Interim genomic prediction considering newly acquired genotypes and phenotypes

J. Vandenplas[1], H. Eding[2] and M.P.L. Calus[1]

[1] Animal Breeding and Genomics, Wageningen UR, P.O. 338, 6700 AH Wageningen, The Netherlands

[2] CRV BV, Wassenaarweg 20, 6843 NW Arnhem, The Netherlands

Abstract

In the context of single-step genomic evaluations, current methods for predicting interim genomically enhanced breeding values (GEBV) for young, genotyped animals only consider new genotypes while ignoring new phenotypes. The aim of this study was to develop a method for predicting interim GEBV for animals associated with genotypes and/or phenotypes not included in a previous single-step genomic evaluation. The method thus developed relies on a Bayesian view of the linear mixed model for pedigree Best Linear Unbiased Prediction (BLUP). Assuming that single nucleotide polymorphism (SNP) effects are known *a priori*, it can be shown that a single-step genomic BLUP is equivalent to a pedigree BLUP with a prior mean based on direct genomic values and a prior covariance structure matrix requiring only pedigree information. Our method was tested on real data extracted from the December 2019 run of the Dutch-Flemish 4-trait evaluation for temperament and milking speed. The initial single-step evaluation included 6 520 406 animals (including 444 genetic groups), 4 147 302 records, and 144 086 genotypes. Four subsequent monthly single-step evaluations were performed by adding genotypes and phenotypes acquired during the corresponding additional period. Interim GEBVs for all animals in the pedigree were also computed for each month using our method and based on the SNP effects estimated by the initial single-step evaluation. For all traits, Pearson correlations between estimated SNP effects obtained from the initial single-step evaluation and from the four subsequent single-step evaluations decreased slightly over time, but were all higher than 0.98. By considering GEBVs obtained from the subsequent monthly single-step genomic evaluations as the references, accuracies of interim GEBVs for animals that were only newly genotyped, only phenotyped, or both, ranged between 0.98 and 1.00 across all traits. The corresponding dispersion biases ranged between 0.99 and 1.01. Therefore, our method results in accurate interim genomic predictions for all groups of animals.

Key words: genomic prediction, interim, indirect, single-step

Introduction

Currently, the single-step genomic evaluation (Legarra et al., 2014) is the method of choice for simultaneously analyzing phenotypic and pedigree information of genotyped and non-genotyped animals with genomic information of genotyped animals. A first approach is the so-called single-step genomic best linear unbiased prediction (ssGBLUP; Aguilar et al., 2010; Christensen and Lund, 2010) that includes genomic information by combining genomic and pedigree relationships into a combined genomic-pedigree relationship matrix. A second approach is the so-called single-step single nucleotide polymorphism (SNP) BLUP (ssSNPBLUP; Fernando et al., 2014; Liu et al., 2014) that fits the SNP effects explicitly as random effects in the model. However, both single-step approaches are still computationally demanding with a large number of genotyped animals, even if several approaches have been proposed in the literature to reduce these computational costs (e.g., Misztal et al., 2014; Mäntysaari et al., 2017; Vandenplas et al., 2020). Therefore, methods were developed to predict interim genomic enhanced breeding values (GEBV) of young animals in a limited

amount of time without performing a full single-step evaluation (Lourenco et al., 2015; Pimentel et al., 2019). These methods rely on the multiplication of the genotypes of young animals by the estimated SNP effects obtained from the previous single-step genomic evaluation. However, these methods do not consider possible newly acquired phenotypes of young animals. The aim of this study was therefore to develop and test a method for the prediction of interim GEBVs for animals with genotypes and/or phenotypes not included in a previous single-step genomic evaluation.

Materials and Methods

Model

In this study we derive our proposed method from the ssSNPBLUP linear equations proposed by Liu et al. (2014). If we assume that estimates of SNP effects $\hat{\mathbf{g}}$ are known before performing a single-step genomic evaluation, then we can assume the following prior multivariate normal (MVN) distributions for the genetic additive effects $\mathbf{u}$:

$$[\mathbf{u}|\hat{\boldsymbol{\mu}}, \mathbf{A}^*] \sim MVN(\hat{\boldsymbol{\mu}}, \mathbf{A}^*\sigma_u^2)$$

with

$$\hat{\boldsymbol{\mu}} = \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1} \\ \mathbf{I} \end{bmatrix} \mathbf{Z}\hat{\mathbf{g}}$$

and

$$\mathbf{A}^{*-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} + \left(\frac{1}{w} - 1\right)\mathbf{A}_{gg}^{-1} \end{bmatrix},$$

where the subscripts n and g refer to ungenotyped and genotyped animals, respectively,

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{nn} & \mathbf{A}_{ng} \\ \mathbf{A}_{gn} & \mathbf{A}_{gg} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} \end{bmatrix}$$ is the inverse of the pedigree relationship matrix partitioned between genotyped and ungenotyped animals, w is the proportion of additive genetic variance explained by the residual polygenic effects, σ_u^2 is the genetic variance, $\mathbf{Z}$ is the genotyped matrix centered with observed allele frequencies, and $\mathbf{I}$ is an identity matrix.

The system of equations associated with these assumptions is written as follows:

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{*-1}\sigma_u^{-2} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} + \mathbf{A}^{*-1}\sigma_u^{-2}\hat{\boldsymbol{\mu}} \end{bmatrix} \quad (1)$$

where $\hat{\boldsymbol{\beta}}$ is the vector of estimated fixed effects, $\mathbf{y}$ is the vector of records, $\mathbf{R}^{-1}$ is the inverse of the residual variance structure matrix, and $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating records to the fixed and additive genetic effects, respectively.

The system of equations (1) is equivalent to a single-step genomic evaluation, either ssGBLUP (Christensen and Lund, 2010) or the ssSNPBLUP proposed by Liu et al. (2014), from which the estimates of SNP effects $\hat{\mathbf{g}}$ would be computed. It can be shown that the system of equations (1) is equivalent to the system of equations (16) of Liu et al. (2014) after some algebra.

By using the estimates of SNP effects from a previous single-step genomic evaluation, the proposed system of equations (1) can be used to compute interim GEBVs for all genotyped and ungenotyped animals by considering previously included and newly acquired genotypes and phenotypes.

Data

The new method was tested on a dataset and associated variance components provided by CRV BV (The Netherlands) and extracted from the December 2019 run of the Dutch-Flemish four-trait evaluation for temperament and milking speed (CRV, 2020a, 2020b). Performances in both countries were considered as different traits, with genetic correlations between Dutch and Flemish traits higher than 0.85. A single record per animal was observed. The four-trait mixed model included random effects (additive genetic and residual), fixed co-variables (heterosis and recombination), and

fixed cross-classified effects (herd × year × season at classification, age at classification, lactation stage at classification, milk yield and month of calving). More details can be found in CRV documents (CRV, 2020a, 2020b).

Single-step evaluations

A total of five ssSNPBLUP evaluations were performed: the initial ssSNPBLUP evaluation that included all phenotypes and genotypes available until month 0, and four subsequent monthly ssSNPBLUP evaluations that included additional data available until the corresponding month (from month 1 until month 4). The initial ssSNPBLUP evaluation aimed to mimic a routine ssSNPBLUP evaluation, and the four subsequent ssSNPBLUP evaluations were considered as reference evaluations for validating our developed interim genomic prediction method.

For the initial ssSNPBLUP evaluation, a total of 4 147 302 records were available, the pedigree included 6 520 406 animals (including 444 genetic groups), and the genotypes of 144 086 animals were available (Table 1). Only the genotypes of the animals with a record included in a single-step evaluation and of their ancestors were considered. After removing non-segregating SNPs and SNPs with a minor allele frequency lower than 0.01, genotypes included 37 995 segregating SNPs.

The four subsequent monthly ssSNPBLUP evaluations were performed by adding to the initial data sets, the phenotypes acquired during the corresponding additional period, as well as the genotypes of the newly phenotyped animals and of their ancestors. The different amounts of genotypes and phenotypes added at each additional period (in comparison to the initial datasets) can be found in Table 1.

Interim genomic prediction

For each additional month (i.e., from month 1 until month 4), interim GEBVs were computed for all animals in the pedigree using the method's system of equations (1) and the estimated SNP effects obtained from the initial ssSNPBLUP evaluation.

Table 1. Amounts of genotypes and phenotypes available for the initial single-step evaluation and added each month to the initial datasets.

		Evaluation at month				
		0	+1	+2	+3	+4
Genotypes		144 086	+2 644	+4 359	+7 492	+9 680
Phenotypes	Trait 1	3 814 020	+8 664	+15 057	+25 448	+32 861
	Trait 2	3 480 757	+6 673	+11 567	+20 140	+25 644
	Trait 3	176 632	+0	+0	+0	+0
	Trait 4	360 453	+1 436	+2 556	+3 356	+4 437

Analysis

For all traits, Pearson correlations between estimated SNP effects obtained from the initial ssSNPBLUP evaluation and from the four subsequent monthly ssSNPBLUP evaluations were computed. Furthermore, for the four interim genomic predictions, we computed (a) accuracy defined as the Pearson correlation between GEBVs of the subsequent ssSNPBLUP and interim GEBVs, (b) level bias defined as the difference between mean GEBV from the subsequent ssSNPBLUP evaluations and the corresponding mean interim GEBV, in genetic standard deviation units, and (c) dispersion bias defined as the slope of the regression of GEBVs from the subsequent ssSNPBLUP evaluations on the interim GEBVs. For the four interim genomic predictions, the different estimators were computed for animals with (a) a genotype acquired since the initial ssSNPBLUP evaluation, (b) a record acquired since the initial ssSNPBLUP evaluation, (c) both, and (d) for

animals without new genotype or phenotype, i.e. all other animals in the pedigree.

The systems of equations for all the ssSNPBLUP evaluations and interim genomic predictions were solved using a two-level preconditioned conjugate gradient (PCG) approach implemented in a Fortran 2003 program (Vandenplas et al., 2019, 2020). For all the systems, the PCG approach iterated until the criterion CK defined by Vandenplas et al. (2021) was ≤ 0.5.

Results & Discussion

On average, for each subsequent monthly ssSNPBLUP evaluation, around 2 400 genotypes (that is about 1.7 % of the amount of genotypes available in the initial single-step evaluation) were added to the genotype set used by the previous monthly ssSNPBLUP evaluation. The number of phenotypes added monthly to each subsequent ssSNPBLUP evaluation and interim genomic prediction varied from 0 (trait 3) to around 8 200 (trait 1; Table 1). The amounts of animals without newly acquired genotype and phenotype, with newly acquired genotype, with newly acquired phenotype, or both, are in Table 2.

For all traits, Pearson correlations between SNP effects estimated by the initial ssSNPBLUP evaluation and by the subsequent ssSNPBLUP evaluations were equal to 0.997 for month 1, 0.995 for month 2, 0.991 for month 3, and 0.988 for month 4. These correlations close to 1 show that the estimated SNP effects are little influenced by the additional phenotypes and genotypes. Further research is needed to analyze the stability of estimated SNP effects across time.

Table 2. Amounts of animals without newly acquired genotype and phenotype, with newly acquired genotype, with newly acquired phenotype, or both.

Animals	Trait	Month			
		1	2	3	4
Without new genotype or phenotype	1	6 511 509	6 504 893	6 494 151	6 486 556
	2	6 512 862	6 507 376	6 497 933	6 491 613
	3	6 517 762	6 516 047	6 512 914	6 510 726
	4	6 516 675	6 5141 19	6 510 443	6 507 521
With a new genotype only	1	233	456	807	989
	2	871	1 463	2 333	3 149
	3	2 644	4 359	7 492	9 680
	4	2 295	3 731	6 607	8 448
With a new phenotype only	1	6 253	11 154	18 763	24 170
	2	4 900	8 671	14 981	19 113
	3	0	0	0	0
	4	1 087	1 928	2 471	3 205
With a new genotype and phenotype	1	2 411	3 903	6 685	8 691
	2	1 773	2 896	5 159	6 531
	3	0	0	0	0
	4	349	628	885	1232

Figures 1, 2, and 3 depict accuracy, level bias, and dispersion bias of interim GEBVs, respectively, for the four subsequent monthly evaluations, and for the animals with newly

acquired genotype, with newly acquired phenotype, both, as well as for animals without new genotype and phenotype. Overall, across all traits, all subsequent evaluations and all groups of animals, the accuracies ranged between 0.98 and 1.00 (Figure 1), the level biases ranged between -0.02 and 0.00 genetic standard deviation (Figure 2), and the dispersion biases ranged between 0.99 and 1.01 (Figure 3). Therefore, the proposed interim genomic prediction method yields highly accurate interim GEBVs for all animals with or without newly acquired genotype or phenotype. The main differences between interim GEBVs and GEBVs from subsequent ssSNPBLUP evaluations were mainly observed for animals with a newly acquired genotype (with or without a new phenotype). For these animals the accuracies decreased over time, while the absolute values of level bias increased.

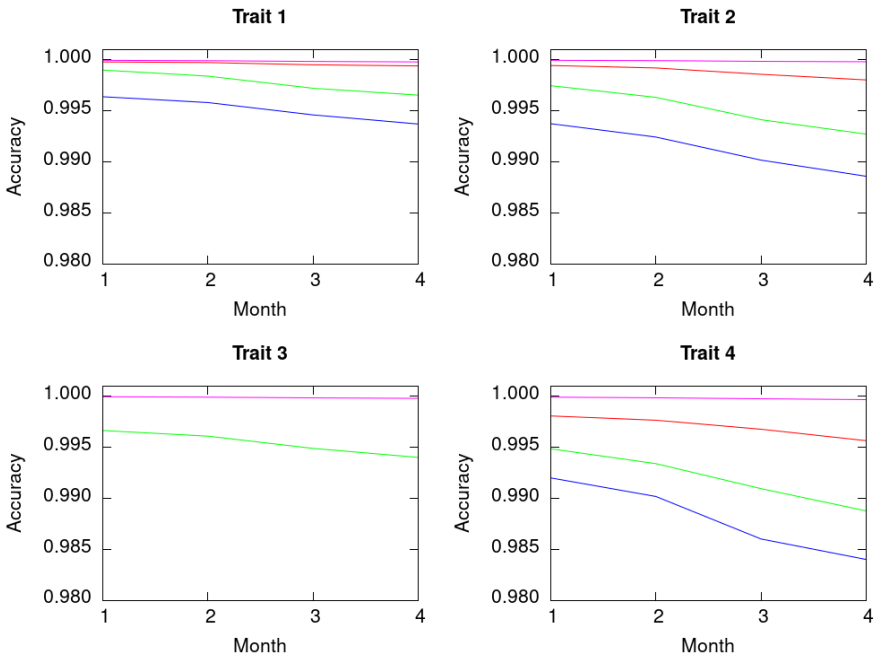

Figure 1. Accuracies of interim GEBVs for animals with newly acquired genotype (green), newly acquired phenotype (red), both (blue), and for animals without new genotype and phenotype (purple).

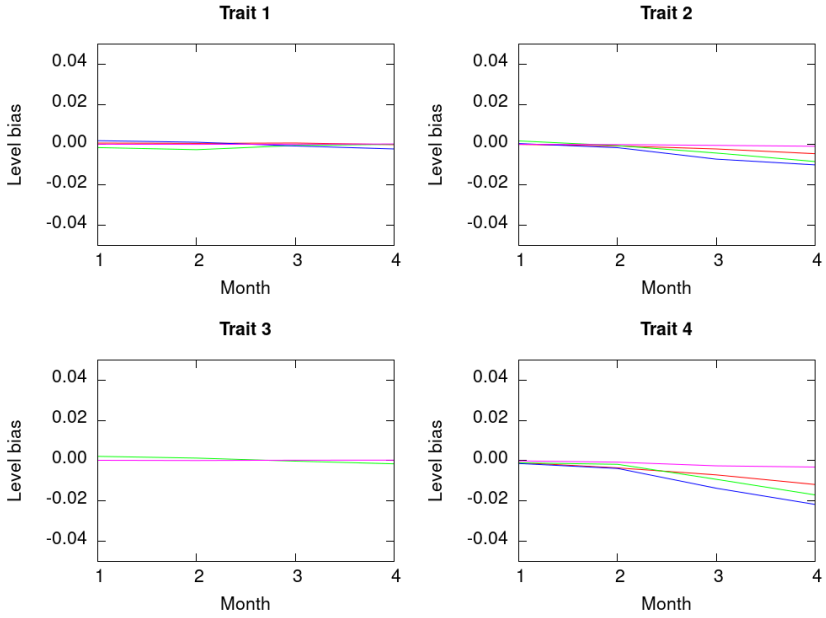

Figure 2. Level biases (expressed in genetic standard deviation) of interim GEBVs for animals with newly acquired genotype (green), newly acquired phenotype (red), both (blue), and for animals without new genotype and phenotype (purple).

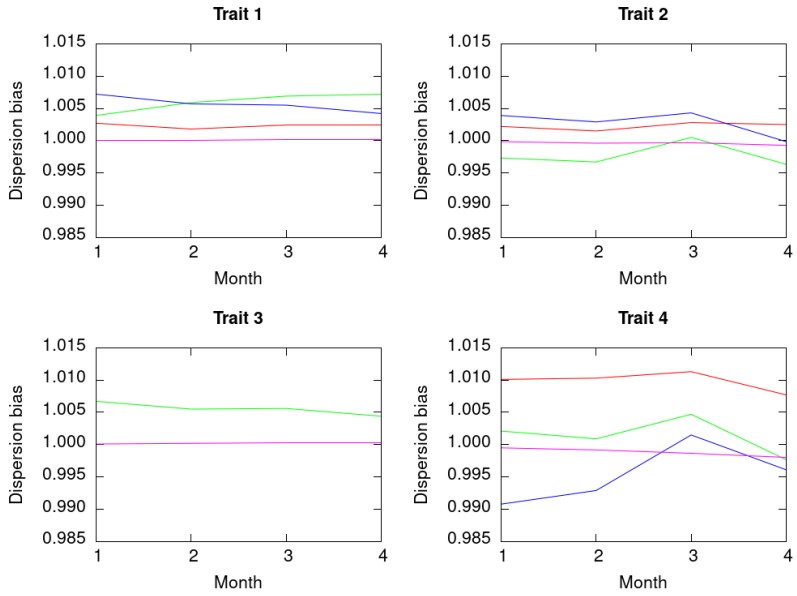

Figure 3. Dispersion biases of interim GEBVs for animals with newly acquired genotype (green), newly acquired phenotype (red), both (blue), and for animals without new genotype and phenotype (purple).

In comparison to interim genomic prediction methods previously proposed in the literature (e.g., Lourenco et al., 2015; Pimentel et al., 2019), our proposed prediction method considers newly acquired phenotypes for computing interim GEBVs, without solving a ssSNPBLUP evaluation. Furthermore, as it can be deduced from the system of equations (1), and based on a computational approach as detailed in Vandenplas et al. (2020), the computational costs (in terms of time and memory) of the proposed interim genomic prediction method are similar to the computational costs of solving a pedigree BLUP with a PCG approach. Therefore, our method could be of interest for evaluating traits

for which phenotypes (or pseudo-phenotypes like deregressed proofs) would be available at the same time as genotypes.

Our method also has the advantages of correctly considering GEBV mean, genetic groups associated with animals not included in the initial ssSNPBLUP evaluation, and the implicit imputation of genotypes. Issues with a mean shift in interim GEBVs and with genetic groups were reported in the literature, together with possible solutions (e.g., Lourenco et al., 2015; Pimentel et al., 2019). Such issues cannot appear with our method because all (old and new) animals in the pedigree are considered simultaneously for the interim prediction. Finally, as mentioned by Pimentel et al. (2019), interim genomic predictions that do not involve the solving of a system of equations derived from a single-step mixed model equations (e.g., Lourenco et al., 2015; Pimentel et al., 2019) can only yield approximate GEBVs, because the newly acquired genotypes not included in a single-step evaluation cannot influence the GEBVs of genotyped and ungenotyped animals (through linear imputation) included in this evaluation (Shabalina et al., 2017; Pimentel et al., 2019). With our method, the newly acquired genotypes will influence the GEBVs of all animals included in the system of equations (1) through the imputation of DGVs of ungenotyped animals. Indeed, as already mentioned, our interim genomic prediction method will result in the same GEBVs as a full single-step evaluation if the estimated SNP effects are the same in both evaluations.

The interest of our proposed interim genomic prediction method over methods previously proposed in the literature (e.g., Lourenco et al., 2015; Pimentel et al., 2019), will depend on several factors, such as the structure of the phenotype and genotype datasets, the amount of animals with newly acquired phenotypes and genotypes, or even the time available for computing interim genomic predictions. Indeed, the computational costs of our method are expected to be higher than previously proposed methods for which one of the main costs is the multiplication of the genotype matrix by the estimated SNP effects. As discussed above, however, our method is more robust than previously proposed interim prediction methods.

Conclusions

Assuming that the SNP effects are known *a priori*, i.e. from a recent single-step genomic evaluation, we developed a method for computing interim genomic predictions for animals with newly acquired genotype, phenotypes, or both. We showed that our method results in accurate interim genomic predictions for all groups of animals.

Acknowledgments

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food Project 16022) and the Breed4Food partners Cobb Europe (Colchester, Essex, United Kingdom), CRV (Arnhem, the Netherlands), Hendrix Genetics (Boxmeer, the Netherlands), and Topigs Norsvin (Helvoirt, the Netherlands). Jeremie Vandenplas thanks Zengting Liu, Esa Mantysaari, Ismo Stranden, and Peter Sullivan for fruitful discussions.

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.

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

CRV Animal Evaluation Unit. 2020a. Statistical indicators, E16: breeding value-Temperament during milking.

https://cooperatiecrv-be6.kxcdn.com/wp-content/uploads/2020/04/E_

16-Gedrag-April-2020-Engels.pdf. Accessed 29 Mar 2021.

CRV Animal Evaluation Unit. 2020b. Statistical indicators, E-15: Breeding value milking speed. https://coope

ratiecrv-be6.kxcdn.com/wp-content/uploads/2020/04/E_15-Melksnelhe

id-April-2020-Engels.pdf. Accessed 29 Mar 2021.

Fernando, R.L., J.C. Dekkers, and D.J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. Genet. Sel. Evol. 46:50.

Legarra, A., O.F. Christensen, I. Aguilar, and I. Misztal. 2014. Single step, a general approach for genomic selection. Livest. Sci. 166:54–65.

Liu, Z., M. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects. J. Dairy Sci. 97:5833–5850.

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.

Mäntysaari, E.A., R.D. Evans, and I. Strandén. 2017. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. J. Anim. Sci. 95:4728–4737.

Misztal, I., A. Legarra, and I. Aguilar. 2014. Using recursion to compute the inverse of the genomic relationship matrix. J. Dairy Sci. 97:3943-3952.

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. 102:3266–3273.

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.

Vandenplas, J., M.P.L. Calus, H. Eding, M. van Pelt, R. Bergsma, and C. Vuik. 2021. Convergence behavior of single-step GBLUP and SNPBLUP for different termination criteria. Genet. Sel. Evol. 53:34.

Vandenplas, J., M.P.L. Calus, H. Eding, and C. Vuik. 2019. A second-level diagonal preconditioner for single-step SNPBLUP. Genet. Sel. Evol. 51:30.

Vandenplas, J., H. Eding, M. Bosmans, and M.P.L. Calus. 2020. Computational strategies for the preconditioned conjugate gradient method applied to ssSNPBLUP, with an application to a multivariate maternal model. Genet. Sel. Evol. 52:24.