Comparing methods for approximating reliabilities in large-scale single-step genomic evaluations

H. Gao[1], I. Strandén[1] and Z. Liu[2]

[1] Natural Resources Institute Finland (Luke), 31600 Jokioinen, Finland

[2] IT Solutions for Animal Production (vit), Heinrich-Schröder-Weg 1, D-27283 Verden, Germany

Corresponding author: hongding.gao@luke.fi

Abstract

Accurate approximation of genomic estimated breeding value (GEBV) reliabilities is vital in single-step genomic prediction as reliable predictions of GEBV facilitate effective selection decisions. However, calculating exact reliabilities by inverting the left-hand side matrix of the mixed model equations is computationally infeasible for large datasets. In this study, we compared two approaches from Luke and Interbull for approximating genomic reliabilities for both genotyped and non-genotyped animals. The Luke approach uses effective record contributions (ERC) derived from the conventional EBV reliabilities as weights to approximate GEBV reliabilities for genotyped animals. A blended approach is used to implicitly account for residual polygenic (RPG) effects. Subsequently, genomic information is propagated to non-genotyped animals using ERC weights derived from the reliabilities of the genotyped animals. In contrast, the Interbull approach requires the derivation of a constant parameter, denoted φ_c, which is the genomic effective daughter contribution (EDC) gain via the Interbull GEBV test. This parameter is used to propagate genomic information to non-genotyped relatives through the pedigree. The final genomic reliabilities are obtained by combining conventional reliabilities with the genomic reliability gain. Notably, accuracy of reliabilities by this method highly depends on the precise estimation and regular updating of φ_c. In addition, this approach requires validation-based adjustments to correct inflated theoretical reliabilities observed in extremely large reference populations. In this study, both approaches were assessed and compared against exact reliabilities using a real dataset from the Finnish Red dairy population under a single trait model. The results demonstrated that the approximated reliabilities from both approaches were in close agreement with the exact reliabilities. Thus, both approaches can offer effective strategies for obtaining the reliabilities of GEBV in practical large-scale single-step evaluations.

Key words: single-step model, GEBV reliability, SNPBLUP, EDC, ERC

Introduction

Single-step methods (Legarra et al., 2009, Christensen and Lund, 2010) allow computing genomic estimated breeding values (GEBV) for both the genotyped and non-genotyped individuals simultaneously. Their adoption in routine genetic evaluations has become increasingly widespread in dairy cattle breeding. Consequently, the accurate computation of GEBV reliabilities has gained importance for supporting effective selection decisions. However, computing exact reliabilities by inverting the left-hand side of the mixed model equations (MME) becomes computationally infeasible for large-scale datasets. Thus, efficient approximation methods are needed.

Several methods for approximating the reliabilities of GEBV have been proposed and implemented (Misztal et al., 2013, Edel et al., 2019, Ben Zaabza et al., 2022, Bermann et al., 2022, Gao et al., 2023). In particular, to ensure the international comparability of national genomic reliabilities, an Interbull working group was established in 2016 to develop a

standardized procedure for estimating GEBV reliabilities in dairy cattle genetic evaluations (Liu et al., 2017). A corresponding guideline targeting large-scale genotyped populations has recently been released (Liu et al., 2024).

In this study, we compared two approaches for approximating genomic reliabilities for both genotyped and non-genotyped animals. The first approach, hereafter referred to as the Luke approach, uses effective record contributions (ERC) as weights within simplified SNPBLUP and PBLUP models to approximate GEBV reliabilities (Gao et al., 2023). The second approach, hereafter referred to as the Interbull approach, combines the genomic reliability gain with the conventional EBV reliability to obtain the final GEBV reliability for all animals (Liu et al., 2024).

Materials and Methods

The Luke approach

This is a three-step approach to approximate GEBV reliabilities in a single-step model that includes a residual polygenic (RPG) effects (Gao et al., 2023).

Step 1: Compute reliabilities of direct genomic values (DGV) for the genotyped animals

A simplified single-trait weighted SNPBLUP without RPG effects was used:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Zg} + \mathbf{e} \quad (1)$$

where $\mathbf{y}$ is an $n \times 1$ vector of (pseudo) phenotypes; μ is the general mean; $\mathbf{1}$ is an $n \times 1$ vector of ones; $\mathbf{Z}$ is an $n \times m$ matrix of SNP marker covariates centered and scaled using VanRaden method 1 (VanRaden, 2008), $\mathbf{g}$ is an $m \times 1$ vector of the SNP marker effects; $\mathbf{e}$ is a vector of residuals. It is assumed that $\mathbf{g} \sim N(\mathbf{0}, \mathbf{I}_m\sigma_u^2)$, and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}_n^{-1}\sigma_e^2)$, where $\mathbf{D}_n$ is a diagonal matrix with elements d_{ii} equal to the ERC_i value for genotyped animal i, computed by reversing the method of Tier and Meyer (2004) using the conventional EBV reliabilities for the genotyped animals, and σ_u^2 and σ_e^2 are the additive genetic and the residual variances, respectively. The MME for model (1) is:

$$\begin{bmatrix} \mathbf{1}'\mathbf{D}_n\mathbf{1} & \mathbf{1}'\mathbf{D}_n\mathbf{Z} \\ \mathbf{Z}'\mathbf{D}_n\mathbf{1} & \mathbf{Z}'\mathbf{D}_n\mathbf{Z} + \lambda\mathbf{I}_m \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{D}_n\mathbf{y} \\ \mathbf{Z}'\mathbf{D}_n\mathbf{y} \end{bmatrix} \quad (2)$$

with $\lambda = \frac{\sigma_e^2}{\sigma_u^2}$. We partitioned and denoted the inverse of the LHS matrix of the MME as $\begin{bmatrix} \mathbf{C}^{\mu\mu} & \mathbf{C}^{\mu\mathbf{g}} \\ \mathbf{C}^{\mathbf{g}\mu} & \mathbf{C}^{\mathbf{gg}} \end{bmatrix}$. The reliability of DGV for genotyped animal i is $r_{g,g,i}^{2*} = 1 - \lambda \frac{\mathbf{Z}_i\mathbf{C}^{\mathbf{gg}}\mathbf{Z}_i'}{\mathbf{G}_{ii}}$, where $\mathbf{Z}_i$ represents row i in $\mathbf{Z}$, and $\mathbf{G}_{ii}$ is the diagonal element i of the genomic relationship matrix $\mathbf{G} = \mathbf{ZZ}'$.

Note that the RPG effects were not explicitly included in model (1) to preserve the dimensionality and computational advantages of SNPBLUP model, particularly in scenarios where the number of individuals (n) greatly exceeds the number of markers (m).

The RPG effects were accounted for by blending the above DGV reliabilities with the traditional EBV reliabilities:

$$r_{g,g,i}^2 = \frac{(1-\omega)\mathbf{G}_{ii}r_{DGV,i}^2 + \omega\mathbf{A}_{22_{ii}}r_{EBV,i}^2}{(1-\omega)\mathbf{G}_{ii} + \omega\mathbf{A}_{22_{ii}}} \quad (3)$$

where $\mathbf{A}_{22}$ is the submatrix of $\mathbf{A}$ corresponding to the genotyped animals, $\mathbf{A}_{22_{ii}}$ is the diagonal element i of the $\mathbf{A}_{22}$ matrix which is equal to $1+F_i$ with F_i equal to the pedigree-based inbreeding coefficient of animal i,; $r_{DGV,i}^2$ is the DGV reliability for animal i and $r_{EBV,i}^2$ is the EBV reliability for animal i. ω is the proportion of the RPG effects.

Step 2: Calculate the genomic ERC for the genotyped animals

The ERC accounting for the genomic information for all genotyped animals can be calculated as:

$$ERC_g = ERC_{conv} + \frac{1-h^2}{h^2}\left(\frac{r_{DGV}^2}{1-r_{DGV}^2} - \frac{r_{EBV}^2}{1-r_{EBV}^2}\right) \quad (4)$$

where ERC_{conv} is the conventional ERC for the genotyped animals. Note that these genomic ERC values are included as weights for the

genotyped animals when computing the GEBV reliabilities for non-genotyped animals in Step 3.

Step 3: Compute reliabilities of GEBV for the non-genotyped animals

A simplified single-trait weighted PBLUP model was used:

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{a} + \mathbf{e} \quad (5)$$

where **y** is a $p \times 1$ vector of pseudo phenotypes with p equal to the number animals in the pedigree; μ is the general mean; **1** is a $p \times 1$ vector of ones; **a** represents a $p \times 1$ vector of additive genetic effects; **e** is a vector of residuals. It is assumed that $\mathbf{a} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$ and $\mathbf{e} \sim N(\mathbf{0}, \mathbf{D}_p^{-1}\sigma_e^2)$, where **A** is the numerator relationship matrix and $\mathbf{D}_p$ is a diagonal matrix with elements of ERC from vector of $\begin{bmatrix}\mathbf{ERC}_{conv}\\ \mathbf{ERC}_g\end{bmatrix}$, and σ_u^2 and σ_e^2 are the additive genetic and residual variances, respectively.

The Interbull approach

This approach is a three-step approach which requires the Interbull GEBV test (Mäntysaari et al., 2010), thus it has been feasible for routine single-step genomic evaluation with millions of genotyped animals (Liu et al., 2024). The approach uses a parameter called genomic effective daughter contribution (EDC) gain (φ_c) for genotyped animals and the propagated EDC (φ_i^{propg}) for non-genotyped animals, to combine the genomic reliability gain with the conventional EBV reliability to obtain the final GEBV reliability.

Step 1: Calculate the genomic EDC gain (φ_c)

This step comprises five sub-steps:

1) compute the DGV reliabilities for all the genotyped animals were computed using the model (1).

2) compute theoretical gain in genomic EDC as:

$$\varphi_i = \frac{1-h^2}{h^2}\left(\frac{r_{DGV}^2}{1-r_{DGV}^2} - \frac{r_{EBV}^2}{1-r_{EBV}^2}\right) \quad (6)$$

we denoted the mean of φ_i as $\bar{\varphi}$.

3) compute φ_i^{propg} using $\bar{\varphi}$ as input to propagate the genomic information from genotyped animals to their non-genotyped relatives via pedigree (VanRaden and Wiggans, 1991, Liu et al., 2004).

4) compute the combined total theoretical EDC. For genotyped animals:

$$\varphi_i^{total} = \varphi_i^{conv} + \varphi_i \quad (7)$$

For non-genotyped animals:

$$\varphi_i^{total} = \varphi_i^{conv} + \varphi_i^{propg} \quad (8)$$

5) convert to the final theoretical GEBV reliability:

$$R_i^2 = \frac{\varphi_i^{total}}{\varphi_i^{total} + \frac{1-h^2}{h^2}} \quad (9)$$

Note that sub-steps 1 through 5 must be applied to both the full and reduced datasets.

6) compute an adjustment factor (*f*) based on the validation bulls:

$$f = \frac{E(\varphi_E)}{\bar{\varphi}_E} \quad (10)$$

where $E(\varphi_E)$ is the expected EDC value:

$$E(\varphi_E) = \frac{1-h^2}{h^2} \times \frac{E(R_E^2)}{1-E(R_E^2)} \quad (11)$$

where

$$E(R_E^2) = \overline{R_L^2} - E(\Delta R^2) \quad (12)$$

where $\overline{R_L^2}$ is the mean reliability of GEBV of the validation bulls from the full dataset, $E(\Delta R^2)$ is the expected change in reliability of GEBV:

$$E(\Delta R^2) = var(\hat{u}_L - \hat{u}_E)/\sigma_u^2 \quad (13)$$

where $\hat{u}_L$ and $\hat{u}_E$ are the GEBV of the validation bulls from the evaluation using the full and reduced datasets, respectively; σ_u^2 is the additive genetic variance. $\bar{\varphi}_E$ is the theoretical EDC value of the validation bulls from the reduced dataset:

$$\bar{\varphi}_E = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-h^2}{h^2} \times \frac{R_{E_i}^2}{1-R_{E_i}^2}\right) \quad (14)$$

7) compute the adjusted genomic EDC gain (φ_i^{adj}) for all the genotyped animals with the *f* factor derived from equation (10):

$$\varphi_i^{adj} = \frac{1-h^2}{h^2}\left(\frac{r_{DGV}^2}{1-r_{DGV}^2} \times f - \frac{r_{EBV}^2}{1-r_{EBV}^2}\right) \quad (15)$$

The constant parameter of φ_c is the mean of the adjusted genomic EDC gain (φ_i^{adj}):

$$\varphi_c = \frac{1}{n}\sum_{i=1}^{n}\varphi_i^{adj} \quad (16)$$

Step 2: Propagate genomic information
This step is the same as sub-step 3) above to obtain φ_i^{propg} for the non-genotyped animals but using φ_c as the input data.

Step 3: Compute the final reliability of GEBV for all animals
For genotyped animals:

$$\varphi_i^{total} = \varphi_i^{conv} + \varphi_c \quad (17)$$

For non-genotyped animals, use the equation (8). The final reliability of GEBV for all the animals can be calculated via equation (9).

Data

To evaluate the approaches, a dataset comprising 47,124 Finnish Red dairy cows with 305-day milk yield records from first lactation was used. The analyses included 19,757 genotyped animals with 46,914 SNPs, and the pedigree encompassed 64,808 animals. The heritability of the trait was set to 0.44, and the proportion of RPG effects was assumed to be 0.30.

Results & Discussion

Reliabilities of the genotyped animals

The mean (SD) reliability of GEBV were 0.66 (0.09), 0.66 (0.09), 0.57 (0.10) from the exact, Luke, and Interbull approach for the genotyped animals, respectively. Figure 1 shows the GEBV reliabilities from the exact method versus those from the Luke method (left panel) and the Interbull method (right panel). Overall, the correlations between Luke/Interbull and exact method were close to one.

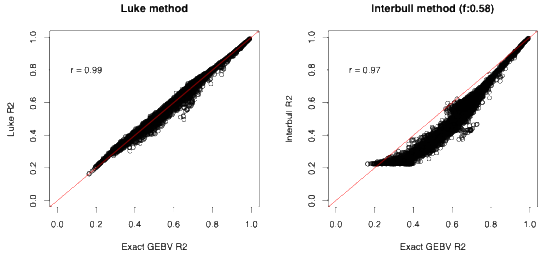

Figure 1. Scatter plot and Pearson's correlation coefficients (r) of the reliabilities of genomic estimated breeding values (GEBV) for genotyped animals via the Luke method (y-axis) versus the exact method (x-axis) (left panel) and via the Interbull method (y-axis) versus the exact method (x-axis) (right panel). The solid red line acts as a reference line with intercept 0 and slope 1

Reliabilities of the non-genotyped animals

The mean (SD) GEBV reliabilities for non-genotyped animals were 0.48 (0.17), 0.44 (0.15), and 0.43 (0.17) using the exact, Luke, and Interbull approach, respectively. Figure 2 presents the GEBV reliabilities from the exact method against those from the Luke approach (left panel) and the Interbull approach (right panel). While the correlations between the Luke/Interbull and exact approaches were slightly lower than those observed for genotyped animals, they remained high overall.

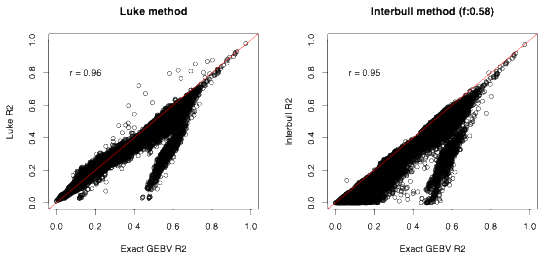

Figure 2. Scatter plot and Pearson's correlation coefficients (r) of the reliabilities of genomic estimated breeding values (GEBV) for non-genotyped animals via the Luke method (y-axis) versus the exact method (x-axis) (left panel) and via the Interbull method (y-axis) versus the exact method (x-axis) (right panel). The solid red line acts as a reference line with intercept 0 and slope 1

In this study, we compared the Luke and the Interbull approaches for approximating GEBV reliabilities. Both approaches computed GEBV reliabilities separately for genotyped and non-genotyped animals and required conventional EBV reliabilities for all animals in the pedigree.

The Luke approach used the information from a PBLUP model to derive ERC which served as weights in a SNPBLUP model that incorporates genomic information when computing GEBV reliabilities for genotyped animals. Similarly, for non-genotyped animals, the genomic information was included indirectly by applying additional weights derived from the genotyped animals within a weighted PBLUP model. An important feature

of this approach is that the models for computing GEBV reliabilities include only a general mean and genetic effects, while the weighting scheme and relationship structure differ between genotyped and non-genotyped groups.

The Interbull approach employed a constant parameter (φ_c) derived from the Interbull GEBV test, to simplify computations in large-scale genotyped populations. The φ_c was propagated to the non-genotyped relatives via pedigree to obtain their respective propagated EDC gain (φ_i^{propg}). The final EDC values were then calculated by combining the conventional EDC with φ_c for genotyped animals and φ_i^{propg} for non-genotyped animals. GEBV reliabilities were subsequently derived from the total EDC using equation (9).

The results showed that the approximated GEBV reliabilities from both approaches were in close agreement with the exact values, supporting their applicability in practical genetic evaluations.

It is important to note that a key feature of the Interbull approach is the derivation and use of the genomic EDC gain parameter (φ_c), which can be repeatedly applied to approximate GEBV reliabilities. However, because φ_c is directly linked to the Interbull GEBV Test, it must be re-estimated and updated each time a new GEBV test is conducted. This feature offers the computational simplicity and efficiency. In contrast, the Luke approach requires precise calculation of ERC weights for each computation of reliabilities, which may increase computational demands.

The RPG effects need to be considered to avoid overestimating the reliability of GEBV; however, these effects were not explicitly incorporated in either approach. The Luke approach employed a blended method to approximate GEBV reliabilities for genotyped animals, thereby retaining the primary advantage of the SNPBLUP model, that is, even as the number of genotyped animals increases, the dimensionality of the coefficient matrix of the MME remains fixed, depending solely on the number of SNPs. The Interbull approach implemented an adjustment factor (*f*) to scale down the theoretical GEBV reliabilities to ensure an appropriate genomic reliability level for young selection candidates.

This study used a relatively small dataset to ensure the feasibility of computing the exact GEBV reliabilities by directly inverting the coefficient matrix of the MME. However, routine single-step genomic evaluations in practice often involve millions of genotyped animals, thus, a larger and more representative datasets might be more appropriate to further evaluate these approaches.

Conclusions

This study compared two approaches for approximating genomic reliabilities for both genotyped and non-genotyped animals. The results demonstrated that both approaches produced reliability estimates in close agreement with the exact reliabilities computed using the full dataset in a ssGBLUP evaluation. Importantly, both methods indirectly accounted for residual polygenic (RPG) effects without explicitly including them in the model. Although the Interbull method relies on the Interbull GEBV test, both approaches offer effective strategies for obtaining GEBV reliabilities in practical large-scale single-step evaluations.

Acknowledgments

The authors are grateful to Viking Genetics (Randers, Denmark) and Nordic Cattle Genetic Evaluation (Aarhus, Denmark) for providing the data.

References

Ben Zaabza, H., M. Taskinen, E. A. Mantysaari, T. Pitkanen, G. P. Aamand, and I. Stranden. 2022. Breeding value reliabilities for multiple-trait single-step genomic best linear unbiased predictor. J Dairy Sci. 105(6):5221-5237.

Bermann, M., D. Lourenco, and I. Misztal. 2022. Efficient approximation of reliabilities for single-step genomic best linear unbiased predictor models with the Algorithm for Proven and Young. J Anim Sci. 100(1).

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 42:2.

Edel, C., E. C. G. Pimentel, M. Erbe, R. Emmerling, and K. U. Gotz. 2019. Short communication: Calculating analytical reliabilities for single-step predictions. J Dairy Sci. 102(4):3259-3265.

Gao, H., A. A. Kudinov, M. Taskinen, T. J. Pitkanen, M. H. Lidauer, E. A. Mantysaari, and I. Stranden. 2023. A computationally efficient method for approximating reliabilities in large-scale single-step genomic prediction. Genet Sel Evol. 55(1):1.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 92(9):4656-4663.

Liu, Z., F. Reinhardt, A. Bunger, and R. Reents. 2004. Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. J Dairy Sci. 87(6):1896-1907.

Liu, Z., I. Strandén, J. Vandenplas, H. Eding, M. Lidauer, K. Haugaard, and P. M. VanRaden. 2024. Guidelines for Approximating Genomic Reliabilities of the Single-Step Genomic Model. Pages 148-160 in Proc. Interbull Bull., Bled, Slovenia.

Ben Zaabza, H., M. Taskinen, E. A. Mantysaari, T. Pitkanen, G. P. Aamand, and I. Stranden. 2022. Breeding value reliabilities for multiple-trait single-step genomic best linear unbiased predictor. J Dairy Sci. 105(6):5221-5237.

Bermann, M., D. Lourenco, and I. Misztal. 2022. Efficient approximation of reliabilities for single-step genomic best linear unbiased predictor models with the Algorithm for Proven and Young. J Anim Sci. 100(1).

Christensen, O. F. and M. S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet Sel Evol. 42:2.

Edel, C., E. C. G. Pimentel, M. Erbe, R. Emmerling, and K. U. Gotz. 2019. Short communication: Calculating analytical reliabilities for single-step predictions. J Dairy Sci. 102(4):3259-3265.

Gao, H., A. A. Kudinov, M. Taskinen, T. J. Pitkanen, M. H. Lidauer, E. A. Mantysaari, and I. Stranden. 2023. A computationally efficient method for approximating reliabilities in large-scale single-step genomic prediction. Genet Sel Evol. 55(1):1.

Legarra, A., I. Aguilar, and I. Misztal. 2009. A relationship matrix including full pedigree and genomic information. J Dairy Sci. 92(9):4656-4663.

Liu, Z., F. Reinhardt, A. Bunger, and R. Reents. 2004. Derivation and calculation of approximate reliabilities and daughter yield-deviations of a random regression test-day model for genetic evaluation of dairy cattle. J Dairy Sci. 87(6):1896-1907.

Liu, Z., I. Strandén, J. Vandenplas, H. Eding, M. Lidauer, K. Haugaard, and P. M. VanRaden. 2024. Guidelines for Approximating Genomic Reliabilities of the Single-Step Genomic Model. Pages 148-160 in Proc. Interbull Bull., Bled, Slovenia.

Liu, Z., P. M. VanRaden, M. H. Lidauer, M. P. Calus, H. Benhajali, H. Jorjani, and V. Ducrocq. 2017. Approximating Genomic Reliabilities for National Genomic Evaluation. Pages 75-85 in Proc. Interbull Bull., Tallinn, Estonia.

Mäntysaari, E., Z. Liu, and P. VanRaden. 2010. Interbull validation test for genomic evaluations. Pages 17-17 in Proc. Interbull bull., Paris, France.

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013. Methods to approximate reliabilities in single-step genomic evaluation. J Dairy Sci. 96(1):647-654.

Tier, B. and K. Meyer. 2004. Approximating prediction error covariances among additive genetic effects within animals in multiple-trait and random regression models. J Anim Breed Genet. 121(2):77-89.

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. J Dairy Sci. 91(11):4414-4423.

VanRaden, P. M. and G. R. Wiggans. 1991. Derivation, calculation, and use of national animal model information. J Dairy Sci. 74(8):2737-2746.