

Challenges in the practical implementation of an interim genomic prediction for young animals in Wallonia, Belgium

K. Wijnrocx^{1*}, R.R. Mota^{1,2}, M. Dufrasne³ and N. Gengler¹

¹ University of Liège - Gembloux Agro-Bio Tech, Passage des Déportés, 2, 5030 - Gembloux, Belgium; ² Council on Dairy Cattle Breeding, Northview Dr 4201, Bowie, MD 20716, USA; ³ Walloon Breeders Association, rue des Champs Elysées 4, 5590 Ciney, Belgium;

*katrien.wijnrocx@uliege.be

Abstract

An official Walloon genomic evaluation for the Holstein population is performed three times a year for several traits, using a single-step GBLUP model, with a “Bayesian” integration comprising both local and foreign breeding values. The need for a system for preliminary evaluation for young animals was expressed by the Walloon breeders. This article gives an overview of methods developed and describes the challenges encountered in the practical implementation, and proposed improvements. Experience identified that a crucial step for the correct calculation of interim values is the estimation of the polygenic term, which requires a deep pedigree knowledge, in order to assure a correct implementation.

Introduction

Currently, official genomic evaluations for the Holstein population in the Southern part of Belgium (Walloon region) are performed for production, conformation, udder health and functional traits since 2015. Moreover, nearly all traits are submitted three times a year to the Interbull Centre (<https://interbull.org/index>, Uppsala, Sweden), which generates a Genomic Multiple Across Country Evaluation (GMACE) with local genomically enhanced breeding values (GEBVs). As outlined by Vandenplas et al. (2014) and Colinet et al. (2018), the Walloon local genomic evaluations are based on a close approximation to single-step GBLUP model (ssGBLUP) using a “Bayesian” integration of local Walloon breeding values, by adding MACE values and, to avoid double-counting, a subtraction of the Walloon contribution from MACE. However, this can generate a delay up to 4 months between the genotyping of an animal, and its official GEBV evaluation. Thus, a system of preliminary genomic evaluations for young animals is of interest for the breeders, to be able make selection decisions in the shortest time possible. Various authors have proposed different solutions to the problem of interim predictions. A study of Lourenco et al. (2015) described an indirect method for the prediction of the GEBV of young animals by using an ssGBLUP analyses and calculation of their initial GEBV through the weighted combination of their direct genomic value (DGV) and parent average (PA), which they adapted by including an algorithm and SNP effects blending and tuning (Lourenco et al., 2018a). This approach however still generated some problems when animals have their own records, and the determination of the weighing factor.

An alternative approach was proposed by Pimentel et al. (2019), in which the GEBVs are directly split into a SNP based part and a polygenic part. Both parts were estimated separately, and then combined afterwards. The SNP part was estimated by using a “SNP model” on the residual GEBVs of the reference animals of the ssGBLUP. These SNP effects were used subsequently to compute DGV, and the “residual” term was the polygenic part. This approach however had some problems to transmit the polygenic term from reference animals to young animals especially when the distance between genotyped animals and candidates is not very close, i.e., direct ancestors are not genotyped. Therefore, Mota et al. (2019) proposed an alternative where he combined the best of both methods to specifically fit to the Walloon

Holstein population to obtain an interim prediction for young animals. In his study, the new method was validated, and a test implementation was done by the Walloon Breeders Association (AWE). However, some inconsistencies arose, resulting in initially low correlations between the predicted interim values, and the calculated values. This research article will through the description of the method by Mota et al. (2019), the implementation by the Walloon Breeders Association and the implemented mitigation strategies show the challenges in this practical implementation.

Materials & Methods

Partitioning of the GEBVs and generation of polygenic terms. The modified approach that was proposed by Mota et al. (2019) followed the derivation of ssGBLUP by Christensen and Lund (2010) to split GEBV into genomic (DGV) and polygenic (PT) parts as described previously by Pimentel et al. (2019). But, in contrast to Pimentel et al. (2019), PT were extracted at first, and only then DGV were calculated. This step can be derived as follows. If we call $\hat{\mathbf{u}}$ the vector of available EBV respectively GEBV each representing the full genetic variance even for the non-genotyped animals $\hat{\mathbf{a}} = \mathbf{Cov}(\mathbf{a}, \mathbf{u})(\mathbf{Var}(\mathbf{u}))^{-1}(\hat{\mathbf{u}} - \mathbf{1}\hat{m})$ can be considered an estimation of the PT of $\hat{\mathbf{u}}$ and $\hat{m}$ being a constant. The matrix $(\mathbf{Var}(\mathbf{u}))^{-1}$ should be equal to $\mathbf{H}^{-1}$ as defined by Legarra et al. (2009) and Christensen and Lund (2010) and the covariance $\mathbf{Cov}(\mathbf{a}, \mathbf{u})$ needs to express the variance ratio between the variance explained by PT, here after called σ_a^2 and the total genetic variance σ_u^2 .

An alternative, approach would be to write this hypothetical mixed model $\hat{\mathbf{u}} = \mathbf{1}\hat{m} + \mathbf{Z}\hat{\mathbf{a}} + \mathbf{e}$. The associated theoretical correct decomposition of the total (co)variances would be $\mathbf{Var}(\mathbf{u}) = \mathbf{Z}\mathbf{Var}(\mathbf{a})\mathbf{Z}' + \mathbf{Var}(\mathbf{e}) = \mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_a^2 + \mathbf{G}\sigma_e^2$. Therefore, the model would be:

$$\begin{bmatrix} \mathbf{1}'\mathbf{G}^{-1}\mathbf{1} & \mathbf{1}'\mathbf{G}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{G}^{-1}\mathbf{1} & \mathbf{Z}'\mathbf{G}^{-1}\mathbf{Z} + \mathbf{A}^{-1}k \end{bmatrix} \begin{bmatrix} \hat{m} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{G}^{-1}\hat{\mathbf{u}} \\ \mathbf{Z}'\mathbf{G}^{-1}\hat{\mathbf{u}} \end{bmatrix} \text{ with } k = \sigma_e^2 / \sigma_a^2 \quad (1)$$

and perfectly equivalent to:

$$\hat{\mathbf{a}} = \sigma_a^2 \mathbf{A}\mathbf{Z}'(\mathbf{Z}\mathbf{A}\mathbf{Z}'\sigma_a^2 + \mathbf{G}\sigma_e^2)^{-1}(\hat{\mathbf{u}} - \mathbf{1}\hat{m}) = \mathbf{A}\mathbf{Z}'(\mathbf{Z}\mathbf{A}\mathbf{Z}' + \mathbf{G}k)^{-1}(\hat{\mathbf{u}} - \mathbf{1}\hat{m}) \text{ with } k = \sigma_e^2 / \sigma_a^2 \quad (2)$$

$$\text{where } \hat{m} = (\mathbf{1}'(\mathbf{Z}\mathbf{A}\mathbf{Z}' + \mathbf{G}k)^{-1}\mathbf{1})^{-1}(\mathbf{1}'(\mathbf{Z}\mathbf{A}\mathbf{Z}' + \mathbf{G}k)^{-1}\hat{\mathbf{u}}) \quad (3)$$

In these equations, $\mathbf{1}$ is a vector of ones, $\mathbf{A}$ is the pedigree-based relationship matrix among known animals, $\mathbf{G}$ represents the hypothetical genomic relationships between all animals; $\hat{m}$ is BLUE estimate of the overall mean; $\hat{\mathbf{a}}$ is the vector of the PT values for all animal, which are to be estimated; σ_e^2 is the residual variance (which is the genetic variance associated with the SNP effects (DGV)); σ_a^2 is the genetic variance (explained by the pedigree and therefore due to PT); $\mathbf{Z}$ is the incidence matrix linking $\hat{\mathbf{a}}$ to $\hat{\mathbf{u}}$. The overall GEBV mean was added to the model to avoid a mean shift in the indirect predictions (Lourenco et al., 2018b). It can be shown that these mixed model equations are a close approximation of the theoretically correct equations deriving polygenic contributions to GEBV. The proportion of additive genetic variance explained by SNPs was defined by our routine genomic evaluations, i.e., currently 60% for production and 65% for other traits, therefore $k = \sigma_e^2 / \sigma_a^2 = 0.60\sigma_u^2 / 0.40\sigma_u^2$ or $0.65\sigma_u^2 / 0.35\sigma_u^2$.

Solving (1) is not trivial therefore we solve these alternative equations:

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k \end{bmatrix} \begin{bmatrix} \hat{m} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\hat{\mathbf{u}} \\ \mathbf{Z}'\hat{\mathbf{u}} \end{bmatrix} \text{ with } k = \sigma_e^2 / \sigma_a^2 \quad (4)$$

where $\mathbf{G}$ was considered equal to an identity matrix and therefore all links between $\hat{\mathbf{u}}$ considered pedigree based and $\hat{\mathbf{a}}$ estimated in an equivalent approach as if this penalized regression $\hat{\mathbf{a}} = \mathbf{AZ}'(\mathbf{ZAZ}' + \mathbf{I}k)^{-1}(\hat{\mathbf{u}} - \mathbf{1}\hat{m})$ was used. It is the use of this approach that allowed the combination of both EBV and GEBV in $\hat{\mathbf{u}}$. After solving this model using all available solutions (i.e., for genotyped and non-genotyped animals) from the ssGBLUP system, the vector $\hat{\mathbf{e}}$ for genotyped animals (i.e., reference animals) is considered equal to the vector of DGV as follows:

$$\mathbf{DGV} = \hat{\mathbf{e}} = \hat{\mathbf{u}} - (\mathbf{1}\hat{m} + \mathbf{Z}\hat{\mathbf{a}}) \quad (5)$$

In this adapted procedure, the generation of PT could be done for newly genotyped animals mixing their EBV with EBV/GEBV from relevant ancestors. As of such, GEBV a posteriori combining PT with SNP effects can be generated, which will potentially also include own records of non-reference animals.

Generation of SNP effects. The second step of the adapted method proposed by Mota et al. (2019) was the estimation of the SNP effects. This step applied the proposal of Lourenco et al. (2015), based on selection index theory, and reduced GEBV into DGV:

$$\hat{\mathbf{g}} = \mathbf{DM}'\mathbf{G}^{-1}(\mathbf{DGV}) \quad (6)$$

where $\hat{\mathbf{g}}$ is a vector for the SNP effects, $\mathbf{D}$ is a diagonal matrix of standard variance weights for SNPs (identity in this study), and $\mathbf{M}$ is a centered genotypes matrix (VanRaden, 2008), $\mathbf{G}^{-1}$ is the inverse genomic relationship matrix ($\mathbf{MDM}'$) between reference animals, and $\mathbf{DGV}$ is a vector of reference population DGV values.

Implementation by the Walloon Breeders Association. Practical implementation of the interim genomic predictions for young animals was performed by AWE using a simplified, approximative procedure, since new genotyped animals are constantly added to the population and performing a full single-step is not feasible on a weekly or monthly basis. AWE prepared the genotype file of these (new) animals, by imputing the 10k genotypes to common SNP between 50k V1 and 50k V2. In the first implement the reduced polygenic term (PT) was obtained using Equation (4) on an extracted pedigree. The interim GEBV for a young animal i can thus be calculated for each trait of interest as follows:

$$\text{Interim GEBV}_i = \text{mean} + \text{DGV}_i + \text{PT}_i = \hat{m} + \hat{a}_i + \mathbf{m}_i\hat{\mathbf{g}} \quad (7)$$

It should be mentioned that $\hat{\mathbf{g}}$ is calculated in the Walloon genomic evaluation centre and transferred to AWE, whereas $\hat{m}$ and $\hat{\mathbf{a}}$ are calculated by AWE for the set of animals of interest; The vector of $\mathbf{m}_i$ representing the centered genotypes for animal i are computed by AWE using the allele frequencies used to generate $\hat{\mathbf{g}}$.

Validation and mitigation. To validate a correct implementation, Pearson correlations were calculated by AWE for a set of 150 Holstein animals without own records in the previous official Walloon evaluation and in which a comparison between calculated interim GEBV and the published GEBV from the current evaluation were made. Importance of deepness of extracted pedigree was identified by recognizing initial limitations to 3 generations as too shallows and mitigated by full extraction as used in generation of SNP effects.

Results

Results of the Pearson correlations for the initial and improved system are shown for three problematic traits (loin strength, udder balance, overall rump) and for milk yield in Table 1.

Correlations for problematic traits were around 0.72, but after correction increased significantly to 0.84-0.87.

Table 1. Pearson correlations between interim and published GEBVs for a subset of 150 young animals (AWE) before and after correction for pedigree depth

Trait	Before correction	After correction
Loin strength	0.72	0.87
Udder balance	0.72	0.84
Overall rump score	0.72	0.87
Milk yield	0.93	0.97

Discussion

This study demonstrated the challenges that arose when practically implementing interim genomic breeding values by AWE using a simplified system. For breeders, this kind of system is of great interest to allow early selection decisions. Our method, after mitigation, correctly implements genetic groups, a problem that was documented in literature (Lourenco et al., 2015; Pimentel et al. 2019), and this as all animals in the pedigree are considered for the interim estimation. However, implementation demonstrated suboptimal results when using shallow pedigree extraction. Improvements increased the correctness of the implementation. This demonstrates the importance of the correct calculation of the PT term. Indeed, crucial for a correct implementation is a compatible $\hat{m}$ and $\hat{a}$, as calculated by the Walloon genomic evaluation centre (Equation 4) on the full set of animals, and $\hat{m}$ and $\hat{a}$ calculated and used by AWE (Equation 7). Correct calculation of $\mathbf{m}_i$ representing the centred genotypes for animal i was ensured by providing the calculated allelic frequencies to AWE.

This research demonstrated that interim GEBVs can be applied successfully after mitigation as some adjustments were needed for the practical implementation by the Walloon Breeders Association, to ensure the correct approximation of the PT part needed to calculate interim GEBVs. However, more research is necessary to further improve the system.

References

- Christensen, O.F., and Lund. M.S. (2010). Genet. Sel. Evol. 42:1. <https://doi.org/10.1186/s12711-014-0052-x>
- Colinet, F.G., Vandenplas, J., Vanderick, S., Hammami, H., Mota, R.R. *et al.* (2018). Animal 12:898–905. <https://doi.org/10.1017/S1751731117002324>
- Legarra, A.; Aguilar, I.; Misztal, I. (2009). J. Dairy Sci. 92: 4656-4663. <https://doi.org/10.3168/jds.2009-2061>.
- Lourenco, D.A.L., Legarra, A., Tsuruta, S., Moser D., S. Miller, *et al.* (2018a). Interbull Bul. 53:48–53. <https://journal.interbull.org/index.php/ib/article/view/1448>
- Lourenco, D.A.L., Tsuruta, S.; Fragomeni, B.O.; Masuda, Y.; Aguilar I. *et al.* (2015). J. Anim. Sci. 93:2653–2662. <https://doi.org/10.2527/jas.2014-8836>
- Lourenco, D.A.L., Tsuruta, S.; Fragomeni, B.O.; Masuda, Y.; Aguilar I. *et al.* (2018b). Proc. of the 10th WCGALP, Vancouver, Canada.
- Mota, R.R., Nader, S.; Vanderick, S.; Colinet, F.G., Gillon, A. *et al.* (2019). Interbull Bul. 55:64–69. <https://journal.interbull.org/index.php/ib/article/view/1472/1586>.
- Pimentel, E.C.G., Edel, C., Emmerling, R. and Götz, K.U. (2019). J. Dairy Sci. 1–8. doi:10.3168/jds.2018-15592. <https://doi.org/10.3168/jds.2018-15592>
- Vandenplas, J., Colinet, F.G. and Gengler, N. (2014). Genet. Sel. Evol. 46:1–15. <https://doi.org/10.1186/s12711-014-0059-3>
- VanRaden, P.M. (2008). J. Dairy Sci. 91:4414–4423. <https://doi.org/10.3168/jds.2007-0980>