Approximate single step genomic prediction for Norwegian Red cattle

J. Jenko[1], I. Stranden[2]
[1] Geno SA, Storhamargata 44, 2317 Hamar, Norway
[2] Natural Resource Institute Finland (Luke), Humppilantie 14, Jokioinen, Finland
Corresponding author: janez.jenko@geno.no

Abstract

The exact single step Genomic Best Linear Unbiased Prediction (ssGBLUP) method has been used for breeding value estimation in the Geno breeding program since 2016. The number of animals with genotype information included in ssGBLUP has increased to over 210,000, making the exact inversion of the genomic relationship matrix computationally demanding. To address this, we tested two alternative approaches on ninety traits used for breeding value evaluation in the Norwegian Red cattle breed. The single step Algorithm for the Proven and Young Genomic Best Linear Unbiased Prediction (ssAPYGBLUP) approach consisted of a core dataset with 16,480 progeny-proven sires and sires of foreign origin, considering a 10% residual polygenic effect. The single step Singular Value Decomposition Genomic Best Linear Unbiased Prediction (ssSVDGBLUP) approach utilized genotypes from 5,186 progeny-proven sires, explaining 90% of genetic variation through chromosome-specific singular values. We compared estimates from these approximate methods to those from ssGBLUP for animals in the pedigree, and young genotyped animals for all the ninety traits. Correlations between ssGBLUP and ssAPYGBLUP estimates ranged from 0.976 to 1.000 for all the individuals in pedigree and from 0.940 to 0.995 for young genotyped individuals. For the ssSVDGBLUP and ssGBLUP approaches, correlations were between 0.971 and 1.000 for animals in the pedigree, and between 0.977 and 0.995 for young genotyped animals. When regressing ssGBLUP estimates to ssAPYGBLUP estimates, the linear regression coefficients were between 0.993 and 1.027 for all animals in the pedigree and between 1.005 and 1.061 for young genotyped animals. For the regression of ssGBLUP estimates to ssSVDGBLUP estimates, the linear regression coefficients were between 0.953 and 1.055 for all animals in the pedigree and between 0.866 and 0.949 for young genotyped animals. This means that predictions for young genotyped animals when using ssSVDGBLUP showed overestimation while predictions from ssAPYGBLUP were slightly underestimated.

Key words: single step genomic prediction, singular value decomposition, algorithm for proven and young, Norwegian Red cattle

Introduction

Single step genomic predictions (ssGBLUP) were implemented in routine evaluation for the estimation of genomic breeding values for Norwegian Red cattle in 2016 (Nordbø et al., 2019). In the beginning, there were about 18,000 genotypes used in the evaluation of genomic breeding values. With genotyping around 35,000 animals annually more than 210,000 genotypes were present in the middle of 2023.

The inverse of the combined pedigree and genomic relationship matrix is calculated prior to the estimation of breeding values and demands a lot of computer memory where the information is stored temporarily. The increase in the number of genotyped animals is increasing computer memory requirements quadratically. This becomes unsustainable in the long term and other solutions must be

applied. One possible solution is to remove genotype information. These could be either genotypes from animals without phenotypic information or genotypes from older animals. Increasing computer memory would be another possible solution but due to a quadratic increase in memory requirements with every genotype added this cannot be a long-term solution.

Application of approximate ssGBLUP methods eg. Algorithm for Proven and Young (ssAPYGBLUP) proposed by Misztal et al. (2014) or Singular Value Decomposition (ssSVDGBLUP) approach proposed by Ødegård et al. (2018) could represent a long-term solution when using single step genomic predictions approach on a large number of genotyped individuals. These approaches decrease computational requirements with approximations which explain only the most important part of genetic variation in the population. The difference between the ssAPYGBLUP and ssSVDGBLUP approaches is that the ssAPYGBLUP algorithm assumes that all genetic variation is explained by the additive genetic effects of the core individuals, while the ssSVDGBLUP approach assumes it is explained by haplotype blocks that segregate among core individuals (Ødegård et al., 2018).

The objective of the current study was to test the ssAPYGBLUP and ssSVDGBLUP approaches for the routine evaluation of breeding values for the Norwegian Red cattle and to compare them to the currently applied ssGBLUP method.

Materials and Methods

We used phenotypes, genotypes, and pedigree information from April 13, 2023, Geno routine evaluation. We estimated breeding values for ninety traits included in the twenty-nine single- or multi-trait mixed model equations. Genomic relationships were estimated with 206,496 genotypes imputed to the in-silico array with 121,740 SNPs and combined with the pedigree information into a single step genomic relationship.

The exact single step approach

The mixed model equations in the ssGBLUP approach combine pedigree and genomic relationship information stored in the matrix H (Christensen and Lund, 2010):

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \lambda\mathbf{H}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \end{bmatrix}$$

where **X** and **W** are incidence matrices for the fixed and random effects, λ is a ratio between the error and additive genetic variances, vectors **b** and **a** are estimates for the fixed and random effects, and **y** is a vector of phenotypes. The inverse of the **H** relationship matrix calculated as:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & (\mathbf{G}_{\mathrm{w}})^{-1} - (\mathbf{A}_{22})^{-1} \end{bmatrix}$$

where **A** is a pedigree relationship matrix, $\mathbf{G}_{\mathrm{w}}$ combines the genomic and pedigree information for genotyped animals with 10% of information coming from genotypes and 90% from pedigree, and $\mathbf{A}_{22}$ is the pedigree-based relationship matrix of the genotyped individuals. A fraction of $\mathbf{A}_{22}$ is added to **G** because the **G** matrix derived using the VanRaden 1 method is often singular while also explaining additive breeding values that cannot be described by the available markers (VanRaden, 2008).

Algorithm for Proven and Young

In the ssAPYGBLUP approach, animals are partitioned into proven (core) and young (non-core) individuals and only the inverse of genomic relationships between the animals in the core is inverted while the estimates from the non-core individuals are calculated recursively. After preliminary analysis where different core assemblies were compared the core used on all the traits contained 16,480 genotyped animals from sires with a

Norwegian herd book number, animals from foreign populations, and animals with sire from foreign populations.

Singular Value Decomposition based models

In the ssSVDGBLUP approach, core animals were used to approximate correlations between markers using the chromosome-specific singular values explaining 90% of genetic variation in the core individuals. Here the core was assembled of 5,186 genotyped sires with a Norwegian herd book number. This core definition was based on a preliminary analysis which showed that differences between various cores and the proportion of genetic variance explained were small when looking at the prediction accuracy and bias while achieving a significant decrease in computational time and memory requirements with a smaller core size and genetic variance explained.

Standardization of breeding values

The obtained estimated breeding values (EBV) from all the tree approaches were standardized (EBVs) using the following equation:

$$EBVs = 100 + k * (EBV-\overline{EBVc})/sd(EBVb))$$

where $\overline{EBVc}$ is the mean EBV of all cows born between April 13, 2015, and April 13, 2020, and sd(EBVb) is the standard deviation of the EBV from the progeny of proven bulls that were born between January 1, 2006, and December 31, 2013.

Results & Discussion

The correlations between the ssGBLUP and ssAPYGBLUP estimates ranged from 0.976 to 1.000 for all the individuals in the pedigree and from 0.940 to 0.995 for the young genotyped individuals. For the ssGBLUP and ssSVDGBLUP approaches, correlations were between 0.971 and 1.000 for animals in the pedigree, and between 0.977 and 0.995 for the young genotyped animals (Table 1).

Table 1: Mean, standard deviation (sd), minimum (min) and maximum (max) correlation between predictions from ssGBLUP and predictions from ssAPYGBLUP (APY) or ssSVDGBLUP (SVD) across ninety traits

	mean	sd	min	max
Individuals in the pedigree				
APY	0.998	0.003	0.976	1.000
SVD	0.997	0.003	0.971	1.000
Young genotyped individuals				
APY	0.983	0.013	0.940	0.995
SVD	0.990	0.004	0.977	0.995

The linear regression coefficients when regressing ssGBLUP estimates to the estimates from ssAPYGBLUP ranged from 0.993 to 1.027 for the individuals in the pedigree and from 1.005 to 1.061 for the young genotyped individuals. Linear regression coefficients when regressing ssGBLUP estimates to the estimates from ssSVDGBLUP ranged from 0.953 to 1.055 for the individuals in the pedigree and from 0.866 to 0.949 for the young genotyped individuals.

Table 2: Mean, standard deviation (sd), minimum (min) and maximum (max) linear regression coefficient when regressing predictions from ssGBLUP to predictions from ssAPYGBLUP (APY) or ssSVDGBLUP (SVD) across ninety traits

	mean	sd	min	max
Individuals in the pedigree				
APY	1.006	0.006	0.993	1.027
SVD	1.004	0.013	0.953	1.055
Young genotyped individuals				
APY	1.029	0.011	1.005	1.061
SVD	0.912	0.022	0.866	0.949

The intercept ranged from -2.929 to 0.928 for the individuals in the pedigree and from -5.848 to -0.397 for young genotyped individuals when regressing ssGBLUP estimates to the estimates from ssAPYGBLUP. When regressing ssGBLUP estimates to the estimates from ssSVDGBLUP, the linear regression coefficient ranged from -5.803 to 4.943 for the individuals in the pedigree and from 4.514 to 13.796 for the young genotyped individuals.

Table 3: Mean, standard deviation (sd), minimum (min) and maximum (max) intercept when regressing predictions from ssGBLUP to predictions from ssAPYGBLUP (APY) or ssSVDGBLUP (SVD) across ninety traits

	mean	sd	min	max
Individuals in the pedigree				
APY	-0.613	0.686	-2.929	0.928
SVD	-0.411	2.030	-5.803	4.943
Young genotyped individuals				
APY	-2.865	1.160	-5.848	-0.397
SVD	8.847	2.378	4.514	13.796

In comparison to the estimates from the ssAPYGBLUP approach, the estimates from the ssSVDGBLUP approach showed on average slightly higher correlation to the estimates from the ssGBLUP approach. This was the case when taking into account animals in the pedigree and even more when looking only at the young genotyped animals. The estimates for the young genotyped animals from the ssSVDGBLUP approach were overestimated in comparison to the estimates from the ssGBLUP approach for all the traits. Just the opposite, but to a smaller extent, was the case with the estimates from the ssAPYGBLUP approach.

The main reason for higher correlations, linear regression coefficients closer to 1 and intercept closer to 0 when analysing all the individuals in the pedigree vs. when analysing only the young genotyped animals is in the historical genetic progress. When analysing all individuals in the pedigree, genetic progress is taken into account into a much larger extent than when considering only the young genotyped animals. As selection candidates are the young genotyped individuals, it is more informative to consider only these animals when comparing different methods.

The computational time and memory requirements for the creation of $\mathbf{G}^{-1}$ were 24h 14min and 670GB, respectively, when using the ssGBLUP approach and 4h 21min and 111GB, respectively, when using the ssAPYGBLUP approach. The $\mathbf{T}_c$ matrix in the ssSVDGBLUP contained 43,917 components spread across 29 chromosomes and approximated the genotype matrix of the core individuals. Computation of $\mathbf{T}_c$ took 2h 3min and 82GB of memory.

Solving the mixed model equations across 29 single or multitrait models using the preconditioned conjugate gradient method took on average 21h 45min with the ssGBLUP approach, 2h 31min with the ssAPYGBLUP approach and 35h 6min with the ssSVDGBLUP approach. The computer memory requirements were low for all three approaches as the relationship matrices were not read into the computer memory during the iteration process.

Overall, this means that the ssAPYGBLUP approach was the fastest and used slightly more computer memory than the ssSVDGBLUP approach. On the other hand, the ssSVDGBLUP approach was slightly faster in comparison to the ssGBLUP approach and used around eight times less memory for the preprocessing of the relationship matrices than the ssGBLUP approach.

Conclusions

The two analysed approximate single step genomic prediction methods showed to be good alternatives to the exact single step genomic prediction method currently used in the Geno breeding program. Further validation studies are required to analyse if the bias observed in the young genotyped individuals is confirmed after animals are phenotyped. However, there are also other approximate single step genomic prediction approaches that need to be tested.

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

Christensen, O.F., and Lund, M.S., 2010. Genomic prediction when some animals are not genotyped. Genetics Selection Evolution. 42, 2.

Misztal, I., Legarra, A., and Aguilar, I., 2014. Using recursion to compute the inverse of the genomic relationship matrix. Journal of Dairy Science 97, 3943–3952.

Nordbø, Ø., Gjuvsland, A.B., Eikje, L.S., and Meuwissen, T., 2019. Level-biases in estimated breeding values due to the use of different SNP panels over time in ssGBLUP. Genet Sel Evol 51, 76.

Ødegård, J., Indahl, U., Strandén, I., and Meuwissen, T., 2018. Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genet Sel Evol 50, 6.

VanRaden, P.M., 2008. Efficient Methods to Compute Genomic Predictions. J. Dairy Sci. 91, 4414–4423.