Proc. Assoc. Advmt. Anim. Breed. Genet. 19:287-290

EFFECT OF RELATIONSHIP AND AGE STRUCTURE BETWEEN TRAINING AND VALIDATION SET ON THE ACCURACY OF GENOMIC BREEDING VALUE PREDICTION USING GENOMIC BLUP

M. Erbe[1], F. Seefried[2] and H. Simianer[1]

[1] Georg-August-University, Department of Animal Sciences, Animal Breeding and Genetics Group, Albrecht-Thaer-Weg 3, 37075 Goettingen, Germany
[2] Vereinigte Informationssysteme Tierhaltung w.V. (VIT), Heideweg 1, 27283 Verden/Aller, Germany

SUMMARY

A dataset of 5698 Holstein Friesian bulls born between 1981 and 2005 was used to study the influence of different relationship levels between a training set and the set of candidates for whom genomic breeding values (GBV) are to be predicted. Traits studied were milk yield and somatic cell score. Different scenarios were modeled while always the GBV of the 500 youngest bulls of the available data set were predicted. The correlation between true breeding value and GBV was used as evaluation criterion. The prediction of the youngest bulls was best when other bulls of the same age or only slightly older or bulls which were especially highly related to the candidates were used to train the model while there was a decrease of accuracy, especially for GBV in somatic cell score, when the oldest bulls formed the training set. Reducing the maximum relationship between all candidates to the training set to less than 0.5 led to a decrease in accuracy. The decrease was even stronger when the maximum relationship was limited to less than 0.25. It seems that accuracy of prediction of GBV depends clearly on the relationship and age structure between the validation and the training set which is in accordance with some previous studies. Therefore, it is implicitly necessary to continuously fill the training sets used for predicting young bulls with new progeny tested bulls to avoid the reduction of maximum relationship.

INTRODUCTION

In the last years, prediction of genomic breeding values has become a popular tool for predicting reliable breeding values of not yet progeny tested bulls of young age, especially in dairy cattle populations. Different studies (e.g. Lund *et al.* 2009; Habier *et al.* 2010) have shown that accuracy of prediction is clearly influenced by the relationship between bulls in the training and in the validation set. Since the methodology of genomic selection is new, there are still enough progeny tested bulls available which are strongly related to the candidates and can be used to train the models. However, in a few years, if genomic selection will be consequently applied, there may be a lack of such animals. It is thus necessary to further investigate how the relationship and age structure influences the accuracy of genomic breeding values of young bulls.

MATERIALS AND METHODS

Data. We used a sample of 5698 Holstein bulls, which were genotyped with the Illumina 50K Single Nucleotide Polymorphism (SNPs) chip. SNPs with a minor allele frequency lower than 1%, with missing position or a call rate lower than 95% were excluded. After filtering, there were 42,551 SNPs remaining for further analyses. Missing genotypes at these SNP positions were imputed using Beagle 3.2 (Browning and Browning 2007).

The bulls were born between 1981 and 2005. The average of the mean pedigree-based relationship between a random bull and all others was 0.093 while the mean of the maximum relationship was 0.459. 1832 bulls had a genotyped father and 1974 had one or both grandsires genotyped. There were 77.2% of bulls having at least 10 half or full sibs. The average inbreeding coefficient

was 0.045. All bulls had pedigree information and breeding values for somatic cell score and milk yield. Average accuracy of the breeding values of the validation bulls was 0.89 and 0.96 for somatic cell score and milk yield, respectively. For bulls in the training sets, it was between 0.92 and 0.96 for somatic cell score and between 0.97 and 0.98 for milk yield in the different scenarios.

Method to predict GBV. Genomic breeding values were predicted using best linear unbiased prediction (BLUP) based on the model

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

where **y** is a vector of quasi-phenotypes (breeding values of milk yield or somatic cell score, respectively) for all bulls in the training set, **1** is a column vector of ones, **μ** is the overall mean, **Z** is the incidence matrix for the random genomic effect, **u** is a vector containing the random genomic effect (i.e. the genomic breeding value) for each animal and **e** is a vector of random error terms. **u** is assumed to be distributed $\mathbf{u} \sim N(0, \mathbf{G}\sigma_g^2)$ and **e** is assumed to follow $N(0, \mathbf{I}\sigma_e^2)$. **G** is a genomic relationship matrix which was built based on all SNPs available after quality control following VanRaden (2008). Variance components were estimated once with the complete data set using ASReml 3.0 (Gilmour et al. 2009) and were then used for all runs.

Validation strategy. The dataset was used for studying the influence of relationship and age structure on prediction of genomic breeding values (GBV). For this, we ran different scenarios with a constant set of candidates (validation set) whose GBV were predicted using different training sets to train the model. Since the usual application of genomic prediction in cattle is the prediction of genomic breeding values for young bulls without phenotypes and not yet progeny tested, we used the 500 youngest bulls in our data set (all born in 2005) as the validation set for all scenarios. For each scenario, 2000 bulls fulfilling scenario specific criteria were chosen from the remaining data set. Prediction was then replicated 10 times in each scenario using always a random sample of 1500 out of the 2000 bulls at a time. As a standard for comparison to all other scenarios the training set comprised first of all completely randomly chosen bulls (**random**). For two further scenarios, the 2000 bulls were the oldest ones (**old**) and the youngest ones (**young**) of the remaining data set. To study the changes in accuracy of prediction when the relationship between training and validation set was reduced, we performed three scenarios where the training set contained only animals with a maximum pedigree-based relationship less than 0.25 ($rel_{max}<0.25$) to all candidates. In the first of these three scenarios, we only controlled the maximum relationship (**<.25**) while in both the others we also controlled the age structure (**<.25y**: youngest bulls with $rel_{max}<0.25$, **<.25o**: oldest bulls with $rel_{max}<0.25$). In one further scenario, a maximum relationship of 0.5 was allowed. (**<.50**) The last scenario (**maxrel**) tried to maximize the relationship between training and validation set by including all available near relatives (i.e. sire, grandsires, full and half sibs) of all candidates to the training set and filling the rest with bulls having a relationship of greater than 0.25 to as many candidates as possible.

Criterion for comparison. For the evaluation of the prediction, the correlation ($r_{GBV,TBV}$) between predicted GBV and true breeding value (TBV) was used. For obtaining $r_{GBV,TBV}$, first Pearson's correlation coefficient between the estimated breeding values (used as phenotypes) and the predicted GBV for the animals in the validation set was calculated in each scenario for each replicate. This correlation coefficient was then divided by the mean accuracy of the estimated breeding values of the animals in the respective validation set. To compare the relationship structure between different scenarios, the maximum and mean relationship of each of the 500 youngest bulls to all animals in the particular training set was calculated as well as the average number of animals in the training set to whom each of the candidates was related with a relationship coefficient greater or equal 0.25.

RESULTS AND DISCUSSION

Results for all scenarios and both traits regarding the mean accuracy of prediction and the key data of the relationship structures are given in Table 1.

Table 1. Accuracy of prediction and relationship measurements in different scenarios and both traits (milk yield and somatic cell score). Results for correlations between predicted genomic breeding values and true breeding values ($\mathbf{r_{GBV,TBV}}$) were averaged over the ten replicates. Relationship criteria were measured between each candidate in the validation set and all animals in the respective training set and then averaged over all 500 candidates and the ten replicates. The last column shows the average number of animals in the training set a candidate is related to with a relationship coefficient greater or equal 0.25.

Scenario	$r_{GBV,TBV}$±s.e. milk yield	$r_{GBV,TBV}$ ±s.e. somatic cell score	Maximum relationship	Mean relationship	No of animals $rel_{max} \geq 0.25$
random	0.630±0.006	0.667±0.004	0.375	0.098	11
old	0.568±0.006	0.563±0.016	0.395	0.094	3
young	0.649±0.005	0.718±0.007	0.334	0.104	25
<.50	0.543±0.006	0.626±0.006	0.318	0.100	9
<.25	0.489±0.009	0.524±0.009	0.223	0.090	0
<.25o	0.534±0.005	0.454±0.011	0.221	0.090	0
<.25y	0.543±0.007	0.573±0.006	0.221	0.090	0
maxrel	0.685±0.005	0.731±0.003	0.430	0.109	28

Boxplots of the accuracy of prediction measured by $r_{GBV,TBV}$ for all scenarios are shown in Figure 1 for milk yield and somatic cell score. For both traits, the prediction was slightly better when random samples of young bulls were used to train the model in comparison to a random sample of bulls regardless of their age. These samples often contain large groups of half sibs of candidates so that the mean and maximum relationship was rather high in comparison to other scenarios. This may explain why prediction was better here.

Including all animals in the training set which were directly related to the candidates (scenario **maxrel**) led only to a slight increase in accuracy for both traits in comparison to the scenario **young**. This was expected due to the fact that relationship between all young Holstein Friesian bulls is quite high on average. Therefore, candidates and bulls in the training sets were related to a large extent even if a random sample of young bulls regardless of the relationship structure was used for the training set.

An unambiguous trend of reduced prediction ability was observed when the relationship between training and validation set was limited to a specific maximum value as well as when the age difference between training and validation set became greater. For somatic sell score, the prediction was lowest when using the oldest available bulls with a maximum relationship of less than 0.25 to every candidate, while for milk it was lowest with a random sample with a maximum relationship restricted to less than 0.25 to every candidate.

We even could find a reduction of accuracy when there were only no more sires (and full sibs) of the candidates in the training set (scenario **<.50**). Lund *et al.* (2009) presented similar tendencies when excluding sires from the training sets for three different traits in a sample of Nordic Holstein bulls. If the maximum relationship was limited to less than 0.25, the reduction in prediction ability was even worse, especially for somatic cell score. This is in accordance with the work of Habier *et al.* (2010) who showed a continuous decrease of accuracy in different traits when reducing the permitted maximum relationship step by step in a limited sample of Holstein Friesian bulls. A limitation of $rel_{max}<0.25$ means that no sires, grandsires, half and full sibs were used to train the mod-

el. From a practical point of view, this is a scenario which would become relevant after only two generations when the breeders fail to rebuild the training sets with enough new progeny tested bulls.

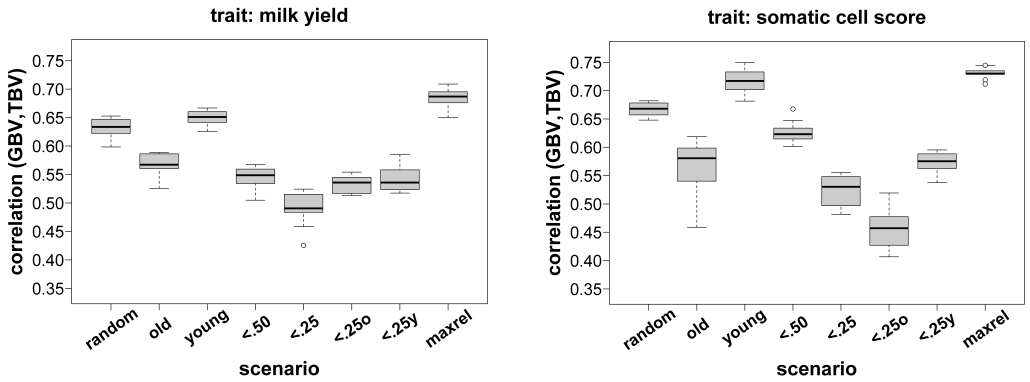

Figure 1. Boxplots of the accuracy of prediction for milk yield and somatic cell score for all scenarios.

CONCLUSIONS

Different training sets were used to train the model and to predict genomic breeding values for the 500 youngest bulls of the available data set. Different levels of relationship and age structure between training and validation set led to differences in accuracy of prediction. Reducing the relationship implicated an apparent decrease of accuracy of prediction. Therefore, in all kinds of validation or cross-validation procedures, relationship and age structure of the sample should be accounted for to ensure fair assessment of the predictive ability.

Concerning practical application of GBV prediction, especially in strongly related samples like progeny tested Holstein Friesian bulls, there seems to be no critical point as long as sires, half or full sibs are included in the training sets. For future prediction, though, a decrease of accuracy is expected when maximum and therefore also mean relationship between the training individuals and the candidates will decrease. If not enough new progeny tested bulls are continuously added to the training set, which may be the case in genomic selection schemes minimising the generation interval (Lillehammer *et al.* 2011), accuracy of prediction will deteriorate in perceivable steps even after only one or two generations.

ACKNOWLEDGMENTS

This research was funded by the German Federal Ministry of Education and Research within the AgroClustEr "Synbreed – Synergistic plant and animal breeding" (Funding ID: 0315526).

REFERENCES

Browning S.R. and Browning B.L. (2007) *Am. J. Hum. Genet.* **81**: 1084.
Gilmour A.R., Gogel B.J., Cullis B.R. and Thompson R. (2009) ASReml User Guide Release 3.0. VSN International Ltd, Hemel Hempstead, UK.
Habier D., Tetens J., Seefried F.-R., Lichtner P. and Thaller G. (2010) *Genet. Sel. Evol.* **42**: 5.
Lillehammer M., Meuwissen T.H.E and Sonesson A.K. (2011) *J. Dairy. Sci.* **94**: 493.
Lund M.S., Su G., Nielsen U.S. and Aamand G.P. (2009) *Interbull Bulletin* **40**: 162.
VanRaden P.M. (2008) *J. Dairy Sci.* **91**: 4414.