A COMPARISON OF METHODS FOR GENOMIC SELECTION IN AUSTRIAN DUAL PURPOSE SIMMENTAL CATTLE

B. Gredler[1], K. G. Nirea[1,2], T. R. Solberg[3], C. Egger-Danner[4], T. Meuwissen[2] and J. Sölkner[1]

[1] University of Natural Resources and Applied Life Sciences Vienna, Department of Sustainable Agricultural Systems, Division of Livestock Sciences, Gregor Mendel Str. 33, A-1180 Vienna
[2] Norwegian University of Life Science, Department of Animal and Aquacultural Sciences, Box 5003, N-1432 Ås
[3] Geno Breeding and AI Association, Box 5003, N-1432 Ås
[4] ZuchtData EDV-Dienstleistungen GmbH, Dresdner Str. 89/19, A-1200 Vienna

SUMMARY

The objective of this study was to compare partial least squares regression (PLSR), multivariate regression analysis using least absolute shrinkage and selection operator (LASSO), a Bayesian approach (BayesC) and an ordinary BLUP method (GS-BLUP) for the estimation of genome-wide breeding values for dual purpose Simmental Fleckvieh in Austria. A five-fold cross validation and a forward prediction were carried out for the traits protein yield, fat percentage, somatic cell count, and non return rate after 56 days in cows. Using cross validation, accuracies of genome-wide breeding values were in the range of 0.30 to 0.74. In the forward prediction, obtained accuracies were between 0.21 and 0.60. BayesC gave slightly better accuracies in forward prediction than the other methods.

INTRODUCTION

The use of molecular markers for improvement of genetic evaluation has been a major issue in animal breeding for many years. High throughput genotyping technologies enable the genotyping of more than 50,000 single nucleotide polymorphisms (SNP). Genomic selection, first introduced by Meuwissen *et al.* (2001), refers to the use of dense markers covering the whole genome to estimate genome-wide breeding values. In this simulation study the authors reported that it was possible to reach accuracies of genome-wide breeding values of 0.85 using markers only. So far, very few studies have reported results of genomic selection using real data. The objective of this study was to carry out a first comparison of methods for the estimation of genome-wide breeding values in dual purpose Simmental cattle in Austria.

MATERIAL AND METHODS

Data. 1,363 dual purpose Simmental (Fleckvieh) bulls, genotyped with the Illumina Bovine SNP50^{TM} Beadchip with a call rate >=95%, were included in the analysis. Bulls were born from 1990 to 2003. The distribution of bulls across birth years is shown in Figure 1. For method validation a five-fold cross validation was carried out. Bulls for the training and test set were randomly sampled across all birth years that 1,091 and 272 bulls were in the training and test set, respectively. Five replicates were carried out resulting in five pairs of training and test sets. In addition, the data set was split into a reference population (training set) including bulls born before 2001 (1,037 bulls) and a test set of bulls born between 2001 and 2003 (326 bulls) for forward prediction.

To be included in the analyses, a minimum minor allele frequency of 1 % was required for each SNP. To test for Hardy-Weinberg equilibrium, the deviation of observed genotype frequencies from expected genotype frequencies based on allele frequencies was calculated. SNP were included if Hardy Weinberg χ^2 values were below 600. In total, 45,519 SNP met all the SNP

selection criteria. As none of the methods described below allows for missing values, missing genotypes were filled in according to allele frequencies by sampling random numbers from a uniform distribution. The phenotypes used were estimated breeding values based on progeny testing obtained from the joint routine genetic evaluation in Austria and Germany for protein yield (Prot-kg), fat percentage (Fat%), somatic cell count (SCC) and non return rate after 56 days (NR56) for cows. Progeny testing involves samples of 50-100 daughters in Austria and Germany.

Model of analysis. In this study, we compared partial least square regression (PLSR), regression analysis using least absolute shrinkage and selection operator (LASSO), a Bayesian approach (BayesC; Meuwissen, 2009) and a BLUP approach as described by Meuwissen *et al.* (2001). For running PLSR and LASSO, the SAS procedures PROC PLS and PROC GLMSELECT were used (SAS, 2007). To assess the accuracy of genomic selection, the correlation between estimated genome-wide breeding values (GEBV) and current estimated breeding values (EBV) based on progeny testing was calculated. The regression coefficient of the current breeding value on the genome-wide breeding value was computed to assess the bias of genome-wide breeding values.

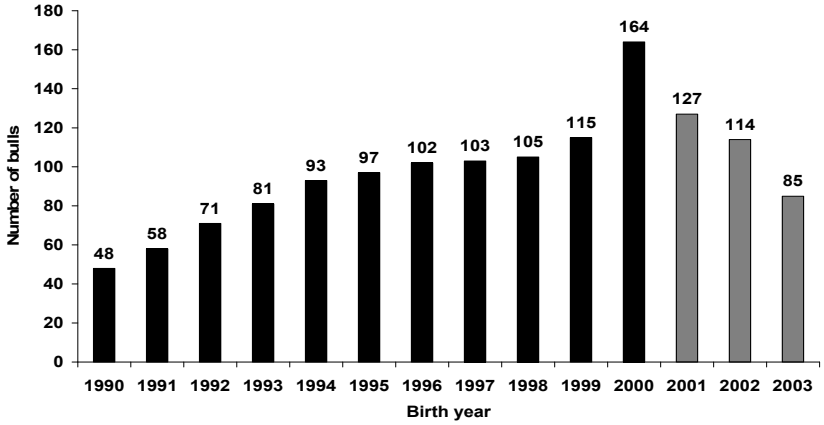

Figure 1. Distribution of bulls across to birth year (black and grey bars represent the number of bulls in the training and test set in forward prediction, respectively).

RESULTS AND DISCUSSION

Accuracies of genome-wide breeding values and regression coefficients for all traits using the 4 different methods applying cross validation are presented in Table 1. Accuracies were in the range of 0.30 to 0.74. BayesC was best to predict genome-wide breeding values for Fat%, whereas GS-BLUP, PLSR and LASSO gave similar, but lower accuracies. The highest accuracies were obtained for Prot-kg applying BayesC, GS-BLUP and PLSR. For the lowly heritable traits NR56 and SCC, all methods except LASSO performed equally in terms of accuracy. Using LASSO, where only a few numbers of SNP were selected to fit the data, accuracies for NR56 and SCC were 0.30 and 0.31, respectively.

Table 1. Accuracy (r) of genome-wide breeding values and regression coefficients (b) of the current estimated breeding value on the genome-wide breeding value with the genomic selection methods BayesC, GS-BLUP, PLSR and LASSO applying cross validation[1]

Trait[2]	BayesC		GS-BLUP		PLSR		LASSO	
	r	b	r	b	r	b	r	b
Fat%	0.64	0.92	0.52	0.93	0.50	0.83	0.56	1.55
Prot-kg	0.72	1.04	0.73	1.18	0.74	1.05	0.54	3.01
NR56	0.49	0.74	0.50	0.81	0.50	0.77	0.30	2.81
SCC	0.52	0.79	0.55	0.90	0.54	0.86	0.31	3.95

[1] Results are arithmetic means of the five-fold cross validation.
[2] Fat% = fat percentage; Prot-kg = protein kilogram; NR56 = non return rate after 56 days; SCC = somatic cell count.

From a practical point of view, animal breeders are more interested in forward prediction, i.e. prediction of genome-wide breeding values for young bulls which were not included in the derivation of the prediction equations. In Table 2, accuracies and regression coefficients for all traits and methods applied are shown for the forward prediction. In general, BayesC slightly outperformed the other methods. For Fat%, accuracy of genome-wide breeding values was 0.60, where all the other methods resulted in accuracies between 0.37 and 0.47. GS-BLUP gave similar accuracies for all traits, where surprisingly the highest accuracy was obtained for the very low heritable trait NR56. The same pattern was observed applying PLSR (Table 2). Lowest accuracies were calculated for Prot-kg, NR56 and SCC with LASSO. Using LASSO only a subset of SNP is included in the model (Tibshirani, 1996). For Fat%, Prot-kg, NR56, and SCC 25, 21, 20, and 21 SNP were selected, respectively. LASSO gave the highest accuracy of 0.47 for Fat% which might be in relation with the polymorphism in the DGAT1 gene which has a large effect on fat percentage (Grisart *et al*. 2004). Similar results were reported by Hayes (2009) where LASSO along with BayesC gave the highest accuracy for fat%. Regression coefficients for all traits and methods were below 1 indicating that genome-wide breeding values were overestimated with all methods used.

So far, only a very few results of genomic selection studies dealing with real data are available. Accuracies in this study were considerably lower compared to other studies. Sölkner *et al*. (2007) reported accuracies for Australian Holstein Friesian bulls in the range of 0.65 to 0.8 for different traits, including fertility, a trait with very low heritability, using different regression methods. Harris *et al*. (2008) have shown reliabilities (r^2) of genome-wide breeding values for young bulls without any daughter information in the range of 0.50 to 0.67 for milk production traits, live body weight, fertility, SCC, and longevity. In that study, Bayesian methods gave also slightly higher reliabilities compared to BLUP and regression methods. Hayes *et al*. (2009) observed reliabilities for Australian Holstein Friesian bulls for different traits between 0.14 and 0.55 using GS-BLUP and a Bayesian method (BayesA). A common finding of these studies was that GS-BLUP gave only slightly worse accuracies compared to Bayesian methods (Hayes *et al.* 2009; VanRaden *et al*. 2009). This is in agreement with the findings in this study, where, compared to BayesC, GS-BLUP resulted in similar accuracies for all traits except for Fat%. The GS-BLUP approach assumes a normal distribution of marker effects with the same variance for each marker (Meuwissen *et al*. 2001) whereas BayesC uses prior information about the distribution of marker effects allowing some markers having a big effect, whilst others having small effects (Meuwissen 2009). From this result, Hayes *et al*. (2009) conclude that the GS-BLUP assumptions, that most traits are influenced by many markers having a small effect and few with moderate to large effects, may be close to the truth.

Table 2. Accuracy (r) of genome-wide breeding values and regression coefficients (b) of the current estimated breeding value on the genome-wide breeding value with the genomic selection methods BayesC, GS-BLUP, PLSR and LASSO applying forward prediction

Trait[1]	BayesC		GS-BLUP		PLSR		LASSO	
	r	b	r	b	r	b	r	b
Fat%	0.60	0.81	0.42	0.78	0.37	0.64	0.47	0.70
Prot-kg	0.46	0.54	0.42	0.59	0.34	0.52	0.21	0.34
NR56	0.52	0.76	0.47	0.71	0.46	0.77	0.25	0.50
SCC	0.43	0.66	0.43	0.73	0.40	0.74	0.23	0.41

[1] Fat% = fat percentage, Prot-kg = protein kilogram, NR56 = non return rate after 56 days, SCC = somatic cell count

CONCLUSIONS

Considering the results for forward prediction, which are most relevant, BayesC turned out to predict the genome wide breeding values slightly more accurately than the other methods in this study. The LASSO method did not predict the genome wide breeding values very well except for Fat%. Results should be interpreted with caution as the analyses were based on a limited number of bulls. Further study is under way with the same methods and number of SNP increasing the number of bulls in the training set.

ACKNOWLEDGMENTS

The authors are grateful to ZuchtData EDV-Dienstleistungen GmbH and the Federation of Austrian Simmental Fleckvieh Cattle Breeders for providing the estimated breeding values and the genotypes.

REFERENCES

Grisart, B., Farnir, F., Karim, L., Cambisano, N., Kim, J.-J., Kvasz, A., Mni, M., Simon, P., Frère, J.-M., Coppieters, W., and Georges, M. (2004) *Proc. Natl. Acad. Sci. USA* **24**: 2398.

Hayes, B. J. (2009) *Symposium Statistical Genetics of Livestock for the Post-Genomic Era, Wisconsin-Madison, USA*

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. and Goddard, M.E. (2009) *J. Dairy Sci*. **92**:433.

Harris, B.L., Johnson, D.L., and Spelman, R.J. (2008) Proc. *Interbull Meeting, Niagara Falls, Canada*

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

Meuwissen, T.H.E. (2009) *Genet. Sel. Evol*. **41**:35.

SAS Institute Inc. (2007) *SAS/STAT® User's Guide, Version 9.2. Cary, NC.*

Sölkner, J., Tier, B., Crump, R., Moser, G., Thomson, P.A. and Raadsma, H. (2007) *Book of Abstracts of the 58th Annual Meeting of the European Association for Animal Production*, p. 161.

Tibshirani, R. (1996) *J. R. Statist. Soc. B* **58**:267.

VanRaden P.M., Van Tassel, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. and Schenkel, F. (2009) *J. Dairy Sci*. **92**:16.