALL	PC_SNP_ALL1	PC_SNP_ALL2
Phenotype			
MSEP	1.55	1.48	1.02
Mean Bias (U_M)	72.2	53.5	56.9
Unequal variances (U_S)	6.9	0.6	1.9
Incomplete covariation (U_C)	21.9	45.9	41.2
Slope bias (U_R)	0.22	11.1	1.1
Random errors (U_D)	27.6	35.4	42.0
r^2	0.57	0.48	0.57
Polygenic breeding value			
MSEP	2.96	2.88	2.72
Mean Bias (U_M)	72.0	75.1	74.6
Unequal variances (U_S)	13.9	8.9	11.9
Incomplete covariation (U_C)	14.1	16.0	13.5
Slope bias (U_R)	0.01	0.00	0.7
Random errors (U_D)	27.9	24.9	24.7
r^2	0.17	0.28	0.33

[1]SNP_ALL= all 5,925 SNPs; PC_SNP_ALL 1= principal components extracted from all SNP genotypes simultaneously and equal contribution of each SNP to the variance of the trait; PC_SNP_ALL 2 principal components extracted from all SNP genotypes simultaneously and contribution of each SNP to the variance of the trait proportional to the eigenvalue.

[2] $U_M + U_S + U_C = U_M + U_R + U_D = 100\%$

The MSEP is always smaller (about half) when GEBV are calculated using phenotypes. Its partition highlights a great relevance of components related to the bias of prediction (i.e., mean bias, inequality of variances) in the approach that directly fits SNP genotypes (about 79%). Methods based on PC extraction are characterized by a prevalence (about 80%) of random terms, measured by the random error and by the incomplete covariation. The use of eigenvalues as variance priors results in the lowest MSEP and, compared with the other PC-based method, in a reduction of the slope bias and the highest relevance of random variation. These differences be clearly seen from the plots of TBV versus GEBV for the PC_SNP_ALL approach using equal (Figure 5a) or eigenvalue-based (Figure 5b) variance.

Figure 5. Regression plots for the scenarios analyzed.



a) Plot of true breeding values versus genomic breeding values predicted using phenotypes when principal components (PC) are extracted from all SNP genotypes simultaneously and variance contribution of the PC scores in the estimation step is assumed equal. b) Plot of true breeding values versus genomic breeding values predicted using phenotypes when principal components are extracted from all SNP genotypes simultaneously and variance contribution of the PC scores in the estimation step is based on their eigenvalues. c) Plot of true breeding values versus genomic breeding values predicted using phenotypes when all SNP genotypes are used as predictors (Continuous line= regression line of true breeding values on genomic breeding values; dotted line= equivalence line, y=x).

The latter shows a regression slope closer to the equivalence line (y = x) and a smaller value for the intercept that indicates a smaller systematic underestimation of TBV. The composition of MSEP becomes very similar across the different methods when polygenic breeding values are used as dependent variables, with a reduced incidence of random components and a larger relevance of unequal variances compared with the phenotype-based estimates (Table 3). Actually, the comparison of plots of TBV versus GEBV estimated with the PC_SNP_ALL approach using phenotypes (Figure 5a) or polygenic breeding values (Figure 5c) clearly shows a reduced range of variability and a higher underestimation (as evidenced by the larger value of the regression intercept) for polygenic breeding value-based GEBV.

An interesting feature of PCA is the possible technical interpretation of extracted variables. Figure 6 reports score averages for the first 2 PC that together explain about 5% of the original variance of the system, calculated for each generation. Averages of the second PC ranged gradually from negative values for the first 3 generations to

positive for the last 3 generations. A possible explanation of the ability of the second PC to distinguish individuals of different generations can be found in its negative correlation with the average observed heterozygosity per animal (−0.26) that tends to decrease from older to younger generations (Figure 7).

Figure 6. Plot of the average scores of the first two principal components for seven generations.

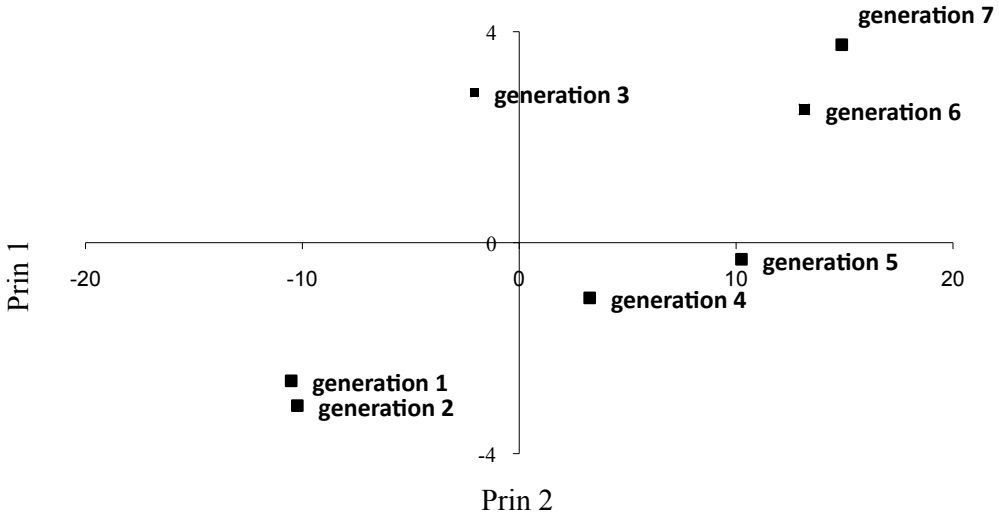
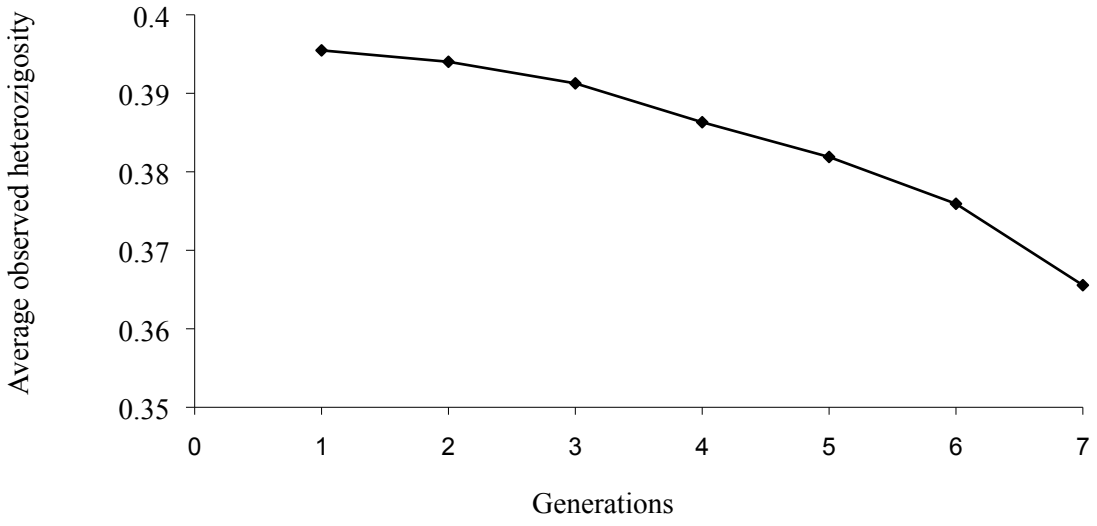


Figure 7. Pattern of the average observed heterozygosity in different generations.



DISCUSSION

The main objectives of this work were to assess the effect of reducing predictor dimensionality in GEBV estimation using PCA and to test the effect of structuring the variance contribution of PC with their eigenvalues.

Principal component analysis allows an efficient description of the correlation matrix of biallelic SNP with a markedly smaller number of new variables (4.7%) compared with the original dimension of the system. Such a huge decrease has a straightforward effect on the calculation speed of GEBV, with a reduction of more than 99% of computing time achieving the same accuracy of predicted GEBV using all SNP. Compared with other methods of reduction of predictors where SNP are selected based on their position along the chromosome (VanRaden et al., 2009) or their relevance with the trait considered (Hayes et al., 2009), the multivariate reduction approach limits the loss of information because each SNP is involved in the composition of each PC.

The GEBV accuracies obtained in the present work agree with a previous report on the use of PCA to estimate GEBV (Solberg et al., 2009) when an equal contribution of each PC to the variance of phenotypes is assumed. This approach follows the common BLUP assumption of equality of variance of predictors, usually criticized for its inadequacy to fit the widely assessed distribution of QTL (i.e., many loci with a small effect and very few with large effect; Hayes and Goddard, 2001). However, when eigenvalues are used as prior of PC variance, accuracies increase by about 10%. These figures highlight the importance of an accurate modeling of the variance structure of random effects in GEBV estimation. Bayesian methods estimate variances of different chromosome segments combining information from prior distribution and data (Meuwissen et al., 2001). These methods usually give the best performance

(accuracies >80%) when simulated data are fitted, whereas results obtained on real data seem to indicate a substantial equivalence with the BLUP approach (Hayes et al., 2009; VanRaden et al., 2009). A common explanation is that, in Bayes method, assumptions on prior distributions of parameters are more difficult to infer when real data are handled. The use of eigenvalues as variance priors relies only on data (i.e., the SNP correlation structure) and does not require assumptions on prior distribution.

A potential drawback in the calculation of GEBV using PCA is represented by PC extraction. In the present work, about 40 min were needed to process an SNP data matrix of 5,865 rows and 5,925 columns. The commercially available SNP panel for cattle has 54,000 marker loci, although about 40,000 are retained on average after editing (Hayes et al., 2009). Such a marked increase of columns, usually not accompanied by a comparable increase of rows (i.e., phenotypic records), may lead to statistical and computational problems if PC are extracted treating all SNP simultaneously. However, results of the present study indicate that PC may be calculated separately for each chromosome, keeping the same GEBV accuracy. It should be remembered that the number of SNP per chromosome is not far from current dairy data (on average 1,200–1,300; Hayes et al., 2009; VanRaden et al., 2009; Wiggans et al., 2009). Thus, PCA carried out on individual chromosomes may be of great interest for real data, also considering the substantial biological orthogonality among chromosomes. The availability of denser marker maps (i.e., 500,000 SNP) will represent a challenge for the method, although the number of PC to be retained does not seem to increase linearly with the number of original variables. Missing genotypes is a potential problem for computation of PCA, which requires data in each cell. Although edits that are normally carried out on SNP data leave only a few missing cells per animal, they are spread across different markers and this may lead to

a severe reduction in the number of records. Missing data can be reconstructed using appropriate algorithms as those described by Gengler et al. (2007) or others implemented in software of common use such as fastPHASE (Scheet and Stephens, 2006) or PLINK (Purcell et al., 2006).

Of particular interest is the difference in GEBV accuracy obtained when using phenotypes versus polygenic breeding values as dependent variable. Polygenic breeding values are phenotypes corrected for additive relationships among animals based on pedigree information. On the other hand, in GEBV predictions the genetic similarity between animals is accounted for by the specific combination of marker genotypes possessed by each individual. Therefore, the use of polygenic breeding values as dependent variable in GEBV prediction may be regarded as redundant in terms of exploitation of genetic relationships. This behavior is particularly evident for the regression using all SNP markers. In this form, the calculation of GEBV is equivalent to the use of an animal model with the additive genetic effect structured by the genomic relationship matrix (Goddard, 2009). Such a double counting of genetic relationship resulted in an evident reduction of the variability of GEBV compared with TBV. From a statistical standpoint, polygenic breeding values are model-predicted values and may not be suitable as a dependent variable in further analyses (Tedeschi, 2006). Results of the present study, although obtained on simulated data, may more accurately reflect the reality of genomic selection programs in cattle. In previous studies, polygenic breeding values were generally the dependent variable. This is because TBV are not available on real data and polygenic breeding values estimated with a high accuracy (>0.90) may represent a sort of golden standard for cross validations. However, the tendency now seems to move toward the use of partially

corrected phenotypes such as deregressed proofs or daughter yield deviations (Hayes et al., 2009; VanRaden et al., 2009).

Finally, an interesting side product of PCA used to reduce the dimensionality of predictors in genome-wide selection is represented by the extraction of synthetic variables that can have a technical meaning. Studies in human and animal genetics have highlighted the role of PC as indicators of population genetic structure. For example, the top eigenvectors of the covariance matrix often show a geographic interpretation (Price et al., 2006; Chessa et al., 2009). Usually, the meaning of the i^{th} PC in terms of relationship with the original variables is inferred from the structure of its eigenvector. In the present study, such an evaluation was not feasible, probably because of both the relatively small amount of variance explained by each PC and the large number of original variables considered (i.e., the 5,925 SNP). However, one of the top PC was able to reflect the genetic variation among generations, although the discrimination between individuals of different generations was rather fuzzy, as expected, given the small amount of variance explained. However, this last point deserves some additional consideration. An assessed criterion in choosing which PC to retain is to look at their eigenvalues. However, sometimes the PC associated with the largest eigenvalue does not have a defined meaning, whereas successive PC characterized by smaller eigenvalues may contain more relevant or biological information (Jombart et al., 2009). In the case of the present work, a meaning of the second PC as indicator of genetic drift, which should be the only reason of variation of genotypic frequencies in the simulated generations (Lund et al., 2009), could be hypothesized.

ACKNOWLEDGMENTS

Research was funded by the Italian Ministry of Agriculture (Rome, Italy), grant SELMOL. The authors thank the organizers of the XII QTL-MAS workshop for providing simulated data. Discussion with P. Ajmone-Marsan (Universita di Piacenza, Italy) is gratefully acknowledged. Helpful comments and suggestions from the 2 reviewers are also acknowledged.

REFERENCES

Aulchenko, Y.S., D.J. de Koning, and C. Haley. 2007. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide predigree-based quantitative trait loci association analysis. Genetics 177:577-585.

Calus, M., 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.

Cavalli-Sforza, L., and M.W. Feldman. 2003. The application of molecular genetic approaches to the study of human evolution. Nat. Genet. 33: 266-275.

Chan E.C.F., R. Hawken, and A. Reverter. 2008. The combined effect of SNP-marker and phenotype attributes in genome-wide association studies. Anim. Genet. 40: 149-156.

Chessa, B., F. Pereira, F. Arnaud, A. Amorim, F. Goyache, I. Mainland, R.R. Kao, J.M. Pemberton, D. Beraldi, M.J. Stear, A. Alberti, M. Pittau, L. Iannuzzi, M.H. Banabazzi, R.R. Kazwala, Y.P. Zhang, J.J. Arranz, B.A. Ali, Z. Wang, M. Uzun, M.M. Dione, I. Olsaker, L.E. Holm, U. Saarma, S. Ahmad, N. Marzanov, E. Eythorsdottir, M.J. Holland, P. Ajmone-Marsan, M.W. Bruford, J. Kantanen, T.E. Spencer, and M.

Palmarini. 2009. Revealing the history of sheep domestication using retrovirus. Science 324: 532.

Dekkers, J.C.M. 2004. Commercial application of marker- and gene-assisted selection in livestock: strategies and lessons. J. Anim. Sci. 82(E. Suppl.):E313-E328.

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

Dempster, A.P., C.M. Patel, M.R. Selwyn, and A.J. Roth. 1984. Statistical and computation aspects of mixed model analysis. Appl. Stat. 33:203-214.

Fernando, R.L., D. Habier, C. Stricker, J.C.M. Dekkers, and L.R. Totier. 2007. Genomic selection. Acta Agr. Scand. A-AN 57: 192-195.

Gengler, N., P. Mayeres, and M. Szydlowski. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. Animal 1: 21-28.

Gianola, D., and van Kaam J.B.C.H.M. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. Genetics 178: 2289-2303.

Goddard, M.E., and B.J. Hayes. 2007. Genomic selection. J. Anim. Breed. Genet. 124: 323-330.

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

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

Harville, D.A. 1997. Matrix algebra from a statistician's perspective. Springer-Verlag, New York, NY.

Hayes, B.J., and M.E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33: 209-229.

Hayes, B.J., A. J. Chamberlain, H.M. McPartlan, I. Macleod, L. Sethuraman, and M.E. Goddard. 2007. Accuracy of marker assisted selection with single markers and markers haplotypes in cattle. Genet. Res. 89: 215-220.

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.

Henderson, C.R. 1985. Best Linear Unbiased Prediction using relationship matrices derived from selected base population. J. Dairy Sci. 68:443-448.

Jombart, T., D. Pontier, and A.B. Dufour. 2009. Genetic markers in the playground of multivariate analysis. Heredity 102: 330-341.

Khatkar, M.S., P.C. Thomson, I.Tammen, and H.W. Raadsma. 2004. Quantitative trait loci mapping in dairy cattle: review and meta-analysis. Genet. Sel. Evol. 36: 163-190.

Konig, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. J. Dairy Sci. 92: 382-391.

Krzanowsky, W.J. 2003. Principles of multivariate analysis. Oxford University Press Inc., New York, NY.

Long, N., D. Gianola, G.J.M. Rosa, K.A. Weigel, and S. Avendano. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. J. Anim. Breed. Genet. 124: 377-389.

Lund, M.S., G. Sahana, D.J. de Koning, G. Su, and Ö. Carlborg. 2009. Comparison of analyses of QTLMAS XII common dataset. I: genomic selection. BMC proc. 3(suppl. 1): S1.

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

Meuwissen, T.H.E. 2009. Accuracy of breeding values of 'unrelated' individuals predicted by dense SNP genotyping. Genet. Sel. Evol. 41: 35.

Morrison, F. 1976. Multivariate statistical methods. McGraw-Hill, New York.

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.

Paschou, P., E. Ziv, E.G. Burchard, S. Choudry, W. Rodriguez-Cintron, M.W. Mahoney, and P. Drineas. 2007. PCA-correlated SNPs for structure identification in worldwide human populations. PLos Genetics 3: 1672-1686.

Price, A.L., N.J. Patterson, R.M. Plenge, M.E. Weimblatt, N.A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genet. 38: 904-909.

Schaeffer, L.R. 2006. Strategy for applying genome-wide selection in dairy cattle. J. Anim. Breed. Genet. 123: 218-223.

Solberg, T.R., A.K. Sonesson, J. Woolliams, and T.H.E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. Genet. Sel. Evol. 41: 29.

Taylor, J.F., B. Bean, C.E. Marshall, and J.J. Sullivan. 1985. Genetic and environmental components of semen production traits of artificial insemination Holstein bulls. J. Dairy Sci.: 2703-2722.

Tedeschi, L.O. 2006. Assessment of adequacy of mathematical models. Agr. Syst. 89: 225–247.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91: 4414-4423.

VanRaden, P.M., C.P. Van Tassell, G.R. Wiggans, T.S. Sonstengard, R.D. Schnabel,J.F. Taylor, and F.S. Schenkel. 2009. Reliability of genomic predictions for north American Holstein bulls. J. Dairy Sci. 92: 4414-4423.

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

CHAPTER V

Effect of prior distributions on accuracy of genomic breeding values for yield dairy cattle traits

E.L. Nicolazzi[1§], A.J. Chamberlain[2], M.E. Goddard[2,3], B.J. Hayes[2]

[1] Istituto di Zootecnica, Facoltà di Agraria, Università Cattolica del Sacro Cuore, 29100, Piacenza, Italy.
[2] Biosciences Research Division, Department of Primary Industries, Victoria, 1 Park Drive, Bundoora, 3083, Australia.
[3] Department of Agriculture and Food Systems, University of Melbourne, Royal Parade, Parkville 3010, Australia.

Submitted to: Genetic Selection Evolution

ABSTRACT

Introduction

The ideal method to estimate genome-enhanced breeding values (DGV) would calculate the conditional mean of the breeding value given the genotype of individuals at each QTL. This conditional mean requires a prior distribution of QTL effects. However, both the QTL and their distribution of effects are unknown, with SNP markers used to track the QTL. In this study we compare accuracies of DGV obtained using three different prior distributions of SNP effects (normal, Student's t and double-exponential) in simulated data, in order to understand the extent of reduction in DGV accuracy when the prior distribution does not match the true distribution of QTL effects. We then apply the methods in a real dataset both to find the prior distribution that is most robust across traits and to make interpretations about the true distribution of QTL effects.

Methods

The simulated dataset was provided by the XII QTL-MAS workshop. Genotypes of 1149 progeny tested Australian Holstein-Friesian bulls were used to test all methods in a real scenario. The traits analyzed were protein yield and fat percentage. Methods using normal and Student's t prior distributions had fixed hyper-parameters, whereas hyper-parameters for double-exponential prior distribution were conditional to the data. Accuracies of DGV and prediction bias were controlled in both datasets.

Results

Using the Student's t distribution for the prior distribution of SNP effects (BayesA) gave the largest estimates of SNP effects in the presence of QTL with large effects in

both simulated and real data, and achieved the best accuracies of DGV in both datasets. The double-exponential distribution (Bayesian LASSO) resulted in higher shrinkage of SNP effect estimates, even when a large true effect was present. As a result, this method obtained lower accuracies than BayesA. The normal distribution (Bayes-BLUP) resulted in the greatest degree of shrinkage of estimated effects, and gave the lowest accuracies.

Conclusions

Hyper-parameters conditional to the data allow a trait-specific shrinkage of the estimates, avoiding the use of fixed parameters that might be sub-optimal for some traits. However, the amount of information of the data analyzed might still be inadequate to estimate these hyper-parameters accurately. A Student's t distribution with fixed hyper-parameters was the best approximation of the QTL distribution for the two dairy traits analyzed.

INTRODUCTION

The recent availability of dense genome-wide SNP panels has allowed implementation of genomic selection (GS) in a number of livestock breeding programs worldwide (e.g. Van Raden et al., 2009; Hayes et al., 2009; Harris et al., 2008; van der Werf, 2009, Legarra et al., 2008, Gonzalez-Recio et al., 2009). The best statistical method for the estimation of direct genomic values (DGV) for selection candidates is still under discussion. Goddard and Hayes (2007) argued that the ideal method to estimate breeding values from genomic data would calculate the conditional mean of the breeding value given the genotype of individuals at each QTL. This conditional mean requires a prior distribution of QTL effects. However, in

practice, both the QTL and their distribution of effects are unknown, with SNP markers being used to track the QTL through linkage disequilibrium. Meuwissen et al. (2001) evaluated Bayesian methods either with a normal (Bayes-BLUP), a Student's t (BayesA), or a 0-t mixture distribution of marker effects (BayesB) in simulated data. Their results indicated that BayesA and BayesB performed better than Bayes-BLUP. However, these results could reflect the simulated distribution of QTL effects they used, which was a small number of QTL of moderate to large effect. On the other hand, results reported from real data show in general small differences in the accuracy of DGV from non-linear Bayesian models and Bayes-BLUP models, depending on the trait analyzed (VanRaden et al., 2009; Hayes et al., 2009; Luan et al., 2009; Gredler et al., 2009; Habier et al., 2010).VanRaden et al. (2009) tested both a Bayes-BLUP and a non-linear model (similar to BayesA) over 27 traits in dairy cattle. They showed that predictions using non-linear model were more accurate for some traits highly affected by QTL of large effects (i.e. fat and protein percentage). However, considering the results obtained across all 27 traits, only 1% average difference was observed between Bayes-BLUP and non-linear models. Also in data from dairy cattle, Hayes et al.(2009) obtained higher accuracies with BayesA in Australian Selection Index (ASI), Australian Profit Ranking (APR), protein yield and protein percentage indexes. Accuracy differences between BayesA and Bayes-BLUP models ranged from a minimum of 2% in APR to a maximum of 7% in protein percentage. Opposite results were obtained for female fertility, where Bayes-BLUP obtained a 4% higher accuracy than BayesA. These results indicate that, when dealing with real data, more flexible (or different) assumptions are needed to account for the different QTL distributions (and heritability) of complex traits. One potential solution is a two or three level hierarchical model, called the Bayesian LASSO (Park

and Casella, 2008), a Bayesian counterpart of the “original” least absolute shrinkage and selection operator (LASSO; Tibshirani,1996). The Bayesian LASSO assumes a double exponential prior distribution for SNP effects. Park and Casella (2008) describe a computationally efficient approach to implement the Bayesian LASSO using a hierarchical approach, whereby SNP effects are sampled from a normal distribution with a SNP specific variance, and these SNP variances are in turn sampled from an exponential distribution. The exponential distribution has a regularization hyper-parameter λ, which is sampled conditional on the data. The regularization hyper-parameter plays a key role in the model, as the degree of shrinkage of the estimates will be determined by the information in the data itself. In simulated data, de los Campos et al. (2009) tested five sets of parameters and two different distributions (gamma and beta distributions) for the prior of the regularization hyper-parameter, observing only small differences in terms of SNP effect estimates. They also analyzed the effect of the inclusion of a polygenic term in the Bayesian LASSO model, on wheat and mice real datasets. Their results (using only beta prior distribution for the regularization parameter) indicated that the inclusion of a polygenic effect increased the predictive ability of the model.

In this paper we compare accuracies of DGV and estimates of SNP effects in both simulated and real scenarios using Bayes-BLUP, BayesA and two Bayesian LASSO with different hyper-parameter prior distributions, as discussed in de los Campos et al. (2009). Since we are comparing accuracies with three different prior distributions of SNP effects (i.e., Normal distribution in Bayes-BLUP, Student’s t in BayesA and double-exponential in Bayesian LASSO), an additional objective is to gain some insight into the distribution of QTL effects for the traits analyzed in the real dataset.

MATERIALS AND METHODS

Methods were tested on both simulated and real data. First, a simulated dataset was used to study the properties of the methods in a scenario with known QTL positions and effects. Then, genotypes, phenotypes and pedigree information of Australian Holstein-Friesian bulls were analyzed for two production traits to test the performance of all methods using real data.

Simulated dataset

The simulated dataset, provided by the XII QTL-MAS workshop (QTLMAS workshop common dataset web-page), comprised 5,865 individuals structured in 7 generations. A total of 46 QTL with additive effect and 2 QTL with epistatic effect were distributed along the simulated genome. QTL effects were drawn from a Gamma distribution using Hayes and Goddard (2001) parameters. Pedigree relationships and genotypes at 6,000 SNP evenly distributed across six chromosomes were available for all individuals, whereas phenotypic information was provided for the first 4 generations only. All 4665 individuals from the first 4 generations were considered as training animals and individuals from the last 3 generations as prediction young animals. Phenotypes were pre-corrected for fixed effects, as in the following model:

$$y_{ipq} = \mu + SEX_{ip} + GEN_{iq} + e_{ipq}$$

where y is the original phenotype for the i^{th}animal; SEX is the fixed effect of sex (p=1,2); GEN is the fixed effect of generation (j=0-6); and e is the random residual (term considered as the corrected phenotype for each animal). True breeding values (TBV), were available for all animals. QTL effects considered in this study where those reported by Crooks et al., 2010. Accuracies were obtained by calculating the

correlation between DGV and TBV (r(DGV,TBV)). Prediction bias was assessed by calculating the regression coefficient of TBV on DGV.

Real dataset

A total of 1250 progeny tested Australian Holstein-Friesian bulls born between 1950 and 2005 were genotyped with the Illumina Bovine SNP50^{TM} chip (54K). SNP were eliminated from the dataset if they had more than 10% of missing genotypes, less than 1% of MAF and extreme values for chi-squared test for Hardy-Weinberg equilibrium. Mendelian inheritance of SNP was investigated and bulls with genotype incompatible with pedigree were eliminated. A total of 1149 animals and 39.048 SNP were retained for the analysis. The 763 older bulls (born between 1950 and 2002) where considered as reference population and the rest as validation population. Dependent variables (e.g. phenotypes) were de-regressed proofs of Australian Breeding Values (ABV), with the de-regression removing contribution from relatives other than daughters (see Hayes et al.,2009 for details). Accuracies were obtained as correlation between DGV and ABV. Regression coefficients of ABV on DGV were calculated to investigate bias of predictions. Traits analyzed were protein yield and fat percentage.

Description of models.

The following model was fitted:

$$y = 1_n' \mu + Xg + Zu + e$$

where **y** is a vector of phenotypes in the reference population; $\mathbf{1_n}$ is a vector of ones; μ is a general mean; **g** is a vector of (random) SNP effects; **X** is the corresponding design matrix with elements of X_{ij}=0,1,2 for genotypes 11, 12 and 22, respectively

for the i^{th}animal and j^{th} SNP; **u** is a vector of polygenic breeding values assumed to be normally distributed, with $u_i \sim N(0,\mathbf{A}\sigma_a^2)$, where **A** is the average relationship matrix and σ_a^2 is the additive genetic variance; **Z** is the corresponding design matrix linking polygenic breeding values to the data; and **e** is a vector of random residuals, with $e_i \sim N(0,\sigma_e^2)$, where σ_e^2 is the residual variance. Direct genomic values were calculated as:

$$DGV_i = \mu + \sum_{j=1}^{z} X_{ij} g_j$$

The Bayes-BLUP method assumed a normal prior distribution of SNP effects (maintaining BLUP infinitesimal assumptions). The variance of this normal distribution was sampled in each iteration of the Gibbs Sampler (Verbyla et al., 2010). BayesA prior structure and (fixed) hyper-parameters followed Meuwissen et al. (2001), thus, degrees of freedom were set to 4.012.

The Bayesian LASSO (B-LASSOgamma) was defined as follows (after de los Campos et al., 2009):

$$likelihood: p(y \mid \mu,g,u_i,\sigma_e^2) = \prod_{i=1}^{n} N(y_i \mid \mu + x_i'g + u_i,\sigma_e^2)$$

$$\begin{aligned} prior: p(\mu,g,u,\sigma_e^2,\sigma_u^2,\tau^2,\lambda^2) &= p(\mu \mid \sigma_\mu^2)p(g \mid \sigma_e^2,\tau^2)p(u \mid \sigma_u^2)p(\sigma_e^2)p(\tau^2 \mid \lambda)p(\lambda^2) \\ &= N(\mu \mid 0,\sigma_\mu^2)\left\{\prod_{j=1}^{z} N(g_j \mid 0,\sigma_e^2\tau_j^2)\right\}N(u \mid 0,A\sigma_u^2) \\ &\times\ \chi^{-2}(\sigma_e^2 \mid dof_e,S_e)\chi^{-2}(\sigma_u^2 \mid dof_u,S_u)\left\{\prod_{j=1}^{z} Exp(\tau_j^2 \mid \lambda)\right\}G(\lambda^2 \mid \alpha_1\alpha_2) \end{aligned}$$

where $N(y_i \mid \mu + x_i'g + u_i,\sigma_e^2)$ is a normal distribution with mean $\mu + x_i'g + u_i$ and variance σ_e^2; $N(\mu \mid 0,\sigma_u^2)$, $N(g_j \mid 0,\sigma_e^2\tau_j^2)$ and $N(u \mid 0,A\sigma_u^2)$ are normal distributions for general mean, SNP effects and polygenic breeding values, with null mean and variances σ_u^2, $\sigma_e^2\tau_j^2$ and $A\sigma_u^2$, respectively; $\chi^{-2}(\sigma_e^2 \mid dof_e,S_e)$ and $\chi^{-2}(\sigma_u^2 \mid dof_u,S_u)$ are scaled inverted chi-

squared distributions with degrees of freedom *dof* and scale parameter *S*, for random residual and polygenic variances, respectively; $Exp(\tau_j^2|\lambda)$ is an exponential distribution for marker variances, controlled by a single parameter λ (the regularization parameter); and $G(\lambda^2 \mid \alpha_1\alpha_2)$ is a Gamma distribution with α_1 and α_2 as shape and rate parameters, respectively. The above conditional distributions have a closed form, thus, a Gibbs Sampler can be used to solve the equations. Shape and scale parameters for the regularization hyper-parameter λ were defined as $p(\lambda^2|P,S) \propto G(\lambda^2 \mid P=0.1, S=1x10^{-04})$.

A modification to B-LASSO*gamma* tested in de los Campos et al. (2009), was a more flexible Beta prior distribution for the hyper-parameter λ. We tested this method as well (B-LASSO*beta)*. This distribution allows setting a relative flat prior in a wider range of values than Gamma distribution. In this case, parameters used for λ hyper-prior distribution were $p(\lambda|max,\alpha_1,\alpha_2) \propto Beta((\lambda/max=400) \mid \alpha_1=1.4, \alpha_2=1.4)$. A Metropolis-Hastings step was required because the Beta distribution is not a conjugate prior. Further details on both Bayesian LASSO methods applied in this study are available in de los Campos et al. (2009).

A total of 20,000 runs of iteration were performed for each method under study. The first 10,000 iterations were discarded as burn-in, and no thinning interval was considered. A residual updating algorithm was implemented to reduce computational time (Legarra and Mistzal, 2008).

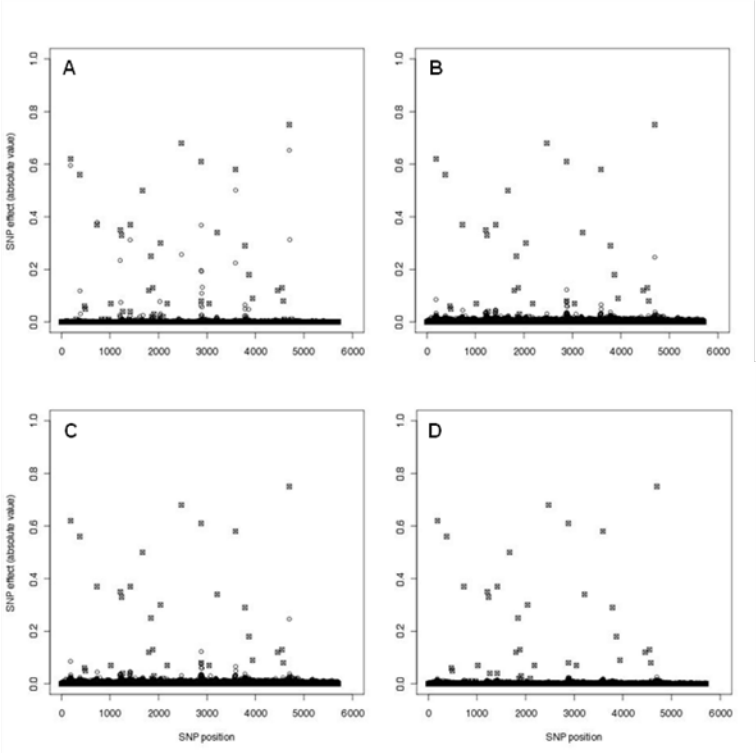
RESULTS

Simulated dataset

The Student's t prior distribution of SNP effects assumed in BayesA allowed to obtain the largest magnitude of effects in the presence of QTL with true moderate to large

effects (Table 1). This prior also resulted in the most accurate estimate of QTL with smaller effects (Figure 1.a). BayesA obtained the highest accuracies of DGV (0.87) and the lowest bias (Table 2).Accuracy of DGV obtained with the two Bayesian LASSO methods were lower than those obtained with BayesA. The degree of shrinkage of marker estimates with the Bayesian LASSO was greater than BayesA in the presence of QTL with large effects, with the largest QTL effects being severely under-estimated (Figure 1b, 1c). In fact, marker effects in these regions were on average 75% lower than those obtained with BayesA, and substantially lower than the true effects of the QTL (Table 1). This led to accuracies 10% lower (and higher bias) than those obtained with BayesA (Table 2). The two distributions of the λ hyper-parameter did not influence the results: negligible differences were observed in posterior estimates of λ from B-LASSO*gamma* and B-LASSO*beta*.

Figure 1. Absolute values of SNP effects in the simulated dataset.



The crossed squares indicate (absolute) simulated true QTL effects and the triangles indicate SNP effects for BayesA (a), B-LASSOgamma (b), B-LASSObeta (c), and Bayes-BLUP (d). On the x-axis, SNP are ordered by position from chromosome 1 to 6.

The Bayes-BLUP method gave accuracies 3% lower than the Bayesian LASSO methods. Marker effect estimates in QTL regions with large effects were always lower from BayesA, but in general they were similar to the estimates obtained in both Bayesian LASSO methods (Table 1). The correlation between DGV from Bayes-BLUP and Bayesian LASSO methods was 0.96, whereas the correlation between DGV from Bayes-BLUP and BayesA was lower at 0.87.

Table 1. Absolute SNP effects in simulated QTL regions with high effect.

Chr.[1]	Position (cM)	QTL[2]	BayesA		B-LASSO *gamma*		B-LASSO *beta*		Bayes-BLUP	
			$\hat{g}$ [3]	Dif.[4]	$\hat{g}$ [3]	Dif.[4]	$\hat{g}$ [3]	Dif.[4]	$\hat{g}$ [3]	Dif.[4]
1	20.00	0.62	0.60	-0.02	0.09	-0.53	0.08	-0.54	0.02	-0.60
1	40.00	0.56	0.12	-0.44	0.02	-0.54	0.03	-0.53	0.01	-0.55
1	77.23	0.37	0.38	0.01	0.04	-0.33	0.04	-0.33	0.01	-0.36
2	27.41	0.35	0.23	-0.12	0.03	-0.32	0.04	-0.31	0.01	-0.34
2	30.00	0.33	0.01	-0.32	0.02	-0.31	0.01	-0.32	0.01	-0.32
2	48.62	0.37	0.31	-0.06	0.05	-0.32	0.05	-0.32	0.02	-0.35
2	74.91	0.50	0.01	-0.49	0.02	-0.48	0.01	-0.49	0.01	-0.49
3	14.91	0.30	0.03	-0.27	0.02	-0.28	0.02	-0.28	0.01	-0.29
3	60.00	0.68	0.26	-0.42	0.03	-0.65	0.03	-0.65	0.01	-0.67
4	3.21	0.61	0.37	-0.24	0.12	-0.49	0.13	-0.48	0.02	-0.59
4	36.93	0.34	0.02	-0.32	0.02	-0.32	0.01	-0.33	0.01	-0.33
4	76.06	0.58	0.50	-0.08	0.07	-0.51	0.07	-0.51	0.02	-0.56
4	96.49	0.29	0.06	-0.23	0.04	-0.25	0.03	-0.26	0.01	-0.28
5	5.15	0.18	0.05	-0.13	0.02	-0.16	0.02	-0.16	0.01	-0.17
5	93.50	0.75	0.65	-0.10	0.25	-0.50	0.24	-0.51	0.03	-0.72

[1] Chromosomes where QTL with large effect were simulated. No QTL was simulated on chromosome 6.
[2] Absolute (true) value of major QTL effects.
[3] Mean posteriors of SNP effects in the region of QTL with moderate to large effects.
[4] Difference between true QTL effect and mean posterior of SNP effect in the region of the QTL.

Table 2. Accuracies, regression coefficients and regularization parameters obtained in the simulated dataset.

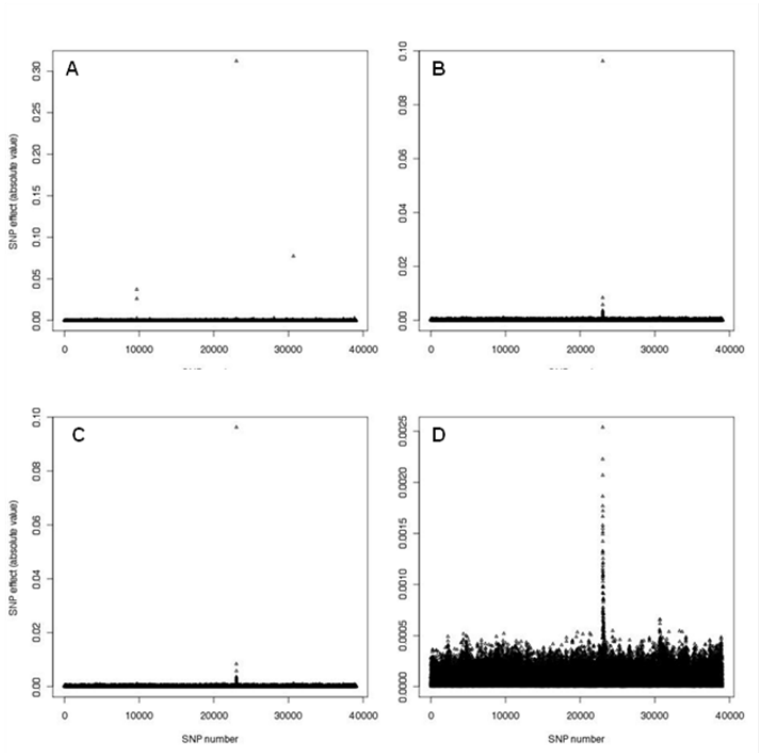
	BayesA	B-LASSO *gamma*	B-LASSO *beta*	Bayes-BLUP
$r_{(TBV,DGV)}$[1]	0.87	0.77	0.77	0.74
$r_{(Phe,DGV)}$[2]	0.432	0.392	0.390	0.373
$b_{(TBV,DGV)}$[3]	1.009	0.867	0.859	1.185
λ[4]	-	98.79 (4.60)	97.45 (4.11)	-

[1] Accuracy in validation population, obtained as correlation between true breeding values and direct genomic breeding values.
[2] Accuracy in validation population, obtained as correlation between pre-corrected phenotypes and direct genomic breeding values.
[3] Regression coefficients of true breeding values on direct genomic breeding values.
[4] Posterior mean (and standard deviation) of the regularization parameter for both Bayesian-LASSO methods.

Real dataset

For fat percentage, BayesA identified three SNPs with large to moderate effects: one on BTA-14 (i.e. in the diacylglycerol-acyltransferase 1 region, DGAT1), one on BTA-20 (i.e. in the growth hormone receptor region, GHR) and one on BTA-5. A mutation with moderate to large effect on fat percentage in the DGAT1 gene has been previously reported (Grisart et al., 2002). Marker effects of flanking SNP in these regions were close to zero (Figure 2). On the other hand, both Bayesian LASSO methods identified 32 SNP in the DGAT1 region (with the highest effect obtaining one third of the effect observed in BayesA) and one SNP in the GHR region (with much lower estimated effect than in BayesA), however failing to identify SNP with relative high effects on BTA-5 (Figure 2b, 2c).

Figure 2. Absolute values of SNP effects in the real dataset for fat percentage.



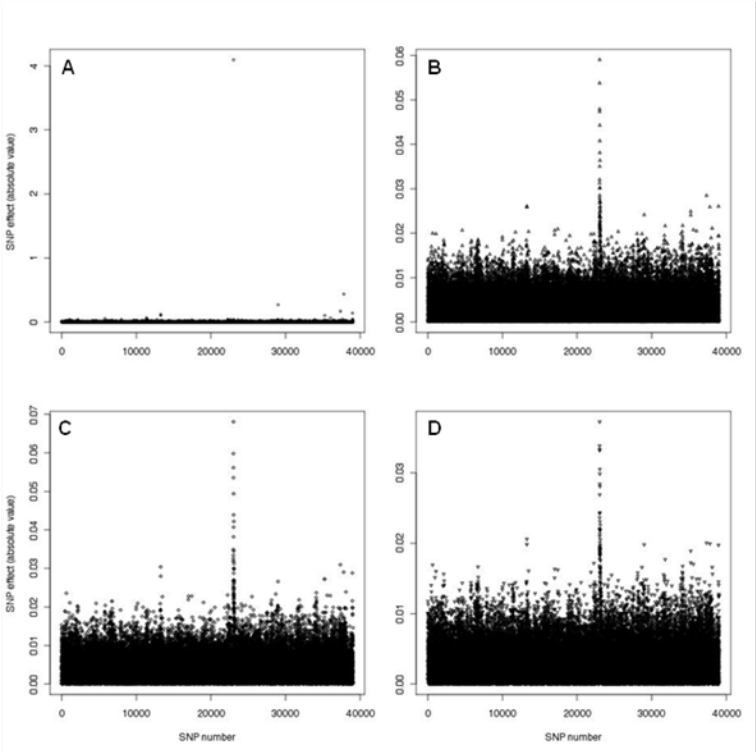
Mean posterior estimates of SNP effects (y-axis) for the 39.048 SNP considered (x-axis, displayed ordered by position from BTA-1 to BTA-29) obtained with BayesA (a), B-LASSOgamma (b), B-LASSObeta (c), and Bayes-BLUP (d).

There are two possible explanations for why BayesA gives a large effect to only one SNP in the DGAT1 region while the Bayesian LASSO methods identify many more SNP. One is that the prior in BayesA is flexible enough to allow all of the mutation effect to be captured by one SNP in high LD, while the prior for the Bayesian LASSO methods shrink the effects so hard that even when this SNP effect is removed, there is some effect of the mutation remaining, and thus this effect is then distributed over other SNP which are in LD with the mutation. An alternative explanation would be that the prior used in BayesA results in poor mixing during Gibbs sampling, such that once the effect of the mutation is allocated to one SNP, the following SNP in the chain never receive an effect (ter Braak et al., 2005). In this case the Bayesian LASSO results would indicate better mixing. If the mutation itself were genotyped and included in the data, we could determine which of the explanations is correct.

The distribution of effects in Bayes-BLUP were similar to those from Bayesian LASSO methods, with a total of 35 SNP with relative high effect in the DGAT1 and GHR regions, and no SNP with relative high effect in BTA-5 (Figure 2d). However, the magnitude of the highest SNP effects was more than ten-fold lower.

The influence of the DGAT1 mutation was also evident on protein yield (Figure 3). Again, BayesA identified only a single SNP with a negative effect in this region, whereas Bayesian LASSO and Bayes-BLUP methods showed a similar pattern grouping a large number of SNPs with relatively large, but smaller than BayesA, effects on BTA-14. In general, for protein yield, SNP estimates of both Bayesian LASSO and Bayes-BLUP showed only a slight difference in terms of magnitude of SNP effects.

Figure 3. Absolute values of SNP effects in the real dataset for protein yield.



Mean posterior estimates of SNP effects (y-axis) for the 39.048 SNP considered (x-axis, displayed ordered by position from BTA-1 to BTA-29) obtained with BayesA (a), B-LASSOgamma (b), B-LASSObeta (c), and Bayes-BLUP (d).

A difference in posterior estimates of the regularization hyper-parameter λ was observed between B-LASSO*gamma* and B-LASSO*beta* for both fat percentage and protein yield (Tables 3, 4). Higher penalization was observed in protein yield (128.72 and 108.47, in B-LASSO*gamma* and B-LASSO*beta*, respectively) rather than in fat percentage (100.48 and 70.04 in the aforementioned methods). Furthermore, posterior estimates of λ were more variable (i.e. higher standard deviation of posterior estimates after burn-in) in protein yield than in fat percentage. These results could reflect the distribution of QTL structure for the traits. The conditional regularization hyper-parameter obtained for fat percentage, is likely to be driven by the few QTL of large effect that affect the trait, resulting in a reduced (and more stable) penalization parameter.

Table 3. Accuracies, regression coefficients and regularization parameters obtained in the real dataset for fat percentage.

	BayesA	B-LASSO*gamma*	B-LASSO*beta*	Bayes-BLUP
$r_{(ABV,DGV)}$[(1)]	0.75	0.71	0.71	0.59
$b_{(ABV,DGV)}$[(2)]	0.943	1.110	1.108	1.395
λ[(3)]	-	100.48 (4.09)	70.04 (6.50)	-

[(1)]Accuracy in validation population as correlation between Australian breeding values and direct genomic breeding values.
[(2)]Regression coefficients of Australian breeding values on direct genomic breeding values.
[(3)]Posterior mean (and standard deviation) of the regularization parameter for both Bayesian-LASSO methods.

Table 4. Accuracies, regression coefficients and regularization parameters obtained in the real dataset for protein yield.

	BayesA	B-LASSO*gamma*	B-LASSO*beta*	Bayes-BLUP
$r_{(ABV,DGV)}$[(1)]	0.52	0.48	0.48	0.47
$b_{(ABV,DGV)}$[(2)]	1.112	1.327	1.241	1.474
λ[(3)]	-	128.72 (15.59)	108.47 (11.57)	-

[(1)]Accuracy in validation population as correlation between Australian breeding values and direct genomic breeding values.
[(2)]Regression coefficients of Australian breeding values on direct genomic breeding values.
[(3)]Posterior mean (and standard deviation) of the regularization parameter for both Bayesian-LASSO methods.

BayesA obtained the best accuracies of DGV (0.75 and 0.52 in fat percentage and protein yield, respectively; Tables 3, 4). Surprisingly, although showing a different distribution of SNP effects, accuracies obtained with both Bayesian LASSO methods were only 4% lower than those obtained in BayesA for both production traits. Regularization hyper-parameter differences in both Beta and Gamma hyper-prior structures did not influence accuracy results. In simulated data, de los Campos et al. (2009) noted that inferences on SNP effects were robust over a large range of values of the regularization hyper-parameters. For example, their results in simulated data showed nearly twofold differences of λ posterior estimates in models using Gamma and Beta prior distributions similar to those reported in this study, although obtaining

similar accuracies. Our results in the dairy data are in agreement with their observations.

Bayes-BLUP obtained the lowest accuracies of DGV (-16% respect to BayesA and -12% respect to both Bayesian LASSO) and the highest bias for fat percentage. For protein yield, however, differences in terms of accuracy of DGV were greatly reduced (-5% with BayesA and -1% with B-LASSO*gamma*).

DISCUSSION

Bayesian LASSO methods have been successfully tested in real data for QTL mapping studies (Yi and Xu, 2008), to test genomic selection in wheat and mice (De los Campos et al., 2009) and to choose sub-sets of SNP for genomic predictions in livestock (Weigel et al., 2009). However, comparisons of performance of the Bayesian LASSO with other methodologies (i.e. BayesA) are only available in simulated data (Cleveland et al., 2010). These authors observed a tendency of the Bayesian LASSO methods to obtain SNP effects of lower magnitude than BayesA where the true QTL effect were moderate to large. The same trend was observed in our study. However, DGV accuracy results in both studies do not agree, most probably because of differences in the model and the simulated dataset used. In Cleveland et al. (2010) the Bayesian LASSO did not include a polygenic term, whereas in our study it was fitted to account for population sub-structure. Furthermore, Cleveland et al. (2010) analyzed the XIII QTLMAS workshop dataset (XIII QTLMAS webpage), which included nearly 1000 animals (structured in 20 full sib families) as training population and 453 markers, whereas the simulation used in this study (XII QTLMAS dataset) was based over nearly 5000 animals and 6000 markers. In addition, the organizers of the XIII QTLMAS workshop provided

longitudinal phenotypic data at five different time-points, whereas phenotypes used as dependent variables had to be predicted at a sixth time-point (meaning that different approaches to predict this values could lead to differences in results). Finally, the distribution of QTL effects in the two simulated datasets were very different, with 3 QTL of medium-large effect and 15 of low effects affecting the parameters used to determine the phenotypes provided in our dataset. Considering all these differences, results obtained in both studies are not comparable. In any case, both simulated datasets are likely to be far from the reality for complex traits, as the number of SNP and QTL was very limited compared to real data. In our case, for example, the XII QTLMAS dataset included four QTL explained more than 50% of the genetic variability of the trait. However, using the simulated data did allow us to study and compare the different performances of all methods in a simple dataset with a known distribution of QTL effects. In this scenario, BayesA more accurately estimated the effects of QTL, in fact obtaining best DGV accuracies and the lowest bias. On the other hand, both Bayesian LASSO methods and Bayes-BLUP resulted in very similar SNP estimates and DGV accuracy results. These results were markedly lower than those obtained with BayesA method, as expected in a trait highly affected by few QTL, as BayesA resulted in much less shrinkage of estimates of large to moderate QTL effects than either Bayes-BLUP or Bayesian LASSO.

In this study we also analyzed milk fat percentage and protein yield in Australian Holstein Friesian dairy cattle, as there is some information about the distribution of QTL effects for these traits. For fat percentage, a mutation in DGAT1 explains up to 30% of the variation of the trait (Grisart et al., 2002). Another interesting QTL for fat percentage is located on BTA-20, in the region of GHR (Viitala et al., 2006). However, no such large or moderate-effect QTL are present for protein yield,

although there is evidence that DGAT1 contributes to the negative correlation between these two traits (Thaller et al.,2003; Cole et al., 2009).Thus, known QTL regions with large effects in the traits analyzed, accuracies and bias results can be used to compare the performance of all methods tested in this study. All methods identified the DGAT1 and GHR regions, although with large differences in the magnitude of the effects. In addition, BayesA identified a third region with a high effect on BTA-5 for fat percentage. This SNP is positioned at 101,042,396bp, which is within a region where two QTL with large confidence intervals have been reported for milk and fat yield (CattleQTLdb; Viitala et al., 2003; Olsen et al 2002).

Using the double exponential prior distribution in Bayesian LASSO resulted in a degree of shrinkage of SNP effect estimates that was much higher than BayesA. However, this method gave only 4% lower accuracies than BayesA for both fat percentage and protein yield. As previously reported, the normal prior distribution used in the Bayes-BLUP method gave the lowest DGV accuracy for fat percentage, compared to methods that allowed SNP effects to assume non-normal distributions (VanRaden et al., 2009; Hayes et al., 2009). Interestingly, estimates of SNP effects from the Bayesian LASSO for protein yield were very similar to Bayes-BLUP for protein yield, a trait not controlled by QTL with large effects (in fact, the accuracies of DGV were only 1% higher, for both B-LASSO*gamma* and B-LASSO*beta*).

CONCLUSION

An appealing feature of Bayesian LASSO methods as described by Park and Casella (2008) is that hyper parameters of the double exponential prior distribution of SNP effects are conditional to the data.This in contrast to methods such as BayesA where the leptokurtosis of the Student t prior must be specified. However, with the limited

amount of data in our study, the Bayesian LASSO methods resulted in strong shrinkage of SNP effect estimates, which in some cases was similar to what observed in Bayes-BLUP. With large datasets there would be more information from the data to condition the hyper-parameters, which may result in more optimal shrinkage. The Bayesian LASSO methods may be particularly useful with next generation SNP-chips (i.e. with many more SNP than analyzed here in much greater linkage disequilibrium with QTL) and larger datasets.

AUTHORS CONTRIBUTIONS

BJH conceived the study. AJC performed all the required lab work. ELN coded the scripts for the analyses and analyzed the data. ELN and BJH wrote the manuscript. MEG, BJH and ELN discussed the results. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

Authors wish to thank Phil Bowman for his help in coding the scripts. Gustavo de los Campos and Cristian Maltecca are acknowledged for their help in setting up the Bayesian Lasso methods. Prof. Paolo Ajmone-Marsan is also kindly acknowledged. This work was carried out while ELN was a visiting scientist at Department of Primary Industries (Victoria, Australia). ELN was funded by AGRISYSTEM Ph.D. fellowship (Ciclo XIII) and by SELMOL project (Italian Ministry of Agricultural Policies).

REFERENCES

CattleQTLdb [http://www.animalgenome.org/cgi-bin/gbrowse/cattle/]

Cleveland, M., S.Forni, N. Deeb, and C. Maltecca. 2010. Genomic breeding value prediction using three Bayesian methods and application to reduced density marker panels. BMC Proceedings, 4(Suppl 1):S6

Cole, J.B., P.M. VanRaden, J.R. O'Connell, C.P. Van Tassel, T.S. Sonstegard, R.D. Schnabel, J.F. Taylor, and G.R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci., 92:2931-2946

Crooks, L., G. Sahana, D. de Koning, M.S. Lund, and Ö. Carlborg. 2009. Comparison of analyses of the QTLMAS XII common dataset. II: genome-wide association and fine mapping. BMC Proceedings, 3(Suppl 1):S2

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J.M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics, 182:375-385.

Goddard, M.E., and B.J. Hayes: Genomic selection. J. Anim. Breed. Genet., 124:323-330

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

Gredler, B., K.G. Nirea, T.R. Solberg, C. Egger-Danner, T.H.E. Meuwissen, and J. Sölkner. 2009. A comparison of methods for genomic selection in Austrian dual purpose Simmental cattle. Pages 569-572 in Proceeding of the Assoc. for the Advancement of Animal Breeding and Genetics, Barossa Valley, South Australia.

Grisart, B., W. Coppieters, F. Farnir, L. Karim, C. Ford, P. Berzi, N. Cambisano, M. Mni, S. Reid, P. Simon, R. Spelman, M. Georges, and R. Snell. 2002. Positional candidate cloning of a QTL in dairy cattle: identification of a missense mutation in the bovine DGAT1 gene with major effect on milk yield and composition. Genome Res., 12:222-231

Habier, D., J. Tetens, F. 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.

Harris, B.L., D.L. Johnson, and R.J. Spelman. 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. Pages 11-16 in Proc. of the 36th ICAR Biennial Session. Niagara Falls, USA.

Hayes, B.J., and M.E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol., 33:209-229.

Hayes, B.J., P.J. Bowman, A.J. Chamberlain, and M.E. Goddard. 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci., 92:433-445.

Legarra, A., and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection. J. Dairy Sci., 91:360-366.

Legarra, A., C. Robert-Granie, E. Manfredi, and J.M. Elsen. 2008. Performance of genomic selection in mice. Genetics, 180:611-618.

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

Meuwissen, T.H.E., B.J. Hayes, M.E., and Goddard. 2001. Prediction of total genetic

values using genome-wide dense marker maps. Genetics, 157:1819-1829.

Olsen, H.G., L. Gomez-Raya, D.I. Vage, I. Olsaker, H. Klungland, M. Svendsen, T. Adnoy, A. Sabry, G. Klemetsdal, N. Schulman, W. Kramer, G. Thaller, K. Ronningen, and S. Lien. 2002. A genome scan for quantitative trait loci affecting milk production in Norwegian dairy cattle. J. Dairy Sci., 85:3124-3130.

Park, T., and G. Casella. 2008. The Bayesian Lasso. J. Am. Stat. Assoc., 103:681-686.

ter Braak, C.J.F., M.P. Boer, and M.C.A.M. Bink. 2005. Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. Genetics. 170:1435-1438.

Thaller, G., W. Krämer, A. Winter, B. Kaupe, G. Erhardt, and R. Fries. 2003. Effects of DGAT1 variants on milk production traits in German cattle breeds. J. Anim. Sci., 81:1911-1918.

Tibshirani, R. 1996 Regression shrinkage and selection via the Lasso. J. R. Stat. Soc., 58:267-288.

van der Werf, J.H.J. 2009. Potential benefit of genomic selection in sheep. Pages 38-41 in Proc. of the Assoc. for the Advancement of Animal Breeding and Genetics, Barossa Valley, South Australia.

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

Verbyla, K., P.J. Bowman, B.J. Hayes, and M.E. Goddard. 2010. Sensitivity of genomic selection to using different prior distributions. BMC Proceedings, 4(Suppl 1):S5

Viitala, S., J. Szyda, S. Blott, N. Schulman, M. Lidauer, A. Mäki-Tanila, M. Georges, and Vilkki J. 2006. The role of the bovine growth hormone receptor and prolactin receptor genes in milk, fat and protein production in Finnish Ayrshire dairy cattle. Genetics, 173:2151-2164

Viitala, S.M., N.F. Schulman, D.J. de Koning, K. Elo, R. Kinos, A. Virta, J. Virta, A. Maki-Tanila, and J.H. Vilkki. 2003. Quantitative trait loci affecting milk production traits in Finnish Ayrshire dairy cattle. J. Dairy Sci., 86:1828-1836

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

XII QTL-MAS workshop data web-page [http://www.computationalgenetics.se/QTLMAS08/QTLMAS/DATA.html]

XIII QTL-MAS workshop web-page [http://www.qtlmas2009.wur.nl/UK/]

Yi, N., and S. Xu. 2008. Bayesian LASSO for quantitative trait loci mapping. Genetics, 179:1045-1055.

CHAPTER VI

Integrating Population Genomics and Genomic Selection

P. Ajmone-Marsan[1], E.L. Nicolazzi[1], R. Negrini[1], N. Macciotta[2], L. Fontanesi[3], V. Russo[3], A.Bagnato[4], E. Santus[5], D. Vicario[6], J.B.C.H.M. van Kaam[7], A. Albera[8], F. Filippini[9], C. Marchitelli[10], G. Mancini[10], A. Nardone[10], A. Valentini[10]

[1]Università Cattolica del Sacro Cuore, Piacenza, Italy;
[2]University of Sassari, Italy;
[3]DIPROVAL, University of Bologna, Italy;
[4]University of Milano, Italy;
[5]ANARB, Italy;
[6]ANAPRI, Italy;
[7]ANAFI, Italy;
[8]ANABORAPI, Italy;
[9]ANABIC, Italy;
[10]Tuscia University, Viterbo, Italy

Published in: Proceedings of the Interbull International Workshop 2010. Bulletin No. 41, 2010. ISSN 1011-6079.

ABSTRACT

The availability of panels of several thousand SNPs ordered on the genome has initiated the era of population genomics, that is the application of genomic approaches to population genetics. One application of population genomics is the investigation of patterns of diversity along the chromosomes in search for signatures left by past and recent selection. These signatures are locus specific and can be identified and distinguished from the genome wide effects caused by genetic drift and demographic events. In this paper we searched for outlier behavior within the 54,001 SNPs of the Illumina Beadchip Array assayed on 2682 bulls belonging to Italian Brown and four other Italian breeds, one dairy (Italian Friesian), one dual purpose (Italian Simmental), and two beef (Marchigiana and Piedmontese) investigated within the Italian SELMOL project on molecular genetics applied to animal breeding. Outlier values of the F_{st}genetic differentiation index averaged along 9 markers sliding windows were searched in pairwise breed comparisons by a permutation strategy. A total of 8944 sliding windows were significant in at least one of the four comparisons that included Italian Brown.

Among these, 869 SNPs were significant in all three comparisons vs. dual purpose and beef breeds. These two subsets of 8944 and 869 SNPs were used in a genomic prediction exercise. The 749 Italian Brown genotyped bulls were divided in training (the 600 older bulls) and prediction (the 149 younger bulls) populations. In all cases DGVs, BayesA estimates of Milk Yield, Protein Percent, Udder Score and Total Economic Index, were not significantly higher than those obtained with a random marker subset of the same size. Selection signatures likely identify genomic regions subjected to historical selection that do not match with those in which genes controlling economic traits are segregating in modern populations. This hampers the use of the

selection signature approach for identifying marker subsets useful in genomic selection.

INTRODUCTION

Population genomics is a term first proposed in a publication on human genetic diseases (Gulcher and Stefansson, 1998) to indicate the use of genomic technologies in population genetics studies. The availability of panels of many thousand and sometimes many hundred thousand SNPs ordered along the genome has recently marked a paradigm shift in the way populations can be investigated. One major advance is the ability to identify genomic regions that are under selection. These can be detected by comparing the distribution of allele frequencies at marker loci within or between populations or groups of populations, in search for markers significantly departing from neutral behavior. The comparison of the distribution of allele frequencies can be either direct or through different statistics, function of allelic or genotypic frequencies, as F_{st}(e.g. The Bovine HapMap Consortium et al., 2009) and linkage disequilibrium (e.g. Ennis, 2007). In addition, specific tests for detecting significant effects have been developed (e.g. Voight et al., 2006). The objectives of this study are i) to detect selection signatures in the Italian Brown cattle breed and ii) to evaluate the performance of markers under selection in the genomic prediction of genetic values of young bulls. To reach these goals we used SNP data produced within the Italian SELMOL project on the application of molecular genetics to animal breeding.

MATERIALS AND METHODS

Animals

A total of 2295 animals from 5 breeds were genotyped with the 54001 SNP markersincluded in the Illumina BovineSNP50 BeadChip: 775 Italian Brown (BRW), 419Italian Friesian (FRI), 379 Piedmontese (PIM), 229 Marchigiana (MCG) and 493 Italian Simmental (SIM).

Genomic data

Following clean up by filtering subjects with more than 5% missing SNPs and SNPs with more than 5% missing typing, the final dataset included 2266 individuals and 45087 SNPs. Among these 43771 were located on the 29 autosomes and BTAX and 1316 remained not anchored to the Btau 4.0 version of the bovine sequence assembly. These latter were excluded from further analyses.

F_{st}

F_{st}index was calculated as $F_{st}=1–H_s/H_t$, where H_s is the Hardy-Weinberg equilibrium (HWE) heterozygosity within subdivisions, averaged across subpopulations and $H_t$is the HWE heterozygosity for the total population, assuming no genetic differentiation among subpopulations. F_{st}values were averaged along sliding windows of nine consecutive SNPs, irrespectively on the relative distance between adjacent markers. Each chromosome contained a number of sliding windows equal to SW*i*=N*i*-(Nsw-1), where SW*i*is the number of sliding windows on chromosome *i*, N*i*is the number of markers on chromosome *i* and Nsw is the number of markers included in the sliding window. In total, 43531 sliding windows were assembled.

Permutations

To estimate the 5% genome-wide significance thresholds of F_{st}values, markers were first randomly shuffled across the genome. Then, the distribution of average F_{st}value of

groups of 9 randomly selected markers was computed. Finally the F_{st}values separating 5% of the distribution were recorded. The highest values among permutation runs were used as F_{st}thresholds to evaluate the significance of F_{st}calculated on markers ordered along chromosomes.

Genomic prediction

Genomic predictions of breeding values (DGVs) were obtained using a BayesA model (Meuwissenet al.,2001). A total of 20.000 runs of iteration were performed on each analysis. First 10.000 iterations were discarded as burn-in and no thinning interval was considered. The model included a polygenic term for taking the population structure into account. Accuracies were obtained as Pearson correlations between DGVs and breeding values obtained from progeny testing (EBVs).

RESULTS

Selection signatures

Average F_{st}values of individual markers varied between 0.034+0.049 in BRW *vs.* PIM to 0.057+0.080 in BRW *vs*. FRI. F_{st}values of sliding windows had same average and smaller values of SD, spanning from 0.023 in BRW *vs*. PIM to 0.035 in BRW *vs*. FRI. Sliding windows spanned on average 483+263 Kb, with a maximum of 1382 Kb and a minimum of 4 Kb. In all comparisons involving Italian Brown 8944 sliding windows had F_{st}significantly higher than the 5% threshold established by the permutation approach. These sliding windows are not equally distributed across breed comparisons (Table 1). Surprisingly the comparison with Italian Friesian was the one in which the highest number of signatures was detected. A remarkable number of signatures was found consistent across all comparisons or across comparisons between BRW and the beef and dual purpose breeds. Selection signatures were not equally distributed across

chromosomes as well. Numbers ranged between 1076 on BTA6 and 4 on BTAX.

Table 1. Selection signatures detected in the comparison between Italian Brown cattle and all four, three beef/dual purpose breeds and each single breed investigated.

Comparison involving Italian Brown	N. Sliding windows with signature (P≤0.05)
Piedmontese	1770
Marchigiana	1795
Italian Simmental	1728
Italian Friesian	1999
Three beef/dual purpose breeds	869
All four breeds	463

Genomic predictions

Table 2 summarizes the correlations between EBVs and DGVs calculated using subsets of markers carrying signatures of selection.

Table 2. Correlation between EBVs and DGVs estimated by different marker subsets in Italian Brown cattle.

Marker Subset (N markers)	Milk Yield	Protein %	Udder Score	ITE
Sel. Sign. all (8944)	0.131	0.423	0.270	0.511
Random (mean) (8944)	0.127	0.407	0.256	0.579
Sel. Sign. Italian Brown specific (869)	0.010	0.105	0.017	0.245
Random (mean) (869)	0.160	0.219	0.291	0.294

Triplicate random sets of markers having the same size of subsets investigated were also used as control. Using 8944 markers, correlations were slightly higher with

markers under selection compared to the average of three runs with random subsets, but always lower than with the random subset giving the highest correlations. With the 869 Italian Brown specific subset correlations were always lower than with random subsets.

DISCUSSION

In this paper we have scanned the genome of the Italian Brown dairy cattle searching for signatures of selection. Among possible indexes, we used $F_{st}$to study selection because it is robust, easy to calculate and widely used for this purpose (e.g. Barendse *et al.,* 2009). Single marker F_{st}values varied substantially even among SNPs very close to each other and had standard deviations even higher than the means. Therefore, we adopted a sliding windows approach to avoid excessive noisiness (Weir et al., 2005). We decided to include in sliding windows an homogeneous number of markers rather than using a predetermined genome size. This to avoid having windows including only one or a few markers. The use of 9 markers was a subjectivechoice but also facilitates the comparison with published data using the same or similar sliding window size (e.g. Stella et al., 2010). On average the 43771 windows spanned genomic regions of 500Kb and among these 1195 regions larger than 1Mb and 118 larger than 2Mb, providing a rather detailed survey of the cattle genome. F_{st}values were calculated in pair wise comparisons in which the dairy Italian Brownwas contrasted with the dairy Italian Friesian, the dual purpose Italian Simmental and the beef Piedmontese and Marchigiana cattle breeds. The highest average F_{st}across markers was found in the BRW *vs.* FRI. This is likely the result of the combined effect of different origin, reduced gene flow and small effective population size of Italian Friesian compared to the beef and dual purpose breeds

investigated. However, ascertainment bias is possibly contributing to this divergence, given that a relevant number of SNPs included in the array have been developed to be highly informative in the Holstein population. A permutation approach permitted the identification of significant selection signatures in each pair wise comparison. In total 8944 were underselection in at least one of the comparisons involving Italian Brown (Table 1). Contrary to expectation, the highest number of signatures was found in the contrast between the two dairy breeds, rather than between Italian Brown and the beef breeds. In pair wise comparisons signatures are due to selection in either breed or to divergent selection in both breeds. Markers having consistent outlier behavior in multiple comparisons involving the same breed are likely to be under selection in that same breed. Using this rationale, we have isolated 869 markers specific to selection in BRW. Markers under selection might include those associated to traits included in selection indexes and hence be informative for genomic prediction of genetic merit. However, DGV of three production traits estimated in young BRW bulls based on markers under selection were no better and often worse than those calculated from an equal number of random markers (Table 2). With the current approach only the historical and strongest effects of selection could be detected, probably on genes close to fixation and having either a qualitative or a major effect on traits that have been selected since Italian Brown breed formation. Therefore, most selection signatures likely correspond to genomic regions subjected to historical selection that do not match with those in which genes controlling economic traits are segregating in modern populations. The selection signature approach is useful in the reconstruction of the interesting process of breed formation, but seems to have little application in the choice of marker subsets that can be profitably used in genomic selection.

REFERENCES

Aulchenko, Y.S., S. Ripke, A. Isaacs, and C.M. van Duijn. 2007. GenABEL: an R package for genome-wide association analysis. Bioinformatics 23:1294-1296.

Barendse, W., B.E. Harrison, R.J. Bunch, M.B. Thomas, and L.B. Turner. 2009. Genome wide signatures of positive selection: the comparison of independent samples and the identification of regions associated to traits. BMC Genomics 10:178.

Bovine HapMap Consortium et al. 2009. Genome-wide survey of SNP variation uncovers the genetic structure of cattle breeds. Science 324:528-532.

Ennis, S. 2007. Linkage disequilibrium as a tool for detecting signatures of natural selection. Methods in Molecular Biology 376:59-70.

Gulcher, J., and K. Stefansson. 1998. Population genomics: laying the groundwork for genetic disease modeling and targeting. Clin. Chem. Lab. Med. 36:523-527.

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.

Stella, A., P. Ajmone-Marsan, B. Lazzari, andP. Boettcher. 2010. Identification of Selection Signatures in Cattle Breeds Selected for Dairy Production. Genetics185:1451-1461.

Voight, B.F., S. Kudaravalli, X. Wen, andJ.K. Pritchard. 2006. A map of recent positive selection in the human genome. PLoS Biology 4:e72.

Weir, B.S., L.R. Cardon, A.D. Anderson, D.M. Nielsen, and W.G. Hill. 2005. Measures of human population structure show heterogeneity among genomic regions. Genome Research 15:1468-1476.

CONCLUSIONS

Genetic evaluation systems are in rapid development worldwide. In most countries, "traditional" breeding programs based on the recording of phenotypes and relationships between animals are currently being integrated and in the future might be replaced by the introduction of molecular information. This thesis stands in this transition period, therefore it covers research on both types of genetic evaluations.

The study in Chapter II treats an important issue of "traditional" genetic evaluations, which is the assessment of international genetic evaluations in predicting domestic breeding values of yield related traits in Holstein bulls. This study showed that international genetic evaluations based on foreign daughters performance are reasonably accurate predictions of bulls' national breeding values in most of the countries analyzed. Among Australia, Canada, France, Italy, the Netherlands and the United States of America, only Italy showed biased results. In the other countries, no strong bias was observed since realized correlations between national and international breeding values were generally similar to their expected values and close to 0.90. It is important to underline that the variations in methodologies in either national or international evaluations during the rather long time period considered (nearly 8 years) might have influenced the results obtained. For example, the Italian results were influenced by a relatively recent change in the predictive model for production traits (from a lactation to a random regression test day model), which heavily affected the correlations between breeding values before and after the change. In addition, the methods used either to edit the data (in order to obtain reliable breeding values) or to assign the foreign bull's country of origin, might have in turn penalized more some of the countries than others (i.e. again the case of Italy that imports large proportions of the semen used in their farms). In spite of these unaccounted effects (and considering the relatively low number of bulls analyzed),

results obtained in this study agreed well with previous studies that compared national and international evaluations, confirming that international genetic evaluations are accurate predictions of domestic breeding values in the countries analyzed. All the software produced in this project (and used for this study) was automated to allow running such analysis to all countries participant to the international genetic evaluation, just by updating a parameter file. Other major group traits (conformation, udder health, longevity, etc.) could be analyzed in the same way.

In Chapter III, two SNP pre-selection methods based on single marker regression for the estimation of genomic breeding values were tested: Bonferroni correction of the significance threshold and permutation test. These methods were chosen by their simplicity of application, and were tested on the XII QTLMAS workshop (simulated) dataset. Although likely rather far from the reality of complex traits (i.e. only 6000 SNP and a simulated QTL structure with six QTL explaining more than 50% of the genetic variance of the trait), this dataset has the advantage of providing a straightforward way to study the behavior of both pre-selection methods in the presence of QTL responsible for both high and low levels of variation.

Bonferroni correction assumes independence of tests. As SNP data are actually (auto)-correlated because of Linkage Disequilibrium, the Bonferroni correction is a highly stringent method. On the other hand, a more relaxed threshold could be used, estimating an approximation of the number of independent “blocks” of SNP along the genome. When using the traditional approach of Bonferroni correction (i.e. considering independence of tests), markers selected obtained better DGV accuracies, compared to those selected by the permutation test, even if permutation test at 0.001 threshold obtained DGV accuracies only 2% less accurate than

Bonferroni correction. Although in this study only 1000 permutations per SNP were run, performing up to 1 million permutations per SNP and choosing the most stringent threshold possible ($P \leq 1 \times 10^{-06}$) only increased the accuracy by 2% (*unpublished results*). This was somehow expected, as the random reduction of 50% of the markers in the permuted dataset gave in turn only a slight decrease in DGV accuracies.

When based upon simulated data, results should be carefully taken into account. The genetic structure of the simulated population, the QTL distribution, the simulated Ne, the heritability of the trait, etc., can heavily influence results. All methods, regardless of their complexity, have to be cross-validated in real data. At the time of this (and next) work, genotypic data were yet not available in Italy. In any case, even if other more advanced approaches can be used to reduce dimensionality of predictors, the methods studied in this paper resulted in a simple and efficient first step in pre-selecting SNP. However, it has to be noted that the methods applied in this and next two chapters were performed over a simulated dataset, with a reduced number of markers, and an unrealistic QTL distribution. Different simulated genetic structures may show different outcomes. For example, a Bonferroni correction over a larger dataset with only small-effect QTL might in turn give a reduced accuracy because of the low amount of markers retained due to the highly stringent threshold applied.

In Chapter IV the same simulated data of the previous study were used to test an alternative approach to reduce the number of variable, a multivariate Principal Component Regression (PCR; Solberg et al., 2010). The novelty introduced in the method, previously proposed by Solberg and coauthors, was the introduction of eigenvalues as variance priors.

Principal Component Regression is an appealing method as it avoids co-linearity between markers by creating a new set of variables that are independent from each other, avoiding at the same time the loss of information as every SNP is part of each principal component. Compared to other methods tested in this study, PCR using eigenvalues as variance priors obtained the best DGV accuracies and the lowest bias. Furthermore, it greatly reduced the computing time required to perform the estimation (the number of variables was reduced by more than 90%). In addition, it allowed the use of principal components to study the genetic structure of the population, by interpreting the orthogonal variables. For example, in this study, the second principal component was able to reflect the genetic variation among generations. As stated previously, the results obtained in a simulated dataset must be cross-validated, as they might be influenced by the simulation method defined by the researcher. In a recent study, this method was tested in three different Italian breeds (Holstein, Brown Swiss and Simmental), confirming good performances, although a low number of animals per breed (1000 or less) was included in the analyses (*unpublished data)*.

The study in Chapter V compared 4 different methods for GS in a simulated and a real dataset. The main objective was to test the performance of the Bayesian LASSO, an appealing method that uses a conditional hyper-parameter of the prior distribution of SNP variances, differently from two other methods tested (BLUP and BayesA), which have fixed hyper-parameters. BayesA obtained the best results in both datasets. In real data, BayesA outperformed all the other methods, especially when analyzing fat percentage, a trait highly affected by a high-effect QTL (DGAT1). Using the double exponential prior distribution, Bayesian LASSO resulted in a degree of shrinkage of SNP effect estimates higher than BayesA. However, this method

gave only 3% lower accuracies than BayesA in both fat percentage and protein yield. The BLUP method obtained the worst results in both traits, although the difference with BayesA in terms of DGV accuracy was large in fat percentage (nearly -15%) but lower in protein yield (nearly -5%). However, our dataset included only ~1000 animals that were further divided in validation and prediction datasets. For this reason, low heritability traits were not tested in the analysis, as the number of animals with phenotypes available for those traits was even lower, possibly further affecting the comparability of results across-traits.

With the limited amount of data in our study, the Bayesian LASSO methods resulted in strong shrinkage of SNP effect estimates, which in some cases was similar to what observed in BLUP. With large data sets there would be more information from the data to condition the hyper-parameters, which may result in more optimal shrinkage, thus, higher accuracies in predicting DGV for high and low heritability traits.

In Chapter VI, selection signatures between Brown Swiss and other 4 dairy, beef or dual purpose Italian breeds were explored with the F_{st} method. Nearly 9000 significant signals were found ($P \leq 0.05$). Contrary to expectation, the highest number of signatures were found in the contrast between the two dairy breeds (Brown Swiss and Holstein Friesian), rather than between Italian Brown and the beef breeds. This could be derived by combined effect of different origin, reduced gene flow and small effective population size of Italian Friesian compared to the other breeds investigated. Furthermore, ascertainment bias effect is not to be excluded. We used different subsets of markers in order to test if this information could be used as a method to reduce the number of variables in the prediction equations for GS. However, comparable or lower DGV accuracies were obtained when using these pre-selected markers, rather than the same number of random markers. An association

between markers carrying significant selection signatures and traits comprised in selection indexes was also tested, showing that only a small proportion of these markers were significantly associated to production traits. With the approach used only the strongest effects of selection could be detected, probably nearby genes having a major effect on traits that have been strongly selected since Italian Brown breed formation and today probably close to fixation. This limits the use of the selection signature approach to reconstruct the selection history of a breed rather than in detecting markers useful for genomic selection.

Genomic information is the present (and future) not only for breeding but for research as well. In the last few years, increasing effort has been input to find signatures of selection in the genome, genomic regions affecting one or more traits, searching genes controlling expression of interesting traits. Further investigation on other issues is ongoing as well. In animal breeding, traditional genetic evaluations have obtained good results for more than 50 years. Now, the introduction of molecular information is expected to further increase this progress. Many countries are introducing genome-enhanced evaluations in their breeding programs, although progeny test is not being abandoned, at least not yet. Possible future developments of this recent convergence between quantitative and molecular genetics can lead to bulls routinely sequenced (next-generation sequencing), in order to capture a greater amount of variability. However, a greater amount of genetic information is not enough. Genetic models will have to evolve in order to capture also the non-additive effects, as dominance deviations and/or epistatic interactions. In any case, the knowledge of the biological processes behind the expression of a phenotype will definitively increase, allowing a whole new series of research programs and analyses.

The results of this thesis indicate that the genetic gains expected from the analysis of simulated data can be obtained on real data. Still, further research is needed to optimize the use of genome-wide information and obtain the best possible estimates for all traits under selection.

OTHER PUBBLICATIONS

PAPERS (or Abstracts of oral presentations in congresses)

Macciotta N.P.P., M. Pintus, G. Gaspa, **E.L. Nicolazzi**, A. Rossoni, D. Vicario, J.B.C.H.M. Van Kaam, A. Nardone, A. Valentini, and P. Ajmone-Marsan. 2010. Use of Principal Component approach for estimating direct genomic breeding values for somatic cell score in cattle. Proceedings of the 9th world congress on genetics applied to livestock production. Leipzig, Germany. 1-6 August 2010, *in press*.

Minozzi G., **E.L. Nicolazzi**, G. Jansen, G. Strozzi, A. Stella, R. Negrini, P. Ajmone-Marsan, and J.L. Williams. 2010. Genome wide scan for somatic cell counts in Holstein Bulls. BMC proceedings, *in press*.

Gaspa G., **E.L. Nicolazzi**, R. Steri, C. Dimauro, and N.P.P. Macciotta. 2009. Effect of estimation approach and number of QTL in accuracies of genomic breeding values for simulated data. J. Anim. Sci. vol. 87; p.315, E-suppl. 2 / J. Dairy Sci. 92 E-suppl 1.

Van Kaam J.B.C.H.M., G.B. Jansen, **E.L. Nicolazzi**, and F. Canavesi. 2009. Genomic information in Genetic Evaluations. Pages 231-234 in Proceedings of the Interbull Meeting. Barcelona, Spain. August 21-24, 2009. Interbull No. 40. Interbull, Uppsala, Sweden.

POSTERS

Lucente, G., R. Negrini, L. Colli, M. Pellecchia, L. Bomba, **E.L. Nicolazzi**, P. Ajmone-Marsan, P. Crepaldi, L. Nicoloso, E. Milanesi, J. A. Lenstra, and RESGEN Consortium. Variabilità genetica di *Bos Taurus* e *Bos Indicus* misurata con marcatori molecolari AFLP. BIOD 2010 meeting. Milano, Italy, June 10-11 2010.

Pellecchia, M., R. Negrini, L. Colli, F. Chegdani, **E.L. Nicolazzi**, P. Ajmone-Marsan, P. Crepaldi, and ECONOGENE Consortium. Capre domestiche e migrazioni dell'uomo: i marcatori AFLP raccontano. BIOD 2010 meeting. Milano, Italy, June 10-11, 2010.

Negrini, R., R. Mazza, L. Colli, **E.L. Nicolazzi**, R. Marino, M. Pellecchia, L. Bomba, A. Stella, J.L. Williams, and P. Ajmone-Marsan. Effect of strategies for SNP discovery on the estimation of population genetics parameters. Plant & Animal Genomes XVIII Conference, San Diego, CA, USA. January 9-13, 2010.

Ajmone-Marsan, P., R. Marino, D. Perini, R. Negrini, **E.L. Nicolazzi**, L. Pariset, A. Valentini, D. Vicario, E. Santus, M. Blasi, L. Fontanesi, F. Schiavini, A. Bagnato, V. Russo, N.P.P. Macciotta, and A. Nardone. 2009. Identification of selection signatures across dairy cattle genome based on the Illumina 54,000 SNP panel. Page 609 in Book of abstracts of the 60th EAAP meeting. Barcelona, Spain, August, 24-27 2009, EAAP, vol. 15.

Pariset, L., A. Caroli, S. Chessa, L. Fontanesi, V. Russo, A. Bagnato, F. Schiavini, A.B. Samore, M. Feligini, I. Bonizzi, D. Vicario, A. Rossoni, S. Sangalli, R. Marino, D. Perini, **E.L. Nicolazzi,** N.P.P. Macciotta, and P. Ajmone-Marsan. 2009. Assessment of 29 candidate genes for milk traits in Italian dairy cattle. Italian Jorunal of Animal Science, vol. 8; p. 226.

Ajmone-Marsan, P., R. Marino, D. Perini, F. Sibella, **E.L. Nicolazzi**, L. Pariset, S. Dall'Olio, L. Fontanesi, A. Bagnato, F. Schiavini, A.B. Samorè, T. Luttmann, E. Santus, M. Blasi, N.P.P. Macciotta, and A. Nardone. 2009. Analysis of BTA6 in Bruna Italiana and Pezzata Rossa cattle assayed with 2,535 SNPs. Italian Journal of Animal Science, vol. 8; p. 183.

ACKNOWLEDGMENTS

Un sentito ringraziamento va al mio tutor, Prof. Paolo Ajmone-Marsan, che per più di 6 anni (tra tesi triennale, quinquennale e, adesso, dottorato) ha dovuto sopportare un "eretico" quantitativo in un folto gruppo di ottimi genetisti molecolari. Grazie per aver creduto in me sin dall'inizio, e per essere stata una più che ottima guida. In questi tre anni ho imparato più di quanto avrei mai sognato, riuscendo così a capire appieno il perché amo questo lavoro.

Un altrettanto sentito ringraziamento va ai miei colleghi del "Lab Animalgenetics": Ric, Lix, Marco, Raf, Franceschita, Fatima, Lorybomb, Giò, Rox (Sel) e Davide (Mol). In questi anni siete stati per me una vera e propria famiglia. So benissimo quanto sia stato fortunato ad incontrarvi lungo la mia strada.

All'inizio del mio dottorato ho avuto il piacere di essere ospite per qualche mese all'Università di Sassari, dove non solo ho incontrato ottimi scienziati, ma anche grandi amici. Ringrazio di cuore tutto il Dipartimento di Scienze Zootecniche, e in modo speciale al Prof. Nicola Macciotta, Giustino Gaspa, Corrado Dimauro, Roberto Steri, e Alberto Atzori. In quei mesi ho imparato davvero tanto, sia dal punto di vista professionale che da quello personale. Ho lasciato una parte del mio cuore con voi, in quella terra magnifica. Aiò, non sono ancora uscito dal tunnel del pane carasau!!!

I have very much appreciated and I am very grateful to the Agrisystem PhD program and the Selmol project for giving me the opportunity of meeting extraordinary scientists and great people around the world. It's impossible to describe what my Aussie experience was like, I guess I'll take that with me as long as I live. Many, many thanks to Ben Hayes, Mike Goddard, Jennie Pryce, Birgit Gredler (from BOKU University, Austria), Bolormaa, Mekonnen and Phil Bowman. Ben, your friendship, ideas, support and supervision were very important for me. I'm still missing those unique and highly scientific lunch times and coffee breaks with you all, mate!!!

I'd like to thank all the Interbull people (Joao, Jette, Hussain, Flavio, Eva, Valentina, Birgit, Mohammad and Worede) for being such a great people and excellent scientists. A special thanks to Flavio and Freddy for the guidance, support and ideas. You both have patiently guided me into the complicated world of international genetic evaluations. Thanks to Jette for her help, time and friendship, and for being the best co-organizer of the extraordinary Laser-game/Elk weekend. I miss lunch times with you as well (I'm talking about food again...as a true Italian!). Finally, I much thank my office-mates, Birgit (& kids), Mohammad and "Eyes and Personality" Worede. You all made my (Swedish) day, everyday.

Un grazie di cuore a Ste, Bia e Mauri. Grazie per avermi guidato nei miei primi passi nel mondo della genetica quantitativa e della programmazione. Con la vostra passione e la vostra amicizia mi avete invogliato a diventare uno di voi, e non avrò mai modo di ripagarvi per questo.

A mi familia, por el apoyo cotidiano... aùn a la distancia. A Teddy,Pochola y Toty, sin los cuales nada de esto hubiese sido posible, por la constante supervision y apoyo. A mis viejos, por todo, absolutamente todo. Ojalà logre ser para mis hijos lo que ustedes representan para mi. A mis hermanos, que aùn metidos en la lucha

argentina tienen siempre una palabra de aliento, un buen consejo o, simplemente, una sonrisa telematica.

To all my friends around the globe: GRACIAS, GRAZIE, THANKS, TACK, OBRIGADO, شكرا. I'll avoid naming you one by one, because I don't want to miss any of you. You know how important you are for me, thank you all.

Agli Ali, ai Topi(ni), a Eugè e ai Mattitty… i migliori compagni di viaggio di sempre, ma soprattutto di questi ultimi 3 anni. Al di là delle distanze, del lavoro e dei neo-impegni familiari, siete sempre stati sempre presenti, sia per me che per la Vale. Avete reso la mia lunga lontananza dal bel paese più leggera per entrambi. Adesso è ora di trovarsi di nuovo per festeggiare questo e altri traguardi!!!

Infine, alla donna cui ho dedicato questa tesi. In copertina dovrebbe esserci anche il tuo nome. Hai fatto sì che questi miei ultimi tre anni di viaggi e distanze fossero uno sforzo sereno, e so fin troppo bene quanto questo sia stato difficile per te. Se oggi sono felice per aver completato un obiettivo vecchio già 10 anni, lo devo a te. Sei molto più speciale di quanto tu non creda, e non vedo l'ora di godermi il resto della mia vita insieme a te.