Nordic Holstein single-step test day model using left truncated genomic data

***M. Koivula*[1]*, U.S. Nielsen*[2]*, I. Strandén*[1]*, G. P. Aamand*[3]**
***and E. A. Mäntysaari*[1]**
[1] Natural Resources Institute Finland (Luke), 31600 Jokioinen, Finland
[2] SEGES Innovation, 8200 Aarhus N, Denmark
[3] NAV Nordic Cattle Genetic Evaluation, Agro Food Park 15, 8200 Aarhus N, Denmark
e-mail: minna.koivula@luke.fi

Abstract

In this study, we present the results from the Nordic Holstein test day (TD) evaluation model using left truncated genomic data in the single-step model (ssGTBLUP). In the truncated genomic data, genotypes of animals born before 2009 were removed from the single-step analyses. It was studied whether the old genotypes can cause bias in the single-step evaluations. The truncated genomic data results were compared to the single-step model with full genomic data using validation where the latest 4 years of phenotypes had been removed. Both analyses used the genomic relationship matrix of VanRaden method 1 and had a 30% residual polygenic proportion (RPG), and an allele frequency of 0.5 for all markers. The results indicate that removing old genotypes reduced the inflation in the young candidate animals considerably, and for validation bulls, the regression (b_1) in predicting the recent GEBV using 4-year-old data improved on average by 11%, and the coefficient of correlation (R^2) on average by 5%. Data truncation had also a positive effect on the differences in the mean Mendelian sampling (MS) terms of young candidate animals. On the other hand, the truncation of genomic data did not affect the GEBVs of the old, reliably evaluated animals – nor the GEBVs of the old animals whose genotypes were removed, as the within birth year correlation between full genomic GEBVs and genomic data cut GEBVs were nearly 1. Truncation of genomic data removed the over-prediction of recent year classes of bulls and reduced the amount of overdispersion in candidate evaluations to a level acceptable in practice.

Key words: genomic selection, data truncation, bias, ssGTBLUP

Introduction

During the last decade, genomic selection has become common in dairy cattle breeding. Since the first papers about single-step genomic evaluation (ssGBLUP) were published (Christensen and Lund, 2010; Aguilar et al., 2010), several alternative ways to overcome the computational challenges of ssGBLUP have been presented (reviewed in Mäntysaari et al., 2020).

It seems that genomic models have a bias problem at least with strongly selected traits. Bias can be defined as a difference between estimated breeding values and modeled phenotypic performances. The bias can be seen in the mean and the variation of young animals. In dairy cattle single-step evaluations, the bias problem has been estimated in several studies (Koivula et al., 2015; Koivula et al., 2018; Oliveira et al., 2019; Tsuruta et al., 2019).

Generally, the bias can be controlled with different adjustments (Koivula et al., 2015; Oliveira et al., 2019), or by using the so-called erosion factor (Croué et al., 2022). In multi-step evaluations, there are more tools for adjusting the bias, because the adjustments, such as the use of lower heritability or scaling the GEBV variance according to the validation accuracy, do not affect the official national EBV.

In this study, we tested left truncated genomic data in the single-step evaluations with

Nordic Holstein test day (TD) data. It was hypothesized that the genotyped animals born far from the current breeding population may exaggerate the genetic trend.

Materials and Methods

Data

The full routine test-day (TD) evaluation data from February 2022 for Holstein were used in the study. The data was obtained from the Nordic Cattle Genetic Evaluation (NAV) and included the official multiple trait milk production evaluation TD records from milk, fat, and protein production. The TD data included 8.8 million cows and 11.4 million animals in the pedigree. To be able to validate the models, a reduced data set was extracted from the full data. In the reduced data, the last four years of phenotypes in the full data were removed.

Holstein genotype data from February 2022 included 384,029 genotyped animals. Until 2019, bulls have been genotyped using Illumina BovineSNP50 and cows with BovineLD Bead Chips with the genotypes imputed to the 50K chip (Illumina, San Diego, CA). Since 2019, both sexes have been genotyped using EuroGenomics MD 80k chip. After applying editing criteria, 46,342 SNP markers on the 29 bovine autosomes were available for the evaluations. Genomic data included 38,628 bulls with Danish, Finnish and Swedish (DFS) origin and 48,068 foreign bulls including the bulls from the Eurogenomics exchange.

The genomic left truncated data was obtained from the full genomic data by removing genotypes of the animals born before 2009. In practice, 33,821 genotypes were removed of which 2,593 were cows or heifers and 31,228 were bulls. The remaining 350,208 genotypes were used in the analyses. Most removed genotypes were Holstein AI bulls, and animals with unknown birth years.

Models

Single-step models were run with ssGTBLUP (Mäntysaari et al., 2017) where the key computations involving the $\mathbf{G}^{-1}$ matrix are replaced by a dense $\mathbf{T}$ matrix of size m by n where n is the number of genotyped animals and m is the number of SNP markers. Two different $\mathbf{G}$ matrices were built for the comparisons. The $\mathbf{T}$ matrices were built as they would yield $\mathbf{G}$ of VanRaden method 1 and 30% residual polygenic proportion (RPG) and were scaled to generate an average diagonal of $\mathbf{G}$ equal to the pedigree-based relationship matrix of the genotyped animals ($\mathbf{A}_{22}$). The genetic groups were accounted for in the single-step models using the so-called partial QP transformation that omitted $\mathbf{G}^{-1}$ in QP (Koivula et al., 2022). The pedigree inbreeding coefficients were accounted for in $\mathbf{A}^{-1}$ and A_{22}^{-1}. The models were 1) ssGTBLUP with full genomic data **(GT_F**), 2) ssGTBLUP with left truncated genomic data **(GT_T**), and 3) animal model BLUP **(AM**) without genomic data.

The models were run with the multiple trait reduced rank random regression TD model (Lidauer et al., 2015). The official estimated breeding values of total 305d lactation yield for milk, protein, and fat were derived from the TD model random regression solutions, and these breeding value estimates were used in the further analyses.

The TD models were solved by MiX99 software (Strandén and Lidauer, 1999). In the new MiX99, the calculations for the $\mathbf{T}$ matrix can be moved from the preprocessing program to the solver, i.e., $(Z'A_{22}^{-1}Z)^{-1}Z'A_{22}^{-1}$. The MiX99 uses preconditioned conjugate gradient (PCG) iteration, and the PCG method was assumed to be converged when $C_r<10^{-7}$. The C_r is defined as a Euclidean norm of the difference between the right-hand side (RHS) of the MME and the one predicted by the current solutions relative to the norm of RHS.

In the validation test, we had 366 DFS candidate bulls with at least 20 daughters with records in the full TD data and no daughters with records in the reduced data. Validation was done with the linear regression (LR) cross-validation method (Legarra and Reverter, 2018). The LR method estimates bias and inflation by comparing predictions based on the reduced and the full data. The coefficient of determination (R^2) corresponds to the reciprocal of the increase in reliability from reduced data evaluations to the full data evaluations.

Results & Discussion

The comparison of GEBVs from the GT_F and GT_T showed that genomic data truncation did not affect the level of GEBVs. Figure 1 shows the correlation between full TD data GEBVs from GT_F and GT_T for DFS bulls selected into AI by bull birth years. The correlation between the full genomic and the truncated genomic data GEBVs were 0.99 – 1.00 for reliably estimated bulls, i.e., with protein GEBV reliability, $r^2 > 0.7$, and 0.98 – 0.99 for bulls with protein $r^2<0.7$.

Table 1 shows the LR validation result from the different models for 366 DFS candidate bulls. Before validation (G)EBVs were centered to the same mean using the mean (G)EBV of HOL cows born in 2007. The b_0 column is the mean difference (kg) between the full and the reduced run (G)EBVs, b_1 is the regression coefficient, and R^2 is the coefficient of determination. The LR validation shows that removing the old genotypes in GT_T had the desired impact. The b_0, indicating the amount of bias, decreased in all traits with genomic data truncation. The b_1 values also improved considerably. For example, for protein, the b_1 increased from 0.77 to 0.88. On average, b_1 improved 11 % with genomic data truncation. Similarly, the R^2 was always higher with the GT_T model than by GT_F, improving on average by 5%, which indicates a better predictive ability of the model with genomic data truncation.

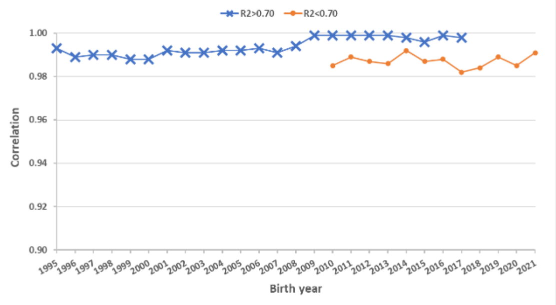

Figure 1. Correlation between the full TD data GEBVs for protein of the selected AI bulls from models using different genomic data. The single-step models are ssGTBLUP with full genomic data (GT_F) and ssGTBLUP with left truncated genomic data (GT_T). For bulls with protein GEBV reliability, r2 >0.7 and for bulls with r2<0.7

Figure 2 shows the genetic trends of protein for DFS Holsteins bulls. After the introduction of genomic selection in 2010, the genetic trend by GT_F was much higher in the reduced data compared to that in the full TD data, whereas with the GT_T the genetic trend in the reduced TD run is similar to that in the full TD data trend. Thus, truncation of genomic data removed the overprediction of the bulls.

Table 1. Linear regression (LR) results for the validation bulls from the ssGTBLUP with full genomic data (GT_F) and ssGTBLUP with left truncated genomic data (GT_T), and from the animal model (AM). The values in the table are: $b_{0=}$ mean(Full_(G)EBV–reduced_(G)EBV), b_1 regression coefficient and R^2 coefficient of determination.

	Model	b_0	b_1	R^2
Milk	AM	-148.09	0.81	0.30
	GT_F	-457.26	0.86	0.64
	GT_T	-148.04	0.93	0.68
Protein	AM	0.26	0.72	0.23
	GT_F	-14.60	0.77	0.59
	GT_T	-3.96	0.88	0.62
Fat	AM	-2.83	0.82	0.36
	GT_F	-21.01	0.79	0.66
	GT_T	-9.53	0.88	0.69

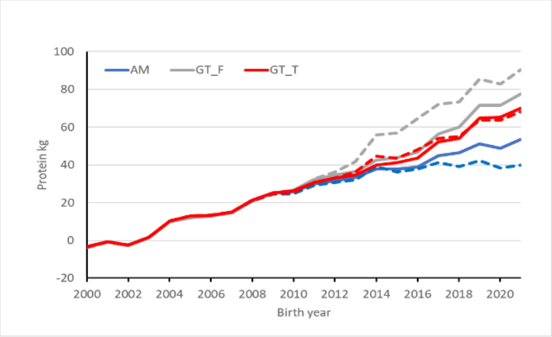

Figure 2. Bull genetic trends for protein (G)EBV. The trend. The models are the animal model (AM), the ssGTBLUP with full genomic data (GT_F), and ssGTBLUP with left truncated genomic data (GT_T). Solid lines are for full data and dashed lines for reduced data trends.

Figure 3 shows the average Mendelian Sampling (MS) term of genotyped bulls by birth years for protein GEBV. The means include all the bulls genotyped and, therefore, it is expected that the mean GEBV would be equal to the parent average. The figure shows that for the youngest age classes, the difference is about 4 kg with GT_F but less than 2 kg for GT_T, and for the bulls born before 2019, MS from GT_T is near the expected zero level like in the animal model. Thus, the genomic data truncation had a positive effect also on the MS term averages. Before 2009, the start of genomic selection, the mean MS terms were below zero probably because of the overprediction of bull dam EBVs.

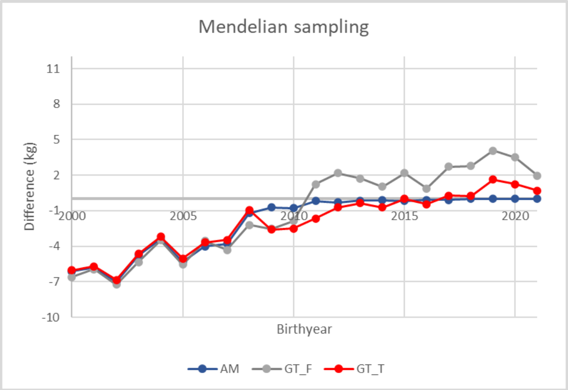

Figure 3. Mendelian sampling term means for protein for all genotyped DFS bulls by birth year calculated from (G)EBV from full TD data. The different models are the animal model (AM), the ssGTBLUP with full genomic data (GT_F) and ssGTBLUP with left truncated genomic data (GT_T).

Based on all comparisons it seems that the use of left truncated genomic data improves the single-step evaluations. The reason might be that the older genotyped animals whose genotypes are removed have weaker connections to the current population than assumed by genomic information. Old genotyped bulls expand the genetic trend of the young animals which becomes too high. Thus, removing the old genotypes in the left truncated data helped to reduce the overvaluation of genomic information. It also seems that the left truncated genomic data works similarly as the erosion factor presented by Boichard (2022) and Croué et al. (2022), where the bias in the young candidate animals is corrected using erosion factor that depends on the distance between the candidate and the reference population.

Conclusions

As a final remark, it seems that the removal of the genotypes of animals born before 2009 led to better validation results compared to the full genomic data, and also the MS term averages were closer to the mean of zero. The genomic data truncation did not much affect the ranking of bulls within birth years. For older and reliably evaluated bulls, the correlation between GEBVs with full genomic data and truncated genomic data was nearly one, and there was no considerable reranking of the candidate bulls. Thus, truncation of genomic data removed the over prediction of recent year classes of bulls and reduced the amount of overdispersion in candidate evaluations to a level acceptable in practice.

Acknowledgements

This work was a part of the Nordic Genomics in Herd project. The Nordic cattle breeding organizations (Viking Genetics, Randers, Denmark; Nordic Cattle Genetic Evaluation (NAV), Aarhus, Denmark), are acknowledged for providing the genotype data and the test day data.

References

Aguilar, I., Misztal, I. Johnson, D-L., Legarra, A., Tsuruta, S., Lawlor, T.J. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci.* 93, 743-752. doi:10.3168/jds.2009-2730.

Boichard, D., Fritz1, S., Croiseau, P., Ducrocq, V., Cuyabano, B., Tribout, T. 2022. Longdistance associations generate erosion of genomic breeding values of candidates for selection. In: Proceedings of the 12th World Congress on Genetics Applied to Livestock Production; July 3 to 8, 2022; Rotterdam (The Netherlands).

Christensen, O.F., Lund M.S. 2010. Genomic prediction when some animals are not genotyped. *Genet. Sel. Evol. 42*, 2. doi: 10.1186/1297-9686-42-2.

Croué, I., Barbat, M., Launay, A., Promp, J., Guillerm, M., Boulesteix, P., Minéry, A., Fritz, S., Tribout, T., Boichard, D. 2022. In France, Single-Step is going live! Interbull open Meeting 2022, Montreal, Canada, May 30 – June 3.

Koivula, M., Strandén, I., Pösö, J., Aamand, G.P., Mäntysaari, E.A. 2015. Single-step genomic evaluation using multitrait random regression model and test-day data. *J. Dairy Sci.*, 98:2775-2784. https://doi.org/10.3168/jds.2014-8975.

Koivula, M., Strandén, I., Aamand, G.P., Mäntysaari, E.A. 2018. Reducing bias in the dairy cattle single-step genomic evaluation by ignoring bulls without progeny. *J. Anim. Breed. Genet*., 135(2), pp.107-115. https://doi.org /10.1111/jbg.12318.

Koivula M., Strandén I., Aamand G.P., Mäntysaari, E.A. Accounting for Missing Pedigree Information with Single-Step Random Regression Test-Day Models. *Agriculture*. 2022; 12(3):388. https://doi.org/10.3390/agriculture12030388

Legarra, A., Reverter, A. 2018. Semi-parametric estimates of population accuracy and bias of predictions of breeding values and future phenotypes using the LR method. *Genet. Sel. Evol. 50*, 53. doi: 10.1186/s12711-018-0426-6.

Lidauer, M., Pösö, J., Pederson, J., Lassen, J., Madsen, P., Mäntysaari, E.A., Nielsen, U., Eriksson, J-Å., Johansson, K., Pitkänen, T., Strandén, I., Aamand, G.P. 2015. Across-country test-day model evaluations for Nordic Holstein, Red Cattle and Jersey. *J. Dairy Sci. 98*, 1296–1309. doi: 10.3168/jds.2014-8307.

Mäntysaari, E.A., Evans, R.D., Strandén, I. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *J. Anim. Sci. 95*, 4728-4737. doi: 10.2527/jas2017.1912.

Mäntysaari, E.A., Koivula, M., Strandén, I. 2020. Symposium review: Single-step genomic evaluations in dairy cattle. *J. Dairy Sci. 103*, 5314-5326. doi: 10.3168/jds.2019-17754.

Oliveira, H. R., Lourenco, D.A.L., Masuda, Y., Misztal, I., Tsuruta, I., Jamrozik, J, Brito,J.L:F., Silva, F.F., Schenkel, F.S.2019. Application of single-step genomic evaluation using multiple-trait random regression test-day models in dairy cattle. J. Dairy Sci, 102:2365-2377. https://doi.org/10.3168/jds.2018-15466.

Strandén, I., Lidauer, M. Solving large mixed models using preconditioned conjugate gradient iteration. J. Dairy Sci. 1999, 82, 2779-2787. doi: 10.3168/jds.S0022-0302(99)75535-9.

Tsuruta, S., Lourenco, D. A. L., Masuda, Y., Misztal, I., Lawlor, T.J. 2019. Controlling bias in genomic breeding values for young genotyped bulls. *J. Dairy Sci.* (online) https://doi.org/10.3168/jds.2019-16789.