Use of a principal component approach for estimating Direct Genomic Breeding Values for Somatic Cell Score in dairy cattle

Nicolo PP Macciotta[*], Maria A Pintus, Giustino Gaspa , Ezequiel L Nicolazzi[Â], Attilio Rossoni[ÿ], Daniele Vicario[§], Jan-Thjis van Kaam[**], Alessandro Nardone[ÀÀ], Alessio Valentini, Paolo Ajmone-Marsan

Introduction

Availability of dense SNP platforms has allowed the prediction of direct genomic breeding values (DGV) as the sum across the genome of marker effects on the trait of interest. A main statistical issue in genomic selection is represented by the large number of predictors (currently around 40K SNPs) and the small number of phenotypes available (few thousands). Such a problem becomes relevant in genomic projects involving different breeds, some of a limited size. An approach that can be used to reduce the number of independent variables when predicting DGV is based on the use of principal component analysis (PCA) (Solberg et al., 2009; Macciotta et al., 2010). Compared to other methods that reduce the number of predictors according to their distribution along the genome (VanRaden et al., 2009) or to their contribution to the phenotypic variance of the trait considered (Meuwissen et al., 2001), the PCA approach modifies the emphasis of each SNP in the extracted variables based on its contribution on the total marker variance. In this work, PCA is used to reduce the number of independent variables in the prediction of DGV for Somatic Cell Score in three dairy cattle Breeds farmed in Italy.

Material and methods

Data. Bulls of three dairy breeds were genotyped with the 54K SNP Illumina beadchip: 863 Holstein, 572 Brown, 479 Simmental. Markers were discarded on the basis of their minor allele frequency (<0.05), deviation from the hardy Weimberg equilibrium (<0.01), mendelian inheritance and absence of heterozygotes. Missing data were replaced by the most frequent allele. Phenotypes were polygenic EBVs published by the different breed associations for somatic cell score. For cross validation purposes, the data set was split into training and validation individuals. Two criteria were used to create the data sets: birth year of bulls or random. For birth year, animals included in the training data were those born before 1998 or

[*] Dipartimento di Scienze Zootecniche, Università di Sassari, Italia
[Â] Istituto di Zootecnica, Università Cattolica del Sacro Cuore, Piacenza, Italia
[ÿ] ANARB, Bussolengo, Italia
[§] ANAPRI, Udine, Italia
[**] ANAFI, Cremona, Italia
[ÀÀ] Dipartimento di Produzioni Animali, Università della Tuscia, Viterbo, Italia

2000. In the case of random choice, two ratio training:prediction were considered: 70%-30% and 90%-10%.

Statistical analyses. PCA was carried out on the matrix of SNP genotypes (n. animals x n. of markers) coded as -1,0, and 1 respectively. The analysis was performed by chromosome and separately for each breed. The number of principal component (PC) to be retained was based of the amount of explained variance (>80%). PC scores were calculated for all animals. Effects of PC on phenotypes were estimated in the training data set with the following mixed linear model

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

where: **y** is the vector of EBV or DYD; **Z** is the (m x k) design matrix of random effects, where each element corresponds to the score of the k-th component of the m-th animal in the training data set; **g** is the vector of solutions for random regression coefficients of PC scores; **e** is the random residual. An equal contribution of each PC to the variance of the trait was assumed. DGV of bulls of the validation data set were calculated by using effects of PC scores $\check{\mathbf{g}}$ estimated in the training data as

$$\mathbf{DGV} = \mu + \mathbf{Z}\check{\mathbf{g}}$$

where **Z** is the matrix of PC score coefficients in the validation data set. Accuracy of DGV prediction was assessed by calculating correlation between DGV and EBV for validation bulls.

Results and discussion

The number of the SNP retained after edits and the corresponding extracted principal components is reported in table 1. It can be observed that the technique was able to reduce of about 95% the dimension of the system. It should be remembered that this huge decrease of predictors does not correspond to the direct elimination of markers, because each SNP enters in the composition of each principal component. The ratio PC extracted/number of SNPs is comparable with those reported for simulated data (Macciotta et al., 2010; Solberg et al., 2009). Slight differences in the number of retained PC between breeds can be ascribed to the number of original variables.

Table 1. Number of SNP retained after edits and number of extracted principal components for each breed.

Breed	SNP	Principal components
Holstein	40,658	2,564
Brown	37,254	2,257
Simmental	40,179	2,476

had a relevant effect in calculation time. DGV accuracies for validation bulls (Table 2), when older animals are used for training, show the highest values (about 0.622) for Holsteins, lowest for Simmental (about 0.35), with Brown in the middle. Besides the different sample size, these figures may reflect differences in the genetic structure of the three breeds. Accuracies here obtained are lower than those published for SCS in US and New Zealand Holsteins (VanRaden et al., 2009; Harris and Johnson, 2010). A part from the estimation approaches and the methods used to calculate reliability, the small sample size considered in this study may be an explanation for such differences. Moreover, it can be also taken into account the range of birth year of bulls: 1979-2004, 1960-2002, and 1972-2002 for Holstein, Brown and Simmental, respectively.
Better results have been obtained, especially for Brown and Simmental, when animals of training and validation data set were taken at random. Values are comparable with those reported for production and udder health traits in Danish and Australian Holsteins (Moser et al., 2009; Su et al., 2010), obtained with more complex estimation methods.

Table 2. Correlations between DGV and polygenic EBV for Somatic Cell Score in the different training/validation scenarios

			by year	
Scenario	Phenotype	Holstein	Brown	Simmental
<1998	EBV	0.62	0.49	0.33
<2000	EBV	0.63	0.46	0.36
			Random	
70:30	EBV	0.61	0.65	0.43
90:10	EBV	0.68	0.69	0.49

Conclusion

In this study suitability of the principal component analysis as an approach to reduce the dimensionality of predictors for calculating direct genomic breeding values for somatic cell score is tested. Accuracies obtained are lower that those obtained on US data when older bulls are used to estimate predictor coefficients for younger bulls. This result was expected, provided the reduced size of the sample of animals considered and the low heritability of the trait. However, correlations between DGV and EBV are comparable with those reported in other studies with larger samples and more complex methods when validation animals are picked up at random. These latter results confirm the usefulness of such a dimension reduction method for genomic selection purposes.

Acknowledgements

This research was funded by the Italian Ministry of Agriculture (grant SelMOL)

References

Harris, B.L., and Johnson, D.L. (2010). *J. Dairy. Sci.*, 93:1243ï 1252.

Meuwissen, T.H.E., Hayes, B.J., and Goddard, M. (2001). *Genetics*, 157:1819ï 1829.

Moser, G., Tier, B., Crump, R.E. et al. (2009). *Gen. Sel. Evol.* 41 :56.

Macciotta, N.P.P., Gaspa, G., Steri, R. *et al.* (2010). *J. Dairy. Sci.*, in press.

Solberg, T.R., Soensson, A.K., Wooliams, J. A. *et al.* (2009) *Gen. Sel. Evol.* 41 :29.

Su, G., Guldbrandtsen, B., Gregersen, V.R. *et al.* (2010). *J. Dairy. Sci.*, 93:1175ï 1183.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R. *et al.* (2009). *J. Dairy. Sci.*, 92:16ï 24.