Including genetic groups as fixed or random effects in large scale single-step genomic predictions

Ø. Nordbø[12], A. B. Gjuvsland[12], L. S. Eikje[1] and I. Strandén[3]

[1] Geno SA, Storhamargata 44, 2317 Hamar, Norway
[2] Norsvin SA, Storhamargata 44, 2317 Hamar, Norway
[3] Natural Resources Institute Finland (Luke), FI-31600 Jokioinen, Finland

Abstract

Genetic groups are often used in genetic evaluations to account for selection that cannot be accounted for by kinship. The genetic groups can be fitted as random or fixed effects, both in traditional BLUP and in single-step genomic predictions (ssGBLUP). In this study, we investigate how inclusion of genetic groups affect the predictive ability and bias in a large-scale three-trait ssGBLUP model for kg of milk, fat percentage, and protein percentage in Norwegian Red cattle. Including genetic groups as random effects rather than fixed effects reduced bias of breeding values. Specifically, for genotyped animals and ungenotyped animals with genotyped offspring and with sparse pedigree data, the bias of the Mendelian term of breeding values was reduced by about 25%, when fitting genetic groups as random compared to fixed.

Key words: Genetic groups, single-step, genomic predictions, ssGBLUP, bias, validation

Introduction

Geno SA has been running single step genomic predictions (ssGBLUP) (Christensen and Lund, 2010; Legarra et al., 2009) for all traits in the routine evaluations of Norwegian Red since 2017. All selection has been based on indexes from these evaluations, performed approximately every second week. For traits with strong genetic progress, genetic groups (Quaas, 1988) have been used to account for selection that cannot be accounted for by kinship.

The genetic groups were initially modelled explicitly as fixed covariate regressions (Nordbø et al., 2019), which is equivalent to QP-transformation of the inverse of the unified pedigree and genomic relationship matrix $\mathbf{H}^{-1}$ (Misztal et al., 2013). However, this approach led to some incorrect extreme EBV levels for 1) genotyped individuals with missing ancestry and 2) ungenotyped animals with missing ancestry and genotyped offspring. For that reason, only the pedigree-based relationship matrix **A** is now augmented to include genetic groups, by using QP transformation on $\mathbf{A}^{-1}$.

By ignoring the genomic contributions in the genetic groups, the abovementioned bias was reduced considerably. However, some of it seems still to remain. To further improve the quality of the genomic predictions, we test in this study whether fitting genetic groups as random effects versus fixed effects could influence the remaining bias.

Materials and Methods

Modelling

To investigate the effect of fitting genetic groups as fixed versus random, we applied a multi-trait repeatability model for lactation production of kg milk, fat percentage and protein percentage. Approximately 8.1 million 305-day records from 1st to 5th lactation, on 3.8 million animals, were included in the data set. Heritabilities and genetic correlations are shown in Table 1.

The following repeatability model was used for prediction of breeding values: $y_{ijklmno} = YML_i + ALM_j + DoL_k + H5Y_l + \beta_m * Het + pe_n + gen_n + hy_o + e_{ijklmno}$.

Fixed effects were YML_i = Year × Month × Lactation number, i=1,…1778, ALM_j = Age at calving × Lactation number × Milking system, j=1,…291, DoL_k = Days open × Lactation number, k=1,…71, and $H5Y_l$ = Herd × 5 years period, l=1,…132 089.

Random effects were pe_n = Permanent environment, gen_n = additive genetics, hy_o = herd × year, o=1,…682 251, and $e_{ijklmno}$ = residual error.

In addition, to correct for heterosis, heterozygosity of SNP-markers (*Het*), was fitted as fixed linear regressions (β_m) within lactation number (*m=1,…5)*. For genotyped animals, we used the proportion of heterozygote SNPs directly (Iversen et al., 2019). For ungenotyped animals, heterozygosity was predicted based on linear regression with inbreeding coefficient as independent variable. The true heterozygosity and inbreeding coefficient of the genotyped animals were used to estimate the regression coefficients.

Table 1. Heritabilities (diagonal) and genetic correlations between Kg milk, protein % and fat % (off-diagonal)

	Kg milk	Protein %	Fat %
Kg milk	0.42		
Protein %	-0.44	0.61	
Fat%	-0.36	0.62	0.36

The phenotypes (*y)* had been pre-corrected for heterogeneous variance due to Age at calving × Lactation number × Milking System.

Pedigree and genotypes

The pedigree consisted of 4.8 million animals. Missing pedigree data were grouped by year of birth and by the following three classes: the missing parent is a missing AI sire, the missing parent is a missing farm bull, or the missing parent is a missing dam. This resulted in 118 genetic groups. In total, about 7% of the animals had missing dam and 14% had missing sire.

In total, imputed genotypes were available for 124 493 animals with 121 740 SNPs. We used equal allele frequencies (0.5) for all SNPs when constructing the genomic relationship matrix $\mathbf{G}$ (Aguilar et al., 2010; VanRaden, 2008), which was subsequently scaled by multiplying a parameter to all matrix elements to make the average diagonal elements equal to 1 (Forni et al., 2011). Then the additive pedigree relationship matrix was weighted by 10% to account for genetic effects that are not captured by the SNPs. The HGINV program (Strandén and Mantysaari, 2016) was used to build the combined inverse pedigree and genomic relationship matrix $\mathbf{H}^{-1}$.

To investigate how genetic groups affected the predictive ability, we made two different prediction models, one where genetic groups were fitted as fixed effects, and one where they were fitted as random effects. In the latter model, they were given the same variance as the additive genetic effect. These two scenarios were compared with each other and with a corresponding best linear unbiased predictions (BLUP) model that excluded genomic information. All predictions were made, using MiX99 (MiX99 Development Team, 2017).

Validations

For validation of genomic predictions, we applied the Interbull's GEBV test (Mäntysaari et al., 2010), where the 5 last years of data were masked before performing genomic predictions. The estimated breeding values (EBV) were then compared with daughter yield deviations (DYD) from the corresponding BLUP using weighted linear regression.

Level-bias (Nordbø et al., 2019), the average change in breeding value when adding an animal's genomic data into $\mathbf{H}^{-1}$, was investigated by masking the genotypes of 1000 young animals. To prevent any potential shift in levels for all animals between subsequent runs, the EBVs were first scaled internally using all

Norwegian proven bulls. The mean change in EBVs of these 1000 young animals between ssGBLUP runs when excluding and including their genomic information was then divided by the genetic standard deviation as in (Nordbø et al., 2019).

Further, we looked at how EBVs deviated from the parental average for groups of animals. The deviation from parental average was divided by the genetic standard deviation to make the term, *MS_{Bias,}* easier to interpret:

$$MS_{Bias} = \text{Mean}\left(\frac{EBV - 0.\ 5(EBV_{sire} + EBV_{dam})}{\sigma_g}\right)$$

The relevant groups consisted of genotyped animals and ungenotyped animals with genotyped offspring, and with zero, one, two, three or four missing grandparents.

Finally, we investigated the genetic trends (average of EBVs of animals, based on birth year). We then looked at genetic trends from the BLUP model versus those from the ssGBLUP models, with full or reduced dataset (masked 5 last years of data).

Results

The reliabilities, R^2 and regression coefficients, β from the GEBV-test were relatively insensitive to the method of including genetic groups (Table 2). Both methods gave satisfactory reliabilities and regression coefficients close to 1 and both models passed the Interbull’s GEBV test.

Table 2. Reliabilities, R^2 and regression coefficients, β from the GEBV-test for three traits and the two different models, where genetic groups were fitted as fixed or random effects

	Kg milk		Protein %		Fat %	
	R^2	β	R^2	β	R^2	β
Fixed	0.573	1.02	0.63	0.96	0.699	1.11
Random	0.574	1.02	0.63	0.96	0.700	1.11

The magnitudes of level-biases were small, and insensitive to the method for fitting genetic groups (Table 3).

Table 3. The mean change in breeding values when including genomic information (level-bias) for three traits and the two different models, where genetic groups were fitted as fixed vs. random effects

	Kg milk	Protein %	Fat %
Fixed	0.02	0.00	0.00
Random	0.02	0.00	0.00

Compared with the measures mentioned above, MS_{Bias} was more sensitive to the method for fitting genetic groups. When fitting genetic groups as random effects instead as fixed, MS_{Bias} for Kg Milk was smaller for the group of genotyped and ungenotyped animals with genotyped offspring (Figure 1). This difference was more prominent when the amount of missing ancestry was larger. For 3 and 4 missing grandparents, MS_{Bias} was reduced by about 25% when fitting groups as random versus fixed. MS_{Bias} for Protein% and Fat% were generally low, about 10% of the values obtained for Kg Milk.

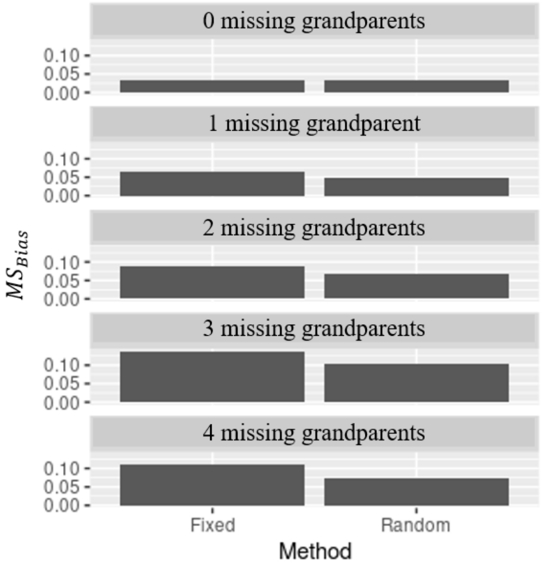

Figure 1. Mean deviation from parental average EBVs (MS_{Bias}) for kg of milk, for genotyped animals and ungenotyped animals with genotyped offspring, and for different number of missing grandparents. To the left genetic groups are fitted as fixed and to the right, as random effects.

Fitting genetic groups as a random rather than fixed made the genetic trend more like the

BLUP trend (Figure 2). In addition, the correspondence between genetic trends based on ssGBLUP predictions with full and reduced datasets improved.

Discussion

In the current study we observed that fitting genetic groups as random effects led to less bias in ssGBLUP predictions when compared to fitting genetic groups as fixed effects. Using random groups reduces the MS_{Bias} term for relevant groups of animals and it gives improved correspondence between genetic trends for BLUP and ssGBLUP, based on both reduced and full data sets. Because of selective genotyping of good animals and preferential treatment of bull-dams (Pedersen et al., 1995), the $\mathrm{MS}_{\mathrm{Bias}}$ should not be expected to be equal to zero. However, one should expect that it should stay at the same level, independent on the amount of missing ancestry. Hence, we conclude that the model where the genetic groups were fitted as random effects were less biased.

The results from Interbull's GEBV-test indicate that fitting genetic groups as random effects gave marginally improved accuracy compared to fitting genetic groups as fixed effects. Regression coefficients were the same for both approaches. Their magnitudes support the hypothesis, together with the good correspondence between genetic trends, that the genomic predictions for milking traits on Norwegian Red are unbiased. Many cows and heifers have been genotyped the last couple of years, so the reference population has grown fast. Random genetic groups will be included in our future model development.

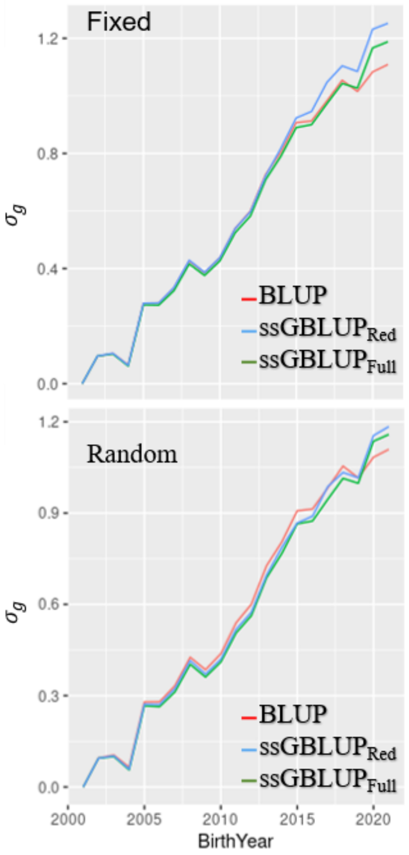

Figure 2. Genetic trend, in terms of genetic standard deviation, σ_g, for kg of milk for evaluation without genomic information (BLUP) vs. single-step evaluations using full ($\mathrm{ssGBLUP}_{\mathrm{Full}}$) and reduced dataset ($\mathrm{ssGBLUP}_{\mathrm{Red}}$). In the upper panel, genetic groups were fitted as fixed effects and in the lower, as random effects.

Acknowledgments

We want to thank the Norwegian Research council for funding this research through the project 309611, «Large scale single step genomic selection in practice».

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. https://doi.org/10.3168/jds.2009-2730

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

Forni, S., Aguilar, I., Misztal, I., 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. Gen. Sel. Evol. 43, 1. https://doi.org/10.1186/1297-9686-43-1

Iversen, M.W., Nordbø, Ø., Gjerlaug-Enger, E., Grindflek, E., Lopes, M.S., Meuwissen, T., 2019. Effects of heterozygosity on performance of purebred and crossbred pigs. Gen. Sel. Evol. 51, 8. https://doi.org/10.1186/s12711-019-0450-1

Legarra, A., Aguilar, I., Misztal, I., 2009. A relationship matrix including full pedigree and genomic information. J. Dairy Sci. 92, 4656–4663. https://doi.org/10.3168/jds.2009-2061

Mäntysaari, E.A., Liu, Z., VanRaden, P., 2010. Interbull validation test for genomic evaluations. Interbull Bulletin 17–17.

Misztal, I., Vitezica, Z.G., Legarra, A., Aguilar, I., Swan, A.A., 2013. Unknown-parent groups in single-step genomic evaluation. J. Anim. Br. Genet. 130, 252–258.

MiX99 Development Team, 2017. MiX99: A software package for solving large mixed model equations. Release XI/2017. Natural Resources Institute Finland (Luke). Jokioinen, Finland.

Nordbø, Ø., Gjuvsland, A.B., Eikje, L.S., Meuwissen, T., 2019. Level-biases in estimated breeding values due to the use of different SNP panels over time in ssGBLUP. Gen. Sel. Evol. 51, 76. https://doi.org/10.1186/s12711-019-0517-z

Pedersen, G.A., Christensen, L.G., Petersen, P.H., 1995. Evaluation of Breeding Value and Selection of Bull Dams in the Danish Dairy Breeds: I. Studies on Realized Efficiency of Bull Dam and Bull Sire Selection. Acta Agric. Scand. A Anim. Sci. 45, 26–31. https://doi.org/10.1080/09064709509410910

Quaas, R.L., 1988. Additive Genetic Model with Groups and Relationships. J. Dairy Sci. 71, 1338–1345. https://doi.org/10.3168/jds.S0022-0302(88)79691-5

Strandén, I., Mantysaari, E., 2016. HGINV program. Natural Resources Institute Finland (Luke).

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