Genomic Selection in Fleckvieh/Simmental – First Results

B. Gredler[1], K.G. Nirea[12], T.R. Solberg[3], C. Egger-Danner[4], T.H.E. Meuwissen[2], 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

Abstract

The objective of this study was to compare partial least squares regression (PLSR), multivariate regression analysis using least absolute shrinkage and selection operator (LASSO), two Bayesian approaches (BayesA, BayesB) and an ordinary BLUP method (GS-BLUP) for the estimation of genome-wide breeding values for dual purpose Simmental Fleckvieh in Austria. A forward prediction and cross validation were carried out for fat percentage, protein yield, 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 0.36 to 0.76. In forward prediction, obtained accuracies were between 0.20 and 0.61.

Keywords: genomic selection, SNP, dual purpose Fleckvieh cattle

1. 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. In Austria a project in collaboration of the Federation of Austrian Simmental Fleckvieh Cattle Breeders, the University of Natural Resources and Applied Life Sciences Vienna and ZuchtData EDV-Dienstleistungen GmbH to develop a genomic breeding value estimation for dual purpose Fleckvieh was established in 2008. 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.

2. Material and Methods

2.1 Data

In total, 1,726 dual purpose Fleckvieh bulls, genotyped with the Illumina Bovine SNP50^{TM} Beadchip with a call rate >=90 % and a minimum reliability of 80 % for the total merit index were included in the analysis. Bulls were born from 1975 to 2004. The distribution of bulls across birth year is shown in Figure 1. For forward prediction the data set was split into a reference population (training set) including bulls born before 2001 and a test set of bulls born between 2001 and 2004. For validation of the methods a cross validation was carried out where bulls were randomly sampled across all birth years. For both, forward prediction and cross validation the training and test set included 1,277 and 449 bulls, respectively. The distribution of bulls in the training and test set for cross validation is shown in Figure 2.

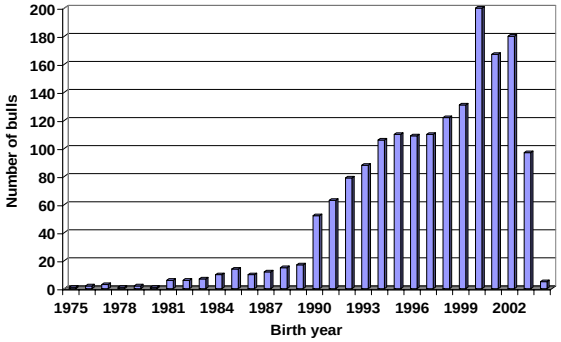

Figure 1. Distribution of bulls across birth year.

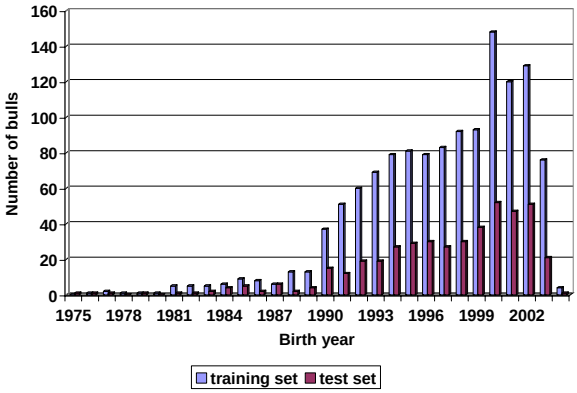

Figure 2. Distribution of bulls across birth year in the training and test set for the cross validation run.

For each SNP a minor allele frequency of 1 % was required. SNPs showing an average GC-Score (i.e. a measure to rank and filter out failed genotypes) of <0.6 over all samples were removed. To test for Hardy-Weinberg equilibrium, the deviation of observed genotype frequencies from expected genotype frequencies based on allele frequencies was calculated. SNPs were included if Hardy Weinberg χ^2 values were below 800. A total of 42,613 SNPs satisfied all selection criteria. Missing genotypes were imputed according to allele frequencies by sampling random numbers from uniform distribution. Breeding values based on progeny testing from the joint Austrian-German genetic evaluation of April 2009 for fat percentage (Fat %), protein yield (Prot-kg) somatic cell count (SCC), and non return rate after 56 days for cows (NR56) were used as phenotypes. Table 1 shows the average number of daughters per bull in the training and test set and reliabilities of conventional breeding values for all traits. As for Fat% and Prot-kg no reliabilities are available the reliability of the milk index is presented instead.

Table 1. Average number of daughters (N-daught.), reliabilities (r^2, %) of milk index (MI), somatic cell count (SCC), and non return rate (NR56) for bulls in test and training set

	Forward prediction				**Cross validation**			
	Test set		**Training set**		**Test set**		**Training set**	
	Mean	Min-Max	Mean	Min-Max	Mean	Min-Max	Mean	Min-Max
N-daught.	577.3	86-15603	1922.9	81-71230	1263.8	86-42795	1681.5	81-71230
r^2 MI	91.3	84-97	94.1	84-99	93.3	84-99	93.4	84-99
r^2 SCC	86.7	76-95	91.1	78-99	89.8	78-99	90.0	76-99
r^2 NR56	51.1	34-73	69.1	45-99	63.9	35-99	64.6	34-99

2.2 Methods

In this study, we compared partial least squares regression (PLSR), regression analysis using least absolute shrinkage and selection operator (LASSO), a GS-BLUP approach (Meuwissen *et al.,* 2001), and two Bayesian approaches BayesA (Hayes *et al.,* 2009) and BayesB (Meuwissen, 2009). The BayesA method used is similar to that described in Meuwissen *et al.* (2001) modified to include a polygenic effect. For the Bayesian methods five replicates were carried out. For running PLSR and LASSO, the SAS procedures PROC PLS and PROC GLMSELECT were used (SAS, 2008). PLSR is used to reduce dimension of SNP data. The optimal number of latent variables is assessed by cross validation such that the covariance of SNP data and phenotypes is maximised. Using the LASSO method only a subset of SNPs is included in the model where cross validation is used to choose the number of SNPs with the highest predictability (Tibshirani, 1996). BayesA, BayesB and GS-BLUP differ in their assumptions about the distribution of SNP effects. GS-BLUP treats all markers alike

assuming that all markers having constant variance (Meuwissen *et al.,* 2001). BayesA assumes that many SNPs have small individual effects and only a few will have large effects (Hayes *et al.,* 2009). The BayesB method implemented here is similar to the BayesB presented by Meuwissen *et al.* (2001) whereas the prior distribution assumed that the majority of markers have a small effect instead of assuming that these markers have no effect (Meuwissen, 2009). To assess the accuracy of genomic selection, the correlation between estimated direct genome-wide breeding values (GEBV) and current estimated breeding values (EBV) based on progeny testing was calculated using bulls in the test set. 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.

3. Results and Discussion

Accuracies of genome-wide breeding values and regression coefficients for all traits using the five different methods applying cross validation are presented in Table 2. Accuracies were in the range of 0.36 to 0.76. BayesB 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 BayesB, GS-BLUP and PLSR. For the lowly heritable traits NR56 and SCC, all methods except LASSO and BayesA 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 0.36 and 0.46, respectively.

Table 2. Accuracy (r) of genome-wide breeding values and regression coefficients (b) of the estimated breeding value on the genome-wide breeding value using different methods applying cross validation

Trait	PLSR		LASSO		GS-BLUP		BayesA		BayesB	
	r	b	r	b	r	b	r	b	r	b
Fat%	0.58	0.92	0.53	0.79	0.59	1.02	0.41	1.06	0.68	0.89
Prot-kg	0.76	1.05	0.58	0.92	0.76	1.08	0.66	0.65	0.74	1.01
SCC	0.58	0.88	0.46	0.80	0.58	0.87	0.52	0.52	0.47	0.77
NR56	0.53	0.85	0.36	0.79	0.52	0.89	0.37	0.39	0.49	0.61

Table 3. Accuracy (r) of genome-wide breeding values and regression coefficients (b) of the estimated breeding value on the genome-wide breeding value using different methods applying forward prediction

Trait	PLSR		LASSO		GS-BLUP		BayesA		BayesB	
	r	b	r	b	r	b	r	b	r	b
Fat%	0.44	0.79	0.45	0.68	0.48	0.86	0.44	0.91	0.61	0.79
Prot-kg	0.37	0.50	0.20	0.28	0.40	0.56	0.32	0.38	0.39	0.51
SCC	0.50	0.92	0.29	0.54	0.5	0.85	0.42	0.43	0.47	0.77
NR56	0.51	0.79	0.37	0.62	0.49	0.77	0.44	0.22	0.49	0.61

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 3, accuracies and regression coefficients for all traits and methods applied are shown for forward prediction. In general, BayesB, GS-BLUP, and PLSR slightly outperformed the other methods. For Fat%, accuracy of genome-wide breeding values was 0.61, where all the other methods resulted in accuracies between 0.44 and 0.48. GS-BLUP gave similar accuracies for all traits (except Prot-kg), where surprisingly the highest accuracy was obtained for the low heritable traits SCC and NR56. The same pattern was observed applying PLSR (Table 3). Lowest accuracies were calculated for Prot-kg and SCC with LASSO indicating that the traits are influenced by more markers than identified with LASSO. For Fat%, Prot-kg,

NR56, and SCC 20, 25, 25, and 29 SNP were selected, respectively. LASSO gave the highest accuracy of 0.45 for Fat% which might be in relation with the polymorphism in the DGAT1 gene which has a large effect on Fat% (Grisart *et al.,* 2004). Similar results were reported by Hayes (2009) where LASSO along with BayesC gave the highest accuracy for Fat%. In general, calculated accuracies from BayesA were slightly lower compared to the other methods. These findings indicate that accuracies of genome-wide breeding values estimated with models not allowing for a polygenic effect are overestimated since the SNPs can predict relationship between individuals as shown by Habier *et al.* (2007).

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 involving other breeds. 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 BayesB, GS-BLUP resulted in similar accuracies for all traits except for Fat%.

4. Conclusions

Considering the results for forward prediction, which are most relevant, BayesB, GS-BLUP and PLSR 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%. However, no clear winner among the methods could be identified for all traits suggesting that a trait by method interaction exists depending on the genetic background of the trait. Results should be interpreted with caution as the analyses were based on a limited number of bulls. Further study is under way with more methods including a polygenic effect and increasing the number of bulls in the training set.

Acknowledgements

The authors wish to thank ZuchtData EDV-Dienstleistungen GmbH and the Federation of Austrian Simmental Fleckvieh Cattle Breeders for providing the breeding values and genotype data.

5. References

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. USA 24,* 2398-2403.

Gredler, B., Nirea, K.G., Solberg, T.R., Egger-Danner, C., Meuwissen, T.H.E. & Sölkner, J. 2009. *Proceedings of the 18th Conference for the Association for the Advancement of Animal Breeding and Genetics,* Barossa Valley, Australia.

Habier, D., Fernando, R.L. & Dekkers, J.C. 2007. The impact of genetic relationship information on genome-assisted breeding values. *Genetics 177,* 2389-2397.

Harris, B.L., Johnson, D.L. & Spelman, R.J. 2008. Genomic selection in New Zealand and the implications for national genetic evaluation. *Proc. Interbull Meeting,* Niagara Falls, Canada

Hayes, B. J. 2009. Genomic selection in the era of the $1000 genome sequence. *Symposium Statistical Genetics of Livestock for the Post-Genomic Era*, Wisconsin-Madison, USA. http://dysci.wisc.edu/sglpge/.

Hayes, B.J., Bowman, P.J., Chamberlain, A.J. & Goddard, M.E. 2008. Invited Review: Genomic selection in dairy cattle: Progress and challenges. *J. Dairy Sci. 92,* 433-443.

Meuwissen, T.H.E. 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,* 1819-1829.

SAS Institute Inc. 2008. *SAS/STAT® User's Guide,* Version 9.2. Cary, NC.

Sölkner, J., Tier, B., Crump, R., Moser, G., Thomson, P.A. & Raadsma, H. 2007. A comparison of different regression methods genome-assisted prediction of genetic values in dairy cattle. *Book of Abstracts of the 58th Annual Meeting of the European Association for Animal Production,* p. 161.

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

.