Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Bias of genetic trend of genomic predictions based on both real and simulated dairy cattle data

P. Ma[1], M.S. Lund[1], U.S. Nielsen[2], G.P. Aamand[3], A.C. Sørensen[1], G. Su[1]
[1]Center for Quantitative Genetics and Genomics, Department of Molecular Biology and Genetics, Aarhus University,
[2]Knowledge Center for Agriculture, [3]NAV Nordic Cattle Genetic Evaluation

ABSTRACT: This study investigated the phenomenon of bias in the trend of genomic predictions and attempted to find the reason and solution for this bias. The data used in this study include Danish Jersey data and simulation data. In Jersey data, the bias was reduced when cows were included in the reference population. In simulated data, there was no bias when the test animals were unselected cows. When the G matrix was derived from genotypes of causal genes, the bias was reduced. The results suggest that the main reasons for causing the bias of the prediction trends are the selection of bulls and bull dams as well as the inaccurate relationship matrix. The possible strategies to eliminate the bias could be to use cow reference and improve genomic relationship matrix.
Keywords: prediction bias; reference population; genomic relationship matrix

Introduction

A difference between trends in genomic breeding values (GEBV) and deregressed proofs (DRP) were observed in our Jersey data, indicating an underestimation of genomic prediction. Some bias may be expected because animals in both the reference and test populations are selected. An alternative explanation could be inaccuracies in the marker derived relationship matrix compared to the relationship matrix at causal loci. It was reported that the marker derived relationship matrix might be different from the causal loci derived relationship matrix as the genomic relationship vary across regions (de Los Campos et al., (2013)). Therefore, the inaccurate relationship matrix could be one of the reasons causing the prediction bias. The other reason could be that dams were not included in the reference. Without dams in the reference population, genetic progress of dams could not be fully accounted for by marker information. The relatedness between reference animals and test animals contributes a major part for predicting GEBV of the test animals (Wientjes et al., (2013)), especially for GBLUP (Habier et al., (2007)). In reality, the bull dams should also contribute for the genetic progress of the test bulls. This might lead to an underestimation of GEBV for the test animals. The objectives of this study were to: 1) investigate the bias based on both real data and simulation data; 2) investigate the influence of cow reference and relationship matrix on the bias; 3) explore possible solutions to reduce this bias.

Material and methods

Data. The Danish Jersey data and simulation data using ADAM (Pedersen et al., (2009)) were used in this study.

The Jersey data consisted of 3,968 individuals. There were 1,255 bulls born from 1981 to 2009 and 2,713 cows born from 2000 to 2011. The bulls were genotyped with Illumina BovineSNP50 BeadChip which includes 54,001 single nucleotide polymorphic (SNPs) and the cows were genotyped with Illumina BovineLD Beadchip which includes 6,909 SNPs.

In the simulation, the genome consisted of 30 chromosomes of 100 cM each. Potential loci in the genome were assumed to be 300,000,000. The ratio of number of markers to number of QTLs was 31. Mutations occurred randomly at the loci at a rate of 1.8×10^{-7} in each meiosis. The effects of quantitative trait loci (QTL) effects were assumed to follow a Normal distribution. Two related traits were simulated with the heritability of 0.3 and 0.04 and with a correlation -0.3. After simulating a historical population for 500 generations with 200 males and 200 females per generation, the current population was simulated for 20 generations. In each generation, observations of animals were sampled according to true breeding value and an independent residual effect. The details could be found in Buch et al. (2012).The selection index was composed of the similar weight of the two traits. In each generation, bull dams were selected according to their parent average estimated breeding values (EBV), and 60 bulls were selected according to parent average and then DYD for these bulls were sampled. Twenty bulls were selected with the highest DYD among these 60 bulls and used as proven bulls. There were around 9,600 cows and 400 bulls in each generation.

Validation. The Jersey data were used to check the impact of adding cows to reference population, while the simulation data were used to investigate the impact of marker or causal loci derived relationship matrix on genomic prediction.

To validate the prediction accuracy and prediction bias, the Jersey bulls were divided to reference and test sets using cutoff date 1st Jan. 2005. The bulls born after that date were used as validation animals. Cows were included in the reference to check whether the relatedness between cows

and test bulls could eliminate the prediction bias. According to the relationship between the reference cows and test bulls, four scenarios were investigated using Jersey data: 1) only bulls (1,030) were used as reference population (Bull); 2) both bulls and cows (2,774) were used as reference population (Cow); 3) bulls and cows (1,850) which were half-sibs or dams of the test bulls were used as the reference population (Dam_sibs); 4) bulls and cows (1,954) excluding dams and half-sibs of the test bulls were used as the reference population (Non_dam_sibs). The reliability and trend of genomic prediction were compared.

In the simulation data, animals from generation 1 to 19 consisting of 1,140 progeny test bulls were used, which correspond to the Jersey Bull reference scenario. Here, 840 bulls from generation 1 to 14 were used as reference set. Three different data sets from generations 15-19 were used as the test set: 1) 300 progeny tested bulls; 2) around 1,200 unselected bulls; 3) around 1,000 randomly selected cows from each generation as the test set. Genomic prediction was performed using both marker derived and causal loci derived genomic relationship matrices. Trends of genomic prediction using the two kinds of genomic relationship matrices were plotted to compare the influence of genomic relationship matrix on prediction bias.

Results and Discussions

The reliability and regression coefficients for each scenario of Jersey data are shown in Table 1. The reliabilities were improved for all traits when the cows were included in the reference population. The reason could be the close relationship between reference and test. However the inflation of regression coefficient also increased for the production traits. These results were consistent with the study from Wiggans et al. (2011). The trends of DRP and GEBV for different scenarios are shown in Figure 1. The bias for the test set was obvious in the scenario Bull reference. The increase in GEBV was smaller than the increase in the DPR while the regression coefficient was smaller than 1 for the test animals. The reason was that the intercept of the regression analysis was bigger than 0, which led to an underestimation of the trend, though there was an inflation of prediction as reflected by the regression coefficients which were lower than one. Therefore, the intercept should be also noticed when the regression coefficients were considered to judge the bias of genomic prediction. The bias was reduced in scenario Cow reference and scenario Dam_sibs reference. However, even though the number of reference animals in scenario Non_dam_sibs reference was more than in scenario Dam_sibs reference, the bias still existed in the scenario Non_dam_sibs reference. The cows in the scenario Dam_sibs were mainly the half sibs of the test bulls, and only 25 were the dams of the test bulls. A possible reason for reduced bias in this scenario could be that information from half sibs captured the genetic progress of the cows in the bull dam generation. Therefore, results from Jersey data suggest that the reduced bias in test bull GEBVs arise from including cows in the reference that are highly related to the test bulls.

Table 1 Reliability (R^2) and regression coefficient (b) in Jersey data

Reference[1]	Protein		Mastitis		Fat		Milk	
	R^2	b	R^2	b	R^2	b	R^2	b
Bull	0.33	0.66	0.37	0.83	0.21	0.67	0.42	0.81
Cow	0.34	0.57	0.44	0.83	0.27	0.55	0.48	0.71
Dam_sibs	0.33	0.61	0.41	0.85	0.24	0.58	0.43	0.75
Non_dam_sibs	0.33	0.65	0.37	0.80	0.24	0.66	0.47	0.81

[1]Bull: the reference population comprised only bulls; Cow: the reference population comprised both bulls and cows; Dam_sibs: the reference population comprised bulls and the cows which were the dams or sibs of the test animals; Non_dam_sibs: the reference population comprised bulls and the cows which were not the dam or sibs of the test bulls.

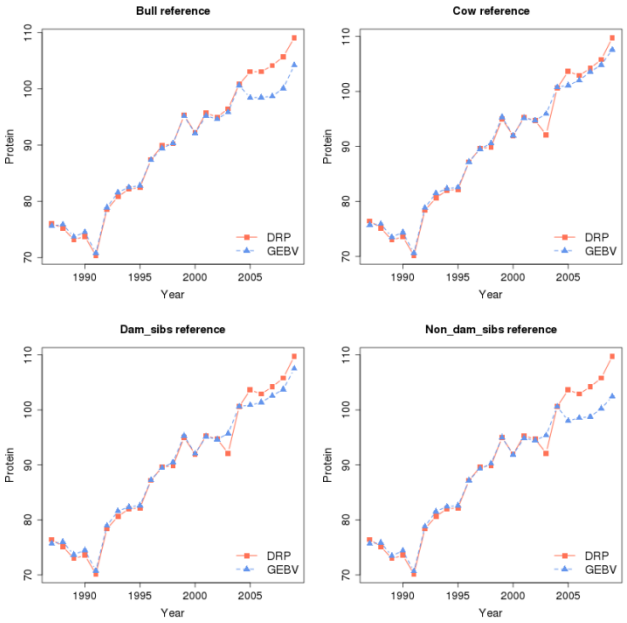

Figure 1. The DRP and GEBV trends in different scenarios in Jersey data

The results from 10 replicates in the simulation data were averaged (Figure 2). Bias was observed when comparing predicted GEBV and TBV of progeny tested bulls. Using a G matrix derived by the causal QTL reduced the bias slightly. As the QTL with large effects contribute more for the G matrix, we put more emphasis on the QTL with larger effect. The QTL genotypes were weighted by the squared QTL effect (g^2) or $2pqg^2$ in the study. The results showed the direct QTL derived G matrix reduced the bias mostly, and the weight using $2pqg^2$ performed similar as the direct QTL derived G matrix. However, using g^2 to weight the QTL genotypes resulted in a higher bias than using markers alone to derive the G matrix. The difference between trends of TBV and GEBV for unselected bulls was smaller than for the progeny test bulls, but still present,

likely because of the selection of the dams of these bulls. However, there was still bias when QTLs were used to derive the G matrix. The results illustrated that the relationship constructed by QTL could reduce the bias.

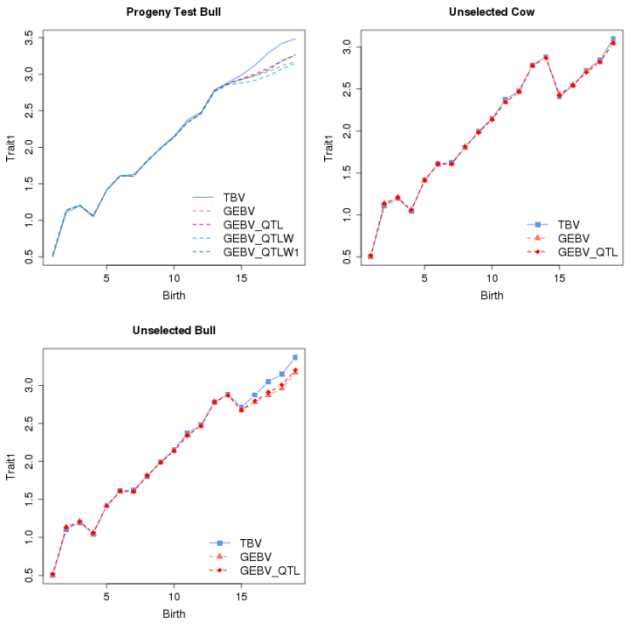

Figure 2. Using different G matrices to predict the GEBV of different test animals. GEBV means using markers to derive the relationship matrix; GEBV_QTL means QTL derived relationship matrix; GEBV_QTLW means using squared QTL effects (g^2) to weight the QTL genotypes; GEBV_QTLW1 means using $2pqg^2$ to weight the QTL genotypes.

The prediction for unselected cows in the test data were not biased, which indicated bias in the prediction trend is associated with selection history of the individuals in the test population.

Conclusion

There is bias in trends of GEBV in both real data and simulation data. The main cause of the bias could be that predictions are validated on selected bulls without dam information in the reference population. The relationship matrix derived from causal QTLs partly reduced the bias. Proper ways to reduce the bias could be to include the cows related to the candidates in the reference and derive more accurate relationship matrix.

Literature Cited

Buch, L.H., Sørensen, M.K., Berg, P. et al. (2012). *J. Anim. Breed. Genet.* 129:138–51.

Habier, D., Fernando, R.L., and Dekkers, J.C.M. (2007). *Genetics*. 177:2389–97.

Kuhn, M.T., Boettcher, P.J., and Freeman, a. E. (1994). *J. Dairy Sci.* 77:2428–2437.

De Los Campos, G., Vazquez, A.I., Fernando, R. et al. (2013). *PLoS Genet.* 9:e1003608.

Pedersen, L.D., Sørensen, A.C. , Henryon, M. et al. (2009). *Livest. Sci.* 121:343–344.

Wientjes, Y.C.J., Veerkamp, R.F. , and Calus, M.P.L. (2013). *Genetics*. 193:621–31.

Wiggans, G.R., Cooper, T.A., VanRaden, P.M. et al. (2011). *J. Dairy Sci. 94*(12), 6188–93.