Università degli Studi di Sassari

SCUOLA DI DOTTORATO DI RICERCA

Scienze dei Sistemi Agrari e Forestali

e delle Produzioni Alimentari

Indirizzo Scienze e Tecnologie Zootecniche

Ciclo XXIV

Development of a multivariate approach to predict Direct Genomic Values in dairy and beef cattle.

dr. Maria Annunziata Pintus

Direttore della Scuola	prof. Giuseppe Pulina
Referente di Indirizzo	prof. Nicolò P. P. Macciotta
Docente Guida	prof. Nicolò P. P. Macciotta
Tutor	dr. Giustino Gaspa

Triennio accademico 2009-2011

Alla mia famiglia

"B' a' cosas chi pro las cumprendere bi chere' tempus e isperienzia; e cosas chi cand' un' at isperienzia non las cumprende' prusu. Cosas chi pro vortuna s' irmenticana e cosas chi pro vortuna s' ammentana; e cosas chi si credene irmenticatas e chi inbezzes una die a s' improvisu torran' a conca."

Mialinu Pira -Sos Sinnos-

RINGRAZIAMENTI

Desidero ringraziare tutti coloro che, in vario modo, hanno contribuito alla realizzazione di questa tesi: ai professori, ricercatori e colleghi del corso di dottorato del Dipartimento di Scienze Zootecniche.

ACKNOWLEDGEMENT

I would like to acknowledge all people who contributed, in different ways, to this PhD thesis: Professors, researches and postgraduate fellows of the Department of Animal Science.

INDEX

CHAPTER 1

GENERAL INTRODUCTION

Selection in livestock is a process that is continuously evolving thanks to scientific achievements of genetics, molecular biology and to the increase of computational resources of machineries. Among livestock species, cattle represents a well defined example of the development of the selection process.

The introduction of the selection index (Hazel 1943) has represented one of the first approaches used to estimate the genetic merit of individuals. Breeding values of quantitative traits were estimated using phenotypes (measured on selection candidate itself and on its relatives) previously adjusted for some fixed effects. The genetic merit was then estimated by maximizing relationships between the calculated index and the true genetic value. Limitations of the genetic index are well known. The method could not account for differences in genetic levels through the years or across herds. For this reason, reliable results could be obtained only for animals that were farmed in the same nutritional and management conditions.

Best Linear Unbiased Predictions (BLUP) allowed to avoid this problem by estimating simultaneously fixed effects and the random genetic additive effect of the bull in a mixed model framework (Henderson 1975). The early applications of BLUP methods evaluated only the male genetic contribution as in the case of *Sire* and *Maternal Grandsire models.* They have represented the most popular genetic evaluation system of almost all countries until the late ‘80s. The main limitation of these earlier versions of the BLUP was that the genetic effect considered was half of the genetic additive effect of daughters because only sires were evaluated. This approach assumed that sires were mated with dams that had equal genetic merit. Thus the genetic merit of dams was not accounted for and it could bias the estimation of sires breeding values, that would be over or under estimated. Moreover, genetic effects of females were not estimated.

This issue was addressed by introducing the *Animal model* which allowed for the simultaneous evaluation of all animals (males and females) within a breed (Mrode 1996). The *Animal model* substituted the *Sire* and *Maternal Grandsire model*. Its adoption implied a huge increase of the number of equations in the mixed model. Actually, its use in the dairy breeds genetic evaluation systems had been feasible possible after the advancements of computer technologies, in the late '80s.

All the above mentioned methods analyzed the total production per lactation, obtained by joining data from test day (TD) record. They were not able to account for environmental effects as climate and feeding that could change along lactation and affect productions in different lactation phases. The *Test Day model* overcame this problem by directly analyzing daily production data (Stanton *et al.* 1992). Among *Test Day models,* the *Random Regression model* is based on the assumption that the shape of lactation curve is different for each cow allowing to evaluate the persistency of lactation and to estimate more accurately the environmental effects that affect the lactation of a cow (Schaeffer *et al.* 2000). This model requires high computational resources to solve equations and store information. Furthermore it is very sensitive to the accuracy of phenotype recording. Finally, different *Random regression models* have been proposed and the large variability of estimation methods often yields different results. Therefore the best *Random regression model* does not exist and each specific situation has to be checked.

MOLECULAR GENETICS

A relevant amount of information on the animal genome of different livestock species is currently available due to the new advancements in molecular techniques. The production of marker maps for different livestock species and the discovery of several chromosomal regions that influence quantitative traits have been followed by several approaches aimed to integrate molecular information in current breeding programs. Marker Assisted Selection programs (MAS) have been carried out to select markers linked to genes of economic interest not identifiable. The association with gene modifications responsible of phenotypic differences between individuals could be a criteria to classify markers (Dekkers 2004) and it affects the success of MAS. This classification allows to identify three kinds of markers:

Direct markers are causative mutations of a gene affecting a quantitative trait and are used in the gene assisted selection (GAS) to calculate molecular score of animals. The inheritance of gene alleles follows the inheritance of marker alleles and this is the best situation for MAS. Unfortunately this kind of markers are the most difficult to find because are the less common.

LD markers are loci in population-wide linkage disequilibrium (LD) with the functional mutation and are used in LD-MAS.

LE markers are loci in population-wide linkage equilibrium with the functional mutation and in LD within family and are used. Selection using this kind of markers is called LE-MAS.

The use of MAS programs yielded good results in France (Guillaume *et al.* 2008) where marker enhanced breeding values were more accurate than those based on pedigree.

However, commercial application of MAS programs has been limited for several reasons. The number of identified genes was very low, thus constraining the application of GAS. Moreover, maps were rather sparse: about 150-200 microsatellite markers were used in the 90's for whole genome scans, with an average distance of 20 cM between adjacent loci (Georges *et al.* 1995). A consequence was that the selection was not directly on the QTL but on the marker in LD with the QTL. Therefore marker effects had to be re-estimated frequently because LD may be different across families and it decreased across generations due to recombination resulting in high genotyping costs.

More recently, the availability of dense maps thousands of single nucleotide polymorphisms (SNP) markers made feasible the development of genome wide association studies (GWA) aimed at finding associations between genomic regions and phenotypes (Van Tassell *et al.* 2008). Moreover, they gave the physical support for the Genomic Selection (GS) (Meuwissen *et al.* 2001).

Genomic selection is a newly developed tool for genetic improvement that allows to estimate Direct Genomic breeding Values (DGV) of farm animals through the use of dense marker maps. GS relies on the segmentation of the entire genome in thousands of intervals between contiguous SNP markers and on estimation of SNP effects on a reference population (with genotype and phenotype) successively used to estimate DGV in a prediction population (without phenotypic information). Accuracy of DGV prediction is commonly measured by the correlation between DGV and phenotype used (estimated breeding values, *EBV*, daughter yield deviation, *DYD* or deregressed proof, *DRPF*) in prediction animals (figure 1).

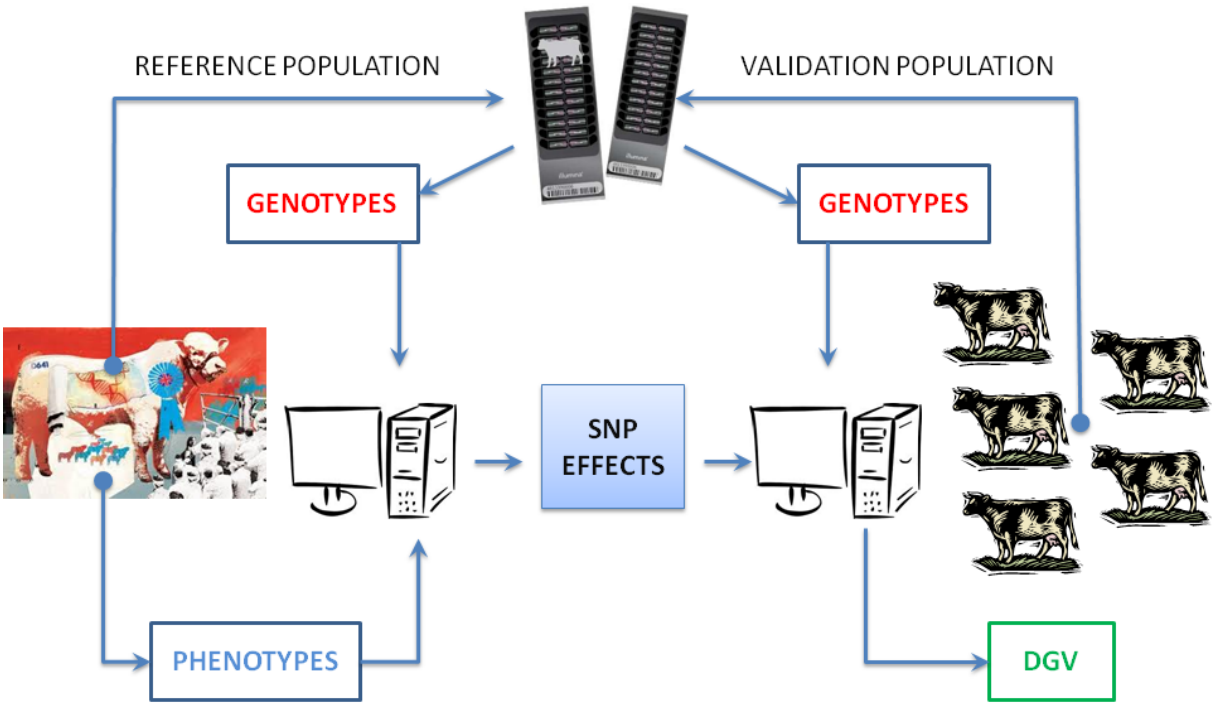

Figure 1. Genomic Selection scheme

The GS is based on the assumption that Quantitative Trait loci (QTL) should be in Linkage Disequilibrium (LD) with at least one marker in the panel used. The higher is the density of maps the lower is the chance of recombination between markers and QTLs. Recently, the advances in molecular technologies made available marker maps dense thousands of markers and relatively not expensive. At present, marker maps are available for human, bovine, porcine, ovine, equine and canine species (Illumina Inc., San Diego, CA). In the future, research aims to produce genome-wide SNP chips for a larger number of species and to augment the density of SNP chips for the species above mentioned. For instance the Illumina *HumanOmni5-Quad* (*Omni5*) BeadChip allows to genotype ~ 4.3 million of markers per sample, in cattle 54K SNP beadchips are available and largely used in different countries and

recently the 800K SNP platform with 777,000 evenly spaced SNPs has been made available, providing genotypic information at much higher density.

The aim of selection is to improve the genetic gain (ΔG), that in animal breeding programs is calculated according to the (Rendel & Robertson 1950) formula:

$$\Delta G = \frac{intensity\ of\ selection * accuracy\ of\ selection * genetic\ standard\ deviation}{generation\ interval}$$

The traditional selective process, based on phenotype recording and pedigree relationships is rather slow in cattle. In dairy cattle, many traits as milk yield and milk quality traits are sex limited and for this reason they can only be recorded on the daughters of bulls under evaluation. This selection scheme is called Progeny Test (PT) and allows to increase ΔG optimizing the accuracy of selection. The time needed to collect information for obtaining the first genetic evaluation of a bull is about 5/7years. Thus a long generation interval characterizes the PT and implies a reduction of ΔG and high costs. Moreover, prediction of breeding values for young animals has a low reliability.

GS could improve conventional selection making the process faster, more reliable and cheaper. GS programs could be useful for traits where the accuracy of selection of conventional breeding schemes is low. Examples are: low heritability traits; traits that are expensive to measure routinely and that require risky challenge testing; traits measured late in life and for which records are not available at time of selection; traits available after death of the animals as carcass quality traits for beef animals. Furthermore GS schemes would allow to reduce the generation interval because the genotype of animals could be known at birth age of animals and even before. The reduction of generation interval allows to improve the ΔG that can doubles if the same accuracy of selection is maintained. Thus GS can be a valid tool to improve conventional selection schemes that have a high accuracy of selection. For example,

progeny testing schemes have a high accuracy of selection but a high generation interval because it takes time to obtain records from daughters and to perform a cycle of selection.

The use of GS could also reduce costs of selection. Proving one bull, including housing and feeding, collection and storage of semen and test mating costs about $50,000 and the total time needed from conception to the first proof is 64 months. If 500 young bulls are tested in one year then the cost for AI will be $25 million per year and assuming that only 20 bulls are returned to service, then the cost for each of those bulls will be $1.25 million. In a GS program the total annual cost would be about $1.95 million considering costs of genotyping 2000 dams and 500 bulls, buying 20 young bulls and keeping them for 3 years. The final cost of an hypothetic GS program is the 7.8% of the cost of the traditional PT scheme so the use of GS in breeding programs can reduce dramatically costs of selection (Schaeffer 2006).

This way to estimate costs of a GS program implies the elimination of traditional PT schemes. The actual reduction of selection costs is smaller because in the practical applications it is still necessary to have phenotypic information for bulls included in the reference population (Goddard & Hayes 2007). Furthermore, not all farmers are ready to accept genomic evaluation without knowing daughter records of genotyped young bulls (Konig *et al.* 2009).

As general conclusion, it can be said that expected advantages of GS should remarkable gains in genetic progress, a more accurate control of inbreeding in the population (Daetwyler *et al.* 2007), more reliable estimation of relationship matrices (VanRaden 2008), and a considerable reduction of costs of selection (Schaeffer 2006; Konig *et al.* 2009).

A KEY POINT OF GENOMIC SELECTION: THE DATA EDITING.

Before to start any kind of analysis, a check of data needs to be done in order to create a dataset that will allow to obtain reliable estimates of DGV. Generally, the first step in GS and genome-wide association studies is represented by a data editing which allows to remove uninformative data and to clean data from scanning errors of machinery used to read the DNA sequence. This step is very important because the elimination of SNPs that do not contribute to the accuracy of estimation reduces computational work and improve stability of estimates of the effects of remaining SNPs (Wiggans *et al.* 2009). In general monomorphic SNPs and polymorphic SNPs with a minimum allele frequency (MAF) below a threshold are excluded from the analysis; this step allows to eliminate monomorphic SNPs that have been genotyped as polymorphic by the machine. The MAF threshold is fixed based on the population size, and it is higher for lower population size. Usually it ranges from 2 to 5% (Hayes *et al.* 2009b; VanRaden *et al.* 2009; Wiggans *et al.* 2009). Moreover heterozygosity of bulls genotypes can be checked to see if there are X-linked SNPs incorrectly assigned. The reason of this check is that bulls should not have heterozygous genotypes for non pseudoautosomal loci because they only have one X chromosome. The number of heterozygous SNPs linked to X chromosome allows to identify errors in the labeling of genotypes and in the sex of the animal. SNPs with a certain number of missing genotypes, parent-progeny conflicts, and which significantly deviate from Hardy-Weinberg equilibrium are usually not included in the analysis as well. In fact genotypes of SNPs that are not in Hardy-Weinberg equilibrium may have some problems in their determination or maybe those SNPs are only duplicates rather than simple genomic loci.

The data editing does not concern only SNPs but also animals can be excluded from the dataset. For example, animals that have a certain percentage of missing SNP genotypes or that

have some parent-progeny conflict are not included in the final dataset. SNP genotypes of each animal are compared with those of its parents to find if there are any conflicts. If the parent has not been genotyped or if a conflict is found, then the SNP genotypes of that animal is compared with those of all other animals to evaluate if a relationship with some other animals can be found. If a duplicate genotype could be consistent with a parent and if its age and sex could be equal to those of a real parent then the animal is considered as a putative parent. An animal with ungenotyped parents is excluded from genomic prediction if a putative parent is found. Animals that have an equal set (highly correlated) of genotypes can also be only duplicates of the same animal that have been included in the platform to check if the machinery reads the DNA in the right way. In that case, after the data editing one of those individual is included in the final dataset and usually it is the one with the lowest number of missing genotypes. Then the missing SNP of that individual are filled with those of the others duplicates or with the most frequent genotype or using specific softwares as PHASE, fastPHASE, and pLINK. Criteria for excluding SNPs and animals from the analysis are not fixed. Anyway, data editing is necessary before to start other analysis to limit errors due to the original data.

ISSUES OF GS

As previously said, GS is a new technique that has been used on simulated data ten years ago (Meuwissen *et al.* 2001) and its use on real data started a few years ago. For this reason, there are still open issues to solve as the number of individuals to be used in the reference population and the best estimation method.

Bulls in the reference population

Different studies have found that number of bulls in the reference population influences considerably accuracies of DGV. The straightforward question is: how many animals should be predictors bulls, and which animals? Calus (2010) indicated as 1,000 the minimum number of animals to include in the reference population. Simulation studies highlighted that to obtain DGV accuracies of about 0.7, few thousands of predictor bulls are needed (Hayes *et al.* 2009c). Furthermore to estimate DGV accuracies with 800 K chip about 30,000 individuals in the reference population are required to obtain acceptable values (VanRaden *et al.* 2011). In practice, different numbers of bulls are genotyped and subsequently used to estimate DGV, ranging from few hundred up to few thousand, as shown in table 1 where the number of bulls genotyped in different countries is reported .

Table 1 Number of bulls genotyped and used to estimate DGV in different studies from different countries

Country	Breed	Authors	N_tot
Australia	Holstein	(Hayes *et al.* 2009b)	798
	Holstein	(Moser *et al.* 2009)	1,945
	Holstein	(Moser *et al.* 2010)	2,624
US,Canada	Holstein	(VanRaden *et al.* 2009)	5,335
US	Holstein	(Long *et al.* 2011) (Weigel *et al.* 2009)	4,703
	Holstein	(Olson *et al.* 2011)	10,585
	BrownSwiss		1,188
	Jersey		2,370
Ireland	Holstein	(Berry *et al.* 2009)	1,209
Germany	Holstein	(Habier *et al.* 2010)	3,863
	Holstein	(Liu *et al.* 2011)	4,908
Canada	Holstein	(Schenkel *et al.* 2009)	4651
			1621
			1584
Italy	Brown Swiss	(Macciotta *et al.* 2010b)	749
	Simmental		479
	Holstein		863
Norway	NorwegianRedCattle	(Luan *et al.* 2009)	500

An option to increase the size of reference population is represented by the multi-breed approach, that consists of using animals from different breeds to build a mixed reference population. In populations that are genetically closer this approach seems to work, improving prediction accuracies. For more divergent populations, higher marker densities are required to achieve comparable accuracies (de Roos *et al.* 2009)

The reference population composition is not less important than its size. So far, not many studies have been carried out to investigate on which animals should be included in the reference population. In dairy cattle, it is very common to use proven bulls for which national breeding values are known (Moser *et al.* 2009; VanRaden *et al.* 2009; Olson *et al.* 2011). Studies have highlighted that when animals in the reference and validation population share their pedigree, reliabilities of estimates are higher (Habier *et al.* 2007). This fact suggests the use of animals that have been largely used as sires in the whole population as a strategy to compose the reference population. However, this may lead to overestimation of GS accuracy

and bad results can be obtained when new young bulls, less related to the reference population, are tested.

Methods

Different statistical methods have been proposed to estimate SNP effects in genomic selection programs. They can be grouped according to many criteria: for example the type of algorithm used for solving the model, or the theoretical assumptions on the underlying genetic model. According to the latter aspect, a distinction can be made between methods that assume an equal contribution of each marker to the genetic variance of the trait and those that assume heterogeneity of variance across chromosome segments. Actually, these two classes of methods are representative of the two main theories on the genetic architecture of complex quantitative traits: the infinitesimal and the finite locus model. The former assumes that the expression of the trait is related to a genetic background with a large number of genes of small (infinitesimal) effect. The latter, supported by the discovery of genes having a major influence on quantitative traits, is based on the assumption that there are very few genes having a large effect and a very large number of genes of small effect. The discovery of QTL of moderate effects highlights flaws in the infinitesimal model. However, also in the finite locus model some drawbacks can be envisaged because it only takes in account single genes of very large effect and not moderate as in the case of QTL (Hayes & Goddard 2001).

The use of genotypic and phenotypic information represents the simplest model to predict DGV. An additive relationship matrix can be derived and polygenic breeding values can be added to the model by pedigree information (Habier *et al.* 2010; Liu *et al.* 2011). In any case, because of the large amount of data to be processed, it's very important to choice an appropriate statistical model and to realize an effective algorithm to solve it. Actually, an estimation of a very large number of effects in a small size dataset has to be performed. For

this reason, methods that allow to handle cases where the number of marker variables greatly exceeds the number of individuals are necessary and they have to be chosen very carefully to avoid the risk of over parameterization.

Least Squares regression

The simplest model to estimate SNPs effects is represented by the least squares regression method which does not make any assumption about their distribution and treats the SNP genotypes as fixed effects. It has been proposed in some studies, but some disadvantages do not suggest its use. Actually the large amount of data to be processed implies some computational drawbacks. The main constraint is that effects of all SNPs cannot be estimated simultaneously, because there are not enough degrees of freedom (Lande & Thompson 1990). A possible solution to these problems could be the use of a stepwise approach that allows to select markers to include in the model and, in a second step, to estimate effects of these SNPs to predict DGVs (Meuwissen *et al.* 2001; Goddard & Hayes 2007; Habier *et al.* 2007; Moser *et al.* 2009). To select important SNPs to include in the model, a threshold is fixed and SNPs that are below it are excluded because they are considered to have zero effect. On the other hand, SNPs that are above the threshold are included in the model because they have a large effect. The choice of a predefined threshold determines the number of SNPs that finally are selected. There are different statistical methods to select SNPs and to fix the threshold. (Meuwissen *et al.* 2001) proposed to calculate a log-likelihood for each chromosome segment and plotted it against the position of the segment. The plot produced several likelihood peaks for chromosome that have been interpreted as an indication of a possible QTL segregating at the midpoint of the chromosome segment. Then authors used a model to simultaneously estimate the effects of the haplotypes at the QTL positions corresponding to a likelihood peak. All other haplotype effects were assumed equal to zero. Habier *et al.* (2007) and Moser *et al.*

(2009) used a stepwise procedure in which markers were included in the model one at a time. In a first step they fitted simple linear regressions and calculated t-statistics for all SNPs loci. Subsequently a threshold was fixed and the marker with the lowest P-value below the threshold was included in the model. The other markers were individually fitted together with the already included marker. Then another marker was added to the model if its P-value was the lowest of the remaining markers and below the threshold. T-statistics for SNPs included earlier were calculated if at least two marker loci were included in the model and the marker locus with the highest P-value above the threshold were excluded from the model. This calculation continued until any SNP locus could be added to the model and any SNP locus in the model could be dropped.

Whatever the method used to select SNPs, the LS approach is developed thorough different steps that could be described as follows.

In the first step a subset of SNP are selected on the basis of their significant association with the phenotype according to the model:

$$\boldsymbol{y} = \mu\mathbf{1}_{\boldsymbol{n}} + \boldsymbol{Q_i g_i} + \boldsymbol{e}$$

where **y** is the phenotype, μ is the general mean, $\mathbf{1_n}$ is a vector of ones and its dimension is the number n of records, $\boldsymbol{Q}_i$ is an incidence matrix that allocates the i^{th} SNP genotype to the phenotypic record, $\boldsymbol{g}_i$ is the vector of effects for the i^{th} SNP and e is the random residual.

In the second step a multiple linear regression is used to regress phenotypes on the previously selected SNP genotypes. Effects of *m* SNP are estimated simultaneously with the model:

$$\boldsymbol{y} = \mu\mathbf{1}_{\mathrm{n}} + \sum_{m} \mathbf{Q_i g_i} + \mathbf{e}$$

Being all other SNPs not considered and set to zero, significant effects are often overestimated (Goddard & Hayes 2007). The magnitude of the overestimation depends on the number of SNP retained. Features of LS make it a not valid tool and results from its application on simulated and real datasets confirmed this conclusion. In fact, accuracies of estimation (correlation between DGV and EBV) are in general low and always lower than those obtained with other methods (Meuwissen *et al.* 2001; Habier *et al.* 2007).

BLUP

R-BLUP model, assuming an equal contribution of each SNP to the genetic variance of the trait, is an alternative to LS to avoid the problem of overestimation and bias of SNP effects. The main difference with LS is that SNP are treated as random effects. The estimates are best linear unbiased predicted (BLUP) if QTL effects are drawn from a normal distribution with constant variance across chromosome segments and all effects could be estimated simultaneously (Goddard & Hayes 2007). The basic model is

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

Where $\boldsymbol{y}$ is the phenotype, $\boldsymbol{X}$ is the incidence matrix of a set of fixed effects b, $\boldsymbol{Z}$ is the incidence matrix that allocates SNP genotype to phenotypic record (it has dimension n individuals x m markers), $\boldsymbol{g}$ is the vector of random SNP effects and **e** is the vector of random residuals. The solution b and g are obtained from the Henderson's mixed model equations (Henderson 1985).

$$\begin{bmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \lambda \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \end{bmatrix}$$

Different ways to model covariance matrices of random effects (**G**) or residuals (**R**) have been proposed. The simplest case considers no interaction between loci. Thus **G** and **R** are assumed

to be diagonal and λ ($\lambda=\sigma_e^2/\sigma_g^2$) can have different values. (Meuwissen *et al.* 2001) proposed σ_g^2 as the total additive genetic variance divided by the number of fitted SNP as $\lambda = \frac{6_e^2}{6_g^2/n}$.

In R-BLUP all random effects have a common variance and SNPs with a large effect tend to be overestimated influencing the accuracy of estimations. The overestimation of SNP effects has a smaller influence on results than in LS and obtained accuracies are always better than those of LS. If marker effects are normally distributed and the variance is constant, R-BLUP performances are similar to those obtained with other methods that are based on the assumption that variance differs between SNPs (Calus 2010). If a polygenic effect for all animals in the population is considered the mixed model becomes:

$$\boldsymbol{y} = \boldsymbol{Xb} + \boldsymbol{Zu} + \boldsymbol{Wg} + \boldsymbol{e}$$

Where $\boldsymbol{y}$ is the vector of phenotypes $\boldsymbol{Z}$ is the incidence matrix that allocates the animal to the phenotypic records and $\boldsymbol{u}$ is the vector of polygenic effects, $\boldsymbol{W}$ is the incidence matrix of marker genotypes and $\boldsymbol{g}$ is the vector of marker effects. Then the solution of this model is

$$\begin{bmatrix} \widehat{\boldsymbol{b}} \\ \widehat{\boldsymbol{u}} \\ \widehat{\boldsymbol{g}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'Z} & \boldsymbol{X'W} \\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \lambda \boldsymbol{A}^{-1} & \boldsymbol{Z'W} \\ \boldsymbol{W'X} & \boldsymbol{W'Z} & \boldsymbol{W'W} + \lambda \boldsymbol{I} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{X'y} \\ \boldsymbol{Z'y} \\ \boldsymbol{W'y} \end{bmatrix}$$

Where $\boldsymbol{A}^{-1}$ is the inverse of additive relationship matrix and $\lambda=\sigma^2_e/\sigma^2_u$.

A Genomic BLUP (G-BLUP) model has been proposed as an improvement of R-BLUP where the genomic relationship matrix (**G**) calculated from marker data replaces the pedigree-based matrix ($\boldsymbol{A}$). The genomic relationship matrix should be more informative than the pedigree-based matrix because it measures the real fraction of alleles shared and not the

expected fraction of alleles shared by descent as in the case of the pedigree-based matrix (VanRaden 2008; Goddard 2009; Clark *et al.* 2011).

Genomic relationship matrix could be calculated in different ways, For example (VanRaden 2008) proposed three methods to use to obtain the genomic relationship matrix. One of this methods uses the formula

$$\boldsymbol{G} = \frac{\boldsymbol{zz'}}{2\sum p_i(i-pi_i)}$$

which makes **G** analogous to the numerator relationship matrix **A** through the division by $2\sum p_i\,(i - p_i)$. In this expression

$$\boldsymbol{Z} = (\boldsymbol{M} - \boldsymbol{P})$$

where ***P*** contains the allelic frequencies of the marker expressed as 2(p_i-0.5), ***M*** is the matrix that specifies which marker alleles each individual inherited. If it is assumed that the parameterization adopted in ***M*** to indicate homozygote, heterozygote and other homozygote is -1,0,1 respectively then the diagonal elements of ***MM'*** matrix count how many homozygous loci for each individuals, and off-diagonals the number of alleles shared by relatives. In this case the equation of mixed model becomes:

$$\begin{bmatrix}\widehat{\boldsymbol{b}}\\ \widehat{\boldsymbol{u}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X'X} & \boldsymbol{X'Z}\\ \boldsymbol{Z'X} & \boldsymbol{Z'Z} + \lambda\boldsymbol{G^{-1}}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{X'y}\\ \boldsymbol{Z'y}\end{bmatrix}$$

where **û** in this case is the DGV equivalent to the DGV obtained from the summation, for all chromosome segment, of marker effects estimated using G-BLUP. Problems inverting the ***G*** matrix could occur because in some cases it could be singular (VanRaden 2008).

Bayesian methods

All markers giving an equal contribution to the genetic variance of the trait could be an unrealistic assumption. For this reason, methods that consider different amounts of variance explained by different loci have been proposed as an alternative to the methods above described. It seems that those methods are based on a more realistic assumption than those that assume that the variance due to each locus is fixed as in the BLUP methods (Meuwissen *et al.* 2001). The superiority of this assumption can be justified by results from QTL mapping experiments. The predicted distribution of QTL is consistent with the hypothesis of few genes of large effect and many genes of small effect. For example, with a meta-analysis approach it has been estimated that the number of QTL affecting a generic quantitative trait ranges from 50 to 100 (Hayes & Goddard 2001). Other examples confirm the theoretical distribution of few genes of large effect and many genes of small effect can be cited. An example is the case of the polymorphism K232A of the diacylglycerol acyl transferase 1 (DGAT1) gene identified on bovine chromosome 14, that explains the 50% of the variance of milk fat content trait (Grisart *et al.* 2004). Even the polymorphism F279Y of growth encountered in the population. In the proposed approach, the hormone receptor gene (GHR) gene identified on bovine chromosome 20 is a good example to cite because it explains about 10% of the variance of trait milk protein content (Blott *et al.* 2003). Bayesian models are usually based on the assumption that different genes contribute in different proportions to the genetic variance of traits. Their use allows to select SNPs having a significant effect on the considered trait. Different approaches can be used also within the Bayesian framework as Bayes A, Bayes B methods (Meuwissen *et al.* 2001), Bayes LASSO (Weigel *et al.* 2009) . The Bayes A approach develops in two submodels that consider the data and the variance of chromosome segments, respectively. The first submodel is similar to a BLUP model, but it allows variance

to differ between segments. Variance is estimated by the latter submodel that combines information from the data with those from the prior distribution of the variances. A Gibbs-Sampling algorithm is used to estimate SNP effects and their variance simultaneously. A Bayes B method has been also developed as an improvement of Bayes A. Bayes B uses a higher density prior and a Metropolis-Hastings algorithm, instead of the Gibbs sampling.(Meuwissen *et al.* 2001). The main difference between the two Bayes methods is that in Bayes A SNP effects can have SNP variances close to zero whereas in Bayes B a certain number of SNP can have a variance equal to zero and are excluded from the analysis allowing to reduce the number of variable that are actually taken in account in the model for predictions.

Even though different methods to estimate DGV have been proposed and are still investigated, BLUP model is adopted by most of countries that are actually running GS programs. The advantage of using BLUP model consist in a limited need of computational resources and time compared to other methods. Furthermore DGV accuracies obtained with BLUP are only slightly lower than those obtained with other approaches in most of traits considered. The infinitesimal assumption of BLUP model can be considered close to the reality for most of traits.

Methods to reduce the number of predictors

One of the most important issues of GS is the huge unbalance between the number of observations (phenotypic information, genotyped individuals) and the number of predictors (genotypic information, SNP markers).

This issue is of particular interest in small populations where the number of markers largely exceeds the number of genotyped bulls, as local or beef breeds (Garrick 2011). However it

affects the estimation process, especially as far as computational time is concerned, also in larger populations. With the advent of the 800K chip, the n>>p problem will interest almost all cattle populations, even those of very large size and wide diffusion as the Holstein cattle.

Different approaches have been proposed to overcome this drawback.

- Methods that directly reduce the number of the original variables as the above mentioned Bayes B (Meuwissen *et al.* 2001)
- Methods of SNP preselection before the estimation step
- Multivariate statistics that substitute the markers with a smaller number of new variables that are linear combination of the original variables

Subsets of SNPs can be selected in different ways. Moser *et al.* (2010) proposed to use the regression coefficients b of Partial Least Square Regression (PLSR) to select relevant SNPs. SNPs were ranked by their absolute value of b and those with the lowest b value were eliminated from the list. The final set of SNPs used in the analysis has been obtained in different steps in which regression coefficients have been recomputed each time because the magnitude of b was influenced by the LD between markers. Authors obtained different subsets with a different number of SNPs selected using four different strategies. Basically they used subsets that contained the highest ranked SNP for each individual trait or evenly spaced SNPs. Results showed that subsets containing ~ 3,000 to 5,000 SNPs provided accuracies of genomic evaluations that were about 90% of those obtained with all available markers. SNPs can also be sorted by their effects and only SNPs with the largest effect are included in the model in order to reduce the number of original variables (Weigel *et al.* 2009).

Anyway apart from the method used to select the subsets of SNPs, either Bayesian or preselection of SNPs, the final dataset is trait dependent because markers included in it are chosen because they have a relevant effect on a specific trait and not on another.

The use of multivariate techniques as Principal Component Analysis (PCA) and Partial Least Squares Regression is a tool to reduce the number of predictors without affecting the accuracy of estimation as found in simulated (Solberg *et al.* 2009; Macciotta *et al.* 2010a) and real (Moser *et al.* 2010; Long *et al.* 2011) data. PCA and PLSR are based on the reduction of the number of predictors in a small number of linear combinations of the predictors that here we call latent components (LC) for PLSR and principal components (PC) for the PCA.

Principal Component analysis

PCA synthesizes information contained in a set of n observed variables ($M_{1,}\ldots$, M_n) by originating a new set of k variables (k<<n) that are orthogonal and named PC ($PC_{1,}\ldots$, PC_k). PCs are calculated from the eigen decomposition of the correlation matrix **M** and each PC is a linear combination of the observed variables

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

where coefficients α_{ij} are the elements of the eigenvector corresponding to jth eigenvalue. Principal components are usually extracted in a descending order of the corresponding eigenvalue that measures the amount of variance of original variables explained by each PC (Macciotta *et al.* 2010a).

The PCA is carried out on the SNP data matrix **M** with *m* rows (number of animals in the dataset) and *n* columns (number of SNP markers) where each element corresponds to the genotype at the j_{th} marker for the i_{th} individual. The sum of PC eigenvalues is used to

determine the number of PC to use for further analysis that is not fixed. A criteria could be to use the number of PC for which the highest correlation between DGV and EBV is obtained (Solberg *et al.* 2009). Furthermore the use of PCA approach in GS allows also to model the variance structure of predictors in the BLUP normal equations by using eigenvalues as variance priors (Macciotta *et al.* 2010a).

Partial Least Square Regression

The PLSR is a very useful statistical technique when the number of predictors largely exceed the number of variables and also when there is a high correlation between predictors as in case of strong collinearity (Dimauro *et al.* 2011). The basic model is:

$$\boldsymbol{Y} = \boldsymbol{XB} + \boldsymbol{E}$$

where $\boldsymbol{Y}$ is an n x m response matrix, $\boldsymbol{X}$ is an n x p design matrix, $\boldsymbol{B}$ is an n x m regression coefficient matrix, and $\boldsymbol{E}$ is an n x m residual term. PLSR consists in the simultaneous decomposition of the matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ into a set of new variables. The extraction of new variables aims to maximize the covariance between $\boldsymbol{X}$ and $\boldsymbol{Y}$ and to minimize the covariance between variables inside each matrix. Extracted new variables account for a decreasing proportion of original variance and are linear combinations of predictors.

The main difference between PLSR and PCA is in the way used to derive synthetic variables. PLSR maximizes the covariance between the set of LC and the response variables (phenotypes) whereas PCA maximizes the proportion of total original variance explained by the set of PC. Therefore PC can be considered trait independent because extracted variables resume all marker information and without consider phenotypes used. On the contrary, LC are derived simultaneously from information both on markers and phenotypes and for this reason they cannot be considered trait independent.

ACCURACIES OF GENOMIC PREDICTIONS

Accuracy of genomic prediction is usually calculated as correlation or square correlation between DGV and the true breeding value (TBV) of the animal. However TBV is available only for simulated data and in real data it is substituted by the national estimated breeding value (EBV), deregressed proof (DRPF) or daughter yield deviation (DYD). To further validate results of estimations, other measurements can be calculated as the bias of estimation measured by the regression coefficient between phenotype and DGV($b_{EBV,DGV}$), or the mean square of prediction (MSEP) and its decomposition.

As previously said, accuracy of prediction is affected by many factors as the number of bulls in the reference population and the number of markers, the statistical model adopted, the level of LD, and the heritability of considered traits. For this reason it's very difficult to compare different studies. Table 2 reports DGV accuracies obtained by several authors using real data on cattle. Methods described above have been used and accuracies have been reported as correlations between DGV and the predicted variable (EBV, DRPF or DYD), except for works where accuracies of blended genomic breeding values were reported instead DGV accuracies. Different number of reference bulls (on average 2,364 ranging from 335 to 8,022) from different breeds were used across and, sometimes, within works. In most of cases bulls were genotyped using a 54K chip and the average number of SNPs retained after data editing was 35,952 (±10,162). In general least square (LS-FR) approach performed worse than all other methods, if the same number of animals and SNPs are considered. The use of Bayesian methods didn't result in higher accuracies except for (Habier *et al.* 2010), were minimum values of accuracies obtained by using BAYES-B were slightly higher than those obtained with BLUP. This fact confirms the substantial equivalence of the two estimation methods in most of traits. In general, among studies where BLUP and Bayesian methods have been used,

lowest accuracies were obtained for lowest number of reference bulls (Luan *et al.* 2009; Macciotta *et al.* 2010b). Therefore, also methods that used a minor number of predictors different from SNPs, as PC, performed better when the reference population was larger. Number of SNPs used seems to not affect accuracies.

Table 2. Accuracy of genomic predictions, obtained by different authors using different number of reference bulls and SNPs and different estimation methods.

Authors	Ref-Bulls[1]	SNPs[2]	Method	Accuracy[3]
(Moser *et al.* 2009)	1,239	7,237	LS-FR	0.43-0.53[a]
			BLUP	0.56-0.71[a]
			BAYES A	0.56-0.71[a]
			PLSR-BLUP	0.55-0.70[a]
			SVR	0.58-0.72[a]
(VanRaden *et al.* 2009)	3,576	38,416	LINEAR(GBLUP)	0.45-0.74[b]
	1,759		NONLINEAR	0.43-0.79[b]
(Hayes *et al.* 2009a)	781 (1068)	39,048	GBLUP	0.49-0.62(0.45-0.62)[c]
			BAYES SSVS	0.47-0.70(0.45-0.70)[c]
			BAYES A	0.47-0.71(0.44-0.69)[c]
(Luan *et al.* 2009)	400	18,991	BLUP	0.15-0.62(0.19-0.61)[d]
			MIXTURE	0.13-0.60(0.19-0.61)[d]
			BAYES B	0.13-0.61(0.19-0.60)[d]
(Weigel *et al.* 2009)	3,305	32,518	BAYES LASSO	0.61(0.43-0.57†;025-0.54‡)[e]
(Habier *et al.* 2010)	~2096	40,588	BLUP	0.44-0.68(0.17-0.38)[f]
			BAYES-B	0.50-0.68(0.29-0.47)[f]
(Macciotta *et al.* 2010b)	604	40,658	PCA-BLUP	0.21-0.61[g]
	524	37,254		0.18-0.54[g]
	335	40,179		0.28-0.46[g]
(Harris & Johnson 2010)	5,212	42,302	BLUP	0.48-0.57(0.51-0.60)[h]
(Olson *et al.* 2011)	8,022	43,382	NONLINEAR	0.26-0.70(0.32-081)[i]
	1,959			0.39-0.68(0.47-0.74)[i]
	1,959			0.10-0.56(0.20-0.63)[i]
(Long *et al.* 2011)	3,305	32,518	PCR	0.68[l]
			PLS	0.67[l]
			Supervised PCR I	0.55[l]
			Supervised PCR II	0.54-0.59[l]
			Sparse PLS	0.55-0.59[l]
(Liu *et al.* 2011)	3,676	45,181	BLUP	0.49-0.77 (0.61-0.70)[m]
(Berry *et al.* 2009)	945	42,598	LINEAR (GBLUP)	0.33-0.83
(Schenkel *et al.* 2009)	1097 [4127]	38416	LINEAR (GBLUP)	0.34-0.72[o] [0.36-0.76][o]

1) number of animals in the reference population only
2) number of SNP after editing procedure (3 chip set 54k, 25k, 9k were used)
3) minimum and maximum DGV accuracies across productive and functional traits and different studies and methods

[a] range of DGV (MBV) accuracy of prediction population for Australian economic (ASI) and protein percentage (PPT) index and
[b] accuracy were expressed as $R^2$by authors (here as $\sqrt{R^2}$) and the range is across production and functional trait
[c]range of DGV accuracy calculated for Australian Holstein when Holstein or(Holstein +Jersey) population in the reference set was used with a multi-breed approach.
[d] range of accuracy for milk production trait estimated using 5 fold cross validation for cohort of animal whose phenotypes were masked on the basis of year of PT or (5 fold cross validation of random animal) to design the reference and prediction population
[e] values of DGV accuracy using whole set of SNPs or (range of accuracy when selecting smaller subsets of SNPs of largest effect†, or evenly spaced in the genome ‡)
[f] minimum and maximum of DGV accuracy for different constrain of additive relationship when building the reference set (DGV due to LD) for milk yield fat yield, protein yield and SCS in German Holstein
[g] range of DGV accuracy for Italian Holstein, Italian Brown Swiss, and Italian Simmenthal building the reference set sorting the bulls by year of birth and using 2,564, 2,257, and 2,476 PC respectively.
[h]range of DGV accuracy in NZ Hostein Holstein and NZ Jersey both not blending the DGV with Parent Average information and (using a blending approach)
[i] range of DGV accuracy for Holstein, Yersey and Brown Swiss with reference animal August 2006 (April 2010) used to compute genomic PTA for validation animals. Accuracy were expressed as $R^2$by authors (here as $\sqrt{R^2}$) and the range is across production and functional trait
[l] DGV accuracy for milk yield in Holstein when using 3000 PCs (PCR), 15 latent components (PLS), ~ 1000 SNPs selected (supervised PCR I), 300 and 500 SNPs selected (supervised PCR II), 272 and 684 SNPs selected (sparse PLS)
[m]minimum and maximum DGV accuracy calculated for different traits (including 20% of residual polygenic variance)
[n]minimum and maximum accuracy of genomic selection(expressed as correlation of EBV on genomic and blended EBVs)
[o]minimum and maximum GPA(Final genotype enhanced PA, computed by an index that combined PA, DGV and the sunset PA, using the respective reliabilities of the three components to determine the appropriate index weights) accuracy calculated for different traits using predictors bulls with domestic proofs only or [official] proof.

OBJECTIVE OF THE THESIS

The overall objective of the present thesis was to develop a method able both to reduce the dimensionality of predictors for the estimation of DGV in cattle populations of limited sizes and to keep the same accuracies of methods that use directly all SNP genotypes available.

In particular, the Principal Component Analysis has been used to reduce the dimensionality of predictors. The method has been tested on three Italian cattle breeds with different production aptitudes, dairy and dual purpose, and population size. The analysis was carried out on dairy, beef and type traits.

REFERENCES

Berry D.P., F. K. & B.L. H. (2009) Genomic Selection in Ireland. *Interbull Bull*.

Blott S., Kim J.J., Moisio S., Schmidt-Kuntzel A., Cornet A., Berzi P., Cambisano N., Ford C., Grisart B., Johnson D., Karim L., Simon P., Snell R., Spelman R., Wong J., Vilkki J., Georges M., Farnir F. & Coppieters W. (2003) Molecular dissection of a quantitative trait locus: A phenylalanine-to-tyrosine substitution in the transmembrane domain of the bovine growth hormone receptor is associated with a major effect on milk yield and composition. *Genetics* **163**, 253-66.

Calus M.P.L. (2010) Genomic breeding value prediction: methods and procedures. *Animal* **4**, 157-64.

Clark S.A., Hickey J.M. & Van der Werf J.H.J. (2011) Different models of genetic variation and their effect on genomic evaluation. *Genetic Selection Evolution* **43:18**.

Daetwyler H.D., Villanueva B., Bijma P. & Woolliams J.A. (2007) Inbreeding in genome-wide selection. *Journal of Animal Breeding and Genetics* **124**, 369-76.

de Roos A.P.W., Hayes B.J. & Goddard M.E. (2009) Reliability of Genomic Predictions Across Multiple Populations. *Genetics* **183**, 1545-53.

Dekkers J.C.M. (2004) Commercial application of marker- and gene-assisted selection in livestock: Strategies and lessons. *J. Anim Sci.* **82**, E313-28.

Dimauro C., Steri R., Pintus M.A., Gaspa G. & Macciotta N.P.P. (2011) Use of partial least squares regression to predict single nucleotide polymorphism marker genotypes when some animals are genotyped with a low-density panel. *Animal* **FirstView**, 1-5.

Garrick D.J. (2011) The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genetics Selection Evolution* **43**.

Georges M., Nielsen D., Mackinnon M., Mishra A., Okimoto R., Pasquino A.T., Sargeant L.S., Sorensen A., Steele M.R., Zhao X.Y., Womack J.E. & Hoeschele I. (1995) Mapping Quantitative Trait Loci Controlling Milk-Production in Dairy-Cattle by Exploiting Progeny Testing. *Genetics* **139**, 907-20.

Goddard M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-57.

Goddard M.E. & Hayes B.J. (2007) Genomic selection. *Journal of Animal Breeding and Genetics-Zeitschrift Fur Tierzuchtung Und Zuchtungsbiologie* **124**, 323-30.

Grisart B., Farnir F., Karim L., Cambisano N., Kim J.J., Kvasz A., Mni M., Simon P., Frere J.M., Coppieters W. & Georges M. (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**, 2398-403.

Guillaume F., Fritz S., Boichard D. & Druet T. (2008) Estimation by simulation of the efficiency of the French marker-assisted selection program in dairy cattle. *Genet Sel Evol* **40**, 91-102.

Habier D., Fernando R.L. & Dekkers J.C.M. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97.

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

Harris B.L. & Johnson D.L. (2010) Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *Journal of Dairy Science* **93**, 1243-52.

Hayes B. & Goddard M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33**, 209-29.

Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K. & Goddard M.E. (2009a) Accuracy of genomic breeding values in multi-breed dairy cattle populations. *Genetics Selection Evolution* **41**.

Hayes B.J., Bowman P.J., Chamberlain A.J. & Goddard M.E. (2009b) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433-43.

Hayes B.J., Visscher P.M. & Goddard M.E. (2009c) Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). *Genetics Research* **91**, 143-.

Hazel L.N. (1943) The Genetic Basis for Constructing Selection Indexes. *Genetics* **28**, 476-90.

Henderson C.R. (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics* **31**, 423-47.

Henderson C.R. (1985) Best Linear Unbiased Prediction Using Relationship Matrices Derived from Selected Base Populations. *Journal of Dairy Science* **68**, 443-8.

Konig S., Simianer H. & Willam A. (2009) Economic evaluation of genomic breeding programs. *Journal of Dairy Science* **92**, 382-91.

Lande R. & Thompson R. (1990) Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. *Genetics* **124**, 743-56.

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

Long N., Gianola D., Rosa G.J.M. & Weigel K.A. (2011) Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*, no-no.

Luan T., Woolliams J.A., Lien S., Kent M., Svendsen M. & Meuwissen T.H.E. (2009) The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* **183**, 1119-26.

Macciotta N.P.P., Gaspa G., Steri R., Nicolazzi E.L., Dimauro C., Pieramati C. & Cappio-Borlino A. (2010a) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *Journal of Dairy Science* **93**, 2765-74.

Macciotta N.P.P., Pintus M.A., Steri R., Pieramati C., Nicolazzi E.L., Santus E., Vicario D., van Kaam J.T., Nardone A., Valentini A. & Ajmone-Marsan P. (2010b) Accuracies of direct genomic breeding values estimated in dairy cattle with a principal component approach. *Journal of Dairy Science* **93**, 532-3.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

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

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

Mrode R.A. (1996) Linear Models for the Prediction of Animal Breeding Values. *CAB International, Wallingford, UK.*

Olson K.M., VanRaden P.M., Tooker M.E. & Cooper T.A. (2011) Differences among methods to validate genomic evaluations for dairy cattle. *Journal of Dairy Science* **94**, 2613–20.

Rendel J. & Robertson A. (1950) Estimation of genetic gain in milk yield by selection in a closed herd of dairy cattle. *Journal of Genetics* **50**, 1-8.

Schaeffer L.R. (2006) Strategy for applying genome-wide selection in dairy cattle. *Journal of Animal Breeding and Genetics* **123**, 218-23.

Schaeffer L.R., Jamrozik J., Kistemaker G.J. & Van Doormaal B.J. (2000) Experience with a test-day model. *Journal of Dairy Science* **83**, 1135-44.

Schenkel F.S., Sargolzaei M., Kistemaker G., Jansen G.B., Sullivan P., Van Doormaal B.J., Van Raden P.M. & Wiggans G.R. (2009) Reliability of genomic evaluation of holstein cattle in canada. *Interbull Bull* **39**.

Solberg T.R., Sonesson A.K., Woolliams J.A. & Meuwissen T.H.E. (2009) Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution* **41**, -.

Stanton T.L., Jones L.R., Everett R.W. & Kachman S.D. (1992) Estimating milk, fat, and protein lactation curves with a test day model. *Journal of Dairy Science* **75**, 1691-700.

Van Tassell C.P., Smith T.P.L., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T., Haudenschild C.D., Moore S.S., Warren W.C. & Sonstegard T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nat Meth* **5**, 247-52.

VanRaden P., O'Connell J., Wiggans G. & Weigel K. (2011) Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 10.

VanRaden P.M. (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* **91**, 4414-23.

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

Weigel K.A., de los Campos G., Gonzalez-Recio O., Naya H., Wu X.L., Long N., Rosa G.J.M. & Gianola D. (2009) Predictive ability of direct genomic values for lifetime net merit of Holstein sires using selected subsets of single nucleotide polymorphism markers. *Journal of Dairy Science* **92**, 5248-57.

Wiggans G.R., Sonstegard T.S., Vanraden P.M., Matukumalli L.K., Schnabel R.D., Taylor J.F., Schenkel F.S. & Van Tassell C.P. (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *Journal of Dairy Science* **92**, 3431-6.

CHAPTER 2

PREDICTION OF DIRECT GENOMIC VALUES FOR DAIRY TRAITS IN ITALIAN BROWN AND SIMMENTAL BULLS BY USING A PRINCIPAL COMPONENT APPROACH.

Submitted to Journal of Dairy Science

ABSTRACT

The huge number of markers in comparison with the phenotypes available represents one of the main issues in genomic selection. In this work, principal component analysis is used to reduce the number of predictors for calculating direct genomic breeding values (DGV). Bulls of two cattle breeds farmed in Italy (749 Brown and 479 Simmental) were genotyped with the 54K Illumina beadchip. After data editing, 37,254 and 40,179 SNP were retained for Brown and Simmental respectively. Principal component analysis carried out on SNP genotype matrix extracted 2,257 and 2,466 new variables, respectively. Bulls were sorted by birth year or randomly shuffled to create reference and prediction populations. The effect of principal component on polygenic EBV in reference animals was estimated with a BLUP model. Results were compared to those obtained by using SNP genotypes as predictors either with BLUP or Bayes_A estimation methods. Traits considered were milk, fat and protein yield, fat and protein percentage, somatic cell score, udder score, and economic index. No substantial differences in correlations between DGV and EBV were observed between the three methods in both breeds. The approach based on the use of principal components showed the lowest prediction bias. The PC method allows for a relevant reduction (>95%) in the number of independent variables when predicting DGV, with a huge decrease in calculation time and without losses in accuracy.

Key words: SNPs, genomic selection, principal component analysis, accuracy

INTRODUCTION

Advancements in genome sequencing technology have been implemented into high throughput platforms able to genotype simultaneously tens of thousands SNP markers distributed across the whole genome of livestock species (Van Tassell *et al*. 2008). Dense marker maps are nowadays used in cattle breeding for genome-wide association studies (Cole *et al*. 2009, Price *et al*. 2006) and for predicting genomic-enhanced breeding values (GEBV) of candidates to become sires and dams in Genomic Selection (GS) programs (Meuwissen *et al*. 2001). The basic frame of genomic selection involves two steps. In the first, effects of chromosomal segments are estimated in a set of reference animals, having known phenotypes and SNP genotypes. Then estimates are used to predict Direct Genomic Values (DGV) in animals for which only marker genotypes are known. DGV are usually blended with other measures of genetic merit as official parent average or pedigree index to obtain the final GEBV (Ducroqc *and* Liu 2009; VanRaden *et al*. 2009). GS programmes have already been implemented in different countries to evaluate young bulls entering progeny testing, achieving reliabilities higher than those of the pedigree index (Hayes *et al*. 2009a, VanRaden *et al*. 2009). Expected benefits of the GS are the reduction of generation intervals, the increase of EBV accuracies for female side of the pedigree and a cost reduction for progeny testing (Konig *et al*. 2009, Schaeffer, 2006).

However, several issues are still to be addressed in GS. Examples are the assessment of the frequency with which marker effects must be re-estimated along generations (Solberg *et al*. 2009), the evaluation of the impact of population structure on estimated effects (Habier *et al*. 2010), the choice of the most suitable mathematical model and dependent variable for the estimation step (Guo *et al*. 2010). Apart from situations in which the number of genotyped animals is quickly approaching or overcoming the number of marker used, as the USA

genomic project (VanRaden *and* Sullivan, 2010), the huge imbalance between predictors and observations still represents the main constraint to the implementation of GS programmes, especially for breeds other than Holstein.

Some authors suggest to combine data from different populations of the same breed or from different breeds in a common reference set, both within and across countries (Boichard *et al*. 2010). Reports on simulated and real data show some increases in DGV accuracies, but results are strongly dependent on the genetic similarity between breeds and on the trait analyzed (de Roos *et al*. 2009, Hayes *et al*. 2009b) and ad hoc models need probably to be developed.

A different strategy is based on the reduction of the number of predictors used in the estimation equations. A straightforward approach is to perform a preliminary selection of markers on the basis of their relationship with the phenotype or of their chromosomal location (Hayes *et al*. 2009a, Moser *et al*. 2010, Vazquez *et al*. 2010). An alternative is represented by the Bayes B method that retains markers with non-zero effect on phenotypes directly during the estimation step (Meuwissen *et al*. 2001, VanRaden, 2008). Other approaches of SNP selection have been proposed mainly for genome wide association analyses (Aulchenko *et al*. 2007, Gianola *et al*. 2006, Gianola *and* van Kaam, 2008, Long *et al*. 2007). In all the above mentioned methodologies, SNP selection is based on their relevance to the considered phenotype. Thus specific sets of markers may be required for different traits.

An alternative to marker selection for reducing predictor dimensionality is represented by their synthesis via multivariate reduction techniques. In particular, principal component analysis (PCA) and Partial Least Squares Regression (PLSR) have been proposed for estimating DGV (Solberg *et al*. 2009). Actually, in the PLSR approach the extraction of latent variables from predictors is carried out by maximizing their correlation with the dependent

variable(s) . Thus the reduction of the system dimension is still based on the magnitude of the predictor effects on the considered trait. On the contrary, the PCA is entirely based on the factorization of the SNP (co)variance (or correlation) matrix. This technique allowed for huge reduction of the number of independent variables (>90%) in the estimation of DGV achieving accuracies comparable to those obtained using SNP genotypes (Macciotta *et al*. 2010, Solberg *et al*. 2009). Compared to other approaches of predictor reduction, PCA limits the loss of information because each SNP is involved in the composition of each PC. Moreover, extracted principal components are orthogonal, thus avoiding multicollinearity problems. The PCA approach allow also to model the variance structure of predictors in the BLUP normal equations by using eigenvalues as variance priors (Macciotta *et al*. 2010). PCA has been also used in Genome-wide association studies to reduce the number of dependent variables (Bolormaa *et al*. 2010).

In this paper, the principal component analysis is used to reduce the number of predictors in the calculation of direct genomic values for dairy traits on real data in Italian Brown and Simmental bulls.

MATERIAL AND METHODS

Data

SNP Genotypes were generated within the SELMOL project funded by the Italian Ministry of Agriculture. A total of 775 Italian Brown and 493 Italian Simmental bulls were genotyped at 54,001 SNP loci with the Illumina Bovine SNP50TM bead-chip. Considering the limited size of the sample, the priority in the edit was to maintain the largest number of bulls as possible. A stringent selection was performed on markers. Edits have been based on the percentage of missing data (<0.025), Mendelian inheritance conflicts, absence of heterozygous loci, MAF (<.05), deviance from Hardy-Weimberg equilibrium (<0.01) (Wiggans *et al*. 2009). Edits on animals were based on the number of missing genotypes (<1,000), and on inconsistencies in the Mendelian inheritance (some father-son pairs were included). An overall accuracy higher than 99% was obtained by double-genotyping some animals. A summary of the initial and final number of bulls and SNPs, together with the magnitude of the different elimination steps is reported in table 1.

Table 1. Number of animals and markers discarded in the different edit steps.

Breed	Repeatability	Mendelian Inheritance	Missing	MAF	HW equilibrium	Final
			Animals			
Brown	17	3	6			749
Simmental	6	2	6			479
			Markers			
Brown		23	1,118	15,046	560	37,254
Simmental		21	999	12.215	587	40,179

In the final data, missing SNP alleles were replaced by the most frequent allele at that specific locus.Phenotypes used were estimated polygenic breeding values (EBVs) provided by national breeders associations. Traits considered were milk, fat and protein yield (kg), fat and protein percentages, somatic cell score, udder score, economic index.

Animals were sorted by year of birth and the dataset split into reference (REF) and prediction (PRED) subsets, comprising older and younger animals, respectively. Three ratios of reference-prediction animals were considered (0.70:0.30, 0.80:0.20, 0.90:0.10). Table 2 reports the number of bulls for REF and PRED data sets for the two breeds.

Table 2. Number of animals in reference and prediction data sets for the three scenarios in the three different breeds.

	70:30		80:20		90:10	
Breed	Reference	Prediction	Reference	Prediction	Reference	Prediction
Brown	524	225	599	150	674	75
Simmental	335	144	383	96	431	48

The distribution of the year of birth bulls in the different breeds is depicted in figure 1.

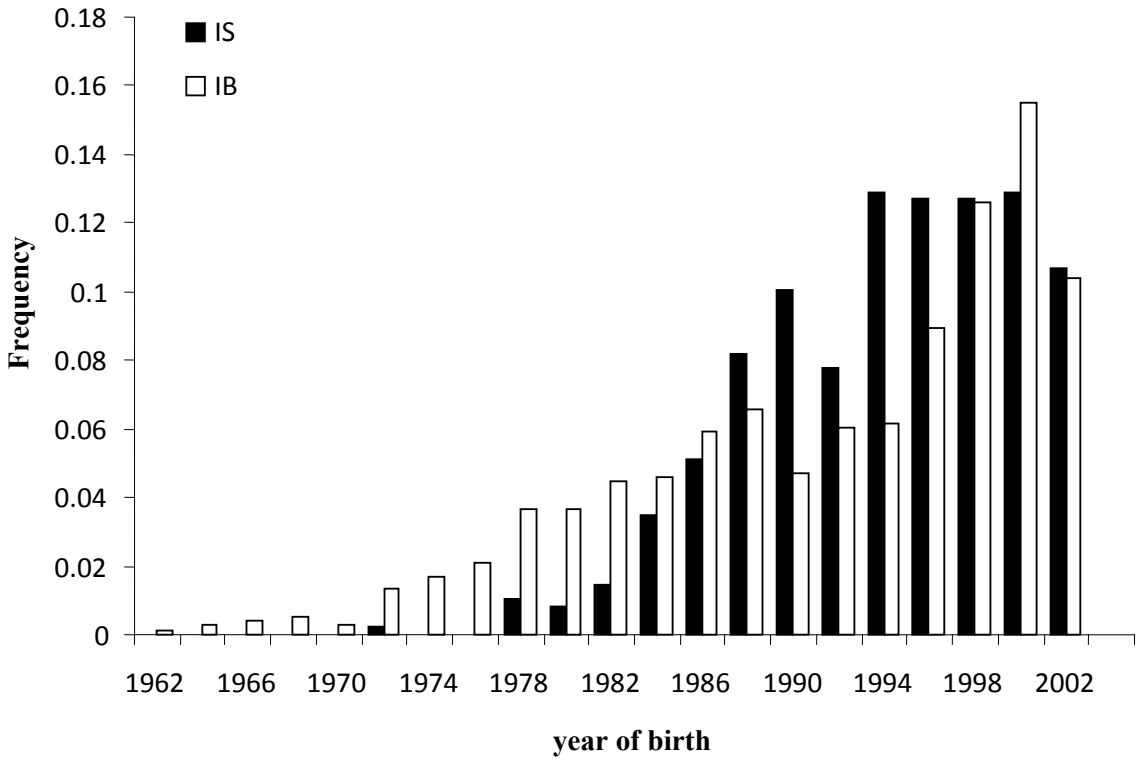

Figure 1.Distribution of number of bulls within year of birth.

Statistical Models

Principal component analysis was used to extract latent variables from the SNP data matrix M with m rows (m= number of individuals in the entire data set, i.e. REF plus PRED) and n columns (n=number of SNP retained after edits). Each element (i,j) corresponded to the genotype at the the jth marker for the ith individual. Genotypes were coded as -1, 0 or 1, where -1 and 1 are the two homozygotes and 0 the heterozygote, respectively. PC extraction was performed separately for each chromosome. In simulated data, the PC extraction on the whole genome simultaneously or separately for each chromosome did not affect DGV accuracy (Macciotta *et al*. 2010). PCA was carried out separately from each breed. The number of Principal Components (PC) retained was based on the sum of their eigenvalues. Scores of the selected PC were calculated for all individuals.

For each breed, the estimation of predictor effects on the REF data set was carried out using a BLUP model (PCA_BLUP)

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

where **y** is the vector of polygenic EBVs, 1 is a vector of ones, μ is the general mean, **Z** is the matrix of PC scores, g is the vector of PC regression coefficients treated as random, and e is the vector of random residuals. Covariance matrices of random PC effects (**G**) and residuals (**R**) were modelled as diagonal $\mathbf{I}\lambda$ and $\mathbf{I}\sigma^2 e$ respectively, where λ is $\sigma^2 e/ (\sigma^2 a/$ n PC) assuming an equal contribution of each latent variable to the additive genetic variance. Variance components were supplied by breed associations. BLUP solutions were estimated using Henderson's normal equations (Henderson, 1985) solved using a LU decomposition.

To evaluate the effect of the PCA reduction of predictors on DGV accuracy, the estimation step was carried out also using also a BLUP model where SNP genotypes were used as predictors (SNP_BLUP). In this case, Z is the matrix of SNP genotypes coded as 0,1 and 2. Mixed model equations were solved using a Gauss-Seidel iterative algorithm.

In order to enlarge the comparison with the most popular methods used to estimate DGV, a Bayesian approach was also tested. Bayesian methods usually outperform BLUP in predicting DGV when simulated data are used. Such a superiority does not seem to be confirmed on real data (Hayes *et al*. 2009a, Moser *et al*. 2009, VanRaden *et al*. 2009). In this paper, a Bayes A model (BAYES_A) that allows for variance to differ across chromosome segments (Meuwissen *et al*. 2001) was fitted:

$$\boldsymbol{y} = 1\boldsymbol{\mu} + \boldsymbol{Zg} + \boldsymbol{Wu} + \boldsymbol{e}$$

where **u** is a vector of polygenic breeding values assumed to be normally distributed, with ui~N(0,**A** σ^2a), where **A** is the average relationship matrix and σ^2a is the additive genetic variance. Prior structure and hyper-parameters where chosen according to Meuwissen *et al*. (2001). A scaled inverted chi-squared prior distribution was assumed for SNP specific variances follow, under the hypothesis that most of markers have nearly zero effects (i.e. markers not linked to any QTL) and only few have large effects. A total of 20,000 iterations were performed, discarding the first 10,000 as burn-in and considering no thinning interval. A residual updating algorithm was implemented to reduce computational time (Legarra *and* Misztal, 2008).

The general mean (**μ**) and the vector (**ĝ**) of the PC or marker effects estimated either with BLUP (SNP_BLUP) or Bayes A (BAYES_A) in the REF population were used to calculate the DGV for the k^{th} animals in the PRED subset for each breed as:

$$\mathrm{DGV_k} = \mu + \sum_{i=1}^{m} \mathbf{z'}_{ik}\, \hat{\mathbf{g}}_i$$

where **z** is the vector of PC scores or marker genotypes and m is the number of PC or markers used in the analysis.

The accuracy of direct genomic values DGV was assessed in PRED individuals by calculating Pearson correlations between EBV and DGV. Bias were assessed by examining regression of EBV on predicted DGV. Goodness of prediction was evaluated also by calculating the mean squared error of prediction (MSEP) and by its partition in different sources of variation related to systematic and random errors (Tedeschi, 2006).

RESULTS

A criterion for choosing the number of principal components to retain is the visual inspection of the eigenvalue pattern. As an example, Figure 2 reports the variance explained by each successive component extracted from SNP located on of BTA6 in the Brown breed.

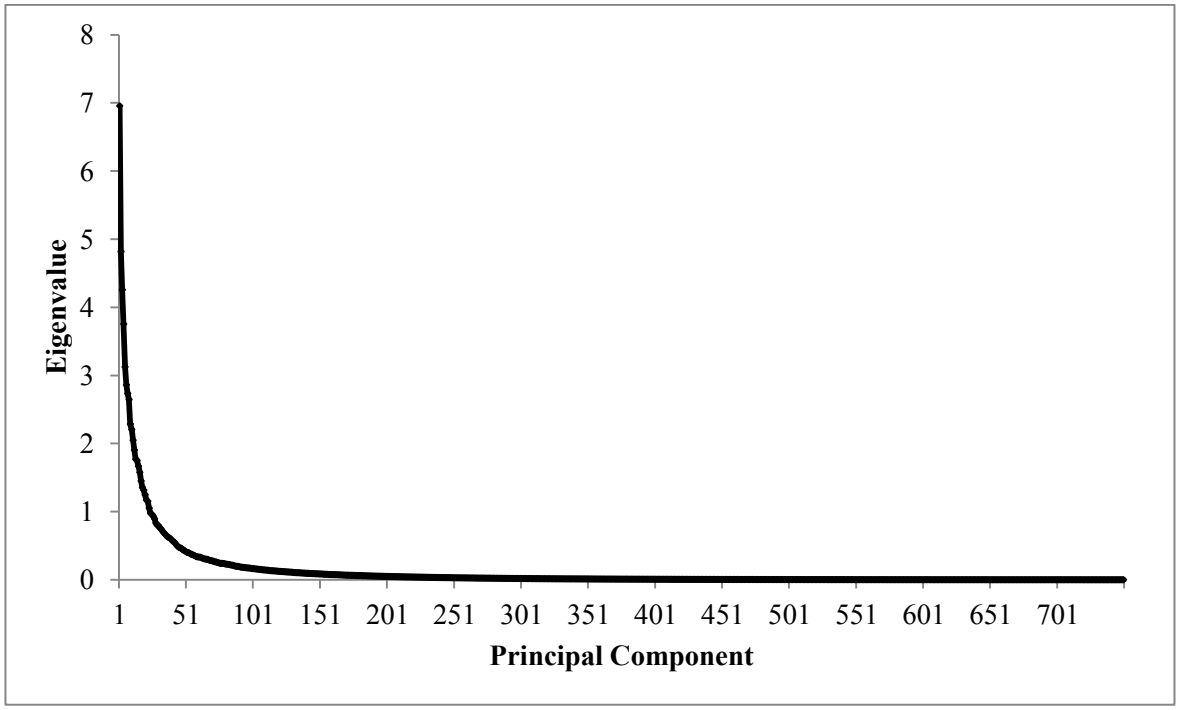

Figure 2. Pattern of the eigenvalues of the correlation matrix of SNP markers for the BTA6 in the Brown breed.

The amount of variance explained is very small also for the top two components (about 7% and 5% for the first and the second, respectively) and it shows a smooth decrease, reaching a plateau at about 100 PCs (86% of variance explained). An empirical threshold between 70% and 80% of the explained variance was considered for retaining a similar number of PC in the two breeds. A large reduction of predictor dimensionality, about 6% of the number of original variables, has been realized (Table 3).

Table 3. Number of retained principal components for the two breeds

		70:30
Breed	Retained PC	Explained variance (%)
Brown	2,257	80
Simmental	2,466	70

The extracted principal components are able to distinguish Brown from Simmental bulls. Individual scores of the first principal component extracted from BTA6, for example, are able to separate the two breeds whereas PC3 highlights a larger heterogeneity within the sample of Italian Brown bulls (Figure 3).

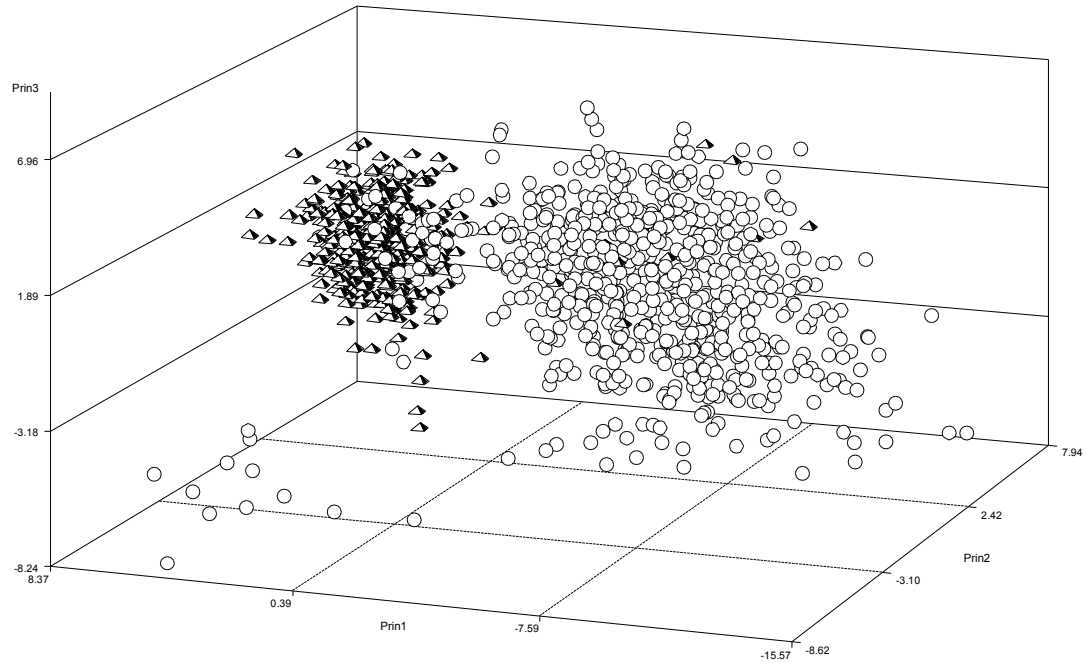

Figure 3. Plot of the first three principal components extracted from BTA6 in the two breeds (Balloons=Brown; Pyramids=Simmental).

The ability of PCA to distinguish between ethnic groups when applied to complex genetic marker patterns has been widely exploited in human genetic studies (Cavalli-Sforza *and* Feldman 2003, Paschou *et al*. 2007). However, considering the large number of original variables, it is rather complicated to give an interpretation of PC by looking at their eigenvectors. Thus their meaning can be inferred from relationships with other variables. For

example, the third principal component extracted from BTA6 in the Brown breed is negatively correlated with the observed average individual heterozigosity (-0.43) and its average score shows a progressive decrease across year of birth of bulls. This result agrees with previous reports on simulated data (Macciotta *et al*. 2010).

Correlations between DGV and EBV for PRED bulls in the different scenarios are reported in tables 5 and 6 for Italian Brown and Simmental, respectively.

Table 4. Pearson correlations between predicted direct genomic breeding values and polygenic breeding values, for different estimation methods, for the PREDICTION animals in the Brown breed.

	Estimation method		
Trait	SNP_BLUP	PC-BLUP	BAYES A
	(training/prediction) 70:30		
Milk yield	0.12	0.19	0.12
Fat yield	0.27	0.35	0.34
Protein yield	0.19	0.23	0.21
Fat percentage	0.40	0.41	0.46
Protein percentage	0.54	0.56	0.56
SCC	0.38	0.44	0.42
Udder score	0.50	0.57	0.57
Economic index	0.27	0.31	0.30
	(training/prediction) 80:20		
Milk yield	0.12	0.20	0.13
Fat yield	0.29	0.34	0.34
Protein yield	0.15	0.20	0.15
Fat percentage	0.41	0.40	0.46
Protein percentage	0.44	0.49	0.52
SCC	0.48	0.52	0.53
Udder score	0.54	0.53	0.57
Economic index	0.29	0.33	0.29
	(training/prediction) 90:10		
Milk yield	0.05	0.16	0.01
Fat yield	0.22	0.27	0.25
Protein yield	0.02	0.12	0.03
Fat percentage	0.30	0.34	0.35
Protein percentage	0.44	0.48	0.50
SCC	0.34	0.28	0.33
Udder score	0.47	0.45	0.54
Economic index	0.25	0.30	0.27

Table 5. Pearson correlations between predicted direct genomic breeding values and polygenic breeding values, for different estimation methods, for the PREDICTION animals in the Simmental breed.

	Estimation method		
Trait	SNP_BLUP	PC_BLUP	BAYES_A
	(training/prediction) 70:30		
Milk yield	0.43	0.43	0.48
Fat yield	0.34	0.36	0.36
Protein yield	0.35	0.37	0.39
Fat percentage	0.20	0.19	0.23
Protein percentage	0.41	0.41	0.43
SCC	0.30	0.34	0.32
Udder score	0.31	0.37	0.35
Economic index	0.13	0.14	0.14
	(training/prediction) 80:20		
Milk yield	0.47	0.46	0.49
Fat yield	0.36	0.37	0.37
Protein yield	0.37	0.41	0.40
Fat percentage	0.17	0.07	0.18
Protein percentage	0.36	0.36	0.38
SCC	0.30	0.33	0.30
Udder score	0.25	0.32	0.31
Economic index	0.05	0.17	0.10
	(training/prediction) 90:10		
Milk yield	0.48	0.51	0.57
Fat yield	0.42	0.43	0.48
Protein yield	0.38	0.46	0.49
Fat percentage	0.17	0.14	0.22
Protein percentage	0.24	0.18	0.27
SCC	0.36	0.30	0.20
Udder score	0.40	0.47	0.30
Economic index	0.19	0.38	0.30

By and large, correlations are low to moderate, as expected due to the reduced size of the reference populations considered. No substantial differences in DGV accuracies across estimation methods were observed although PC_BLUP and BAYES_A performed slightly better than SNP_BLUP. Moreover enlarging the ratio REF:PRED size seems to reduce DGV accuracy, at least for the Brown bulls.

In particular, correlations ranged from 0.01 to 0.57 for Italian Brown (Table 4). Lowest DGV accuracies (<0.20) were obtained for yield traits, in particular for milk and protein. Highest accuracies were observed for protein percentage, somatic cell count and udder conformation (on average 0.50, 0.33 and 0.54 respectively). Similar values for protein percentage were reported by Moser *et al*. (2009) and Hayes *et al*. (2009) on Australian Holsteins and Jerseys using different approaches and a comparable size of reference population. Best results in genomic predictions for protein percentage and udder traits have been also observed on US Holsteins (VanRaden *et al*. 2009).

DGV accuracies obtained for the Simmental bulls ranged from 0.17 to 0.57 (Table 5). In particular, accuracy for milk yield was more than two times compared to the Brown breed (on average 0.49 across all scenarios and methods) and yield traits had higher values compared to composition traits. For some scenarios, accuracies for protein yield were similar to those recently reported for Fleckvieh cattle (Gredler *et al*. 2010). Intermediate accuracies were obtained for somatic cell count and udder score (0.31 and 0.35 on average, respectively). Again PC_BLUP and BAYES_A slightly outperformed the SNP_BLUP approach.

When a small population of genotyped animals is considered, as in the case of the present study, different reference and prediction data sets can be obtained by randomly picking up animals from the original archive (Luan *et al*. 2009). Another strategy is to create different cohorts of animals based on birth year and using all of them either in the reference or in the

prediction population (Su *et al*. 2010). In table 6 are reported DGV accuracies for milk yield in the two breeds considered in this work, obtained by creating REF and PRED data set by randomly extracting bulls from the whole data. Each scenario has been replicated 5 times. For brevity, only results for the PCA_BLUP approach are reported.

Table 6. Pearson correlations between predicted direct genomic breeding values and polygenic breeding values for milk yield in the two breeds using Principal component scores as predictors when reference and prediction populations are created by picking up animals randomly.

REF: PRED	Breed			
	Brown		Simmental	
	Mean	sd	Mean	Sd
70:30	0.81	0.02	0.64	0.20
80:20	0.82	0.02	0.64	0.20
90:10	0.84	0.04	0.65	0.17

It can be clearly seen that accuracies increase dramatically, reaching values commonly reported for GS programmes carried out on large populations (VanRaden *et al*. 2009). These results do not agree with previous reports of Luan *et al*. (2009) for Norwegian Red Bulls, who did not find substantial differences in DGV accuracies of PRED animals obtained by randomly shuffling the original data set or by sorting bulls according to their progeny testing year. In the present work, similar improvement of DGV accuracies have been obtained for the other traits and for all statistical approaches.

Tables 7 and 8 show the decomposition of the mean squared error of prediction MSEP for milk yield, protein percentage, udder score, and SCC in the two breeds.

Table 7. Mean squared error of prediction (MSEP) and its decomposition (%), regression coefficients (bEBV,DGV) of polygenic breeding values on direct Genomic Breeding, and coefficient of determination (r2) for some dairy traits the Brown PREDICTION animals (scenario 70:30) using principal components scores (PC_BLUP), SNP genotypes (ALL_SNP) or Bayes (BAYES_A) estimation method.

Milk								
	MSEP	UM	US	UC	UR	UD	$b_{EBV,DGV}$	r^2
PC_BLUP	243286,39	0,33	0,02	0,66	0,20	0,47	0,23	0,04
ALL_SNP	325988,80	0,60	0,12	0,28	0,04	0,36	0,27	0,01
BAYES_ A	231731,51	0,42	0,12	0,46	0,08	0,50	0,23	0,01
Protein%								
PC_BLUP	0,01	0,00	0,12	0,88	0,03	0,97	0,79	0,31
ALL_SNP	0,01	0,00	0,72	0,28	0,11	0,90	2,20	0,29
BAYES_A	0,01	0,01	0,47	0,53	0,02	0,97	1,29	0,29
SCS								
PC_BLUP	136,15	0,20	0,08	0,73	0,07	0,74	0,62	0,20
ALL_SNP	143,49	0,25	0,38	0,37	0,00	0,75	1,12	0,14
BAYES_A	131,36	0,22	0,25	0,54	0,00	0,78	0,86	0,18
Udder score								
PC_BLUP	118,26	0,02	0,16	0,82	0,01	0,97	0,86	0,32
ALL_SNP	160,85	0,15	0,57	0,28	0,06	0,79	1,91	0,25
BAYES_A	123,40	0,03	0,51	0,46	0,05	0,92	1,48	0,33

UM = Mean Bias; US = Unequal variances; UC = Incomplete covariation; UR = Slope bias;

UD = Random errors

Note that UM+ US+ UC= UM+ UR+ UD=100%

Table 8. Mean squared error of prediction (MSEP) and its decomposition (%), regression coefficients (bEBV,DGV) of polygenic breeding values on direct Genomic Breeding, and coefficient of determination (r2) for some dairy traits the Simmental PREDICTION animals (scenario 70:30) using principal components scores (PC_BLUP), SNP genotypes (ALL_SNP) or Bayes (BAYES_A) estimation method.

Milk								
	MSEP	UM	US	UC	UR	UD	$b_{EBV,DGV}$	r^2
PC_BLUP	181129,19	0,19	0,11	0,70	0,05	0,76	0,66	0,19
ALL_SNP	240138,40	0,42	0,28	0,30	0,00	0,58	1,15	0,18
BAYES_ A	213680,68	0,41	0,29	0,30	0,01	0,59	1,22	0,22
Protein%								
PC_BLUP	0,01	0,04	0,32	0,64	0,00	0,96	0,86	0,17
ALL_SNP	0,01	0,03	0,76	0,22	0,06	0,91	2,43	0,17
BAYES_A	0,01	0,01	0,72	0,27	0,06	0,94	2,10	0,19
SCS								
PC_BLUP	119,24	0,00	0,13	0,87	0,10	0,91	0,52	0,11
ALL_SNP	112,03	0,01	0,55	0,44	0,00	0,99	1,04	0,09
BAYES_A	172,97	0,01	0,45	0,55	0,00	1,00	0,88	0,10
Udder score								
PC_BLUP	70,76	0,05	0,17	0,79	0,06	0,90	0,62	0,14
ALL_SNP	77,58	0,13	0,55	0,32	0,00	0,87	1,26	0,09
BAYES_A	76,53	0,16	0,44	0,41	0,00	0,85	1,08	0,12

UM = Mean Bias; US = Unequal variances; UC = Incomplete covariation; UR = Slope bias;

UD = Random errors

Note that UM+ US+ UC= UM+ UR+ UD=100%

In general, the method that fits principal component scores is characterized by the lowest values of MSEP. The partition of MSEP highlights further differences: in particular, the PC_BLUP approach shows the lowest values for components related to prediction bias (i.e. mean bias and inequality of variances) and highest for incomplete covariation, which is an element of random errors. These results are in agreement with previous reports on simulated data (Macciotta *et al*. 2010). As far as differences between traits are concerned, protein percentage is characterized by a reduced relevance of the mean bias and a higher weight of the unequal variance term. Regression coefficients ($b_{dgv\text{-}ebv}$) are always lower than one for the PC_BLUP indicating an underprediction of EBV for high values and overprediction for low values, respectively. Largest bias is highlighted for the SNP_BLUP method.

DISCUSSION

In this paper, bull direct genomic breeding values for some dairy traits have been estimated using a principal component approach. The PC based method has been also compared to some of the most popular methods used for predicting DGV, i.e. BLUP regression using marker genotypes and the Bayes A.

The reduction of predictor dimensionality aims at simplifying data handling and at reducing computational burdens while retaining most of the information. Although the BLUP methodology formally solves the issue of lack of degrees of freedom that affects Least Squares method when applied to the estimation of a large number of marker effects (Lande *and* Thompson, 1990, Meuwissen *et al*. 2001) the curse of dimensionality represents the most important theoretical constraint for GS implementation. This problem is enhanced when a small number of genotyped animals is available, as in the case of this study. Actually, PCA does not completely address such an issue because of the data structure. The SNP correlation matrix is singular and therefore the number of eigenvalues different from zero is equal to the number of animals (i.e. the rows) minus one (Bumb, 1982; Patterson *et al*. 2006). However, PC extraction has been carried out separately by each chromosome. Thus the gap between predictors and observations has been reduced and the number of components retained (on average 75 and 82 per chromosome in Brown and Simmental, respectively) was markedly smaller than the number of markers and of animals.

In agreement with previous findings on simulated data, PCA has been able to efficiently describe the correlation matrix of SNP genotypes (around 80-70% of explained variance) with approximately 6% of the original variables. Such a reduction had a straightforward impact on calculation time. The PC_BLUP approach required about 2 minutes using a personal computer with a 2.33 GHz Quad core processor and 3.25 Gb of RAM. On the other hand, on

average from 6 to 9 hours were needed for the SNP_BLUP and Bayes_A approaches using a Linux server with 4 x 4 quad core processors and 128 Gb RAM. PC extraction required approximately one hour and a half, but it has to be done just one time at the beginning of the work. Although calculation speed is not usually considered a technical priority for GS, compared for example to genotyping costs, it is likely to become more relevant due to the recent development of a larger (800K) SNP platform and to the upcoming very low cost sequencing technologies.

Of great interest is that such a huge reduction of calculation time has not been followed by a loss in DGV accuracy. The substantial equivalence of the PC_BLUP approach with the other two methods considered in the present paper confirms previous findings obtained with another multivariate dimension reduction technique, the Partial Least Squares Regression (Moser *et al*. 2010, Moser *et al*. 2009). The reduction of the predictor dimensionality obtained by selecting subsets of SNPs based on their chromosomal location or on their relevance to the trait usually resulted in a decrease of DGV accuracy (VanRaden *et al*. 2009, Vazquez *et al*. 2010). Actually, compared to subset SNP selection, the multivariate reduction has the advantage of not discarding any marker and of using uncorrelated predictors. The latter feature is confirmed by the observed lower bias of the PCA method found in this study compared to the BLUP_SNP method. Moreover, it is of interest to notice that both the SNP selection or the multivariate reduction seem to indicate a rather optimum number of predictors around 2,000 variables.

The equivalence between methods characterised by different theoretical foundations suggests further considerations. The BLUP assumption of an equal effect of all markers on the variance of the trait is commonly considered rather inadequate to fit the assessed distribution of QTLs, i.e many loci with a small effect and a few with large effects (Hayes *and* Goddard,

2001). On other hand, the superiority of the Bayes approach that fits heterogeneous variances across chromosome segments is marked in simulations but not in real data; (Hayes *et al*. 2009a, VanRaden *et al*. 2009). Genome Wide association studies on human height suggest that genetic variation is explained by many loci of small additive effects (Visscher *et al*. 2007). Moreover, a superior predicting ability of GEBVs for models that assume a heavy-tailed distribution of gene effects compared with finite locus models has been recently reported (Cole *et al*. 2009). Thus also BLUP methodology, even though not very accurate in terms of description of gene effect distribution, may offer robust DGV estimates (Goddard, 2009) with reasonable accuracies.

A possible criticism to the use of PCA is the lack of biological meaning of extracted variables. Such a feature is rather in contrast with the general aims of the use of molecular markers in animal breeding, i.e. the overcome of the black-box approach of traditional quantitative genetics. However, even though a clear interpretation based on eigenvectors is not feasible, some results obtained in this work are worth to be mentioned. The extracted PC scores have been able to cluster animals of the two breeds, confirming the ability of this statistical technique to capture genetic variation across and within populations (Jombart *et al*. 2009, Price *et al*. 2006). Moreover, a relationship between one of the extracted PC and the average individual heterozygosity has been evidenced, similarly to previous reports for simulated data. It is interesting to notice that, in the case reported for BTA6, it was not the first extracted component to show the relationship with heterozygosity but the third. This is also a distinguishing common feature of PCA: the first extracted component seldom contains biologically relevant information whereas these may be retrieved in components associated to smaller eigenvalues (Jombart *et al*. 2009). By and large, obtained DGV accuracies were rather low, as expected due to the reduced size of the sample of bulls considered and to their

distribution across years of birth. Composition traits showed higher accuracies compared to yield traits, whereas udder score and SCC had intermediate values. These results, in agreement with previous findings (Hayes *et al*. 2009a, VanRaden *et al*. 2009), may reflect some variation in the genetic determinism of the traits (Cole *et al*. 2009). In particular, genes with large effects for fat and protein percentages have been discovered (Cohen-Zinder *et al*. 2005, Cole *et al*. 2009, Grisart *et al*. 2002). Thus, considering that genomic predictions work by tracking the inheritance of causal mutations (VanRaden *et al*. 2009), the method may be more efficient for traits where few loci affect a large proportion of the genetic variance.

In general, the Brown breed showed higher variation in DGV accuracy across traits compared to the Simmental. Moreover, a relevant difference in accuracy between the two breeds was observed for milk yield and protein percentage. These figures may reflect, at least in part, the different selection background. The Brown is a dairy breed that has been intensively selected for dairy traits with a strong emphasis on protein yield and content (Samore *et al*. 2010). The Simmental is a dual purpose breed and young bulls that are first subjected to performance test for beef traits and then are progeny tested for milk yield.

CONCLUSIONS

Principal Component Analysis was effective in reducing the number of predictors needed for calculating direct genomic values for dairy traits in Brown and Simmental bulls. Such a reduction did not affect DGV accuracy and allowed for a relevant decrease of calculation time. The obtained accuracies, although moderate to low mainly due to the size of the sample of animals considered, highlighted some differences between traits ad breeds. Results of the present work suggest the PC approach as a possible alternative for predicting DGV, especially for populations of limited size.

ACKNOWLEDGMENTS

Research funded by the Italian Ministry of Agriculture, grant SELMOL.

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 Pedigree-Based Quantitative Trait Loci Association Analysis. Genetics 177(1):577-585.

Boichard, D., V. Ducrocq, S. Fritz and J. J. Colleau, 2010 Where is dairy cattle breeding going? A vision of the future. Interbull Workshop on the Use of Genomic Information in Genetic evaluations. Paris, March 4-5, 2010

Bolormaa, S., J. E. Pryce, B. J. Hayes, and M. E. Goddard. 2010. Multivariate analysis of a genome-wide association study in dairy cattle. Journal of Dairy Science 93(8):3818-3833.

Bumb, B., 1982 Factor analysis and development. Journal of Development Economics. 11: 109-112.

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

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(7):936-944.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and Location of Genetic effects for Dairy traits (vol 92, pg 2931, 2009). Journal of Dairy Science 92(7):3542-3542.

de Roos, A. P. W., B. J. Hayes, and M. E. Goddard. 2009. Reliability of Genomic Predictions Across Multiple Populations. Genetics 183(4):1545-1553.

Ducrocq, V., and Liu, Z. (2009). Interbull Bulletin. 40:172-177

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-Assisted Prediction of Genetic Value With Semiparametric Procedures. Genetics 173(3):1761-1776.

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(4):2289-2303.

Goddard, M. 2009. Genomic selection: prediction of accuracy and maximisation of long term response. Genetica 136(2):245-257.

Gredler, B. ,H. Schwarzenbacher, C. Egger-Danner, C. Fuerst, R. Emmerling, and J. Sölkner. 2010. Accuracy of genomic selection in dual purpose Fleckvieh cattle using three types of methods and phenotypes. Proc. 9th World Congr. Genet. Appl. Livest. Prod. Article n. 0907

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(2):222-231.

Guo, G., M. Lund, Y. Zhang, and G. Su. 2010. Comparison between genomic predictions using daughter yield deviation and conventional estimated breeding value as response variables. Journal of Animal Breeding and Genetics:no-no.

Habier, D., J. Tetens, F. R. Seefried, P. Lichtner, and G. Thaller. 2010. The impact of genetic relationship information on genomic breeding values in German Holstein cattle. Genet. Sel. Evol. 42.

Hayes, B. and M. E. Goddard. 2001. The distribution of the effects of genes affecting quantitative traits in livestock. Genet. Sel. Evol. 33(3):209-229.

Hayes, B. J., P. J. Bowman, A. C. Chamberlain, K. Verbyla, and M. E. Goddard. 2009a. Accuracy of genomic breeding values in multi-breed dairy cattle populations. Genet. Sel. Evol. 41.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009b. Invited review: Genomic selection in dairy cattle: Progress and challenges. Journal of Dairy Science 92(2):433-443.

Henderson, C. R. 1985. Best Linear Unbiased Prediction Using Relationship Matrices Derived from Selected Base Populations. Journal of Dairy Science 68(2):443-448.

Jombart, T., D. Pontier, and A. B. Dufour. 2009. Genetic markers in the playground of multivariate analysis. Heredity 102(4):330-341.

Konig, S., H. Simianer, and A. Willam. 2009. Economic evaluation of genomic breeding programs. Journal of Dairy Science 92(1):382-391.

Lande, R. and R. Thompson. 1990. Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits. Genetics 124(3):743-756.

Legarra, A. and I. Misztal. 2008. Technical note: Computing strategies in genome-wide selection. Journal of Dairy Science 91(1):360-366.

Long, N., D. Gianola, G. J. M. Rosa, K. A. Weigel, and S. Avendaño. 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. Journal of Animal Breeding and Genetics 124(6):377-389.

Luan, T., J. A. Woolliams, 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(3):1119-1126.

Macciotta, N. P. P., G. Gaspa, R. Steri, E. L. Nicolazzi, C. Dimauro, C. Pieramati, and A. Cappio-Borlino. 2010. Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. Journal of Dairy Science 93(6):2765-2774.

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(4):1819-1829.

Moser, G., M. Khatkar, B. Hayes, and H. Raadsma. 2010. Accuracy of direct genomic values in Holstein bulls and cows using subsets of SNP markers. Genet. Sel. Evol. 42(1):37.

Moser, G., B. Tier, R. E. Crump, M. S. Khatkar, and H. W. Raadsma. 2009. A comparison of five methods to predict genomic breeding values of dairy bulls from genome-wide SNP markers. Genet. Sel. Evol. 41.

Paschou, P., E. Ziv, E. G. Burchard, S. Choudhry, W. Rodriguez-Cintron, M. W. Mahoney, and P. Drineas. 2007. PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations. Plos Genet 3(9):e160.

Patterson, N., A. L. Price, and D. Reich. 2006. Population Structure and Eigenanalysis. Plos Genet 2(12):e190.

Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904-909.

Samore, A. B., R. Rizzi, A. Rossoni, and A. Bagnato. 2010. Genetic parameters for functional longevity, type traits, somatic cell scores, milk flow and production in the Italian Brown Swiss. Ital J Anim Sci 9(2):145-152.

Schaeffer, L. R. 2006. Strategy for applying genome-wide selection in dairy cattle. Journal of Animal Breeding and Genetics 123(4):218-223.

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

Su, G., B. Guldbrandtsen, V. R. Gregersen, and M. S. Lund. 2010. Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. Journal of Dairy Science 93(3):1175-1183.

Tedeschi, L. O. 2006. Assessment of the adequacy of mathematical models. Agr Syst 89(2-3):225-247.

Van Tassell, C. P., T. P. L. Smith, L. K. Matukumalli, J. F. Taylor, R. D. Schnabel, C. T. Lawley, C. D. Haudenschild, S. S. Moore, W. C. Warren, and T. S. Sonstegard. 2008. SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. Nat Meth 5(3):247-252.

VanRaden, P. M. 2008. Efficient Methods to Compute Genomic Predictions. Journal of Dairy Science 91(11):4414-4423.

VanRaden, P. M. and P. G. Sullivan. 2010. International genomic evaluation methods for dairy cattle. Genet. Sel. Evol. 42:-.

VanRaden, P. M., C. P. Van Tassell, G. R. Wiggans, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and F. S. Schenkel. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. Journal of Dairy Science 92(1):16-24.

Vazquez, A. I., G. J. M. Rosa, K. A. Weigel, G. de los Campos, D. Gianola, and D. B. Allison. 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. Journal of Dairy Science 93(12):5942-5949.

Visscher, P. M., S. Macgregor, B. Benyamin, G. Zhu, S. Gordon, S. Medland, W. G. Hill, J. J. Hottenga, G. Willemsen, D. I. Boomsma, Y. Z. Liu, H. W. Deng, G. W. Montgomery, and N. G. Martin. 2007. Genome partitioning of genetic variation for height from 11,214 sibling pairs. Am J Hum Genet 81(5):1104-1110.

Wiggans, G. R., T. S. Sonstegard, 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. Journal of Dairy Science 92(7):3431-3436

.

CHAPTER 3

USE OF DIFFERENT STATISTICAL MODELS TO PREDICT DIRECT GENOMIC VALUES FOR PRODUCTIVE AND FUNCTIONAL TRAITS IN ITALIAN HOLSTEINS

ABSTRACT

A relevant issue in genomic selection is the huge unbalance between number of markers and phenotypes available. In this work, principal component analysis is used to reduce the number of predictors for calculating direct genomic breeding values for production and functional traits. 2,093Italian Holstein bulls have been genotyped with the 54K Illumina beadchip and 39,555 SNP markers were retained after data editing. Principal Components were extracted from SNP matrix and 15,199 were used as predictors. Bulls born before 2001 were included in the reference population, younger animals were the validation population. A BLUP model was used to estimate the effect of principal components on Deregressed Proof for 35 traits and results were compared to those obtained by using SNP genotypes as predictors either with BLUP or Bayes_A models. Correlations between DGV and DRPF did not substantially differ among the three methods except for milk fat content. The lowest prediction bias was obtained for the method based on the use of principal components. Regression coefficients of DGV on DRPF highlighted a difference between methods being lower than one for the approach based on the use of principal components and higher than one for the other two methods. The use of principal components resulted in a reduction of predictors (about 38% of the original variables) and of computational time that was about the 9% of the time needed to estimate SNP effects with the other two methods. Accuracies of genomic predictions were in most of cases slightly higher than those of traditional pedigree index.

Key words: genomic selection, accuracy, principal component, SNPs.

INTRODUCTION

Genomic Selection (GS) allows for an early prediction of the genetic merit of selection candidates by combining genotypes of biallelic SNP markers and phenotypes (Meuwissen *et al.* 2001). In GS programs, the effect of a large number of SNP on the considered trait is estimated in a reference (REF) population and then used to predict Direct Genomic Values (DGV) in a test (TEST) population where only marker information are available (Meuwissen *et al.* 2001).

The switch from traditional to GS breeding programmes should be justified by a higher reliability of DGV predictions compared to parent average (PA). Actually, DGV accuracy is primarily influenced by the REF population size and, to a lesser extent, by the estimation method. Early simulation studies highlighted that few thousands of animals are needed in order to obtain DGV accuracies of 0.7 (Hayes *et al.* 2009b) and that about 30,000 unrelated individuals should be considered as REF to estimate DGV with the 800K chip (Meuwissen 2009). Such figures are rather difficult to achieve in practice, also in the case of major cosmopolite breeds and large international GS projects. Even in the USA, where the Holstein population is larger than in other countries, the REF population size in December 2010 was 16,293 (Wiggans *et al.* 2011). Actually most studies on Holstein cattle have dealt with REF populations of about one (Berry *et al.* 2009) or few thousands of animals,(Schenkel *et al.* 2009; VanRaden *et al.* 2009; Habier *et al.* 2010; Su *et al.* 2010; Liu *et al.* 2011).

The increase of REF population size just by new genotyping is still rather expensive. This situation will be further enhanced by the use of denser SNP platforms (i.e. 800K) or the whole genome sequence. Cooperation across countries represents a cheaper way to enlarge the number of genotyped animals. Some experience has already been done. For example,

United States, Canada, Italy and Great Britain shared their data (Olson *et al.* 2011; VanRaden *et al.* 2011) and in Europe the EuroGenomics project allowed Germany, France, The Netherlands, Denmark, Finland and Sweden to join their datasets and obtain a REF population of about 18,000 bulls (VanRaden *et al.* 2011). Similar experiences have occurred also in other breeds, as the Brown Swiss with the Intergenomics project.

Apart from the mathematical structure of the algorithm, differences between methods used to predict DGV depends on the assumptions on marker effect distribution. The BLUP approach fits an equal contribution of each SNP to the genetic variance of the trait (Meuwissen *et al.* 2001). It is equivalent to the use of an animal model with the additive genetic effect structured by the genomic relationship matrix (Bolormaa *et al.* 2010). On the other hand, Bayesian methods allow genetic variance to differ across chromosome segments, assuming that few SNPs have a large effect and many SNPs have a small effect on the genetic variance of the trait, respectively (Meuwissen *et al.* 2001; Hayes *et al.* 2009a; Su *et al.* 2010). Both approaches may implement a mixed inheritance by including a polygenic effect structured by pedigree relationship matrix to explain a part of the genetic variance (Berry *et al.* 2009; Habier *et al.* 2010). In early studies developed on simulated data, Bayesian methods usually outperformed BLUP, (Meuwissen *et al.* 2001; Clark *et al.* 2011) On real data, such differences are no longer detectable except for traits for which the existence of few genes with a larger effect has been detected (Hayes *et al.* 2009a; VanRaden *et al.* 2009).

A further issue on GS is represented by the adoption of techniques for reducing the huge unbalance between the number of phenotypes and genotypes available. It represents a basic requirement in the implementation of GS program in populations of limited size. However, reduction of predictor dimensionality may also be useful for large populations, as the Holstein breed, with the perspective of using a 800K SNP chip or the complete sequence

in the near future. SNP pre-selection based on the relevance to the trait or the use of dimension reduction multivariate methods as principal component analysis (PCA) and partial lest squares regression represent the two main strategies adopted to address this issue (Moser *et al.* 2009; Solberg *et al.* 2009; Macciotta *et al.* 2010; Moser *et al.* 2010; Vazquez *et al.* 2011). Compared to SNP pre-selection, PCA reduction does not discard any SNP and the reduced panel of predictors is independent from the trait considered.

In this work, DGV of different production and functional traits for a sample of Italian Holstein bulls obtained by joining data generated into two GS research projects are calculated by using different type of predictors, i.e. the SNP genotypes or the scores of a reduced number of principal components. Moreover, also the assumptions on predictor effect are compared by using a Bayesian or a BLUP method.

MATERIALS AND METHODS

Data

Genotypes of 2,093 Italian Holstein bulls were generated in two Italian research projects: the SELMOL and the PROZOO. Birth years of bulls ranged from 1979 to 2007, with an average number of 72 animals per year. Bulls born before or after 2001 were included in the REF and TEST populations, respectively. Distribution of REF and TEST bulls across birth years is illustrated in Figure 1.

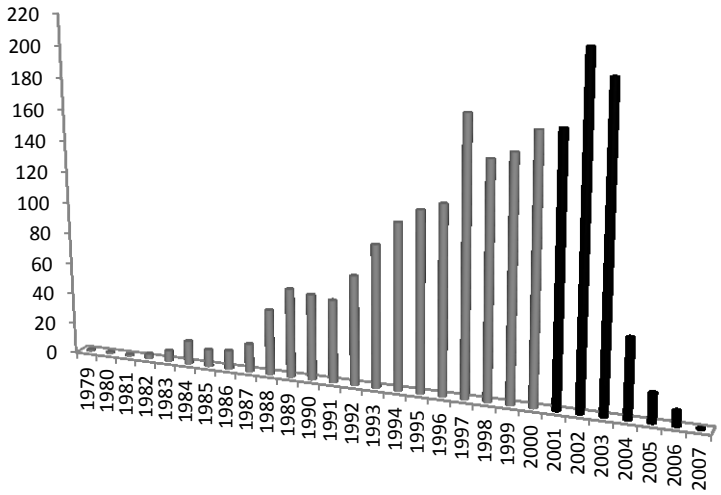

Figure 1: distribution of number of bulls per birth year in the reference and test population.

Animals were genotyped using the BovineSNP50BeadChip (Illumina, San Diego, CA). A data editing has been performed. SNP were discarded based on missing data (>0.025), minor allele frequency (>0.05), existence of Mendelian inheritance conflicts, absence of heterozygous genotypic class, deviance from Hardy-Weimberg equilibrium (<0.01 bonferroni corrected) (Wiggans *et al.* 2009). Markers retained after edits were 39,555. Missing SNP alleles were replaced by the most frequent allele at that specific locus. A total of 86 bulls were

discarded: 48 samples were replicates or had inconsistent mendelian inheritance information, whereas 38 samples had low overall call rate (>1000 missing SNPs).

Phenotypes were Deregressed EBV (DRPF) provided by the Italian Holstein Association ANAFI. Thirty-five productive and functional traits have been considered (Table 1).

Table 1. Pearson correlations between predicted DGV and DRPF, for different estimation methods, for the test animals.

TRAIT	SNP-BLUP	PC-BLUP	Bayes_A	PI
PFT	0.38	0.42	0.39	0.41
Milk Yield	0.39	0.43	0.46	0.45
Fat Yield	0.37	0.42	0.49	0.34
Protein Yield	0.36	0.39	0.38	0.40
Fat %	0.40	0.47	0.64	0.45
Protein %	0.48	0.53	0.55	0.50
SCC	0.52	0.54	0.52	
Longevity	0.31	0.37	0.31	
Fertility	0.26	0.28	0.28	
Type	0.50	0.51	0.51	0.43
Overall Conformation Score	0.41	0.42	0.40	
Overall Udder Score	0.46	0.49	0.46	0.41
Overall Feet & Leg Score	0.35	0.34	0.36	
Stature	0.44	0.48	0.46	0.50
Strength	0.33	0.36	0.35	0.13
Body Depth	0.36	0.40	0.37	0.46
Angularity	0.43	0.44	0.44	0.41
Rump Angle	0.48	0.53	0.49	0.43
Rump Width	0.42	0.42	0.43	0.54
Rear leg side view	0.34	0.35	0.34	0.39
Foot Angle	0.37	0.38	0.37	0.35
Rear leg rear view	0.34	0.32	0.34	
Locomotion	0.44	0.44	0.45	
Fore Udder Attachment	0.43	0.45	0.44	0.38
Rear Udder Attachment Height	0.43	0.46	0.44	0.39
Rear Udder Attachment Width	0.25	0.25	0.26	0.30
Udder Cleft	0.40	0.41	0.41	0.41
Udder Depth	0.41	0.45	0.42	0.37
Front Teat Placement	0.42	0.41	0.41	0.26
Teat Length	0.31	0.33	0.32	0.20
Rear Teat Placement	0.36	0.35	0.36	
Direct Calving Ease	0.05	0.05	0.05	
Maternal Calving Ease	0.05	0.05	0.05	
Production Persistency	0.25	0.27	0.30	
Maturity rate	0.33	0.33	0.34	

Phenotypes were not available for all bulls, thus small differences in sizes of REF and TEST populations across traits have occurred. On average, sizes of REF and TEST populations were of 1,424 and 634 bulls, respectively.

Methods

Methodologies used to calculate DGV differ in the dimensionality of predictors (SNP genotypes vs. PC scores) and in the assumptions on marker effect distributions (BLUP vs Bayes).

Reduction of predictor dimensionality by Principal Component Analysis

PCA has been used to extract latent variables from the SNP matrix (n x m) (where n=total number of animals, and m=number of SNPs retained after edits). Genotypes were coded as -1 and 1 for homozygotes and 0 for heterozygote, respectively. PC were extracted separately for each chromosome for computational reasons. Results obtained on simulated data reported the same DGV accuracy for PCA carried out on the entire genome or separately per chromosome (Macciotta *et al.* 2010). The number of components to retain was based on the amount of original variance explained, calculated as sum of eigenvalues. In particular, five thresholds of explained variance were considered with a corresponding number of extracted variables ranging from about 2,600 to 15,200. Component scores for each animal were used as predictors in the further steps of DGV calculation and validation.

BLUP

The effect of predictors, either SNP (SNP_BLUP) or principal component scores (PC_BLUP), on phenotypes of the REF bulls was estimated with the following mixed linear model

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{Zg} + \boldsymbol{e} \qquad [1]$$

where **y** is the vector of Deregressed EBV, **1** is a vector of ones, μ is the general mean respectively, **Z** is the matrix of SNPs genotypes or PC scores, **g** is the vector of their effects

treated as random, and **e** is the vector of random residuals. Covariance matrices of random effects (**G**) and residuals (**R**) were modelled as diagonal $\mathbf{I}\lambda$ and $\mathbf{I}\sigma^2_e$respectively, where λ is $\sigma^2_e/(\sigma^2_a$ /n PC) assuming an equal contribution of each predictor to the additive genetic variance. Additive genetic σ^2_a and residual σ^2_e variances for all traits were provided by the Holstein association. BLUP solutions were estimated using Henderson's normal equations (Henderson 1985) and mixed model equations were solved using a Gauss-Seidel iterative algorithm.

BAYES_A

A Bayes A method (BAYES_A) that assumes that most of markers have nearly zero effects (i.e. markers not linked to any QTL) and only few have large effects was fitted to the REF data set with the same structure used in model [1]. Prior distributions and parameters where chosen according to Meuwissen *et al.* (2001). Twenty thousand iterations were performed, discarding the first 10,000 as burn-in and considering no thinning interval. Computational times were reduced by using a residual updating algorithm to solve the model (Legarra & Misztal 2008).

DGV estimation

DGVs in the TEST population were calculated using the general mean (μ) and the vector ($\hat{\mathbf{g}}$) of the solution of predictors effects estimated with BLUP or BAYES_A in the previous step as:

$$DGV_k = \mu + \sum_{i=1}^{m} \mathbf{z}'_{ik}\, \hat{\mathbf{g}}_i$$

where **z** is the vector of PC scores or marker genotypes and *m* is the number of PC or markers used in the analysis.

The accuracy of direct genomic values DGV was assessed in TEST individuals by calculating Pearson correlations between DRPF and DGV. Bias were assessed by examining regression of DRPF on predicted DGV. Goodness of prediction was evaluated also by calculating the mean squared error of prediction (MSEP) and by its partition in different sources of variation related to systematic and random errors (Tedeschi 2006). Moreover, the accuracy of genomic predictions was compared to the realized accuracies of 2005 pedigree indexes (PI) in TEST individuals for some traits. PI from 2005 were chosen because nearly all animals in the TEST population were not progeny tested at that time.

RESULTS

The effect of different thresholds of explained variance used in PC extraction on the DGV accuracy for seven traits in TEST bulls is reported in Figure 2.

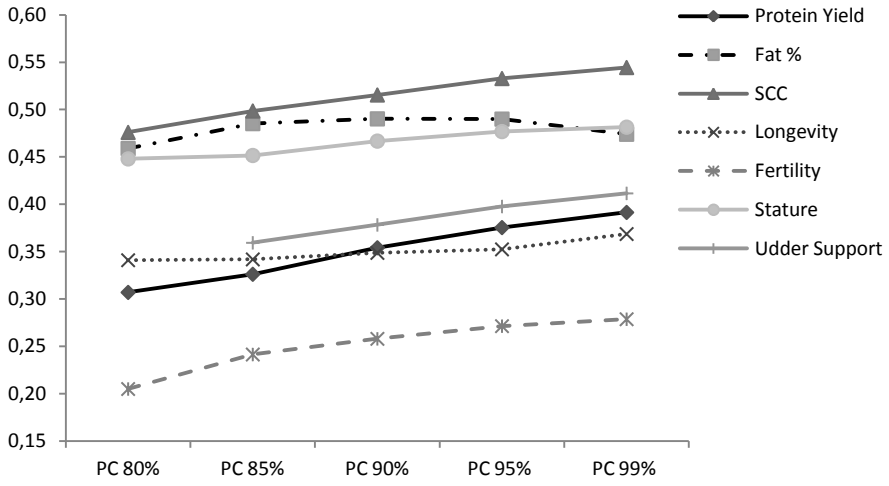

Figure 2 Pearson correlations between predicted direct genomic breeding values and deregressed proof, for the PC-BLUP method using a different number of PC, for the TEST animals.

Basically, correlations between DGV and DRPF exhibit a slight linear increase for larger amounts of extracted components. This behavior can be observed for almost all traits except milk and fat percentage. Thus the value of explained variance further considered in the study was 99%, with a corresponding number of 15,199 extracted components.

Pearson correlations between predicted DGV and DRPF in TEST bulls for the different estimation methods are reported in Table 1. Values are low to moderate and differences between traits and, to a lesser extent, methods can be observed. Smallest accuracies were obtained for fertility traits, especially calving ease, below 0.10. Milk composition traits, as protein and also somatic cell count showed highest values, ranging from 0.40 up to 0.64. Also some conformation traits as type, udder score and rump angle showed accuracies around 0.50. Yield traits had intermediate values of correlations (about 0.40-0.45).

Slight differences in $r_{DGV,DRPF}$ between methods can be observed (Table 1). In general, accuracies of PC_BLUP and BAYES_A (for 21 and 12 traits out of 35, respectively) were slightly higher than those of BLUP method that uses SNP genotypes as predictors. On average, the maximum and the minimum value of accuracy for each trait differed of about 4%. A relevant exception is represented by fat percentage where BAYES_A markedly outperformed the other methods, yielding an accuracy greater than about 25% and 15% compared to the other approaches. Such a better performance, even though of a reduced magnitude, con be observed also for fat yield.

Comparison between accuracies of genomic predictions and of pedigree indexes shows a slight superiority for most of traits for genomic predictions

Table 2 shows the coefficient of determination (R^2), mean squared error of prediction and its decomposition of DGV calculated with the three methods for some selected traits: protein yield, fat percentage, somatic cell count, longevity, fertility, stature and udder support.

Table 2. Mean squared error of prediction (MSEP) and its decomposition (%), and coefficient of determination (r^2) of Deregressed Proof on direct Genomic Breeding values for some traits in the TEST animals using different estimation method.

Protein Yield	**r2**	**MSEP**	**mean bias**	**unequal variances**	**incomplete (co)variation**	**Systematic bias**	**Random errors**
PC_BLUP	0.15	312.93	0.24	0.10	0.66	0.06	0.70
SNP_BLUP	0.13	370.90	0.38	0.20	0.42	0.01	0.61
Bayes_A	0.14	356.88	0.36	0.19	0.45	0.01	0.63
Fat %							
PC_BLUP	0.22	0.04	0.00	0.26	0.74	0.01	0.99
SNP_BLUP	0.16	0.05	0.00	0.53	0.47	0.01	0.99
Bayes_A	0.42	0.03	0.00	0.20	0.80	0.00	1.00
Somatic Cell Count							
PC_BLUP	0.29	25.27	0.01	0.27	0.72	0.00	0.99
SNP_BLUP	0.27	27.00	0.00	0.55	0.45	0.04	0.96
Bayes_A	0.29	26.49	0.00	0.54	0.46	0.04	0.96
Longevity							
PC_BLUP	0.14	63.34	0.23	0.17	0.61	0.03	0.74
SNP_BLUP	0.10	61.46	0.20	0.42	0.38	0.00	0.80
Bayes_A	0.09	61.46	0.19	0.53	0.28	0.01	0.80
Fertility							
PC_BLUP	0.08	80.86	0.08	0.23	0.69	0.05	0.87
SNP_BLUP	0.07	81.95	0.13	0.45	0.42	0.00	0.87
Bayes_A	0.07	82.37	0.14	0.49	0.37	0.00	0.86
Stature							
PC_BLUP	0.23	1.58	0.21	0.27	0.52	0.00	0.79
SNP_BLUP	0.20	2.04	0.33	0.42	0.25	0.02	0.65
Bayes_A	0.20	1.98	0.32	0.41	0.27	0.02	0.66
Udder support							
PC_BLUP	0.17	1.80	0.11	0.21	0.69	0.02	0.87
SNP_BLUP	0.16	1.99	0.20	0.42	0.39	0.00	0.80
Bayes_A	0.16	2.00	0.21	0.43	0.37	0.01	0.79

The PC_BLUP method showed the lowest values of MSEP across all the considered traits. Moreover, as far as the decomposition of the MSEP is concerned, for almost all traits this approach was characterized by the lowest incidence of components related to prediction bias, i.e. mean bias (on average 13% of the MSEP) and inequality of variances (22%), and highest for incomplete covariation (66%) and random error (85%), i.e. the sources of random variation. SNP_BLUP and BAYES_A had basically the same composition of the MSEP. Less defined is the pattern across traits. Protein yield, for example, had the highest value for mean

bias but the lowest for inequality of variance. In any case, fat percentage and somatic cell count showed the largest incidence of random variation.

Regression coefficients ($b_{DGV,DRPF}$) of DGV on DRPF are shown in Figure 3.

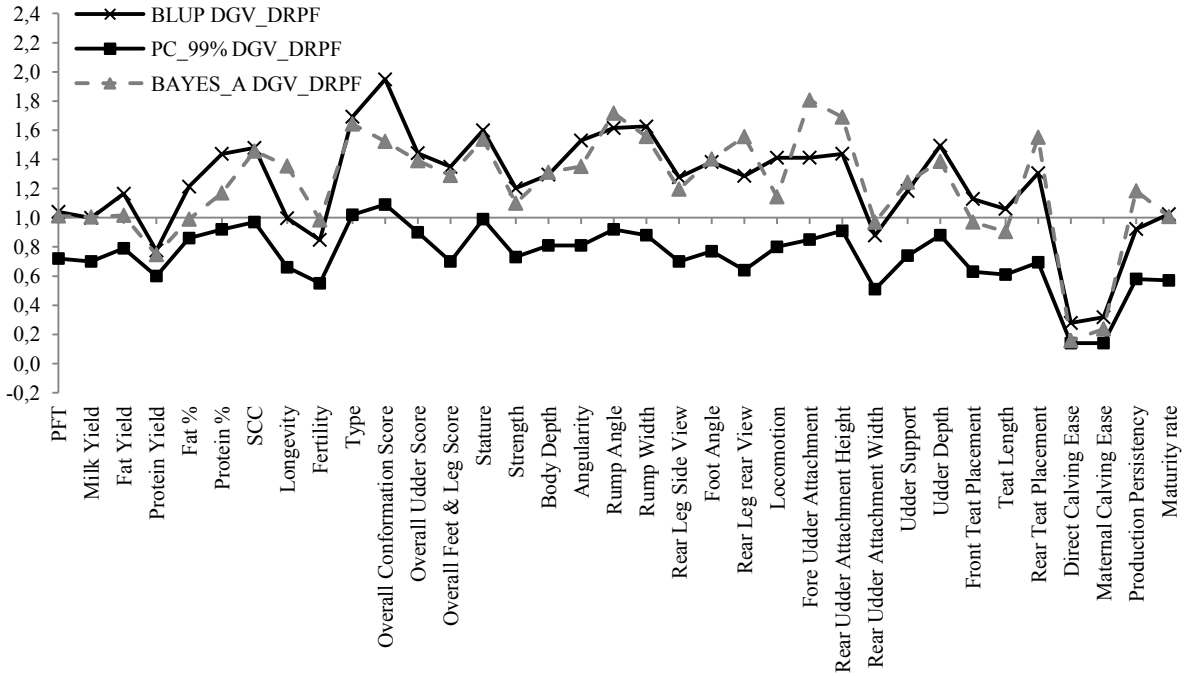

Figure 3. Regression coefficients ($b_{DRPF,DGV}$) of Deregressed Proof on direct Genomic Breeding Values estimated with PC_BLUP, SNP_BLUP and BAYES_A methods, and on Parent Average for all traits considered in test animals

A relevant difference between methods can be observed. Regression coefficient values are lower than one in almost all traits for the PC_BLUP method (on average 0.74±0.21), indicating that positive values of DGV overpredict DRPF and vice versa for negative DGV values. On the contrary, all methods that use directly SNP genotypes showed ($b_{DGV,DRPF}$) almost always greater than one (except for calving ease): 1.23±0.35, 1.22±0.37, for SNP_BLUP and BAYES_A, respectively. Moreover, among all methods, the PC_BLUP showed the lowest degree of accuracy. A definite pattern across traits could not be identified,

except for the very low values for calving ease and the rather high (>1.30) for some conformation traits.

DISCUSSION

As expected, due to the limited size of the reference population, prediction accuracies for direct genomic values were low to moderate. For example, squared correlations reported for US Holstein (VanRaden *et al.* 2009) obtained by using a REF population of 3,576 bulls are on average 0.2 higher than those reported in the present work for a set of 23 common traits. Similar differences can be observed with reliabilities reported by Su et al. (2010) on a 3,330 Danish Holsteins. In VanRaden *et al.* (2009), the R^2 for Net merit has been calculated also with REF population sizes of 1,151 and 2,130. Values were similar to those here reported, i.e. 0.12 and 0.17 vs 0.16, respectively. Accuracies obtained in the present work were similar to those reported by (Moser *et al.* 2010) with a REF population of 1,847 bulls. All the above mentioned figures confirm the importance of the number of the genotyped animals in the realized accuracy of genomic predictions. In any case accuracies of DGV were equal or in many cases higher than realized accuracies of traditional pedigree indexes.

The reduction of predictor dimensionality by principal component analysis did not affect DGV predictions compared to methods that use directly all SNP genotypes available. In most of cases the PC_BLUP approach gave the best accuracies even if differences with the other methods were rather small. Such results confirm previous reports on simulated (Solberg *et al.* 2009; Macciotta *et al.* 2010) and real data (Long *et al.* 2011). The reduction performed in this study was of a lower magnitude compared to some of the above mentioned researches, being the number of PC to be retained not fixed a priori but based on the test of different thresholds of explained variance (PC were about 38% of the original variables). However, the effect on calculation speed was still evident. The average computation time for the PC_BLUP method was about 2 hours (from 50 min to 4 h depending on the trait), whereas 18 hours

(from 9 h to 29 h) were needed on average with the SNP_BLUP and BAYES_A approaches using a Linux server with 4 x 4 quad core processors and 128 Gb RAM.

DGV predictions obtained with the PC_BLUP methods were quite always characterized by the lowest bias. This result has been also confirmed by the decomposition of the mean squared error of prediction, that highlighted a larger incidence of the random variation for the PC-based method compared to the other approaches Moreover, the comparison between the two BLUP-based methods showed slight better accuracies for the PC_BLUP than for the SNP_BLUP (magnitude of difference was always lower than 8%).These results may be ascribed to better numerical properties of the extracted variables compared to the direct use of SNP genotypes. Actually principal components are uncorrelated and this feature prevents problems of multicollinearity that are likely to occur because of linkage disequilibrium between loci when dense marker genotypes are used as predictors (Long *et al.* 2011).

As far as the effect of the assumption on marker effect distribution is concerned, BAYES_A yielded substantially the same accuracies of BLUP methods for almost all traits. These figures do not agree with simulation studies were Bayesian statistics performed better than BLUP methods (Meuwissen *et al.* 2001; Habier *et al.* 2007). On the other hand, they are similar to those obtained for real data (Moser *et al.* 2009; VanRaden *et al.* 2009; Su *et al.* 2010). A relevant exception is represented by the behavior of milk fat percentage. For this trait, the accuracy of the BAYES_A method was markedly higher (>30%) than in BLUP methods. A possible explanation can be found in the genetic structure of the trait. It is well known that fat content is largely influenced by single genes with major effect as the DGAT1 (Grisart *et al.* 2004). Previous studies reported that methods that assume heterogeneity of variance across chromosome segments usually perform better than those that assume an equal contribution of

all markers to the genetic variation in case of traits influenced by few genes.(VanRaden *et al.* 2009; Hayes & Goddard 2010).

Some differences across traits were evidenced, although no definite trend between categories (e.g. yield, conformation, udder, etc.) was observed. Highest values were observed for milk composition, for some conformation and yield traits. Lowest values were found for calving ease, fertility and most of conformation traits. Such different behaviour between traits is in agreement with reports on North American (Schenkel *et al.* 2009; VanRaden *et al.* 2009; Olson *et al.* 2011) and German (Liu *et al.* 2011) Holsteins. These figures seems to be related, even if roughly, to the heritability of the trait even if in some exception can be observed, as for somatic cell count. Liu et al. (2011), partially explained the smallest genomic accuracies for traits with low heritability as a consequence of the lowest accuracies of their conventional EBV in the REF population.

CONCLUSIONS

In this work direct genomic breeding values of Italian Holstein bulls for productive and functional traits have been calculated using different methods and types of predictors. Realized accuracies of genomic predictions are low to moderate, conforming the relevant importance of the size of the REF populations. However, DGV accuracies were similar or, in many cases, slightly higher than those of pedigree indexes. The use of dimension reduction techniques did not result in a decrease of accuracy of genomic prediction compared to methods that use all SNP available. Assumptions on marker effect had a relevant influence in the efficiency of the genomic selection for traits that are known to be affected by a limited number of genes with a large effect.

ACKNOWLEDGMENTS

Research funded by the Italian Ministry of Agriculture (grant SELMOL) and by the Fondazione CARIPO (grant PROZOO)

REFERENCES

Berry D.P., F. K. & B.L. H. (2009) Genomic Selection in Ireland. *Interbull Bull*.

Bolormaa S., Pryce J.E., Hayes B.J. & Goddard M.E. (2010) Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science* **93**, 3818-33.

Clark S.A., Hickey J.M. & Van der Werf J.H.J. (2011) Different models of genetic variation and their effect on genomic evaluation. *Genetic Selection Evolution* **43:18**.

Grisart B., Farnir F., Karim L., Cambisano N., Kim J.J., Kvasz A., Mni M., Simon P., Frere J.M., Coppieters W. & Georges M. (2004) Genetic and functional confirmation of the causality of the DGAT1 K232A quantitative trait nucleotide in affecting milk yield and composition. *Proc Natl Acad Sci U S A* **101**, 2398-403.

Habier D., Fernando R.L. & Dekkers J.C.M. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97.

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

Hayes B. & Goddard M. (2010) Genome-wide association and genomic selection in animal breeding. *Genome* **53**, 876-83.

Hayes B.J., Bowman P.J., Chamberlain A.J. & Goddard M.E. (2009a) Invited review: Genomic selection in dairy cattle: Progress and challenges. *Journal of Dairy Science* **92**, 433-43.

Hayes B.J., Visscher P.M. & Goddard M.E. (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). *Genetics Research* **91**, 143-.

Henderson C.R. (1985) Best Linear Unbiased Prediction Using Relationship Matrices Derived from Selected Base Populations. *Journal of Dairy Science* **68**, 443-8.

Legarra A. & Misztal I. (2008) Technical note: Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360-6.

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

Long N., Gianola D., Rosa G.J.M. & Weigel K.A. (2011) Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*, no-no.

Macciotta N.P.P., Gaspa G., Steri R., Nicolazzi E.L., Dimauro C., Pieramati C. & Cappio-Borlino A. (2010) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *Journal of Dairy Science* **93**, 2765-74.

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

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

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

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

Olson K.M., VanRaden P.M., Tooker M.E. & Cooper T.A. (2011) Differences among methods to validate genomic evaluations for dairy cattle. *Journal of Dairy Science* **94**, 2613–20.

Schenkel F.S., Sargolzaei M., Kistemaker G., Jansen G.B., Sullivan P., Van Doormaal B.J., Van Raden P.M. & Wiggans G.R. (2009) Reliability of genomic evaluation of holstein cattle in canada. *Interbull Bull* **39**.

Solberg T.R., Sonesson A.K., Woolliams J.A. & Meuwissen T.H.E. (2009) Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution* **41**, -.

Su G., Guldbrandtsen B., Gregersen V.R. & Lund M.S. (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *Journal of Dairy Science* **93**, 1175-83.

Tedeschi L.O. (2006) Assessment of the adequacy of mathematical models. *Agricultural Systems* **89**, 225-47.

VanRaden P., O'Connell J., Wiggans G. & Weigel K. (2011) Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 10.

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

Vazquez A.I., Rosa G.J.M., Weigel K.A., de los Campos G., Gianola D. & Allison D.B. (2011) Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins (vol 93, pg 5942, 2010). *Journal of Dairy Science* **94**, 537-.

Wiggans G.R., Sonstegard T.S., Vanraden P.M., Matukumalli L.K., Schnabel R.D., Taylor J.F., Schenkel F.S. & Van Tassell C.P. (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *Journal of Dairy Science* **92**, 3431-6.

Wiggans G.R., Van Raden P.M. & Cooper T.A. (2011) The genomic evaluation system in the United States: Past, present, future *Journal of Dairy Science* **94**, 3202-11.

CHAPTER 4

USE OF PRINCIPAL COMPONENT APPROACH TO PREDICT DIRECT GENOMIC BREEDING VALUES FOR BEEF TRAITS IN ITALIAN SIMMENTAL CATTLE

ABSTRACT

In the current study, principal component (PC) analysis was used to reduce the number of predictors in the estimation of direct genomic breeding values (DGV) for meat traits in a sample of 479 Italian Simmental bulls. SNP marker genotypes were determined with the 54K Illumina beadchip. After edits, 457 bulls and 40,179 SNPs were retained. PC extraction was carried out separately for each chromosome and 2,466 new variables able to explain 70% of total variance were obtained. Bulls were divided into reference and validation population. Three scenarios of the ratio reference:validation were tested: 70:30, 80:20, 90:10. Effect of PC scores on polygenic EBVs was estimated in the reference population with a BLUP model. Traits analyzed were daily live weight gain, size score, muscularity score, feet and legs score, beef index (economic index), calving ease direct effect, and cow muscularity. Accuracy was calculated as correlation between DGV and polygenic EBV in the validation bulls. Muscularity, feet and legs, and the beef index showed the highest accuracies calving ease the lowest. In general, accuracies were slightly higher when reference animals were selected at random and the best scenario was 90:10.

Key Words: genomic selection, meat trait, principal component analysis

INTRODUCTION

In the last years, the development of high density SNP platforms has had a relevant impact in animal breeding and genetics studies for several livestock species. Genotypes of thousands of marker loci are currently used in dairy cattle to search for genomic regions associated with yield and functional traits; (Cole *et al.* 2009; Bolormaa *et al.* 2010) and for predicting genomic enhanced estimated breeding values (GEBV) in genomic selection (GS) programmes. In beef cattle most of studies have dealt with genome-wide scans for association between SNP polymorphisms and beef and functional traits such as residual feed intake, average daily gain, hip height, and carcass traits (Bolormaa *et al.*) or to detect signature of selection able to discriminate between beef and dairy cattle (Hayes *et al.* 2009a). Actually, less pressure has been put on the implementation of GS programs, even though this technology may represent a valuable option also for beef cattle, allowing to increase breeding value accuracy and to enlarge breeding goals by including traits that are difficult or expensive to measure routinely.

Possible constraints to the application of GS in beef cattle are the number of genotyped animals (Garrick 2011), usually smaller than in dairy cattle, and the genotyping costs. To handle the latter issues some authors suggested to develop a smaller SNP chip specific for beef trait (Rolf *et al.* 2011). The former issue, one of the most relevant for GS and frequent also in some situations in dairy cattle (breeds of limited size, beginning of programmes), can be addressed by using strategies able to reduce predictor dimensionality. Multivariate reduction techniques as principal component analysis (PCA) and partial least squares regression, have been suggested for reducing the number of predictors in DGV calculations both for simulated and actual data (Moser *et al.* 2009; Solberg *et al.* 2009; Long *et al.* 2011). In particular, PCA allowed for a considerable reduction (>90%) of the number of independent

variables in DGV estimation with accuracies similar to those obtained using directly all SNP genotypes available both on simulated and real data. (Solberg *et al.* 2009; Macciotta *et al.* 2010a).

Aim of this work was to develop a methodology for calculating DGV for beef traits in the dual purpose Italian Simmental cattle breed. PCA was used to reduce the number of predictors. Moreover, the method was compared with two other approaches commonly used to predict DGV in genomic selection programs that use directly SNP genotypes as predictors.

MATERIALS AND METHODS

A total of 465 Italian Simmental bulls were genotyped at 54,001 SNP loci using the Illumina Bovine SNP50TM bead-chip (Illumina, San Diego, CA). Animals with more than 1,000 missing genotypes and with inconsistencies in the mendelian inheritance were excluded from the analysis. The selection of SNP was more conservative. Edits were based on the number of missing records (> 0.025), mendelian inheritance conflicts, absence of heterozygous individuals, minor allele frequency (> 0.05), deviance from Hardy-Weimberg equilibrium (P < 0.01) (Wiggans *et al.* 2009). After editing, 8 animals (2 for mendelian inheritance conflicts, 6 for missing genotypes) and 13,822 SNP (21 SNP for mendelian inheritance conflict, 999 SNP with missing exceeding the threshold, 12,215 SNP with MAF≤ 0.05 and 587 were not in HW equilibrium) were discarded. Final number of bulls and SNP used were 457 and 40,179 respectively. Missing SNP were replaced with the most frequent allele at that specific locus.

Phenotypes used were polygenic EBV provided by Simmental national breeders associations (evaluation of December 2009). Seven traits were considered: average daily weight gain (ADWG, kg/d), size score (SS), muscularity score (MS), feet and legs score (FLS), beef index (BI = 0.40*ADWG + 0.10*SS + 0.40*MS + 0.10*FLS), calving ease direct effect (CED), cow muscularity score(CWM). In table 1 are listed the basic statistics about EBV used and their mean reliability.

Table 1. Heritability of average daily weight gain (ADWG), feet and leg score (FLS), Calving Ease direct (CED), Beef Index (BI), Muscularity Score (MS), Size Score (SS) and Cow Muscularity (CWM). Mean and standard deviation of EBV used as phenotypes and their average reliability

Trait	h^2	Mean EBV[a] ± SD	Mean Reliability ± SD
ADWG[b]	0.35	104.08 ± 6.57	0.43 ± 0.12
SS[b]	0.32	103.07 ± 6.45	0.43 ± 0.12
MS[b]	0.61	106.45 ± 9.17	0.60 ± 0.16
FLS[b]	0.25	104.72 ± 7.31	0.42 ± 0.12
BI[c]	-	104.99 ± 6.29	0.43 ± 0.12
CED[d]	0.05	99.13 ± 6.98	0.59 ± 0.17
CWM[d]	0.36	100.76 ± 9.10	0.71 ± 0.21

a) all trait are reported as standardized breeding values with mean 100 and genetic standard deviation 12
b) Ebv estimated in performance test
c) combinded index of ADWG, SS, MS and FLS
d) Ebv estimated in progeny test

EBV for CED and CWM were derived from progeny test whereas the other traits were measured on performance test. The scale of EBV analyzed were equivalent for different traits (standardized with mean 100 and genetic standard deviation 12).

Animals were sorted by year of birth (range 1972-2002) and the whole dataset was split into two subsets, reference (REF) and validation (VAL), containing the oldest and youngest animals, respectively. Different sizes of REF population were tested. Bulls born before 1999, 2000 or 2001 were included in the reference population (Figure 1), corresponding to the ratios REF/VAL of 70:30, 80:20 and 90:10 respectively.

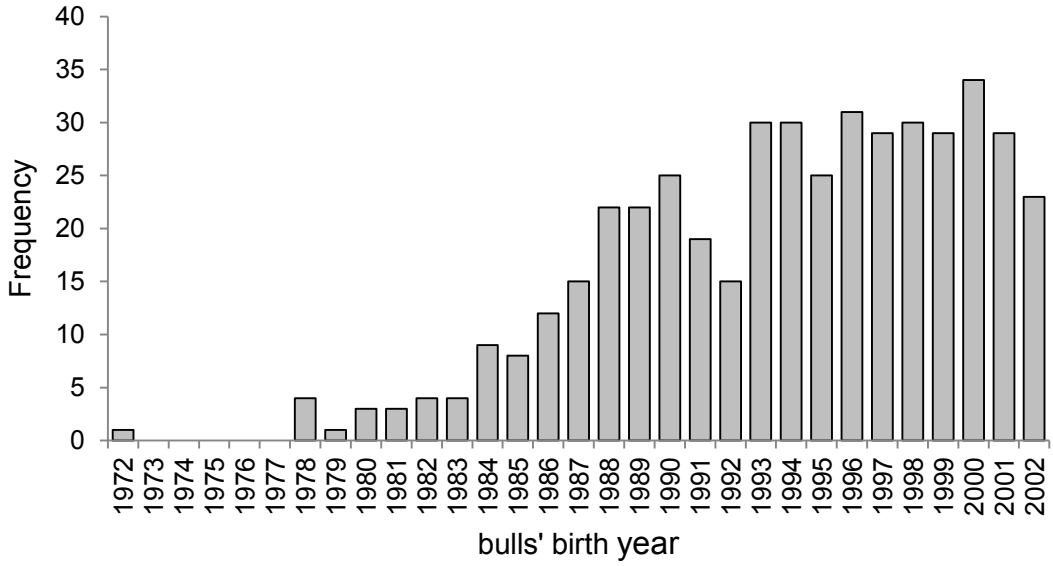

Figure 1. Distribution of bulls by birth's year.

Statistical models

PC-BLUP

Data matrix $\mathbf{M}_{nxm}$ of marker genotypes was set up (n=total number of individuals, m = number of marker genotypes). Each element m_{ij} corresponded to the genotype at the j-th marker for the i-th individual. Genotypes were coded as -1, 0 or 1, where -1 and 1 are the two homozygotes and 0 the heterozygote, respectively, according to the parameterization of (Solberg *et al.* 2009). The PC extraction was carried out separately by chromosome. The number of PC retained was based on the percentage of variance explained by PC (Macciotta *et al.* 2010a). Scores of the selected PC were calculated for all individuals.

The estimation of effects of the PC on the REF data set was carried out using a BLUP model.

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e} \qquad [1]$$

where **y** is the vector of polygenic EBVs, **1** is a vector of ones, μ is the overall mean, **Z** is the matrix of PC scores, **g** is the vector of PC regression coefficients treated as random, and **e** is

the vector of random residuals. Two different assumption on the distribution of PC effect were adopted.

In the first hypothesis (PC-BLUP) random PC effects (**g**) were assumed identically and normally distributed with $g_i \sim N(0, \mathbf{I}\sigma_{gi}^2)$ where $\sigma_{gi}^2=\sigma_a^2/k$ (k=number of PC retained).Random residuals were assumed normally distributed with $e_i \sim N(0, \mathbf{I}\sigma_e^2)$. In the second approach (PC-BLUP_EIGEN), the (Co)variance matrices of random PC effects (G) and residuals (R) were modeled as diagonal $\mathbf{I}\sigma^2_{gi}\lambda_j$ and $\mathbf{I}\sigma_e^2$ respectively. In particular, the contribution of each j-th principal component to the genetic variance was assumed to be proportional to its corresponding eigenvalue (λ_i) $\sigma_{gi}^2 = (\sigma_a^2/k)*\lambda_j$ (Macciotta *et al.* 2010a). Variance components were supplied by breed associations. BLUP mixed model equations were solved by using Gauss-Seidel iterative method.

To evaluate the effect of the reduction of predictor dimensionality on DGV accuracy by PCA, DGV were calculated also with other two approaches that uses directly all markers available (R-BLUP and BAYES A), but with different theoretical assumptions on the distribution of marker effects.

R-BLUP.

In this model, marker effects were estimated using the same BLUP structure of [1]. In this case, **Z** is the design matrix of SNP genotypes – coded as 0,1 and 2 according to the number of the second allele. Marker effects were assumed to be sampled from the same normal distribution. (Co)variance matrix of SNP effects (**G**) was modelled as diagonal $\mathbf{I}\sigma_{gi}^2$, where $\sigma_{gi}^2=\sigma_a^2/n$ SNP. Mixed model equations were solved using a Gauss-Seidel iterative algorithm until convergence.

BAYES A.

A Bayes A model (BAYES A) that allows for variance to differ across chromosome segments (Meuwissen *et al.* 2001) was fitted:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{Wu} + \mathbf{e} \qquad [2]$$

where **W** is the incidence matrix that allocate the animal with their phenotypic record and **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 numerator relationship matrix and σ_a^2 is the additive genetic variance. The other symbols were the same as in [1]. Prior structure and hyper-parameters were chosen according to (Meuwissen *et al.* 2001). A scaled inverted chi-squared prior distribution was assumed for SNP specific variances, under the hypothesis that most of markers have nearly zero effects and only few have large effects. A total of 20,000 iterations were performed, discarding the first 10,000 as burn-in and considering no thinning interval. A residual updating algorithm was implemented to reduce computational time (Legarra & Misztal 2008).

DGV estimation and accuracy assessment.

The overall mean (μ) and the vector ($\hat{\mathbf{g}}$) of the PC score (or marker effects) estimated with the three above described methods were used to calculate the DGV for VAL bulls according to the formula [3]:

$$\hat{\mathbf{y}} = \mu + \sum_{k=1}^{n} \mathbf{Z}'_k \hat{\mathbf{g}} \qquad [3]$$

Where $\hat{\mathbf{y}}$ is the vector of DGV, **Z** is the matrix of PC scores (or marker genotypes) for validation bulls and *n* is the number of PC or markers used in the analysis.

The accuracy of the genomic prediction in the validation set was evaluated through analysis of Pearson correlation between EBV and DGV. Bias was assessed by examining regression coefficient of EBV on predicted DGV, and 95% confidence interval for b estimates have been

calculated. Mean squared error of prediction (MSEP) and its partition in different sources of variation related to systematic and random errors (Tedeschi 2006)were used to evaluate the goodness of prediction.

RESULTS

Accuracy of genomic prediction

The number of principal components to retain was assessed based on the pattern of DGV accuracies for increasing amounts of explained variance (Figure 2).

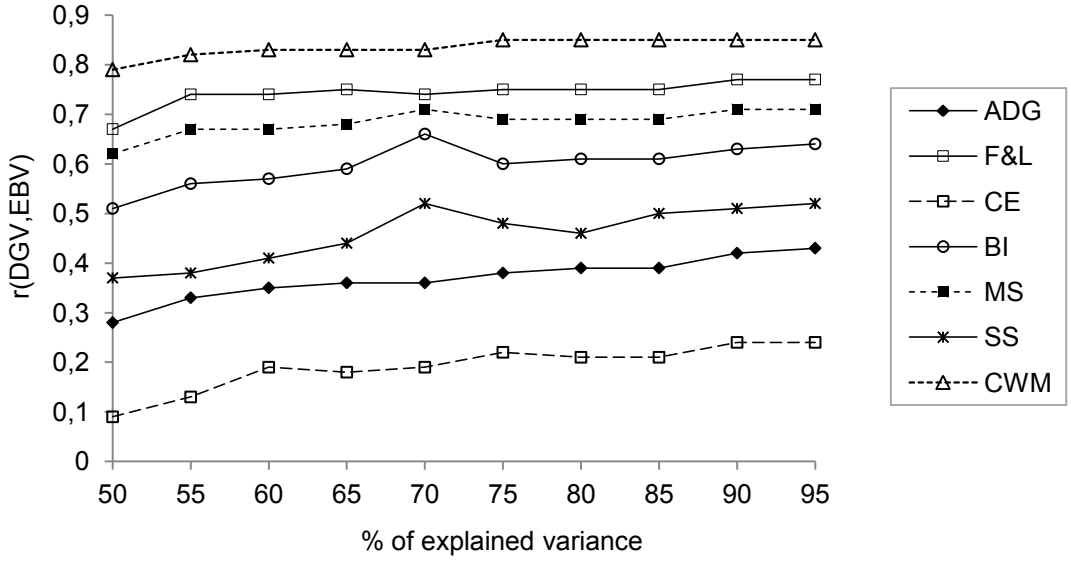

Figure 2. Number markers and number of PC components retained by chromosome.

A slight increase of DGV accuracy can be observed when the proportion of explained variance rose from 0.50 to 0.95 with a peak at 0.70 for some traits. This value, that corresponded to 2,466 extracted PCs was further used in the study. This figure minimized the computational demand of DGV estimation without losing in accuracy. The distribution of extracted PC by chromosome basically was proportional to the number of markers. (Figure 3).

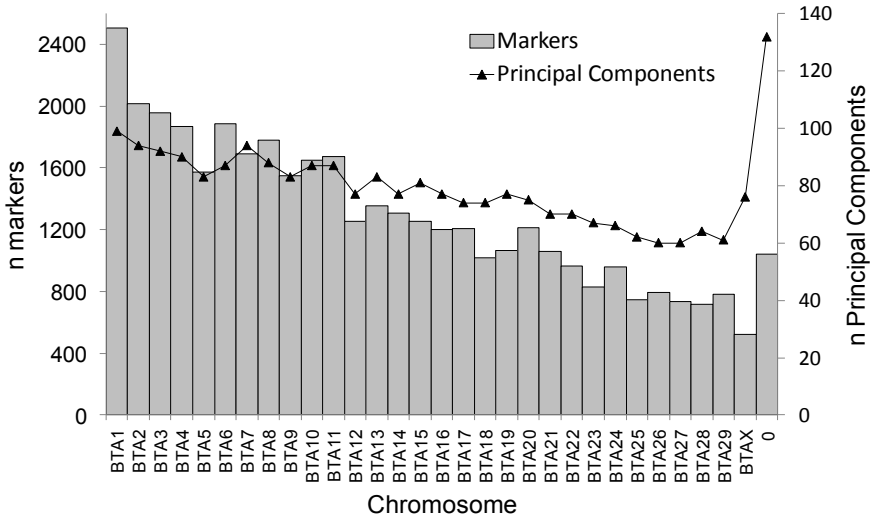

Figura 3. Pattern of DGV correlation function of % of variance explained by the PC of 7 meat traits (ADWG=average daily weight gain, FLS=Feet and leg score, CED=calving ease direct effect, MS=muscularity score, SS=Size Score, CWM=cow muscularity).

Table 2 reports the Pearson correlation coefficients between DGV and polygenic EBV across four different estimation methods and for different ratios REF:VAL.

Table 2. Correlation coefficient between DGV on EBV of average daily weight gain (ADWG), feet and leg score (FLS), Calving Ease direct (CED), Beef Index (BI), Muscularity Score (MS), Size Score (SS) and Cow Muscularity (CWM) for three estimation methods tested and 3 composition ratios of reference/validation set.

Trait[1]	PC-BLUP	PC-BLUP_EIGEN	R-BLUP	BAYES A	AVERAGE
	REF:VAL 70:30				
ADWG	0.39	0.39	0.43	0.41	0.41
SS	0.43	0.44	0.49	0.50	0.47
MS	0.73	0.67	0.73	0.73	0.72
FLS	0.72	0.73	0.70	0.72	0.72
BI	0.63	0.59	0.67	0.67	0.64
CED	0.23	0.27	0.18	0.23	0.23
CWM	0.80	0.73	0.80	0.81	0.79
	REF:VAL 80:20				
ADWG	0.36	0.35	0.45	0.39	0.39
SS	0.47	0.47	0.53	0.53	0.50
MS	0.67	0.64	0.70	0.72	0.68
FLS	0.74	0.70	0.74	0.76	0.74
BI	0.57	0.54	0.66	0.64	0.60
CED	0.23	0.27	0.20	0.20	0.23
CWM	0.85	0.84	0.83	0.85	0.84
	REF:VAL 90:10				
ADWG	0.53	0.51	0.58	0.54	0.54
SS	0.53	0.53	0.61	0.60	0.57
MS	0.81	0.79	0.78	0.81	0.80
FLS	0.85	0.84	0.79	0.83	0.83
BI	0.74	0.71	0.75	0.76	0.74
CED	0.24	0.34	0.22	0.27	0.27
CWM	0.83	0.81	0.81	0.83	0.82

Accuracies were moderate to high except for calving ease, which showed lowest values across all different validation sets and estimation methods (on average 0.24). In particular, highest accuracies were obtained for traits related to muscularity: averaging values across estimation methods accuracies of 0.73 and 0.82 were found for MS and CWM, respectively 0.76 (FLS) and 0.66 (BI). ADWG and SS showed moderate values (0.45 and 0.51, respectively).

In general, for larger ratios REF:VAL, the DGV accuracy tended to increase in almost all traits the best accuracy was obtained with a ratio REF:VAL 90:10 (Table 2). A slight effect of

the estimation method can be observed, even though without a clear pattern. R-BLUP performed best for average daily weight gain (accuracy of 0.49 averaged across REF:VAL ratios) compared to the other methods. A similar pattern can be observed for BI, due to the relevance of ADWG in its composition. As far as the SS is concerned, the two methods that used all the markers available were equivalent and showed better average accuracies than the PC based approaches (average values of 0.54 vs 0.48 respectively). No substantial differences can be observed for the other traits.

The use of eigenvalues of SNP covariance matrix as prior variance, as in the PC-BLUP_EIGEN approach, did not result in higher DGV accuracy, except for CED. For this trait, accuracy rose from 4% to 10% passing from REF:VAL 70:30 to 90:10. In general, for the other traits the PC-BLUP_EIGEN performed the same or slightly worse than PC-BLUP and the maximum difference between methods was 7%.

Accuracies obtained with methods that used simultaneously all markers as predictors were substantially equivalent. Basically, slightly higher accuracies were found using BAYES A whereas the maximum difference between the two SNP based methods was 6%. The mean accuracy averaging the values across traits and sizes of reference population was 0.60 (PC-BLUP), 0.59 (PC-BLUP_EIGEN) , 0.61 (R-BLUP) and 0.62 (BAYES A).

Bias and goodness of prediction assessment.

Regression coefficients between EBV and DGV were quite variable across methods (figure 4).

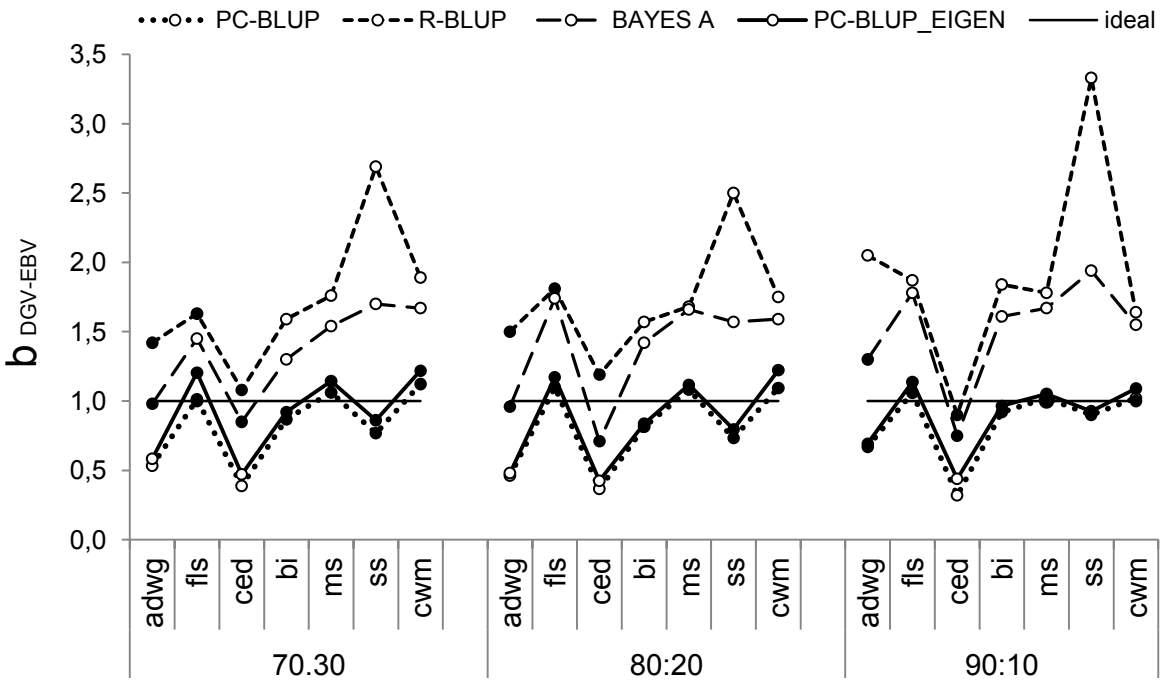

Figura 4. Pattern of regression coefficient of DGV vs EBV of 7 meat traits (ADWG=average daily weight gain, FLS=Feet and leg score, CED=calving ease direct effect, MS=muscularity score, SS=Size Score, CWM=cow muscularity).

In particular, PC-BLUP and PC-BLUP_EIGEN estimates showed the smallest regression coefficients which, in most of cases were lower than 1 (on average 0.82±0.27 and 0.89±0.28 respectively) but not significantly different from 1 (α<0.05) (figure 4). On the contrary, the methods that use SNP genotypes showed regression coefficients higher than 1 (on average 1.78±0.54 R-BLUP and 1.42±0.36 BAYES A) indicating that positive values of DGV underpredict EBV and vice versa for negative DGV values. Conversely to all other traits, the effect on prediction bias of CED was less defined: regression slopes tended to be closer to one only for the SNP genotypes methods whereas became worst for the PC based approaches. Furthermore, figure 4 shows the lowest variability of the regression coefficients of PC based

approaches across different traits in all REF:VAL ratios. Moreover, the PC-based estimates were less inflated than SNP based estimates, in particular PC-BLUP-EIGEN performed slightly better than PC-BLUP, especially when the reference population was larger (REF:VAL 90:10).

Table 3 reports the mean squared error of prediction of DGV calculated with the four methods for all traits and its decomposition.

Table 3. Mean squared error of prediction (MSEP) and its decomposition EBV on DGV for beef traits in the validation bulls using different estimation method.

Trait	MSEP[1]	RMSEP	MB	UV	IC	SB	RE
ADWG							
PC-BLUP	44.68	6.68	0.33	0.05	0.63	0.08	0.60
PC-BLUP_EIGEN	41.04	6.41	0.30	0.08	0.63	0.06	0.65
BLUP	38.79	6.23	0.33	0.39	0.28	0.01	0.66
BAYES A	41.14	6.41	0.37	0.26	0.38	0.00	0.64
SS							
PC-BLUP	43.71	6.61	0.09	0.21	0.71	0.02	0.90
PC-BLUP_EIGEN	42.42	6.51	0.08	0.27	0.66	0.01	0.92
BLUP	44.92	6.70	0.08	0.72	0.20	0.10	0.82
BAYES A	42.93	6.55	0.11	0.57	0.33	0.05	0.85
MS							
PC-BLUP	63.15	7.95	0.23	0.17	0.61	0.00	0.77
PC-BLUP_EIGEN	61.84	7.86	0.10	0.28	0.63	0.01	0.90
BLUP	59.66	7.72	0.06	0.57	0.38	0.17	0.79
BAYES A	58.70	7.66	0.10	0.47	0.44	0.11	0.79
FLS							
PC-BLUP	40.01	6.33	0.33	0.11	0.56	0.00	0.67
PC-BLUP_EIGEN	34.50	5.87	0.22	0.25	0.54	0.03	0.76
BLUP	39.73	6.30	0.18	0.46	0.37	0.11	0.72
BAYES A	40.75	6.38	0.27	0.35	0.39	0.07	0.67
BI							
PC-BLUP	36.25	6.02	0.36	0.08	0.56	0.01	0.64
PC-BLUP_EIGEN	32.76	5.72	0.25	0.15	0.61	0.00	0.75
BLUP	29.93	5.47	0.23	0.42	0.35	0.08	0.70
BAYES A	31.86	5.64	0.31	0.28	0.41	0.03	0.66
CED							
PC-BLUP	49.13	7.01	0.02	0.14	0.85	0.13	0.86
PC-BLUP_EIGEN	46.54	6.82	0.02	0.17	0.82	0.09	0.89
BLUP	44.79	6.69	0.04	0.69	0.28	0.00	0.97
BAYES A	43.44	6.59	0.03	0.55	0.43	0.00	0.98
CWM							
PC-BLUP	42.02	6.48	0.01	0.23	0.77	0.02	0.98
PC-BLUP_EIGEN	55.16	7.43	0.02	0.33	0.66	0.04	0.96
BLUP	58.39	7.64	0.03	0.64	0.33	0.27	0.70
BAYES A	51.04	7.14	0.01	0.59	0.41	0.23	0.77

1) MB = Mean Bias; UV = Unequal variances; IC = Incomplete covariation; SB = Slope bias; RE = Random errors. Note that MB + UV+ IC= MB + SB + RE = 1

MSEP did not show large variation among traits excepted for MS (average of 32.7) that experienced the lower figure and BI with the highest MSEP (average of 60.8). Within traits, MSEP of DGV obtained with PC based approaches were on average higher than those of DGV calculated with SNP based approaches. Exceptions were observed for SS, FLS and CWM. PC-BLUP_EIGEN showed MSEP always lower than PC_BLUP except for CWM. In any case, MSEP differences among methods were rather small. On the other hand, larger differences in the MSEP composition can be highlighted. In general, mean bias was not very high (highest average value has been found for ADWG 0.33) and for some traits was close to zero. Systematic bias was very low for all traits being the maximum obtained for CWM (27% and 23% of the MSEP for BLUP and BAYES A respectively). A large incidence of random errors can be observed among traits with values ranging from 63.75% (ADGW) to 92.5% (CED). Methods that use PC as predictors showed the lowest incidence of components related to prediction bias, as inequality of variance, for all traits. Furthermore, sources of random variation as incomplete co-variation and sometimes, random errors, showed higher figures for PC based compared to SNP based methods.

DISCUSSION

In this paper, principal component analysis was used for reducing predictor dimensionality and computational demand in the calculation of DGV for beef traits in a dual purpose cattle breed. Possible effects of such a reduction on the accuracy of predicted DGV have been tested by comparing the PC results with other methods that use all predictors available.

The number of PC retained was about 6% of the number of original marker genotypes corresponding to around 2,500 PC. The magnitude of the reduction is similar to this reported for milk yield on US Holsteins (Long *et al.* 2011). The reduction of computational times is not the only advantage of using PC. Problems of multicollinearity that may occur when high number of SNP are used as predictors could be prevented through the use of PC that are uncorrelated variables. The option of extracting PC from SNP data by chromosome allows to work with full rank (co)variances matrix (Dimauro *et al.* 2011) reducing both multicollinearity and bias of estimates in comparison to techniques that use simultaneously all genotypes available as predictors without losses in accuracy.

In general, DGV accuracies here obtained were high to moderate, with the exception of CED that showed rather low values. Results are rather difficult to compare with other studies being most of literature on beef cattle related to genome wide association analysis. Thus results from dairy cattle studies were considered for traits as SS and CED even if different estimation methodologies and population sizes.

Values of reliability found in literature were converted into simple correlation using $\sqrt{r^2}$. For SS, the accuracies found in the present study were similar to those reported by Olson *et al*. (2011) for Brown Swisss and Jersey bulls using BAYES B. (Liu *et al.* 2011) reported a value of 0.71 in German Holstein (n=5025). The values of accuracy for SS presented in these works

were relatively in agreement with results of other studies in which the reference populations were from 1 to 15 fold larger

More emphasis has been put also analyzing CED. Calving ease and growth traits are particularly important in beef cattle selection in particular for their unfavorable genetic correlation.

Selection for CED aims to prevent distocya (Johanson & Berger 2003). CED experienced lower values in all different validation set and for all methods. Although CED is a low heritability trait (h^2 = 4.9% in Italian Simmental), for which the environmental component is by far the most important in phenotypic expression, the genetic component can be exploited for selective purposes, and calving ease may take a great advantage by application of GS procedure. The values of CED accuracy are by far the lowest in the Simmental dataset. These results are not fully comparable with results found in other breeds. (Garrick 2011) using around 2000 Angus bulls found that DGV accuracy using BAYES C ranged from 0.47 to 0.63 according to the composition of the reference population. (Luan *et al.* 2009) in a study on Norwegian Red Cattle with a similar population size (n=500 bulls) found superiority of using BAYES B (0.43) against BLUP (0.41). Accuracies of 0.38 were found for Piedmontese breed using both PC BLUP or SNP BLUP methods (Ajmone *et al.* 2010). All these values were higher than results presented in this paper. Moreover (Olson *et al.* 2011) found values of accuracy ranging from 0.26 to 0.32 in Holstein, and from 0.10 to 0.39 in Brown Swiss using BAYES B. The latter correlations were surprisingly close to the values of accuracy in Italian Simmental, nonetheless the differences in population size.

DGV accuracy across estimation methods

No large differences in DGV accuracies can be highlighted among methods (average 4%). The maximum difference (9%) has been found for CED when PC-BLUP_EIGEN has been used which accuracy was higher. In general the reduction of predictors dimensionality slightly

affected the accuracy of DGV. PC-BLUP and PC-BLUP_EIGEN performed weakly worse than BAYES A or R-BLUP for some traits.

The above mentioned slight differences are more than counterbalanced by the huge save in calculation times provided by the PCA based approaches. In fact, about 1 minute was needed for PC-BLUP estimation using a personal computer with a 2.33 GHz Quad core processor and 3.25 Gb of RAM. On the other hand, on average from 5 to 8 hours were needed for the SNP_BLUP and BAYES A approaches using a Linux server with 4 x 4 quad core processors and 128 Gb RAM. Methods used in this research basically differed in two aspects. One deals with the dimensionality and the type of predictors, that differentiates the PC-BLUP and PC-BLUP_EIGEN from the other two approaches.

Results obtained in this paper confirm the substantial equivalence between the direct use of SNP genotypes and the use of reduced number of multivariate synthetic variables as DGV predictors already observed on simulated (Solberg *et al.* 2009; Macciotta *et al.* 2010a) and real data for milk traits (Macciotta *et al.* 2010b; Long *et al.* 2011). The second point regards the distribution of predictor effects. Two methods, PC-BLUP and R-BLUP, assume an equal contribution of each predictor (SNP or PC score) on the variance of the trait whereas the BAYES A and PC-BLUP_EIGEN relies on a heterogeneity of variance across predictor effects. The use of different estimation methods based on different assumption did not come up in very different results. The theoretical assumption of BLUP methods is general in contrast with theoretical distribution of QTL effects proposed in literature (Hayes & Goddard 2001; Chamberlain *et al.* 2007). Although BLUP is not very accurate to reproduce the true distribution of gene effects or to dissect the genetic architecture of a complex quantitative trait, anyhow offer robust estimates of breeding values with reasonable accuracies (Goddard 2009). Early results on simulated data have highlighted the net superiority of the BAYES method over the BLUP approach (Meuwissen *et al.* 2001; Habier *et al.* 2007), confirming the

suitability of the finite locus model. However, researches on real data for dairy traits have substantially underlined a substantial equivalence between methods (Moser *et al.* 2009; VanRaden *et al.* 2009; Su *et al.* 2010). This basic equivalence has been confirmed also in this work. Given the present results, the PCA based methods can be proposed for calculating DGV for beef traits.

Among the factors that affect the DGV accuracy the size of references population and the heritability of the traits are those ones mainly involved. The lower the heritability the larger the references population needs to be (Hayes *et al.* 2009b). This can partially explain the lower values of DGV accuracy find for CE in the present work. Simulated results showed how the heritability of the trait affect positively the estimation accuracy (Calus & Veerkamp 2007; Kolbehdari *et al.* 2007) as confirmed also by theoretical expectations (Daetwyler *et al.* 2008). The combination of low heritability and reduced population size may be able to explain the results presented here on CED accuracy.

The increase of the size of the reference population has been widely reported to improve the accuracy of genomic prediction (Meuwissen *et al.* 2001). Moreover, (Liu *et al.* 2011) presented results where a positive effect of population size on the reliability of genomic indices can be highlighted in German Holstein. Results presented by (VanRaden *et al.* 2009; Olson *et al.* 2011) confirm this general pattern. Also in this paper, a larger size of the REF population resulted in an increase of correlations between DGV and EBV even if the increase in reference population were moderate. These results found their theoretical justification in the reduction of the statistical asymmetry of data matrix due to the increased sample size.

Within each trait a large variability of *b* coefficient among different methods can be envisaged in all ratios of references/validation bulls. It is worth to notice a reduced variability of regression coefficients among different traits when PC based approach were applied. Furthermore estimates of DGV obtained using PC based methods showed the lowest bias, and

the decomposition of mean square of prediction confirmed it. In fact approaches based on the use of PC as predictors showed a larger incidence of random variation compared to the other two methods.

Examples of possible scenarios of application of genomic prediction in beef industry have been provided by (Garrick 2011). The majority of beef traits are recorded in performance test. Only few traits are recorded late in life or imply the slaughtering of the animals. In Particular, in double purpose breed, like Simmental, genomic selection may be helpful to select for both dairy traits (progeny testing) and to increase the accuracy of beef traits (mainly did in performances).

CONCLUSIONS

Although no clear and unambiguous pattern of DGV accuracy across traits or methods have been detected in the present work, it seems that the differences in accuracy are mainly related to the trait analyzed, size and structure of training population rather than model used to develop the prediction equations. However, the increase of the magnitude of accuracy found in the present study may not be exclusively due to the ratio REF:VAL. Other possible interpretation of the presented DGV accuracy may be the effects of the relatedness between reference and validation bulls which affects the accuracy as shown by (Habier *et al.* 2010) that split the observed accuracy into two component, one related to LD and the other due to the relatedness of bulls in training and prediction population. Being around 69 the number of sire-son pairs a possible effect of the relatedness might be envisaged. A high number of phenotypic records are needed to achieve reasonable accuracy as to overcome the curse of dimensionality and GS implementation.

This fact is emphasized when a small number of genotyped bulls is available, as in the case of this study. Although PCA does not completely address of such an issue, among methods PCA yielded similar results in comparison to other methods but with reduced calculation speed and less estimation bias. Both issues are particularly relevant for small population size.

REFERENCES

Bolormaa S., Pryce J.E., Hayes B.J. & Goddard M.E. (2010) Multivariate analysis of a genome-wide association study in dairy cattle. *Journal of Dairy Science* **93**, 3818-33.

Calus M.P.L. & Veerkamp R.F. (2007) Accuracy of breeding values when using and ignoring the polygenic effect in genomic breeding value estimation with a marker density of one SNP per cM. *Journal of Animal Breeding and Genetics* **124**, 362-8.

Chamberlain A.J., McPartlan H.C. & Goddard M.E. (2007) The number of loci that affect milk production traits in dairy cattle. *Genetics* **177**, 1117-23.

Cole J.B., VanRaden P.M., O'Connell J.R., Van Tassell C.P., Sonstegard T.S., Schnabel R.D., Taylor J.F. & Wiggans G.R. (2009) Distribution and Location of Genetic effects for Dairy traits (vol 92, pg 2931, 2009). *Journal of Dairy Science* **92**, 3542-.

Daetwyler H.D., Villanueva B. & Woolliams J.A. (2008) Accuracy of predicting the genetic risk of disease using a genome-wide approach. *Plos One* **3**, e3395.

Dimauro C., Cellesi M., Pintus M.A. & Macciotta N.P.P. (2011) The impact of the rank of marker variance–covariance matrix in principal component evaluation for genomic selection applications. *Journal of Animal Breeding and Genetics*, no-no.

Garrick D.J. (2011) The nature, scope and impact of genomic prediction in beef cattle in the United States. *Genetics Selection Evolution* **43**.

Goddard M. (2009) Genomic selection: prediction of accuracy and maximisation of long term response. *Genetica* **136**, 245-57.

Habier D., Fernando R.L. & Dekkers J.C.M. (2007) The impact of genetic relationship information on genome-assisted breeding values. *Genetics* **177**, 2389-97.

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

Hayes B. & Goddard M.E. (2001) The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution* **33**, 209-29.

Hayes B.J., Chamberlain A.J., Maceachern S., Savin K., McPartlan H., MacLeod I., Sethuraman L. & Goddard M.E. (2009a) A genome map of divergent artificial selection between Bos taurus dairy cattle and Bos taurus beef cattle. *Animal Genetics* **40**, 176-84.

Hayes B.J., Visscher P.M. & Goddard M.E. (2009b) Increased accuracy of artificial selection by using the realized relationship matrix. (vol 91, pg 47, 2009). *Genetics Research* **91**, 143-.

Johanson J.M. & Berger P.J. (2003) Birth Weight as a Predictor of Calving Ease and Perinatal Mortality in Holstein Cattle. *Journal of Dairy Science* **86**, 3745-55.

Kolbehdari D., Schaeffer L.R. & Robinson J.A.B. (2007) Estimation of genome-wide haplotype effects in half-sib designs. *Journal of Animal Breeding and Genetics* **124**, 356-61.

Legarra A. & Misztal I. (2008) Technical note: Computing strategies in genome-wide selection. *Journal of Dairy Science* **91**, 360-6.

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

Long N., Gianola D., Rosa G.J.M. & Weigel K.A. (2011) Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. *Journal of Animal Breeding and Genetics*, no-no.

Luan T., Woolliams J.A., Lien S., Kent M., Svendsen M. & Meuwissen T.H.E. (2009) The Accuracy of Genomic Selection in Norwegian Red Cattle Assessed by Cross-Validation. *Genetics* **183**, 1119-26.

Macciotta N.P.P., Gaspa G., Steri R., Nicolazzi E.L., Dimauro C., Pieramati C. & Cappio-Borlino A. (2010a) Using eigenvalues as variance priors in the prediction of genomic breeding values by principal component analysis. *Journal of Dairy Science* **93**, 2765-74.

Macciotta N.P.P., Pintus M.A., Steri R., Pieramati C., Nicolazzi E.L., Santus E., Vicario D., van Kaam J.T., Nardone A., Valentini A. & Ajmone-Marsan P. (2010b) Accuracies of direct genomic breeding values estimated in dairy cattle with a principal component approach. *Journal of Dairy Science* **93**, 532-3.

Meuwissen T.H.E., Hayes B.J. & Goddard M.E. (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819-29.

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

Olson K.M., VanRaden P.M., Tooker M.E. & Cooper T.A. (2011) Differences among methods to validate genomic evaluations for dairy cattle. *Journal of Dairy Science* **94**, 2613-20.

Rolf M.M., Taylor J.F., Schnabel R.D., McKay S.D., McClure M.C., Northcutt S.L., Kerley M.S. & Weaber R.L. (2011) Genome-wide association analysis for feed efficiency in Angus cattle. *Animal Genetics*, no-no.

Solberg T.R., Sonesson A.K., Woolliams J.A. & Meuwissen T.H.E. (2009) Reducing dimensionality for prediction of genome-wide breeding values. *Genetics Selection Evolution* **41**, -.

Su G., Guldbrandtsen B., Gregersen V.R. & Lund M.S. (2010) Preliminary investigation on reliability of genomic estimated breeding values in the Danish Holstein population. *Journal of Dairy Science* **93**, 1175-83.

Tedeschi L.O. (2006) Assessment of the adequacy of mathematical models. *Agricultural Systems* **89**, 225-47.

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

Wiggans G.R., Sonstegard T.S., Vanraden P.M., Matukumalli L.K., Schnabel R.D., Taylor J.F., Schenkel F.S. & Van Tassell C.P. (2009) Selection of single-nucleotide polymorphisms and quality of genotypes used in genomic evaluation of dairy cattle in the United States and Canada. *Journal of Dairy Science* **92**, 3431-6.

CHAPTER 5

CONCLUSIONS

The problem of the dimensionality of predictors in the calculation of direct genomic values for dairy and beef traits has been approached with a multivariate dimension reduction method, the Principal Component Analysis. In the different experimental contributions, the n>>p problem was particularly relevant, due to the small size of the samples of animals analysed compared to researches carried out in other countries. Actually, this is a common issue of GS, especially at the start of programmes or for breeds of limited size.

In general, the principal component analysis was effective in reducing the number of predictors by a factor of ranging from 94% to 60%. The more evident consequence has been a huge reduction in computing time. Moreover, the relevant decrease of the number of independent variables used in the estimation was not accompanied by a reduction in the accuracy of DGV. These results, obtained in different traits (production, functional) confirmed what already observed on simulated data.

A possible criticism to the use of the PCA-based approach, or other dimension reduction techniques, is that at present the main constraint for the implementation of GS programmes is represented by genotyping costs rather than the availability of computing resources. It may be argued that saving computational time is not a trivial issue, especially if the recent availability of high density SNP platforms and of the whole genome sequence are considered. The handling of such amount of information, and the relevant problem of multicollinearity between strictly adjacent SNP would presumably result in a requirement of a synthesis (Meuwissen & Goddard 2010; VanRaden *et al.* 2011)

DGV accuracies with the PCA approach did not substantially differ from those obtained with two of the most popular estimation methods used in GS, i.e. BLUP regression and the Bayes A. In particular, the PCA was almost always slightly better than the BLUP and equal to Bayes A except from some composition and beef traits. These results are of great importance and, if confirmed on larger data sets, may support the use of this method in current GS programmes.

The approach here proposed has not a strong theoretical foundation in terms of underlying genetic model. Actually, the use of PCA can be considered as a sort of return to the black-box approach of the quantitative genetics. It is a matter of fact that the extracted new variables do not possess a definite genetic meaning. On the other hand, they are able to detect differences in the genetic structure of animals, evidencing both between and within breeds variation, as reported for the BTA6 in the study on Italian Brown and Italian Simmental bulls. Moreover, the use of eigenvalues as variance priors offers an opportunity to assign a relative weight to each component based on their relevance to the (co)variance structure of SNP. Finally, the absence of an underlying genetic model for the PCA (Jombart *et al.* 2009) may be not regarded necessarily as a weak point. If several experimental evidences on causal mutations of genes with a relevant effect on phenotypic expressions may indicate an inadequacy of the infinitesimal model, results on the genetic dissection of some complex traits as height in humans (Visscher *et al.* 2007) raised some points about the finite loci model.

Accuracies of direct genomic values obtained in the different experimental contributions were low to moderate, for dairy and functional traits, and high for some beef traits. The results were somewhat expected, considering the small size of the populations of genotyped animals in the different studies. In any case, the effect of the size of the reference population in the accuracy of DGV has been evidenced, especially if results obtained on Holsteins (that had the larger population size) are compared to those for Brown and Simmental. Differences observed between breeds for the same traits may have different explanations. First of all the sampling effect. However, also differences in breeding goals (dairy for Holstein and Brown, dual purpose for Simmental), selection pressure, demographic (effective population) and in genetic structure (linkage disequilibrium) are among the possible causes.

Differences between traits do not seem to follow a defined pattern. In some cases, higher accuracies have been found for high heritability traits, in agreement with reports of other

authors. But in other cases, as for SCS in Holstein, this relationship did not hold. The GS seems to be more effective, as highlighted by composition traits in Brown and Holstein, for traits that are known to be affected by a few genes with a large effect. The results obtained for fat percentage in Holsteins are worth to be mentioned. In this breed the SNP markers located in the region of BTA14 were segregating (whereas in the other two breeds were monomorphic). Only in this case, the Bayesian approach gave accuracies markedly higher than BLUP based methods.

Of particular interest are results on meat traits. Until now, the impact of GS on beef cattle has been less dramatic than in dairy breeds. Beef breeding has shortest cycles and therefore the effect of an anticipated genetic evaluation has less importance. However, especially for dual purpose breeds, the availability of DGV with good accuracy for beef traits could be an opportunity for limiting the number of animals subjected to performance test and, therefore, for reducing the costs of selection schemes.

Finally, the method is easy to implement also in the routine of breed associations. The principal component extraction, carried out by chromosome, takes about 20 minutes of computing time. It may be performed with commercially available softwares or implemented in the genetic evaluation pipelines as fortran routines. Also the optimization of the PC number, that may be different across traits, based on the amount of variance explained can be easy implemented. The method can be further modified by the inclusion of a polygenic effect able to explain a fixed quota (for example 10%) of the original variance.

The PCA approach can be therefore suggested as a valid alternative to the direct use of SNP genotype in the calculation of direct genomic values in GS programmes. Further investigations could deal with its use on populations of larger size, the optimisation of the extraction method, and the application of higher density platform and the whole genome sequence.

REFERENCES

Jombart T., Pontier D. & Dufour A.B. (2009) Genetic markers in the playground of multivariate analysis. *Heredity* **102**, 330-41.

Meuwissen T. & Goddard M. (2010) Accurate Prediction of Genetic Values for Complex Traits by Whole-Genome Resequencing. *Genetics* **185**, 623-U338.

VanRaden P., O'Connell J., Wiggans G. & Weigel K. (2011) Genomic evaluations with many more genotypes. *Genetics Selection Evolution* **43**, 10.

Visscher P.M., Macgregor S., Benyamin B., Zhu G., Gordon S., Medland S., Hill W.G., Hottenga J.J., Willemsen G., Boomsma D.I., Liu Y.Z., Deng H.W., Montgomery G.W. & Martin N.G. (2007) Genome partitioning of genetic variation for height from 11,214 sibling pairs. *American Journal of Human Genetics* **81**, 1104-10.