Meta-model for genomic relationships of metafounders applied on large scale single-step random regression test-day model

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

Abstract

In this study traditional genetic groups and metafounders were compared in analysis of a random regression TD model with ssGTBLUP. The compared models were 1) ssGTBLUP with QP transformation of genetic groups and a genomic relationship matrix **G** built using base population allele frequencies for the markers, 2) ssGTBLUP with QP transformation of genetic groups and **G** with the same allele frequency of 0.5 for the markers, 3) ssGTBLUP with the metafounder (MF) approach and **G** with the same allele frequency 0.5. All models used VanRaden method 1 in **G** and had a 30% residual polygenic proportion (RPG). The **G** matrix in cases 1) and 2) was scaled to have average diagonal equal to the pedigree-based relationship matrix $\mathbf{A}_{22}$ of genotyped animals. Models 2) and 3) gave very similar results in terms of overprediction. Also, it seems that ssGTBLUP is quite robust to allele frequency used in the **G** matrix. However, the MF approach might be more efficient in reducing bias. In conclusion, both the QP transformation and the MF approach can be implemented in large-scale ssGTBLUP evaluation.

Key words: genomic selection, metafounder, allele frequency, genetic groups

Introduction

During the last decade, genomic selection has become common in dairy cattle breeding (VanRaden, 2020). 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 the ssGBLUP have been presented (reviewed in Mäntysaari et al., 2020).

An important issue in the single step evaluations is the manner the genetic groups for the unknown parents are included in the model, or how the pedigree and genomic relationship matrices relate to them. If unknown parent groups (UPG) are included in the full pedigree **A** matrix but not accounted in $\mathbf{A}_{22}$ or **G**, it may cause problems in convergence of the iterative PCG solver (Matilainen et al., 2018). In many cases, this problem can be solved by properly accounting the contributions of the genotyped animals to the genetic groups.

There are also other ways to make **A** and $\mathbf{A}_{22}$ compatible with **G**. The metafounder (MF) approach was proposed by Legarra et al. (2015) to achieve the compatibility in the pedigree and genomic relationship matrices. The MF are like UPG but allow a related base population with nonzero inbreeding coefficients (e.g., Legarra et al., 2015). However, the large number of UPG in dairy cattle evaluation models can make implementation of the MF approach challenging.

A base population allele frequency (AF) has been recommended to be used when computing the genomic relationship matrix (VanRaden, 2008). If AF are not properly estimated, biased relationships and subsequently biased genomic breeding values may result (Aguilar et al., 2010; Christensen and Lund, 2010). Thus, AF should be estimated from the unselected base population.

Base population AF can be estimated from genotyped animals that have missing parents or parents that have genotyped ancestors (details Gengler et al., 2007). However, if there are many genotyped females in the most recent years, the AF of the youngest females will dominate observed data AF. AF can also be estimated from different base populations, e.g., groups of animals with unknown parents. This requires different groups (e.g., breed or breed-origin and the birth decade) to be defined into pedigree. Then AF of these groups can be estimated with the method by McPeek et al. (2004) applied for several base population groups as has been implemented in Bpop (Strandén and Mäntysaari, 2020). The simplest option is to assume that the base population AF is 0.5. This approach assumes that the base population is infinite number of generations back in time.

In this study the different options available to model missing information in pedigree and the genomic relationship matrices in random regression TD model with the ssGTBLUP were compared.

Materials and Methods

Data

Analyses used data from the official Nordic Holstein milk production evaluations. The official multiple trait milk production evaluation includes TD records from milk, fat, and protein production. The full routine evaluation data from November 2020 for the Holstein were obtained from the Nordic Cattle Genetic Evaluation (NAV). For the production traits, the TD data included 8.5 million cows and 10.9 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 observations in the full data were removed.

Holstein genotype data included 274 145 genotyped animals. Bulls were 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 evaluation.

Models

Single-step models were run with ssGTBLUP that uses Woodbury matrix identity based inverse genomic relationship with a **T** matrix (Mäntysaari et al. 2017). Three different **G** matrices were built for the comparisons. All models used VanRaden method 1 in **G** and had 30% residual polygenic proportion (RPG) but differed in the used AF and the scaling method. Tested models were 1) ssGTBLUP with a full QP transformation of genetic groups, and the **G** matrix built using base population AF of the birth decade group 1980 (GT_1980), 2) ssGTBLUP with full QP transformation of genetic groups, and the **G** matrix with AF 0.5 for all markers (GT_0.5). The third option was 3) ssGTBLUP with metafounder approach and the **G** matrix used AF 0.5 with MF inbreeding coefficients accounted in $\mathbf{A}^{\Gamma}$ and $\mathbf{A}_{22}^{\Gamma}$ (GT_MF). In addition, animal model without genomic information was used to show changes in predictions due to genomic information. The pedigree inbreeding coefficients were accounted in $\mathbf{A}^{-1}$ and $\mathbf{A}_{22}$ and in construction $\mathbf{A}^{\Gamma}$ and $\mathbf{A}_{22}^{\Gamma}$ for GT_MF.

Efficient method to include UPG into the mixed model equations is to use QP transformation (Quaas and Pollak, 1981). The full QP transformation (Matilainen et al. 2018) for ssGTBLUP has been described in Koivula et al., (2021). For the GT_1980, the AF were calculated simultaneously for six base groups by decades starting from 1960s using Bpop program (Strandén and Mäntysaari, 2020). From these base group AF estimates, the AF for the decade 1980s were used in **G** -matrix. The **G** matrices in cases 1) and 2) were scaled to have average diagonal equal to the pedigree-based relationship matrix of genotyped animals ($\mathbf{A}_{22}$).

The MF approach needs a covariance matrix for the metafounders. This is challenging to estimate for populations with many metafounders, long pedigree and many young genotyped animals (Kudinov et al., 2020). Thus, we first defined less genetic groups than in the original NAV evaluation. The new groups were based on breed (HOL, RDC, JER, other) and country of origin within Holstein (DNK, SWE, FIN, red and other). Within each of these 8 sources, UPG were further grouped by decade of birth and by selection path when appropriate. Thus, the original number of UPGs 446 was reduced to 176.

The same 176 UPGs were used as meta-founders in the MF model. The MF approach assumes that the meta-founders have defined self-relationships and relationships defined by a $\boldsymbol{\Gamma}$ -matrix. The relationships in Γ are to be "genomic compliant" so that the genotyped animals in future generations descending from the metafounders will have a pedigree relationship matrix ($\mathbf{A}_{22}^{\boldsymbol{\Gamma}}$) that matches the genomic relationships in **G.** Using the literature values (Kudinov et al., 2020) we first defined a covariance function model (Kirkpatrik et al., 1994) for the 8 base meta-founder groups. In addition, each of the groups were assumed to have a breed specific time trend in self-relationships. With known regression coefficients and design matrices the CF model can be used to describe any size $\boldsymbol{\Gamma}$ matrices, e.g. $\boldsymbol{\Gamma}_{176}=\Phi_{176}\boldsymbol{K}\Phi_{176}'$ Here we choose heuristic values for K based on expectations from numerous descriptive analyses, but for more formal analyses see Kudinov et al. (2021). After solving the $\boldsymbol{\Gamma}_{176}$-matrix, we attained $\boldsymbol{\Gamma}$ compliant inbreeding coefficients and could make the computations involving the inverses $\mathbf{A}^{-\boldsymbol{\Gamma}}$ for all the pedigree animals and $\mathbf{A}_{22}^{-\boldsymbol{\Gamma}}$ for the genotyped animals when solving iteratively the mixed model equations.

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

In validation test, we had 524 candidate bulls with at least 20 daughters with records in the full data and no daughters with records in the reduced data. Validation was done with linear regression cross-validation method (named LR by Legarra and Reverter, 2018). The method estimates bias and inflation by comparing predictions based on the reduced and the full data. The coefficient of determination (R^2) from the LR validation can be interpreted as a reciprocal of the increase in reliability from reduced data evaluations to the full data evaluations.

Results & Discussion

Table 1 shows the LR validation result from the different models for 524 Danish, Finnish and Swedish (DFS) Holstein bulls in the validation. To correct the level differences between models, the EBVs and GEBVs were standardized according to mean of cows born 2007. The b_0 column is the mean difference (kg) between the full and reduced run (G)EBVs. The values show that with the MF model the difference was slightly smaller than with the QP models. The regression coefficients (b_1) and coefficient of correlation (R^2) values were similar in the QP models. However, it seems that the MF model had slightly better results both in terms of bias and reliability.

Table 1. Bull LR validation (Bulls=524) results. Regression coefficients (b_1) and coefficient of correlation (R^2) from the animal model (EBV) and different single-step models. b_0=mean(Full_(G)EBV – reduced_(G)EBV). The single-step models are ssGTBLUP with QP and allele frequency 1980 (GT_1980), ssGTBLUP with QP and allele frequency 0.5 (GT_0.5) and ssGTBLUP with metafounders (GT_MF).

	Model	b_0	b_1	R^2
Milk	EBV	-101.7	0.84	0.32
	GT_1980	-311.8	0.87	0.67
	GT_AF	-319.8	0.87	0.67
	GT_MF	-272.3	0.89	0.68
Protein	EBV	0.80	0.74	0.24
	GT_1980	-10.81	0.82	0.63
	GT_AF	-11.10	0.81	0.63
	GT_MF	-9.71	0.83	0.64
Fat	EBV	-2.18	0.73	0.23
	GT_1980	-15.81	0.82	0.64
	GT_AF	-16.16	0.82	0.64
	GT_MF	-14.67	0.85	0.65

Figure 1 shows the means of Mendelian sampling (MS) terms of bulls by birth year for protein in the full data model. Difference is plotted for DFS bulls, including also young bulls that have no daughters. The bulls born after 2016 have only genomic information. The figure shows that for the youngest age classes the difference is about 6 kg for the GT_1980 model, and 5 kg with other models, so it seems that the model with AF 1980 gives a bit higher bias for the youngest bulls. Before the start of genomic selection, MS term means were quite stable. It is a little below zero presumable because of overprediction f bull dam EBVs. After the genomic selection started to affect, the mean also started to increase.

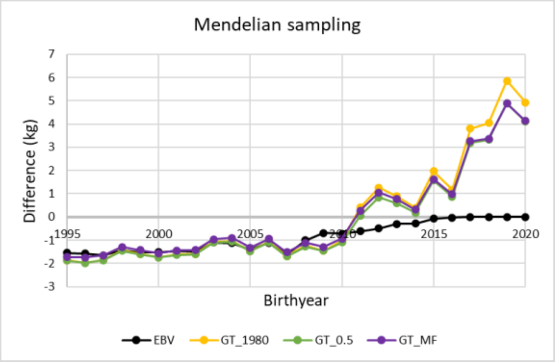

Figure 1. Mendelian sampling term in protein for all DFS bulls by birth year. The trend for protein (G)EBV from the full data model. The models are the animal model (EBV), ssGTBLUP with QP and allele frequency 1980 (GT_1980), ssGTBLUP with QP and allele frequency 0.5 (GT_0.5) and ssGTBLUP with metafounders (GT_MF).

Figure 2 shows the genetic trends of protein for DFS Holsteins Bulls. Solid lines are from the full data runs and dashed lines from the reduced data runs. Except for the clearly lower trend for animal model EBVs, the ssGTBLUP model trends are quite similar. For the models using QP, only the youngest bull groups in the full data in the model using AF 1980 seems to give slightly higher trend than the AF 0.5 model. In the reduced runs, the trends are similar. When the trends from the GT_1980 model are compared with GT_0.5 and the MF model trends, they appear very similar in the full data. However, for the MF model in the reduced data set, the trend differs slightly from the other single-step models, indicating lower overprediction in the MF model compared to the QP models.

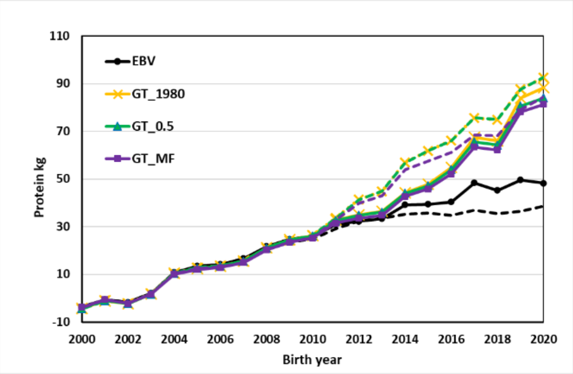

Figure 2. Genetic trends for bulls by birth year. The trend for protein (G)EBV. The models are the animal model (EBV), ssGTBLUP with QP and allele frequency 1980 (GT_1980), ssGTBLUP with QP and allele frequency 0.5 (GT_0.5) and ssGTBLUP with metafounders (GT_MF). Solid lines are for full data and dashed lines for reduced data trends.

Based on all comparisons it seems that the traditional genetic group model and the MF model both are feasible options for handling genetic groups in single step evaluations. Also, it seems that ssGTBLUP is quite robust to AF used in the **G** matrix. However, a **G** matrix using base population AF is theoretically more correct, but on the other hand requires more work because AF has to be estimated. Also, it appears that the traditional genetic group model has similar inflation and prediction ability as the MF model.

Single step MF model might be a more sophisticated way to combine pedigree and genomic information because both the genomic and the pedigree-based relationship matrices are modified according to genomic information. Moreover, it seems that the MF model does not increase the trend of young, genotyped animals as much other single step methods tested. The MF model also gives marginally better validation results compared to the other methods. However, the MF model requires some additional estimation in order to build the **Γ**-matrix.

Conclusions

As a final remark, it seems that both the traditional genetic groups and MF approach can be implement in ssGTBLUP. With large genomic data both methods are computationally efficient. However, it seems that the MF approach might be more efficient in reducing bias.

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.

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.

Gengler, N., Mayeres, P., Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal 1*, 21-28.https://doi.org/10.1017/S1751731107392628.

Kirkpatrick, M., Hill, W.G, Thompson, R. 1994. Estimating the covariance structure of traits during growth and ageing, illustrated with lactation in dairy cattle. *Genet. Res. 64*, 57–69. https://doi.org/10.1017/S0016672300032559.

Koivula, M., Strandén, I., Aamand, G.P., Mäntysaari, E.A. 2021. Practical implementation of genetic groups in single-step genomic evaluations with Woodbury matrix identity based genomic relationship inverse. *J. Dairy Sci., 104*, 10049-10058. https://doi.org/10.3168/jds.2020-19821.

Kudinov, A.A., Mäntysaari, E.A, Aamand, G.P, Uimari, P., Strandén, I. 2020. Metafounder approach for single-step genomic evaluations of Red Dairy cattle. *J. Dairy Sci. 103*, 6299-6310. doi: 10.3168/jds.2019-17483

Kudinov, A.A., Koivula, M., Strandén, I., Aamand, G.P., Mäntysaari, E.A. 2021. Single-step genomic predictions of a minor breed concurrently with a main breed large national genomic evaluation. *Interbull Bulletin 56*, in press.

Legarra, A., Christensen, O.F., Vitezica, Z.G., Aguilar, I., Misztal. I. 2015. Ancestral relationships using metafounders: finite ancestral populations and across population relationships. *Genetics, 200*, 455-468. doi: 10.1534/genetics.115.177014.

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.

Matilainen, K., Strandén, I., Aamand, G. P., Mäntysaari, E.A. 2018. Single step genomic evaluation for female fertility in Nordic Red dairy cattle. *J. Anim. Breed. Genet. 135*, 337-348. doi: 10.1111/jbg.12353.

McPeek M. S., Wu, X, Ober, C. 2004. Best linear unbiased allele-frequency estimation in complex pedigrees. *Biometrics 60*, 359–367. doi: 10.1111/j.0006-341X.2004.00180.x.

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.

Quaas, R. L., Pollak, E.J. 1981. Modified equations for sire models with groups. *J. Dairy Sci., 64*, 1868–1872. doi: 10.3168/jds.S0022-0302(81)82778-6

Stranden, I., Mäntysaari, E.A. 2020. Bpop: an efficient program for estimating base population allele frequencies in single and multiple group structured populations. *Agr. Food Sci. 29*, 166–176. doi: 10.23986/afsci.90955.

Tijani, A., Wiggans, G.R., Van Tassell, C.P., Philpot. J,C., Gengler, N. 1999. Use of (Co)Variance Functions to Describe (Co)Variances for Test Day Yield. *J. Dairy Sci. 82*, 226.e1-226.e14. https://doi.org/10.3168/jds.S0022-0302(99)75228-8.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91*, 4414-4423. doi: 10.3168/jds.2007-0980.

VanRaden, P.M. 2020. Symposium review: How to implement genomic selection. *J. Dairy Sci. 103*, 5291–5301. doi: 10.3168/jds.2019-17684.