Strategy to stabilize genomic breeding values under an evolving sire and cow reference population in the single-step evaluation system of Walloon region of Belgium

***R.R. Mota*[1]*, S. Nader*[2]*, S. Vanderick*[1]*, F.G. Colinet*[1]*, A. Gillon*[3]*, P. Mayeres*[3]*, N. Gengler*[1]**

[1] *University of Liège – Gembloux Agro-Bio Tech, 5030 Gembloux, Belgium*

[2] *Irish Cattle Breeding Federation - ICBF, Highfield House, Shinagh, Bandon, Co. Cork, Ireland*

[3] *Walloon Breeding Association, 5190 Ciney, Belgium*

Abstract

The national genomic evaluations for production, conformation, udder health and functional traits in Wallonia are official since 2015. Nearly all evaluated traits are submitted to Interbull three times a year which give gaps of four months between official genomic estimated breeding values (**GEBV**). Generating reliable GEBV is a major challenge. With our small population, changes in the reference population size can be extremely important. Currently, approximately 12 000 Holstein genotypes are available, with 8 500 being actually used in the evaluations. However, through projects and intensive testing an increase of at least 20% per year is expected. The question is now how can we stabilize GEBV when the reference population is constantly moving? This research here is associated to the derivation of an interim computational method intended to help breeders to make early decisions. This implementation consisted in: GEBV partition into polygenic (**PT**) and direct genomic (**DGV**) values; SNP effect estimation from DGV; GEBV prediction for new animals by combining DGV generated from SNP effects and PT. We also investigated the hypothesis that a core group of animals would be sufficient to estimate GEBV for other non-reference animals. To test this, a list of 648 genotyped animals having official GEBV from the last run was used as validation. Interim GEBVs were generated (by summing up PT, DGV and mean trait), and correlated with their official GEBV. Correlations between official and interim GEBVs were 0.92, 0.93, 0.93, 0.93, 0.94, 0.91, 0.91, 0.94 and 0.94 for milk yield, fat yield, fat percentage, protein yield, protein percentage, somatic cell score, longevity, direct calving ease and maternal calving ease, respectively. Relative mean differences were up to 6%. These results are a first indication that we could develop a stable reference population, generate high quality SNP effects and generate appropriate GEBV reflecting potentially own records for non-reference population animals. The last point is joint with efforts to generate appropriate reliabilities based on the approach promoted by Interbull.

Key words: early decisions, GEBV partition, indirect predictions, young animals

Introduction

The genomic evaluations for production, conformation, udder health and functional traits in the Walloon region of Belgium are official since 2015. Moreover, nearly all traits nationally evaluated are submitted to Interbull Centre (https://interbull.org/index, Uppsala, Sweden), which allows us to participate in the Genomic Multiple Across Country Evaluation (**GMACE**) with local genomically enhanced breeding value (**GEBV**). Our local single step GBLUP (**ssGBLUP**) model is using a "Bayesian" integration of local breeding values, by adding MACE values and subtracting Walloon contribution to MACE to avoid double-counting (Vandenplas et al., 2014; Colinet et al., 2018). These evaluations are submitted three times a year to Interbull (https://interbull.org/ib/servicecalendar). This gives us gaps of four months between official genomic breeding values. For example, an animal genotyped a week after the evaluation deadline has to wait for at least seven weeks to have a GEBV.

Generating reliable GEBV is a major challenge. With our rather small population, changes in the size of the reference population can be, relatively spoken, extremely important. Currently (April 2019), approximately 12 000 Holstein genotypes are available, with 8 500 used in the evaluations, but through projects and intensive testing an increase of at least 20% per year is expected. From that, two main questions have then risen up: 1) How can we provide GEBV more often to help breeders in early decisions? 2) How can we stabilize these estimated GEBV when the reference population is constantly moving? This is an issue that we think to be very important in our type of ssGBLUP implementation due to its complexity (Aguilar et al., 2010; Colinet et al., 2018). Fortunately, given the recent massive development of ssGBLUP methods in many circumstances, this problem of interim predictions is not ignored. Lourenco et al. (2015) proposed an ssGBLUP exploiting indirect prediction for young animals in order to calculate initial GEBV for young animals through weighted combination of their direct genomic values (**DGV**) and parent average (**PA**). Although these authors showed correlations between indirect predictions and GEBV higher than 0.99, issues such as computational costs and average differences, which make values not comparable, have emerged. They later presented a solution for both situations, based on a core algorithm and SNP effects blending and tuning, respectively (Lourenco et al., 2018a). Unfortunately, it remained rather problematic especially considering the values of the weights and when newly genotyped animals have their own records.

More recently, Pimentel et al. (2019) proposed an alternative approach that directly decomposed the GEBV into SNP based and polygenic parts. They developed strategies to estimate those parts separately and to combine them later. The SNP part was obtained by using a "SNP model" on GEBV of reference animals from the ssGBLUP run, but discounting for the fact that markers only explain a part of the genetic variance. With these SNP effects they computed DGV, here expressed as genomic part of GEBV. The difference between GEBV and DGV was called by these authors as polygenic term (**PT**). In their approach Pimentel et al. (2019) computed these PT for reference animals but struggled to develop a strategy to propagate the polygenic term from reference to young animals.

This research is associated to the derivation of an interim computational method to help breeders to make early decisions. Our implementation consisted in to take the best of both aforementioned ideas to develop and test the implementation of a novel indirect prediction algorithm in the context of the Walloon genomic evaluation system. We also investigated the hypothesis that a core group of animals would be enough to estimate GEBV for non-reference animals.

Materials and Methods

GEBV partition and generation of polygenic terms for animals not in the ssGBLUP

We followed the derivation of ssGBLUP by Christensen and Lund (2010) to split GEBV into genomic (DGV) and polygenic (PT) parts as proposed by Pimentel et al. (2019). These ideas are also in line with alternative ssGBLUP derivations splitting genomic from pedigree contributions. We modified Pimentel et al. (2019) by first extracting PT, and then computed DGV. The PT values for reference animals were calculated by solving the following set of mixed model equations:

$$\begin{bmatrix} \mathbf{1}'\mathbf{1} & \mathbf{1}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{1} & \mathbf{Z}'\mathbf{Z}+\mathbf{A}^{-1}k \end{bmatrix} \begin{bmatrix} \hat{\mathrm{m}} \\ \hat{\mathbf{a}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\hat{\mathbf{u}} \\ \mathbf{Z}'\hat{\mathbf{u}} \end{bmatrix}, \text{ with}$$

$$k = \frac{\sigma_e^2}{\sigma_a^2},$$

in which **1** is a vector of ones; **A** is the pedigree-based relationship matrix among animals; $\hat{\mathrm{m}}$ is the estimated overall mean; **â** is the vector of to be estimated PT values; σ_e^2 is the residual variance, i.e., the variance associated to SNP effects (DGV); σ_a^2 is the additive genetic variance explained by pedigree and therefore due to PT; **Z** is the incidence matrice linking **â** to **û**. The overall GEBV mean was added to the model in order to avoid 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. Thus,

$$k = \frac{\sigma_e^2}{\sigma_a^2} = \frac{0.60\sigma_u^2}{0.40\sigma_u^2} \text{ or } k = \frac{\sigma_e^2}{\sigma_a^2} = \frac{0.65\sigma_u^2}{0.35\sigma_u^2}.$$

After solving this model using all available solutions (i.e. for genotyped and un-genotyped animals) from the ssGBLUP system, the vector $\hat{\mathbf{e}}$ for genotyped animals (i.e., reference animals) is considered equal to the vector **DGV** as follows:

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

where all terms were described above.

Traditional BLUP solving software can be used for GEBV partition. We used the blupf90 software (http://nce.ads.uga.edu/). Please note a major difference with Pimentel et al. (2019) which is that we are establishing and solving the mixed model equations for all animals involved in our ssGBLUP routine runs. This will allow the inclusion of additional animals that were genotyped outside of the routine ssGBLUP scheme. Indeed, generation of PT could be done for newly genotyped animals mixing their EBV with EBV/GEBV from relevant ancestors. This strategy can generate appropriate GEBV a posteriori combining these PT with SNP effects, reflecting potentially also own records for non-reference population animals. This later feature is possible because we are inverting the order of computational steps and we can obtain PT easily for non-genotyped supplemental animals.

Generating SNP effects and defining reference populations

The next step is the estimation of required SNP effects to further calculate DGV for young animals. In preliminary tests we found that, despite Pimentel et al. (2019) findings, the proposal by Lourenco et al. (2015), which is based on selection index theory, worked well in our setting where we reduced first GEBV into DGV. Thus, we can use the vector of DGV values as input in the software postGSf90 (http://nce.ads.uga.edu/) to estimate SNP effects as follows:

$$\hat{\mathbf{g}} = \mathbf{D}\mathbf{M}'\mathbf{G}^{-1}(\mathbf{DGV}),$$

in which $\hat{\mathbf{g}}$ is a vector of SNP effects, **D** is a diagonal matrix of standard variance weights for SNPs (identity in this study), and **M** is a centered genotypes matrix (VanRaden, 2008), $\mathbf{G}^{-1}$ is the inverse genomic relationship matrix ($\mathbf{MDM}'$) between reference animals, and **DGV** is vector of reference population DGV values. This step should allow the definition of a stable reference population by limiting voluntarily the DGV used to ones that are considered stable. This should also allow the generation of high quality SNP effects.

Validation

To validate our method, we calculate the DGV for validation animals as:

$$\mathbf{DGV}_v = \mathbf{M}_v\hat{\mathbf{g}},$$

in which $\mathbf{DGV}_v$ and $\mathbf{M}_v$ are direct genomic values and a centered genotypes matrix for those animals not included in SNP effect estimation above, respectively. Because we based our DGV on strict partitioning of GEBV, we can combine $\mathbf{DGV}_v$ and $\mathbf{PT}_v$ for validation animals to estimate their approximate GEBV simply as follows:

$$\mathbf{GEBV}_v = \mathbf{DGV}_v + \mathbf{PT}_v,$$

where $\mathbf{GEBV}_v$ is the approximate GEBV for validation animals, $\mathbf{PT}_v$ is the polygenic term of validation animals estimated in the GEBV partition section and $\mathbf{DGV}_v$ was previously described.

In order to investigate the efficiency of the method (reference population) as well as the ability to predict future GEBVs (validation population), correlations between official and interim GEBV predictions (**r_{pa}**) were calculated.

Finally, a relative mean difference between official and approximated GEBVs for both populations (reference and validation) were calculated. This was done by using the equation below:

$$\text{RMD} = \frac{(\bar{\text{X}}\text{GEBV}_{off} - \bar{\text{X}}\text{GEBV}_{int})}{\text{SD}_t}$$

in which RMD is the relative mean difference between official $\bar{\text{X}}\text{GEBV}_{off}$ and interim $\bar{\text{X}}\text{GEBV}_{int}$ genomic breeding values and SD_t is the official standard deviation SD of the trait t.

Genotypic, phenotypic and pedigree data

The implementation of indirect prediction for new genotyped animals was validated by using data from the April 2019 official run submitted to Interbull. Official GEBVs (please see https://www.elinfo.be for publication bases) were used as input values, i.e., phenotypes (n=9 378), for five production traits, i.e., milk yield (**MY**), fat yield (**FY**), protein yield (**PY**), fat percentage (**FP**) and protein percentage (**PP**), and somatic cell score (**SCS**), longevity (**LONG**), direct (**DCE**) and maternal calving ease (**MCE**), The pedigree extracted from the genotyped animals included 40 740 animals.

Currently in Wallonia, there are 12 399 animals combined in a list of 32 867 SNPs. However, samples with call rates <0.95, minor allele frequencies (MAF) <0.05, or those with highly significant deviations from Hardy-Weinberg equilibrium ($P<10^{-7}$) were removed. At the same time, old genotyped animals, i.e., animals born more than 15 years ago, and animals not sufficiently Holsteins (i.e., bulls not having at least 87.5% Holstein and cows without three generations of sires having at least 87.5% Holstein) were also not considered in evaluations. After usual edits on SNPs as well as on animals, 8 506 (68.6%) animals (3 119 bulls and 5 387 cows) and 30 290 (92.2%) markers remained for further analyses. The genotyped population was then divided into reference (n=7 858) and validation (n=648).

Results & Discussion

Correlations between the official and interim GEBVs were always higher than 0.99 for reference, and 0.91 for validation populations (Table 1). This leads us to believe that, as soon as a threshold number of reference animals is achieved, the composition of the reference population at each routine run seems to have major influence rather than the size of the reference population (Lourenco et al., 2015). In summary, our results endorse that GEBV can be efficiently approximated through indirect predictions even with a non-large genotyped reference population.

No important GEBV mean shift was observed (Table 1). We think that this was because our method followed the idea of Pimentel et al. (2019), in which the mean was simultaneously estimated by solving the mixed model equations. In order to provide more informative values, the relative RMD between official and interim GEBVs for both populations have been checked. In general, RMD were higher for validation (range 0.00%-3.90%) vs. (range 0.01%-5.90%) animals (Table 1). Higher RMD for validation animals were expected due to its size (n=648), and by the fact that these genotypes were not accounted in the ssGBLUP relationship matrix to generate SNP effects. As reported by Shabalina et al. (2017), one single animal included in the system modifies the aforementioned matrix, and is able to influence not only its own prediction but of other animals. Fortunately, the validation animals relatively small variance testifies that this method works even for a single animal.

Moreover, few animals, in both populations had high GEBV values variance (results not shown). In general, reference animals with higher differences tended to be foreign bulls with smaller genetic links to our population. On the other hand, for validation animals no clear pattern from the sire line was observed: daughters of the same genotyped bull, i.e., well transmitted information to young animals via SNP effects, and presented high or even null differences. However, as stressed by Pimentel et al. (2019) well transmitted contributions depends upon dams as well. According to these authors, dam contribution to the breeding value of a young animal would only be well transmitted if the dam is genotyped and has its own records. Hence, we believe that dams with missing genotypes/phenotypes or even with low reliabilities may have contributed for some differences between official and interim breeding values for some animals. Even though, we believe ssGBLUP indirect prediction for young animals can be well approximated by partitioning the use of SNP effects and the polygenic term.

Table 1. Average ± standard deviations, mean relative difference (*RMD*, %), and correlation (*r*) between official and interim genomic breeding values for reference and validation populations

		Reference Population		
Traits	*Official*	*Interim*	*RMD*	*r*
MY	439.0±432.7	440.4±423.4	0.27	0.99
FY	16.9±16.5	17.0±16.2	0.23	0.99
FP	-0.01±0.2	-0.01±0.2	0.00	0.99
PY	13.2±14.1	13.3±13.8	0.44	0.99
PP	-0.02±0.1	-0.02±0.1	0.00	0.99
SCS	103.2±11.8	102.8±11.6	3.90	0.99
LONG	105.0±9.5	104.7±9.5	3.20	0.99
DCE	100.1±8.3	100.0±8.2	0.60	0.99
MCE	102.3±10.45	102.28±10.33	0.30	0.99
		Validation Population		
MY	593.2±311.2	595.5±301.7	0.43	0.92
FY	26.1±13.0	26.6±12.9	2.82	0.93
FP	0.03±0.1	0.04±0.1	0.05	0.93
PY	20.0±10.7	20.3±10.1	1.81	0.93
PP	-0.01±0.1	-0.01±0.1	0.01	0.94
SCS	109.5±8.6	109.8±8.6	3.20	0.91
LONG	113.8±6.0	114.2±5.9	4.10	0.91
DCE	103.2±6.3	102.7±6.2	5.20	0.94
MCE	110.2±8.1	110.8±7.8	5.90	0.94

MY: milk yield; FY: fat yield; FP: fat percentage; PY: protein yield; PP: protein percentage; SCS: somatic cell score; LONG: longevity; DCE: direct calving ease; MCE: maternal calving ease.

Conclusions

Our results are a first indication that we could 1) develop a core, high quality, stable reference population, 2) generate high quality SNP effects to estimate DGV and 3) generate appropriate GEBV from DGV and PT reflecting potentially own records for non-reference population animals. The last point is joint with efforts to generate appropriate reliabilities based on the approach promoted by Interbull.

Acknowledgements

The authors acknowledge the support of the National Fund for Scientific Research (Brussels, Belgium) also through grants PDR T.0095.19 “DeepSelect” and T.1053.15 “Genodomics”. We gratefully acknowledge the financial and technical support provided by the Walloon Breeding Association (AWE, Ciney, Belgium). Computational resources have been provided by the Plateau de Calcul et Modélisation informatique (CAMI) of the University of Liège, Gembloux Agro-Bio Tech and the Consortium des Equipements de Calcul Intensif (CECI) of the Federation Wallonia-Brussels (Brussels, Belgium), funded by the National Fund for Scientific Research (Brussels, Belgium) funded under grant 2.5020.11.

References

Aguilar, I., I. Misztal, D.L. Johnson, A. Legarra, S. Tsuruta, and T.J. Lawlor. 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., and M.S. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:1. https://doi.org/10.1186/s12711-014-0052-x

Colinet, F.G., J. Vandenplas, S. Vanderick, H. Hammami, R.R. Mota, A. Gillon, X. Hubin, C. Bertozzi, and N. Gengler. 2018. Bayesian single-step genomic evaluations combining local and foreign information in Walloon Holsteins. Animal 12:898–905. https://doi:10.1017/S1751731117002324

Lourenco, D.A.L., A. Legarra, S. Tsuruta, D. Moser, S. Miller, and I. Misztal. 2018a. Tuning Indirect Predictions Based on SNP Effects from Single-Step GBLUP. Interbull Bul. 53:48–53. https://journal.interbull.org/index.php/ib/article/view/1448

Lourenco, D.A.L., S. Tsuruta, B.O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, J.K. Bertrand, T.S. Amen, L. Wang, D.W. Moser, and I. Misztal. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. J. Anim. Sci. 93:2653–2662. https://doi.org/10.2527/jas.2014-8836

Lourenco, D.A.L., S. Tsuruta, B.O. Fragomeni, Y. Masuda, I. Aguilar, A. Legarra, S. Miller, D.W. Moser, and I. Misztal. 2018b. Single-step genomic BLUP for national beef cattle evaluation in US: from initial developments to final implementation. 10th World Congr. Genet. Appl. to Livest. Prod. http://www.wcgalp.org/system/files/proceedings/2018/single-step-genomic-blup-national-beef-cattle-evaluation-us-initial-developments-final.pdf

Pimentel, E.C.G., C. Edel, R. Emmerling, and K.-U. Götz. 2019. Technical note: Methods for interim prediction of single-step breeding values for young animals. J. Dairy Sci. 1–8. doi:10.3168/jds.2018-15592. https://doi.org/10.3168/jds.2018-15592

Shabalina, T., E.C.G. Pimentel, C. Edel, L. Plieschke, R. Emmerling, and K.-U. Götz. 2017. Short communication: The role of genotypes from animals without phenotypes in single-step genomic evaluations. J. Dairy Sci. 100:8277–8281. doi:10.3168/jds.2017-12734. https://doi.org/10.3168/jds.2017-12734

Vandenplas, J., F.G. Colinet, and N. Gengler. 2014. Unified method to integrate and blend several, potentially related, sources of information for genetic evaluation. Genet. Sel. Evol. 46:1–15. doi:10.1186/s12711-014-0059-3. https://doi.org/10.1186/s12711-014-0059-3

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