

**This is an electronic reprint of the original article.**

**This reprint *may differ* from the original in pagination and typographic detail.**

**Author(s):** Jeremie Vandenplas, Jan ten Napel, Saeid Naderi Darbaghshahi, Ross Evans, Mario P. L. Calus, Roel Veerkamp, Andrew Cromie, Esa A. Mäntysaari & Ismo Strandén

**Title:** Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates

**Year:** 2023

**Version:** Published version

**Copyright:** The Author(s) 2023

**Rights:** CC BY 4.0

**Rights url:** <http://creativecommons.org/licenses/by/4.0/>

**Please cite the original version:**

Vandenplas, J., ten Napel, J., Darbaghshahi, S.N. et al. Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genet Sel Evol* 55, 37 (2023). <https://doi.org/10.1186/s12711-023-00808-z>


All material supplied via *Jukuri* is protected by copyright and other intellectual property rights. Duplication or sale, in electronic or print form, of any part of the repository collections is prohibited. Making electronic or print copies of the material is permitted only for your own personal use or for educational purposes. For other purposes, this article may be used in accordance with the publisher's terms. There may be differences between this version and the publisher's version. You are advised to cite the publisher's version.

RESEARCH ARTICLE

Open Access



# Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates

Jeremie Vandenplas<sup>1\*</sup>, Jan ten Napel<sup>1</sup>, Saeid Naderi Darbaghshahi<sup>2</sup>, Ross Evans<sup>2</sup>, Mario P. L. Calus<sup>1</sup>, Roel Veerkamp<sup>1</sup>, Andrew Cromie<sup>2</sup>, Esa A. Mäntysaari<sup>3</sup> and Ismo Strandén<sup>3†</sup>

## Abstract

**Background** Single-step genomic best linear unbiased prediction (ssGBLUP) models allow the combination of genomic, pedigree, and phenotypic data into a single model, which is computationally challenging for large genotyped populations. In practice, genotypes of animals without their own phenotype and progeny, so-called genotyped selection candidates, can become available after genomic breeding values have been estimated by ssGBLUP. In some breeding programmes, genomic estimated breeding values (GEBV) for these animals should be known shortly after obtaining genotype information but recomputing GEBV using the full ssGBLUP takes too much time. In this study, first we compare two equivalent formulations of ssGBLUP models, i.e. one that is based on the Woodbury matrix identity applied to the inverse of the genomic relationship matrix, and one that is based on marker equations. Second, we present computationally-fast approaches to indirectly compute GEBV for genotyped selection candidates, without the need to do the full ssGBLUP evaluation.

**Results** The indirect approaches use information from the latest ssGBLUP evaluation and rely on the decomposition of GEBV into its components. The two equivalent ssGBLUP models and indirect approaches were tested on a six-trait calving difficulty model using Irish dairy and beef cattle data that include 2.6 million genotyped animals of which about 500,000 were considered as genotyped selection candidates. When using the same computational approaches, the solving phase of the two equivalent ssGBLUP models showed similar requirements for memory and time per iteration. The computational differences between them were due to the preprocessing phase of the genomic information. Regarding the indirect approaches, compared to GEBV obtained from single-step evaluations including all genotypes, indirect GEBV had correlations higher than 0.99 for all traits while showing little dispersion and level bias.

**Conclusions** In conclusion, ssGBLUP predictions for the genotyped selection candidates were accurately approximated using the presented indirect approaches, which are more memory efficient and computationally fast, compared to solving a full ssGBLUP evaluation. Thus, indirect approaches can be used even on a weekly basis to estimate GEBV for newly genotyped animals, while the full single-step evaluation is done only a few times within a year.

<sup>†</sup>Jeremie Vandenplas and Ismo Strandén contributed equally to this work

\*Correspondence:

Jeremie Vandenplas

jeremie.vandenplas@wur.nl

Full list of author information is available at the end of the article



## Background

The original single-step genomic best linear unbiased prediction (ssGBLUP) genomic evaluation studies [1, 2] presented a theoretically well-justified model for genetic evaluation, which allows the inclusion of pedigree and phenotypes of genotyped and non-genotyped animals. Practical implementation of ssGBLUP has met both computational and modelling challenges (e.g., [3, 4]). Several approaches have been presented to allow efficient modelling and solving of ssGBLUP (e.g., [5–8]). In practice, estimated breeding values for dairy and beef cattle are computed three to four times for a trait or trait group during a year, but genotypes for newly born animals become available almost continuously throughout the year. Because the computation of the full data ssGBLUP predictions is demanding, computationally efficient calculation of breeding values for these newly genotyped animals is desirable. We will use the term genotyped selection candidate for a newly genotyped animal without progeny and own phenotype and for which we would like to compute genomic predictions.

Legarra and Ducrocq [9], Fernando et al. [6], Liu et al. [7], and Taskinen et al. [10] presented single-step approaches, which are hereinafter denoted ssSNPBLUP where the mixed model equations (MME) include single nucleotide polymorphism (SNP) effects. Solutions of the SNP effects allow a simple approach to estimate GEBV for genotyped selection candidates without the need to solve the full updated MME. Lourenco et al. [11] presented a similar approach for ssGBLUP where the SNP effects are estimated using a formula derived from a GBLUP model. Pimentel et al. [12] presented and tested approximate approaches for the prediction of selection candidates in ssGBLUP, which were based on SNP-BLUP or GBLUP and facilitated by including the so-called residual polygenic (RPG) effect. Liu et al. [13] presented formulas for predicting breeding values of genotyped selection candidates when solutions from ssSNPBLUP were available and the model had an RPG effect.

Prediction of breeding values of genotyped selection candidates can be integrated with the existing computational approaches used for solving the full data ssGBLUP. Mäntysaari et al. [8] presented an efficient computational approach with T-factoring named ssGTBLUP for single-step genomic evaluations. This approach assumes that the genomic relationship matrix has the form  $\mathbf{G} = \mathbf{Z}\mathbf{Z}' + \mathbf{C}$ , where  $\mathbf{Z}$  is a centered and scaled genotype marker matrix and  $\mathbf{C}$  is a non-singular easily invertible regularization matrix. The main computational step in solving the MME of ssGBLUP by an iterative method is the calculation of  $\mathbf{G}^{-1}$  times a vector product in each iteration. In ssGTBLUP, this product was shown to require two products involving a rectangular matrix of size  $m$  by  $n$  where  $m$  is

the number of SNPs and  $n$  is the number of genotyped animals. Consequently, computational work increases linearly with the number of genotyped animals  $n$  instead of quadratically as in regular ssGBLUP. According to Mäntysaari et al. [3], absorption of the SNP effects in the MME of the ssSNPBLUP model proposed by Liu et al. [7] leads to ssGTBLUP where  $\mathbf{C}$  is the pedigree-based relationship matrix among genotyped animals multiplied by the proportion of RPG effect.

In this study, we present a unified model for ssGTBLUP [8] and ssSNPBLUP [7], which extends the general regularization matrix  $\mathbf{C}$  used in ssGTBLUP to be integrated in ssSNPBLUP. Furthermore, we present efficient indirect approaches for both ssGTBLUP and ssSNPBLUP models that allow the prediction of GEBV for the genotyped selection candidates under different regularization matrices. We investigate and compare the performance of the ssGTBLUP, ssSNPBLUP and indirect approaches by using a multi-trait model with more than 2.6 million genotyped animals.

## Methods

### ssGTBLUP and ssSNPBLUP

We investigated two computational approaches for single-step genomic evaluation. First, we use a general notation to describe the ssGTBLUP approach proposed by Mäntysaari et al. [8]. Second, ssGTBLUP is used to derive the ssSNPBLUP approach by Liu et al. [7]. In spite of apparent differences in the MME of these approaches, we show how, in theory, they are computationally similar.

### The ssGTBLUP approach

A standard univariate mixed model for ssGBLUP can be written as:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \begin{bmatrix} \mathbf{W}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_g \end{bmatrix} \begin{bmatrix} \mathbf{u}_n \\ \mathbf{u}_g \end{bmatrix} + \mathbf{e}, \quad (1)$$

where  $\mathbf{y}$  is the vector of records,  $\mathbf{b}$  is the vector of fixed effects,  $\mathbf{u}_n$  is the vector of additive genetic effects for the non-genotyped animals,  $\mathbf{u}_g$  is the vector of additive genetic effects for the genotyped animals, and  $\mathbf{e}$  is the vector of residuals. The matrices  $\mathbf{X}$ ,  $\mathbf{W}_n$ , and  $\mathbf{W}_g$  relate records in  $\mathbf{y}$  to the corresponding effects.

We assume normally distributed additive genetic effects  $\mathbf{u}' = [\mathbf{u}'_n \ \mathbf{u}'_g]$  with a mean zero and a covariance structure matrix  $\mathbf{H}$ ,  $\begin{bmatrix} \mathbf{u}_n \\ \mathbf{u}_g \end{bmatrix} \sim MVN(\mathbf{0}, \mathbf{H}\sigma_u^2)$ , where  $\sigma_u^2$  is the additive genetic variance and  $\mathbf{H}$  is the additive genetic covariance matrix defined below. Without loss of generality, we assume an independent and identical normal distribution for the residual effects  $\mathbf{e} \sim MVN(\mathbf{0}, \mathbf{I}\sigma_e^2)$ ,

where  $\mathbf{I}$  is the identity matrix and  $\sigma_e^2$  is the residual variance.

In ssGTBLUP [8], the genomic relationship matrix is assumed to have the form  $\mathbf{G}_C = \mathbf{G}_m + \mathbf{C}$ , where  $\mathbf{G}_m$  is a function of genomic data and  $\mathbf{C}$  is an invertible regularization matrix. We define the genomic part as having the form  $\mathbf{G}_m = \mathbf{Z}\mathbf{B}\mathbf{Z}'$ , where  $\mathbf{Z} = (\mathbf{M} - \mathbf{P})$  is an  $n$  by  $m$  matrix of centered marker genotypes with  $n$  being the number of genotyped animals and  $m$  being the number of SNPs,  $\mathbf{M}$  is an  $n$  by  $m$  matrix of SNP genotypes,  $\mathbf{P}$  is an  $n$  by  $m$  centering matrix and  $\mathbf{B}$  is an  $m$  by  $m$  diagonal scaling matrix. The centering matrix has often the form  $\mathbf{P} = 2\mathbf{1}_n\mathbf{p}'$  where the vector  $\mathbf{p}$  has  $m$  allele frequencies. The marker genotype matrix  $\mathbf{M}$  has counts of the first allele, such that the homozygous genotype for the second allele has a value of 0, the heterozygous genotype has 1, and the homozygous genotype for the second allele has 2. It is recommended to use base population allele frequencies in the vector  $\mathbf{p}$  [14]. In VanRaden's [14] Method 1, the scaling matrix  $\mathbf{B}$  is equal to  $\mathbf{B}_{VR} = \mathbf{I}_k^{-1}$  with the scaling constant  $k = 2\sum_{i=1}^m p_i(1 - p_i)$ . The linear system of MME for ssGTBLUP is [8]:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'_n\mathbf{W}_n & \mathbf{X}'_g\mathbf{W}_g \\ \mathbf{W}'_n\mathbf{X}_n & \mathbf{W}'_n\mathbf{W}_n + \mathbf{H}^{nn}\lambda & \mathbf{H}^{ng}\lambda \\ \mathbf{W}'_g\mathbf{X}_g & \mathbf{H}^{gn}\lambda & \mathbf{W}'_g\mathbf{W}_g + \mathbf{H}^{gg}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_n \\ \hat{\mathbf{u}}_g \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'_n\mathbf{y} \\ \mathbf{W}'_g\mathbf{y} \end{bmatrix}, \tag{2}$$

where  $\lambda = \frac{\sigma_e^2}{\sigma_u^2}$ ,  $\mathbf{X}' = [\mathbf{X}'_n \ \mathbf{X}'_g]$ ,  $\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{H}^{nn} & \mathbf{H}^{ng} \\ \mathbf{H}^{gn} & \mathbf{H}^{gg} \end{bmatrix}$   
 $= \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_C^{-1} - \mathbf{A}_{gg}^{-1} \end{bmatrix}$ ,  $\mathbf{A}^{-1}$  is the inverse of the pedigree-based relationship matrix, and  $\mathbf{A}_{gg}$  is the pedigree-based relationship matrix among the genotyped animals. Matrix  $\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} \end{bmatrix}$  is denoted similarly as for  $\mathbf{H}^{-1}$ . The inverse genomic relationship matrix can be expressed using the Woodbury matrix identity [8] as:

$$\mathbf{G}_C^{-1} = (\mathbf{Z}\mathbf{B}\mathbf{Z}' + \mathbf{C})^{-1} = \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{Z}\mathbf{K}^{-1}\mathbf{Z}'\mathbf{C}^{-1},$$

where  $\mathbf{K} = \mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z} + \mathbf{B}^{-1}$  is a symmetric positive definite matrix.

Solving MME (2) iteratively using the preconditioned conjugate gradient (PCG) approach requires computing the product of the MME coefficient matrix times a vector, say  $\mathbf{v}$ . When the number of genotyped animals is

large, most of the computing time for this product is due to  $\mathbf{G}_C^{-1}\mathbf{v}$ . Thus, it is important that the product  $\mathbf{C}^{-1}\mathbf{v}$  is fast, especially for many genotyped animals. Mäntysaari et al. [8] presented two previously proposed forms for  $\mathbf{C}$  that allow fast computation of  $\mathbf{C}^{-1}\mathbf{v}$ . First,  $\mathbf{C} = \varepsilon\mathbf{I}$ , where  $\varepsilon$  is a small number (e.g.,  $10^{-2}$ ). Second,  $\mathbf{C} = w\mathbf{A}_{gg}$  where  $w$  is the proportion of polygenic variance not accounted for by the markers, i.e., the RPG proportion. When  $\mathbf{C} = \varepsilon\mathbf{I}$ , the  $\mathbf{K}$  matrix becomes  $\mathbf{K} = \frac{1}{\varepsilon}\mathbf{Z}'\mathbf{Z} + \mathbf{B}^{-1}$ , with  $\mathbf{B} = \mathbf{B}_{VR} = \mathbf{I}_k^{-1}$ . When  $\mathbf{C} = w\mathbf{A}_{gg}$ , the  $\mathbf{K}$  matrix becomes  $\mathbf{K} = \frac{1}{w}\mathbf{Z}'\mathbf{A}_{gg}^{-1}\mathbf{Z} + \mathbf{B}^{-1}$ . Furthermore, the proportion  $(1 - w)$  of additive genetic variance accounted for by the markers must be included in the scaling matrix  $\mathbf{B}$ , i.e.,  $\mathbf{B} = (1 - w)\mathbf{B}_{VR} = \mathbf{I}_k^{-1-w}$  when  $\mathbf{G}_m$  is computed following VanRaden's [14] Method 1.

The first proposition of the ssGTBLUP approach [8] (hereafter called original ssGTBLUP approach) was derived to give the product  $\mathbf{G}_C^{-1}\mathbf{v}$  of form  $\mathbf{G}_C^{-1}\mathbf{v} = (\mathbf{C}^{-1} - \mathbf{T}'_C\mathbf{T}_C)\mathbf{v}$ , where  $\mathbf{T}_C = \mathbf{L}_C^{-1}\mathbf{Z}'\mathbf{C}^{-1}$ , with the lower triangular matrix  $\mathbf{L}_C$  being the Cholesky decomposition of  $\mathbf{K}$ , i.e.,  $\mathbf{L}_C\mathbf{L}'_C = \mathbf{K}$ . Using double-precision arithmetic, software that rely on this original ssGTBLUP approach will use  $8nm$  bytes for storing  $\mathbf{T}_C$  in memory.

However, instead of computing the product  $\mathbf{G}_C^{-1}\mathbf{v} = (\mathbf{C}^{-1} - \mathbf{T}'_C\mathbf{T}_C)\mathbf{v}$ , a computationally more efficient approach proposed by Mäntysaari et al. [3] can be to use the original form explicitly, as follows (hereafter called component-wise ssGTBLUP approach):

$$\mathbf{G}_C^{-1}\mathbf{v} = \mathbf{C}^{-1}\mathbf{v} - \mathbf{C}^{-1}\left(\mathbf{Z}\left(\mathbf{L}_C \setminus \left\{ \mathbf{L}'_C \setminus \left[ \mathbf{Z}'\left(\mathbf{C}^{-1}\mathbf{v}\right) \right] \right\}\right)\right),$$

where the matrix times vector products are calculated from the innermost brackets to outward, and the backslash ( $\setminus$ ) is an operator indicating that the system of equations is solved by forward or backward substitutions. This approach allows the computations that involve  $\mathbf{Z}$  to use  $\mathbf{M}$ , which can be efficiently stored in a compressed form to take less memory than  $\mathbf{T}_C$  [3].

**From ssGTBLUP to ssSNPBLUP**

The MME (2) can be reformulated using an equivalent model by appending the vector of estimated SNP marker effect solutions  $\hat{\mathbf{g}}$  to the vector of solutions of MME (2) [7]. An extended MME form can be written as:

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'_n\mathbf{W}_n & \mathbf{X}'_g\mathbf{W}_g & \mathbf{0} \\ \mathbf{W}'_n\mathbf{X}_n & \mathbf{W}'_n\mathbf{W}_n + \mathbf{A}^{nn}\lambda & \mathbf{A}^{ng}\lambda & \mathbf{0} \\ \mathbf{W}'_g\mathbf{X}_g & \mathbf{A}^{gn}\lambda & \mathbf{W}'_g\mathbf{W}_g + (\mathbf{A}^{gg} - \mathbf{A}_{gg}^{-1} + \mathbf{C}^{-1})\lambda & -\mathbf{C}^{-1}\mathbf{Z}\lambda \\ \mathbf{0} & \mathbf{0} & -\mathbf{Z}'\mathbf{C}^{-1}\lambda & \mathbf{K}\lambda \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_n \\ \hat{\mathbf{u}}_g \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'_n\mathbf{y} \\ \mathbf{W}'_g\mathbf{y} \\ \mathbf{0} \end{bmatrix}, \tag{3}$$

where the vector  $\hat{\mathbf{g}}$  has the estimated SNP effect solutions. The extended MME (3) is the same as that derived for the ssSNPBLUP when the RPG regularization matrix  $\mathbf{C} = w\mathbf{A}_{gg}$  is used.

We can denote the inverse of the covariance structure in MME (3) as:

$$\mathbf{H}_L^{-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} & \mathbf{0} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} - \mathbf{A}_{gg}^{-1} + \mathbf{C}^{-1} & -\mathbf{C}^{-1}\mathbf{Z} \\ \mathbf{0} & -\mathbf{Z}'\mathbf{C}^{-1} & \mathbf{K} \end{bmatrix}.$$

Following Liu et al. [7], we have  $\text{Var} \begin{pmatrix} \mathbf{u}_n \\ \mathbf{u}_g \\ \mathbf{g} \end{pmatrix} = \mathbf{H}_L \sigma_u^2$

with  $\mathbf{H}_L$  defined as:

$$\begin{aligned} \mathbf{H}_L &= \begin{bmatrix} \mathbf{A}_{nn} + \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}(\mathbf{G}_C - \mathbf{A}_{gg})\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{G}_C & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{ZB} \\ \mathbf{G}_C\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{G}_C & \mathbf{ZB} \\ \mathbf{BZ}'\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{BZ}' & \mathbf{B} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_{nn} + \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}((\mathbf{G}_m + \mathbf{C}) - \mathbf{A}_{gg})\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}(\mathbf{G}_m + \mathbf{C}) & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{ZB} \\ (\mathbf{G}_m + \mathbf{C})\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & (\mathbf{G}_m + \mathbf{C}) & \mathbf{ZB} \\ \mathbf{BZ}'\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{BZ}' & \mathbf{B} \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{G}_m\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{G}_m & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{ZB} \\ \mathbf{G}_m\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{G}_m & \mathbf{ZB} \\ \mathbf{BZ}'\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{BZ}' & \mathbf{B} \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{A}_{nn} - \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \\ &+ \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{C}\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{C} & \mathbf{0} \\ \mathbf{C}\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}. \end{aligned}$$

This is the same definition of the (co)variance structure matrix used in Christensen and Lund [2] when  $\mathbf{C} = w\mathbf{A}_{gg}$  in  $\mathbf{H}_L$ . When  $\mathbf{C} = \mathbf{0}$  in  $\mathbf{H}_L$ , we have the same definition of the (co)variance structure matrix that is used in Legarra et al. [15] and Fernando et al. [6]. Thus, on the one hand, changes in the regularization matrix  $\mathbf{C}$  affect the part of the genetic covariance not including the SNP effects. On the other hand, changes in the scaling matrix  $\mathbf{B}$  affect only the marker effect (co)variances. Furthermore, it should be noted that the upper left two by two block of matrices in  $\mathbf{H}_L$  is the same as the  $\mathbf{H}$  of ssGBLUP. This further illustrates the auxiliary nature of the vector of SNP effects  $\mathbf{g}$  in the described ssSNPBLUP. In other words, the breeding values in ssGBLUP and ssSNPBLUP have the same covariance structure, which have been augmented with

the marker covariances in ssSNPBLUP. It is worth noting that MME (3) cannot be derived for  $\mathbf{C} = \mathbf{0}$  because  $\mathbf{C} = \mathbf{0}$  is not invertible.

### Indirect prediction of GEBV for genotyped selection candidates

Computation of GEBV for the genotyped selection candidates, i.e., genotyped animals without own and progeny records, using solutions from the previous ssGBLUP evaluation facilitates earlier obtention of selection candidate predictions than waiting for the next full data genetic evaluation. Furthermore, to reduce the computational costs of the single-step genomic evaluations, it can be of interest [16] to ignore the genotypes of animals

without own and progeny records and to predict indirectly their GEBV afterwards by using the solutions of the latest single-step genomic evaluation.

Computation of indirect GEBV predictions can use the decomposition of GEBV  $\mathbf{u}' = [\mathbf{u}'_n \ \mathbf{u}'_g]$  corresponding to the (co)variance structure matrix  $\mathbf{H}_L$ , as follows:

$$\begin{bmatrix} \mathbf{u}_n \\ \mathbf{u}_g \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{Zg} \\ \mathbf{Zg} \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{d}_n \\ \mathbf{d}_g \end{bmatrix}, \tag{4}$$

where

$$\begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{Zg} \\ \mathbf{Zg} \end{bmatrix} \sim MVN \left( \mathbf{0}, \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{G}_m\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{G}_m \\ \mathbf{G}_m\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{G}_m \end{bmatrix} \sigma_u^2 \right),$$

$\boldsymbol{\epsilon} \sim MVN \left( \mathbf{0}, (\mathbf{A}_{nn} - \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn}) \sigma_u^2 \right)$  is the vector of imputation residuals, and



$\mathbf{d} = \begin{bmatrix} \mathbf{d}_n \\ \mathbf{d}_g \end{bmatrix} \sim MVN\left(\mathbf{0}, \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{C}\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1}\mathbf{C} \\ \mathbf{C}\mathbf{A}_{gg}^{-1}\mathbf{A}_{gn} & \mathbf{C} \end{bmatrix} \sigma_u^2\right)$  is a vector that corresponds to the part of the genetic effects not explained by the genomic data.

If  $\mathbf{C} = \varepsilon\mathbf{I}$ , the  $\mathbf{d}$  vector is usually neglected. However, when  $\mathbf{C} = w\mathbf{A}_{gg}$ , the  $\mathbf{d}$  vector corresponds to the RPG effects [2] which account for the genetic variation that is not accounted for by the markers. Thus, for the genotyped animals, the GEV can be decomposed into two components: the direct genetic value due to the marker effects,  $\mathbf{Z}_g\mathbf{g}$ , and the estimated breeding value due to the RPG effects  $\mathbf{d}_g$  [2, 12, 17].

Based on Eq. (4), the GEV of genotyped selection candidates can be predicted using already computed GEV ( $\hat{\mathbf{u}}$ ) and SNP solutions ( $\hat{\mathbf{g}}$ ) from ssSNPBLUP. GEV of the genotyped selection candidates consist of two components: the direct genetic value due to the marker effects (i.e.,  $\mathbf{Z}_c\hat{\mathbf{g}}$ ) and the estimated breeding value due to the RPG effects (i.e.,  $\hat{\mathbf{d}}_c$ ). Thus, the GEV of the genotyped selection candidates can be computed as  $\hat{\mathbf{u}}_c = \mathbf{Z}_c\hat{\mathbf{g}} + \hat{\mathbf{d}}_c$ . When ssGT-BLUP has been used, the marker solutions  $\hat{\mathbf{g}}$  can be easily calculated in a post-processing step after solving the MME (2) by using the Eq. (17) in Liu et al. [7]:

$$\hat{\mathbf{g}} = \mathbf{K}^{-1}\mathbf{Z}'\mathbf{C}^{-1}\hat{\mathbf{u}}_g. \tag{5}$$

The computation of GEV for the selection candidates ( $\hat{\mathbf{u}}_c$ ) is straightforward when the model has no RPG effects (i.e.,  $\mathbf{d}_c = \mathbf{0}$ ) or if the effect of the regularization matrix  $\mathbf{C}$  can be ignored (i.e.,  $\mathbf{d}_c \approx \mathbf{0}$ ). For these cases, the GEV of the genotyped selection candidates can be directly computed using their centered genotypes and the estimated marker effect solutions, as  $\hat{\mathbf{u}}_c = \mathbf{Z}_c\hat{\mathbf{g}}$ .

### Efficient computation of the $\mathbf{d}_c$ effects

While the direct genetic values  $\mathbf{Z}_c\hat{\mathbf{g}}$  due to the marker effects can be easily computed for the genotyped selection candidates, the estimated breeding values due to the RPG effects  $\hat{\mathbf{d}}_c$  are not directly available from the solutions of the latest single-step evaluation, and must therefore be computed. When  $\mathbf{C} = w\mathbf{A}_{gg}$ , Liu et al. [13] showed that the RPG effects for the genotyped selection candidates can be computed as:

$$\hat{\mathbf{d}}_c = w\mathbf{A}_{cg}\mathbf{G}_C^{-1}\hat{\mathbf{u}}_g = \mathbf{A}_{cg}\mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g, \tag{6}$$

where  $\mathbf{A}_{cg}$  is the pedigree-based relationship matrix between the genotyped selection candidates and the genotyped animals already included in the latest single-step evaluation, and the RPG effects for the genotyped animals are computed as:  $\hat{\mathbf{d}}_g = \hat{\mathbf{u}}_g - \mathbf{Z}_g\hat{\mathbf{g}}$ .

The computation of  $\hat{\mathbf{d}}_c$  using Eq. (6) can be done using the Eq. (22) of Fernando et al. [6] as follows:

$$\begin{bmatrix} \mathbf{s}_o \\ \hat{\mathbf{d}}_c \end{bmatrix} = -\begin{bmatrix} \mathbf{A}^{oo} & \mathbf{A}^{oc} \\ \mathbf{A}^{co} & \mathbf{A}^{cc} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{A}^{og} \\ \mathbf{A}^{cg} \end{bmatrix} \hat{\mathbf{d}}_g,$$

where  $\begin{bmatrix} \mathbf{A}^{oo} & \mathbf{A}^{og} & \mathbf{A}^{oc} \\ \mathbf{A}^{go} & \mathbf{A}^{gg} & \mathbf{A}^{gc} \\ \mathbf{A}^{co} & \mathbf{A}^{cg} & \mathbf{A}^{cc} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{oo} & \mathbf{A}_{og} & \mathbf{A}_{oc} \\ \mathbf{A}_{go} & \mathbf{A}_{gg} & \mathbf{A}_{gc} \\ \mathbf{A}_{co} & \mathbf{A}_{cg} & \mathbf{A}_{cc} \end{bmatrix}^{-1}$  is the inverse of the pedigree relationship matrix partitioned among the genotyped animals ( $g$ ) included in the latest single-step genomic evaluation, the selection candidates ( $c$ ) and the non-genotyped ancestors ( $o$ ) of the genotyped animals and selection candidates, and  $\mathbf{s}_o$  is the vector of estimated breeding values due to the RPG effects for non-genotyped ancestors of the genotyped animals and selection candidates.

To avoid the solving of a system involving the matrix  $\begin{bmatrix} \mathbf{A}^{oo} & \mathbf{A}^{oc} \\ \mathbf{A}^{co} & \mathbf{A}^{cc} \end{bmatrix}$ , the computation of  $\hat{\mathbf{d}}_c$  using Eq. (6) can be done efficiently in two steps. First, calculate  $\mathbf{x} = \mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$  for which efficient sparse matrix computations can be used [18]. Second,  $\hat{\mathbf{d}}_c = \mathbf{A}_{cg}\mathbf{x}$  can be computed using the algorithm of Colleau [19] that performs the full pedigree-based relationship matrix times vector product, i.e.

$$\begin{bmatrix} \mathbf{s}_n \\ \hat{\mathbf{d}}_g \\ \hat{\mathbf{d}}_c \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{nn} & \mathbf{A}_{ng} & \mathbf{A}_{nc} \\ \mathbf{A}_{gn} & \mathbf{A}_{gg} & \mathbf{A}_{gc} \\ \mathbf{A}_{cn} & \mathbf{A}_{cg} & \mathbf{A}_{cc} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{x} \\ \mathbf{0} \end{bmatrix}.$$

The computation of  $\hat{\mathbf{d}}_c = \mathbf{A}_{cg}\mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$  can be further simplified by splitting it into separate computations for two sets of animals. In the computation of  $\hat{\mathbf{d}}_c = \mathbf{A}_{cg}\mathbf{x}$  using the algorithm of Colleau [19], it can be noted that  $\hat{\mathbf{d}}_c$  depends only on the estimated breeding values due to the RPG effects of the genotyped animals and all their ancestors. Thus, first, the vector of estimated breeding values due to RPG for the non-genotyped ancestors of the genotyped animals,  $\hat{\mathbf{d}}_{anc_g}$ , can be computed as:

$$\hat{\mathbf{d}}_{anc_g} = -(\mathbf{A}^{anc_g,anc_g})^{-1}\mathbf{A}^{anc_g,g}\hat{\mathbf{d}}_g,$$

where the matrix  $\begin{bmatrix} \mathbf{A}^{anc_g,anc_g} & \mathbf{A}^{anc_g,g} \\ \mathbf{A}^{g,anc_g} & \mathbf{A}^{gg} \end{bmatrix}$  is the inverse of the pedigree relationship matrix among the genotyped animals included in the latest single-step genomic evaluation and all their ancestors ( $anc_g$ ). Second, the vector of estimated breeding values due to RPG for the genotyped selection candidates,  $\hat{\mathbf{d}}_c$ , can be computed by calculating the parent average of the estimated breeding values due to RPG from the oldest to the youngest animal included in the pedigree of the genotyped selection candidates.

The presented formulas are based on derivations from the MME of the ssSNPBLUP model and therefore yield exact solutions. We introduce a regression-based approach for the RPG part of GEV  $\hat{\mathbf{d}}_c$  that allows an even simpler computational approach than the one

presented. The RPG term  $\widehat{\mathbf{d}}_c$  can be estimated by the mean of parent RPG effect values. The parent average uses  $\widehat{\mathbf{d}}_g$  for the genotyped animals but values from a regression equation are used for the non-genotyped parents. Thus, for the genotyped parents, the RPG part for the genotyped reference animals is computed as  $\widehat{\mathbf{d}}_g = \widehat{\mathbf{u}}_g - \mathbf{Z}\widehat{\mathbf{g}}$ . For the non-genotyped parents,  $\widehat{\mathbf{d}}_n$  is approximated by  $\widehat{\mathbf{d}}_n = \widehat{a} + \widehat{b}\widehat{\mathbf{u}}_n$  where the coefficients  $\widehat{a}$  and  $\widehat{b}$  are estimated by linear regression of  $\widehat{\mathbf{d}}_g$  on  $\widehat{\mathbf{u}}_g$ , that is  $\widehat{\mathbf{d}}_g = \widehat{a} + \widehat{b}\widehat{\mathbf{u}}_g$ . Thus,  $\widehat{\mathbf{d}}_c$  is approximated by a parent average using values in the vector  $\widetilde{\mathbf{d}} = [\widetilde{\mathbf{d}}_n' \ \widetilde{\mathbf{d}}_g']'$ . Then, GEBV of selection candidates can be calculated as  $\widehat{\mathbf{u}}_c = \mathbf{Z}_c\widehat{\mathbf{g}} + \widetilde{\mathbf{d}}_c$ .

**Consideration of other effects in the indirect prediction of GEBV**

In addition to the direct genetic values due to the marker effects ( $\mathbf{Z}_c\widehat{\mathbf{g}}$ ) and the estimated breeding values due to the RPG effects ( $\widehat{\mathbf{d}}_c$ ) when  $\mathbf{C} = w\mathbf{A}_{gg}$ , the GEBV may also include other effects, such as the contributions of a covariate that models the difference from the pedigree base to the genomic base (hereinafter called J-factor; [20–22]), or the contributions of genetic groups in the model. For example, if a J-factor is fitted in the model, Eq. (4) for GEBV becomes:

$$\begin{bmatrix} \mathbf{u}_{j,n} \\ \mathbf{u}_{j,g} \end{bmatrix} = \begin{bmatrix} -\mathbf{A}_{ng}\mathbf{A}_{gg}^{-1} \\ -\mathbf{I} \end{bmatrix} \mathbf{1}\mu + \begin{bmatrix} \mathbf{A}_{ng}\mathbf{A}_{gg}^{-1} \\ \mathbf{I} \end{bmatrix} \mathbf{Z}\mathbf{g} + \begin{bmatrix} \boldsymbol{\epsilon} \\ \mathbf{0} \end{bmatrix} + \begin{bmatrix} \mathbf{d}_n \\ \mathbf{d}_g \end{bmatrix},$$

where  $\mathbf{1}$  is a vector of 1s, and  $\mu$  is the covariate that models the difference from the pedigree base to the genomic base. The GEBV for the genotyped selection candidates can therefore be computed as:

$$\widehat{\mathbf{u}}_{j,c} = -\mathbf{1}_c\widehat{\mu} + \mathbf{Z}_c\widehat{\mathbf{g}} + \widehat{\mathbf{d}}_c,$$

**Table 1** Number, means and standard deviations (SD) of records, and heritabilities for the six traits

Trait	Number of records	Mean	SD	$h_d^2$	$h_m^2$
1	2,099,743	1.30	0.60	0.13	0.04
2	7,085,063	1.20	0.50	0.07	0.02
3	984,905	1.39	0.71	0.16	0.08
4	6,255,183	1.26	0.57	0.14	0.08
5	2,280,213	3.16	0.68	0.19	0.05
6	232,582	41.51	7.94	0.14	0.03

<sup>1</sup>  $h_d^2$  = heritability of the direct additive genetic effect

<sup>2</sup>  $h_m^2$  = heritability of the maternal additive genetic effect

with  $\widehat{\mu}$  being the solution estimated in the latest single-step genomic evaluation. This approach can also be used for genetic groups.

It is worth noting that our proposed indirect approach is similar to previously proposed indirect approaches (e.g., [11, 12, 23]). However, our proposed approach is different in the sense that all GEBV components are computed without approximation for the indirect computation of GEBV for selection candidates. Further details on the similarities and differences between our approach and previously proposed approaches can be found in the "Discussion" section.

**Data and models**

The single-step genomic evaluations and indirect prediction approaches were tested using data from the routine six-trait calving-difficulty evaluation for Irish dairy and beef cattle performed by Irish Cattle Breeding Federation (ICBF; Ireland) in March 2022. The single-step genomic evaluations were based on the same multi-trait animal model and variance components as the current official routine breeding value evaluation described in more detail in Evans et al. [24].

After extraction and editing, the data file included 16.59 million data records (across the 6 traits), and the pedigree included 26.46 million animals. The number of records per trait is in Table 1. The genotypes of 2.61 million genotyped animals included 47,006 SNPs on 29 bovine autosomes, with a minor allele frequency higher or equal to 0.01. The genotype data was from 30 different arrays ranging in size from 3 to 850K SNPs. However, 91% were from the International Beef and Dairy (IDB) customised chip with an array density between 50 and 54K and within those IDB chips 55% were Illumina Bead Chip technology (Illumina, San Diego, USA) and the remaining 45% were ThermoFisher Scientific Microarray technology (ThermoFisher Scientific Waltham, MA, USA). Missing SNP genotypes were imputed using FImpute [25] to a 50K SNP set based on version 3 of the IDB chip. Among the 2.61 million genotyped animals, 457,171 genotyped animals were without their own and progeny records. The remaining 2.16 million genotyped animals had either their own records or descendants with records.

The six-trait linear mixed effects model included random effects (additive direct and maternal genetic, contemporary group, and residual effects), fixed covariables for direct breed proportion (n=22), dam breed proportion (n=22), specific heterosis coefficients (n=13), age of dam (primiparous), age nested with parity (multiparous), and fixed cross-classified effects for birth year and sex of calf. For all single-step genomic evaluations, an additional J-factor, which is

a fixed covariable that models the difference from the pedigree base to the genomic base, was fitted separately for the direct and maternal genetic effects of each trait [26]. A single J-factor was fitted for all breeds following Aldridge et al. [27]. The genotype matrix was centred using observed allele frequencies computed over all breeds.

**Study design**

**Single-step genomic evaluations**

The ssGTBLUP (hereinafter called ssGTABLUP when an RPG effect is fitted) and ssSNPBLUP models used an RPG effect. The RPG proportion was equal to  $w = 0.20$ . The two models were used to compute GEBV from a full and a reduced dataset. First, the models were solved with the full dataset that included all phenotypic and genomic information, i.e., 2.61 million genotypes (i.e., including genotypes of the selection candidates). Second, both models were solved with a reduced dataset that did not include genotypes of the selection candidates, i.e., there were 2.16 million genotypes. Thus, the reduced dataset analysis included all phenotypic information, but only the genotypes of animals with their own records or with descendants (across all generations) with records. For both datasets, the pedigree was extracted for the phenotyped animals and the selected set of genotyped animals.

All models were solved with the software MiXBLUP 3.0 [28] using the solver hplblup, which used the PCG method for solving the MME. The convergence efficiency of the PCG method relies mainly on the so-called preconditioner  $\mathbf{P}$ . In this study, for both ssGTABLUP (MME (2)) and ssSNPBLUP (MME (3)), the submatrix of  $\mathbf{P}$  corresponding to the fixed effects,  $\mathbf{P}_{ff}$ , was equal to  $\mathbf{P}_{ff} = \mathbf{X}'\mathbf{X} + \text{diag}(\mathbf{X}'\mathbf{X}) * 10^{-4}$  with  $\text{diag}(\mathbf{X}'\mathbf{X})$  corresponding to the diagonal elements of  $\mathbf{X}'\mathbf{X}$ . This addition to the diagonal elements ensures that  $\mathbf{P}_{ff}$  is positive definite, as required for its Cholesky decomposition. The submatrix of  $\mathbf{P}$  corresponding to the random effects,  $\mathbf{P}_{rr}$ , included for both ssGTABLUP and ssSNPBLUP a block-diagonal matrix with blocks corresponding to equations for different traits within a level (e.g., an animal). While the original and component-wise ssGTABLUP approaches have the same coefficient matrix (MME (2)), the block-diagonal matrix of  $\mathbf{P}_{rr}$  corresponding to  $\mathbf{u}_g$  were different for these two approaches because the contributions of the diagonal elements of  $\frac{1}{w^2}\mathbf{A}_{gg}^{-1}\mathbf{Z}\mathbf{K}^{-1}\mathbf{Z}'\mathbf{A}_{gg}^{-1}$  (being a term of  $\mathbf{G}_a^{-1} - \mathbf{A}_{gg}^{-1} = \left(\frac{1}{w} - 1\right)\mathbf{A}_{gg}^{-1} - \frac{1}{w^2}\mathbf{A}_{gg}^{-1}\mathbf{Z}\mathbf{K}^{-1}\mathbf{Z}'\mathbf{A}_{gg}^{-1}$ ) were not computed and, thus, were not available for inclusion in for the component-wise ssGTABLUP approach. For ssSNPBLUP, it is worth noting that, for the SNP effects, the  $i$ -th diagonal element of  $\mathbf{Z}'\mathbf{A}_{gg}^{-1}\mathbf{Z}$  of  $\mathbf{K} = \frac{1}{w}\mathbf{Z}'\mathbf{A}_{gg}^{-1}\mathbf{Z} + \mathbf{B}^{-1}$  was approximated by  $2p_i(1 - p_i)n$  [29], and that a second-level

diagonal preconditioner was also included, as in Vandenplas et al. [29].

The software MiXBLUP supports reading genomic information in the Plink 1 binary form [30], and for both the full and reduced datasets the genotypes were provided in this form. Both the original and component-wise approaches for solving ssGTABLUP with an RPG effect are implemented in MiXBLUP. Briefly, for the original ssGTABLUP approach, the solver hplblup requires a matrix equal to  $\mathbf{T}_a = \frac{1}{w}\mathbf{L}^{-1}\mathbf{Z}'\mathbf{A}_{gg}^{-1}$  with  $\mathbf{L}$  being the Cholesky decomposition of  $\mathbf{K} = \frac{1}{w}\mathbf{Z}'\mathbf{A}_{gg}^{-1}\mathbf{Z} + \mathbf{B}^{-1}$ . The  $\mathbf{T}_a$  matrix was computed using double-precision arithmetic with the program calc\_grm [31], i.e.,  $\mathbf{T}_a$  used  $8nm$  bytes. In the solving phase,  $\mathbf{T}_a$  was stored in the random access memory (RAM) to allow efficient parallel computations using multi-threading. For the component-wise ssGTABLUP approach, the solver hplblup only requires  $\mathbf{L}$  and  $\mathbf{M}$ , both stored in memory. The marker matrix  $\mathbf{M}$  is stored in RAM using the Plink 1 binary form that requires  $nm/4$  bytes [26].

For both ssGTABLUP approaches, the SNP effects  $\hat{\mathbf{g}}$  were computed by the solver after the end of the PCG iterative process using  $\mathbf{T}_a$  as  $\hat{\mathbf{g}} = \mathbf{L}^{-1}\mathbf{T}_a\hat{\mathbf{u}}_g$  (derived from Eq. (5)) for the original approach, and as  $\hat{\mathbf{g}} = \frac{1}{w}\mathbf{L}^{-1}\mathbf{L}^{-1}\mathbf{Z}'\mathbf{A}_{gg}^{-1}\hat{\mathbf{u}}_g$  for the component-wise approach. This strategy allows the computation of SNP effects in less time than needed for one PCG iteration.

Similarly to the component-wise ssGTABLUP approach, the ssSNPBLUP approach allows direct use of the marker matrix  $\mathbf{M}$  for the multiplication of the coefficient matrix by a vector in MME (3). Consequently, for solving the ssSNPBLUP model, the marker matrix  $\mathbf{M}$  was also stored in the RAM using the Plink 1 binary form [26].

The convergence criterion for the PCG iteration was the relative difference between the left- and right-hand sides of the MME:

$$C_r = \sqrt{\frac{(\mathbf{C}_{MME}\mathbf{s}^{[k]} - \mathbf{r}_{MME})'(\mathbf{C}_{MME}\mathbf{s}^{[k]} - \mathbf{r}_{MME})}{\mathbf{r}_{MME}'\mathbf{r}_{MME}}}$$

where  $\mathbf{C}_{MME}$  is the coefficient matrix of the MME,  $\mathbf{s}^{[k]}$  is the vector of solutions at round  $k$ , and  $\mathbf{r}_{MME}$  is the right-hand side vector. For all evaluations, convergence was assumed to be reached when  $C_r < 10^{-7}$ .

**Indirect approaches**

Four approaches for computing indirectly GEBV for the genotyped selection candidates using solutions from the reduced data single-step evaluations were compared with those calculated in the full data single-step evaluations. The indirect prediction approaches were:



**Table 2** Computational statistics of ssGTABLUP and ssSNPBLUP using the full and reduced datasets

Computational statistic	Original ssGTABLUP		Component-wise ssGTABLUP		ssSNPBLUP	
	Full	Reduced	Full	Reduced	Full	Reduced
T <sub>a</sub> matrix (or its components)	1004	832	1000	829	–	–
RAM	31.0	23.6	22.3	18.1	–	–
<sup>a</sup> Time (h)						
Neq	372	366	372	366	372	367
Number of PCG iterations	488	595	478	584	828	872
$\lambda_{min}(10^{-5})$	4.77	3.13	4.98	3.24	3.42	2.75
$\lambda_{max}$	2.98	2.98	2.98	2.98	3.59	3.34
RAM <sup>2</sup> (GB)	1015	845	102	95	86	79
Time/iteration (sec)	172.4	143.4	64.9	51.2	65.4	52.7
<sup>b</sup> Time (h)	26.4	25.9	9.9	9.5	16.4	13.7
Total time (h)	62.2	52.8	37.0	32.5	21.9	18.8

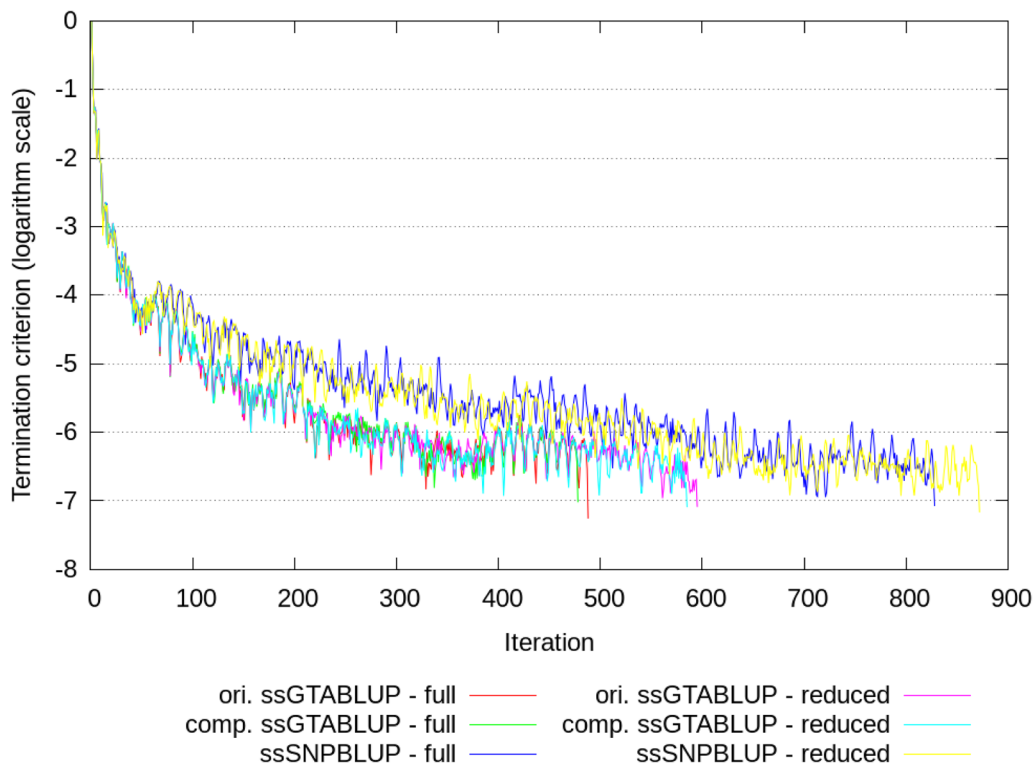
Neq = number of equations in millions; RAM = software peak random access memory (RAM) defined as the peak resident set size (VmHWM) obtained from the Linux/proc virtual file system;  $\lambda_{min}$  = smallest effective eigenvalue of the preconditioned coefficient matrix;  $\lambda_{max}$  = largest effective eigenvalue of the preconditioned coefficient matrix; Time/iteration = Average wall clock time per PCG iteration (expressed in seconds); Total time = wall clock time of the MiXBLUP software expressed in hours

<sup>a</sup>Time = wall clock time of the program calc\_grm expressed in hours

<sup>b</sup>Time = wall clock time of the solver expressed in hours

(1) the parent average (PA): mean of parent GEBV computed in a previous single-step approach; (2) the direct genomic values (DGV) computed as  $\mathbf{Z}_c \hat{\mathbf{g}}$ , using estimated SNP effects  $\hat{\mathbf{g}}$  from the reduced data single-step

evaluation; (3) the regression approach (REG): GEBV computed as  $\tilde{\mathbf{u}}_{j,c} = -\mathbf{1}\hat{\mu} + \mathbf{Z}_c \hat{\mathbf{g}} + \tilde{\mathbf{d}}_c$  with  $\tilde{\mathbf{d}}_c$  being approximated using the regression approach; and (4) exact computation of GEBV (that is, of genomic and



**Fig. 1** Convergence according to the termination criteria used (y axis in in  $\log_{10}$  units) during PCG iteration for ssSNPBLUP and the original (ori.) and component-wise (comp.) ssGTABLUP approaches with the reduced and full datasets

residual polygenic values; GRV): the GEBV computed as  $\hat{\mathbf{u}}_{j,c} = -\mathbf{1}\hat{\mu} + \mathbf{Z}_c\hat{\mathbf{g}} + \mathbf{A}_{cg}\mathbf{A}_{gg}^{-1}\hat{\mathbf{d}}_g$  with  $\hat{\mathbf{d}}_g = \hat{\mathbf{u}}_g - \mathbf{Z}_g\hat{\mathbf{g}}$ .

The PA approach is the simplest approach and can be considered as a reference. The DGV, REG, and GRV approaches were implemented in a Fortran 2018 program called indirectpred. This program requires the genotypes, the inbreeding coefficients, the SNP effect solutions, and the GEBV solutions for all animals included in the reduced data single-step evaluation, as well as the genotypes and the pedigree of the genotyped selection candidates.

GEBV computed indirectly for the genotyped selection candidates were compared with the GEBV computed from the full data single-step evaluations for both ssSNPBLUP and ssGTABLUP. For each trait and for both direct and maternal genetic effects, we calculated (1) the Pearson correlations between GEBV from the full data evaluations and the indirect GEBV hereinafter also called accuracy, (2) dispersion biases as the regression coefficients of the regression of GEBV from the full data evaluations on the indirect GEBV, and (3) level biases for each trait  $j$ , defined as the average of  $(\hat{\mathbf{u}}_{j,c} - \hat{\mathbf{u}}_{j,c,full})/\sigma_j$  where  $\hat{\mathbf{u}}_{j,c}$  and  $\hat{\mathbf{u}}_{j,c,full}$  are the indirect GEBV and the full data GEBV for the genotyped selection candidates, respectively, and  $\sigma_j$  is the genetic standard deviation of trait  $j$ .

All computations for ssSNPBLUP, ssGTABLUP and indirect approaches, were performed on a computer with 2.9 TB RAM and running RedHat 7.7 (x86\_64) with four Intel Xeon Gold 6242 (2.80 GHz) processors, each having 16 cores. The number of OpenMP threads used for all computations was equal to 10. All reported times are indicative, because they may have been influenced by other jobs running simultaneously on the computer.

## Results

### Performances of ssSNPBLUP and ssGTABLUP

The Pearson correlations for all traits between GEBV for ssSNPBLUP and ssGTABLUP were higher than 0.997, and all regression coefficients of the regression of GEBV for ssSNPBLUP on GEBV for ssGTABLUP were between 0.991 and 1.009, for both the reduced and full datasets. Thus, the GEBV for all direct and maternal traits of the ssSNPBLUP and ssGTABLUP evaluations were (almost) the same after convergence was reached for both the reduced and full datasets. Comparing GEBV obtained with the full dataset and the reduced dataset for animals present in both datasets resulted in Pearson correlations higher than 0.993 and regression coefficients of the regression of GEBV for the full dataset on GEBV of the reduced dataset between 0.991 and 1.014, for both ssGTABLUP and ssSNPBLUP (results not shown).

Computational statistics of ssGTABLUP and ssSNPBLUP using the full and reduced datasets are in Table 2.

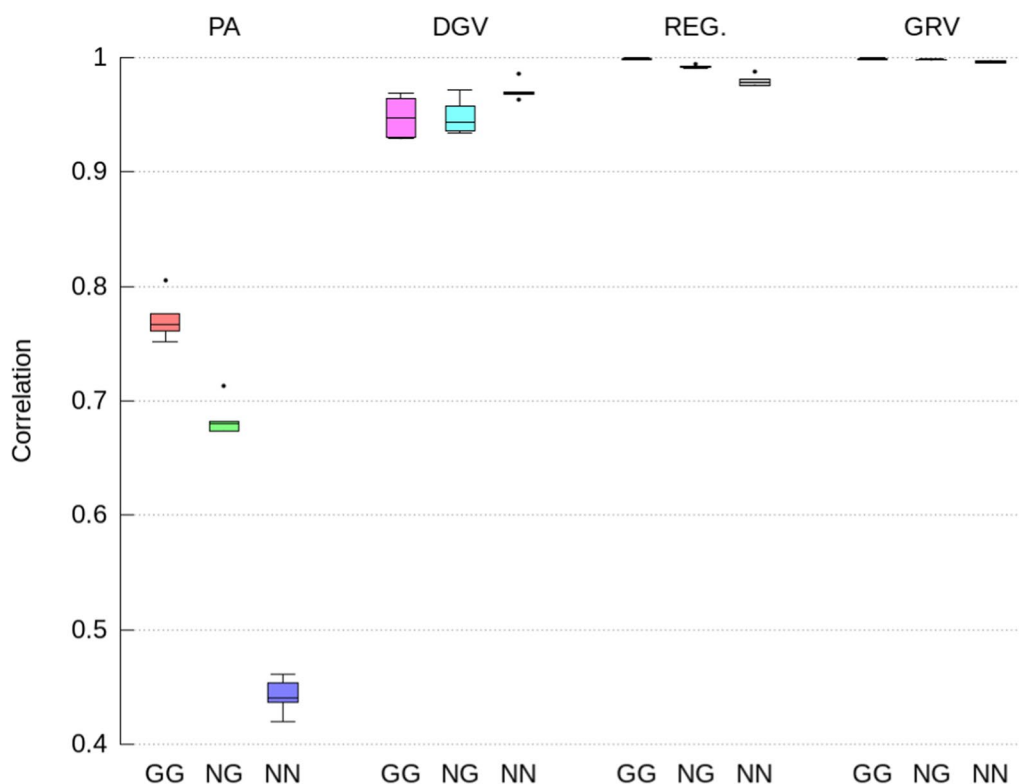
**Table 3** Computational statistics of indirect predictions of direct and maternal GEBV

Model	Approach	RAM (GB)	Total Time (h)
ssGTABLUP	REG	44	0.49
	GRV	50	0.55
ssSNPBLUP	REG	44	0.48
	GRV	50	0.52

REG. GEBV with approximated residual polygenic effects, GRV exact computation of GEBV. RAM = software peak random access memory (RAM) defined as the peak resident set size (VmHWM) obtained from the Linux/proc virtual file system

The MME of ssGTABLUP and ssSNPBLUP evaluations included about 372 million and 366 million equations with the full and reduced datasets, respectively. When analysing the full dataset, the solver required 1015 GB of RAM and 26 h with 488 iterations for the original ssGTABLUP, 102 GB of RAM and 10 h with 478 iterations for the component-wise ssGTABLUP, and 86 GB RAM and 16 h with 828 iterations for ssSNPBLUP (Fig. 1 and Table 2). Each iteration required on average 172 s for the original ssGTABLUP, and 65 s for both the component-wise ssGTABLUP and ssSNPBLUP. Analysing the reduced dataset resulted in a reduction of the computing time per iteration of about 17% for the original ssGTABLUP, and between 20 and 21% for the component-wise ssGTABLUP and ssSNPBLUP, although the number of iterations to reach convergence increased for both approaches with the reduced dataset. The estimated effective smallest eigenvalues were around  $10^{-5}$  for all systems of equations, while the estimated effective largest eigenvalues were equal to 2.98 for both ssGTABLUP, and a bit larger for ssSNPBLUP (i.e., 3.34 and 3.59 for the reduced and full dataset, respectively; Table 2).

Our implementation of the original ssGTABLUP requires the computation of an additional matrix based on the genomic and pedigree information, that is  $\mathbf{T}_a$ . The computation of  $\mathbf{T}_a$  was performed with the program calc\_grm and required 1004 GB and 31 h for the full dataset, and 832 GB and 24 h for the reduced dataset (Table 2). Our implementation of the component-wise ssGTABLUP requires the computation of  $\mathbf{L}$ . Like the computation of  $\mathbf{T}_a$ , the computation of  $\mathbf{L}$  was performed with the program calc\_grm and required 1000 GB and 22 h for the full dataset, and 829 GB and 18 h for the reduced dataset (Table 2). Finally, the complete evaluation that included, among others, the editing and renumbering of all files, the computation of pedigree-based inbreeding coefficients, the computation of the additional matrices for the two ssGTABLUP approaches, and solving of the MME, required 62 and 53 h for the full and reduced original ssGTABLUP, respectively, 37 and 33 h for the full and



**Fig. 2** Pearson correlations for direct GEBV computed from the full ssSNPBLUP and from the indirect prediction approaches for genotyped selection candidates with both parents genotyped (GG), with only one parent genotyped (NG), and no parents genotyped (NN). Indirect prediction approaches are: (1) PA: mean of parent GEBV; (2) DGV: direct genomic values; (3) REG: GEBV with approximated residual polygenic effects; and (4) GRV: exact computation of GEBV

reduced component-wise ssGTABLUP, respectively, and 22 and 19 h for the full and reduced ssSNPBLUP, respectively. Thus, the reduction in total computing time was around 15% for both ssGTABLUP and ssSNPBLUP when analysing the reduced instead of the full data (Table 2). However, the reduction for the solver step for both ssGTABLUP approaches was almost nil, due to the increase in the number of iterations to reach convergence.

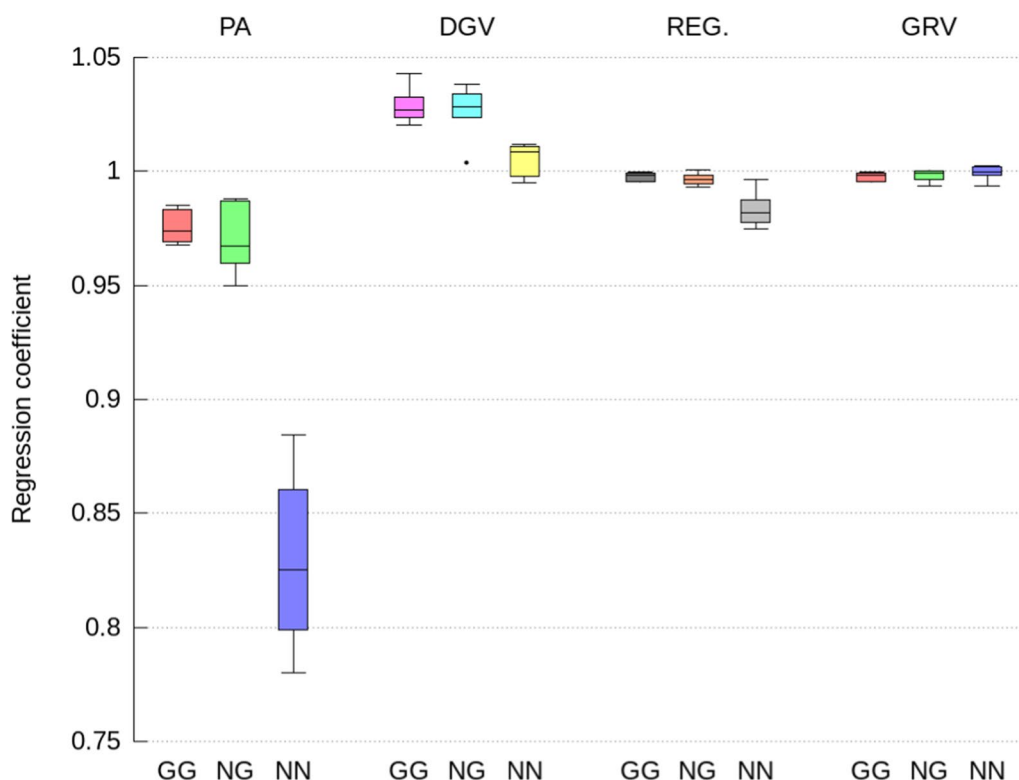
**Accuracies and bias of the different indirect prediction approaches**

The indirect predictions of direct and maternal GEBV for all genotyped selection candidates required about 50 GB RAM and 0.5 h for both the original ssGTABLUP and ssSNPBLUP (Table 3). Results for the component-wise ssGTABLUP are equivalent to those obtained with the original ssGTABLUP, because both ssGTABLUP rely on the same MME, and thus are not presented.

Accuracies, dispersion biases, and level biases of the indirect GEBV for the selection candidates with both, only one and no parents genotyped, were computed for both the direct and maternal genetic effects and all

traits separately (Figs. 2, 3, 4) and (see Additional file 1: Tables S1 to S6). Because ssGTABLUP and ssSNPBLUP arrived at (almost) the same GEBV, the results for indirect GEBV were also (almost) the same for ssGTABLUP and ssSNPBLUP. Therefore, this section will only present results for the indirect GEBV computed from ssSNPBLUP. All results for both ssGTABLUP and ssSNPBLUP are in Additional file 1: Tables S1–S6.

The indirect GEBV approximated by the GRV approach were associated with the highest accuracies (i.e., correlations higher than 0.996 on average), the lowest level biases, and no over- or under-dispersion (i.e., regression coefficient close to 1.0), across all traits, all genetic effects, and all categories of selection candidates (Figs. 2, 3, 4, 5, 6 and 7). In comparison, the REG approach resulted in the same accuracy and bias as the GRV approach for the selection candidates with both parents genotyped, as expected. However, for the selection candidates with only one or no genotyped parents, indirect GEBV computed with the REG approach for both genetic effects were slightly less accurate (correlations between 0.980 and 0.992 on average), with some dispersion bias (regression coefficients between 0.983



**Fig. 3** Regression coefficients of direct GEBV computed from the full ssSNPBLUP on GEBV computed from the indirect prediction approaches for genotyped selection candidates with both parents genotyped (GG), with only one parent genotyped (NG), and no parents genotyped (NN). Indirect prediction approaches are: (1) PA: mean of parent GEBV; (2) DGV: direct genomic values; (3) REG: GEBV with approximated residual polygenic effects; and (4) GRV: exact computation of GEBV

and 0.997), and with some level bias between  $-0.014$  and  $0.018$  points of genetic standard deviation (Figs. 2, 3, 4, 5, 6 and 7).

Approximating the GEBV of selection candidates with the DGV approach resulted in indirect GEBV associated with an average accuracy of 0.948 or higher across all groups of selection candidates (Figs. 2 and 5). In addition to being less accurate than the GRV and REG approaches, the indirect GEBV computed with the DGV approach showed also more level bias, which was almost equal to 0.2 points of genetic standard deviation across all traits (Figs. 4 and 7), and more dispersion, as shown by averaged regression coefficients between 1.005 and 1.032 (Figs. 3 and 6). Finally, the PA approach was the least accurate and showed the highest dispersion and level biases (Figs. 2, 3, 4, 5, 6 and 7).

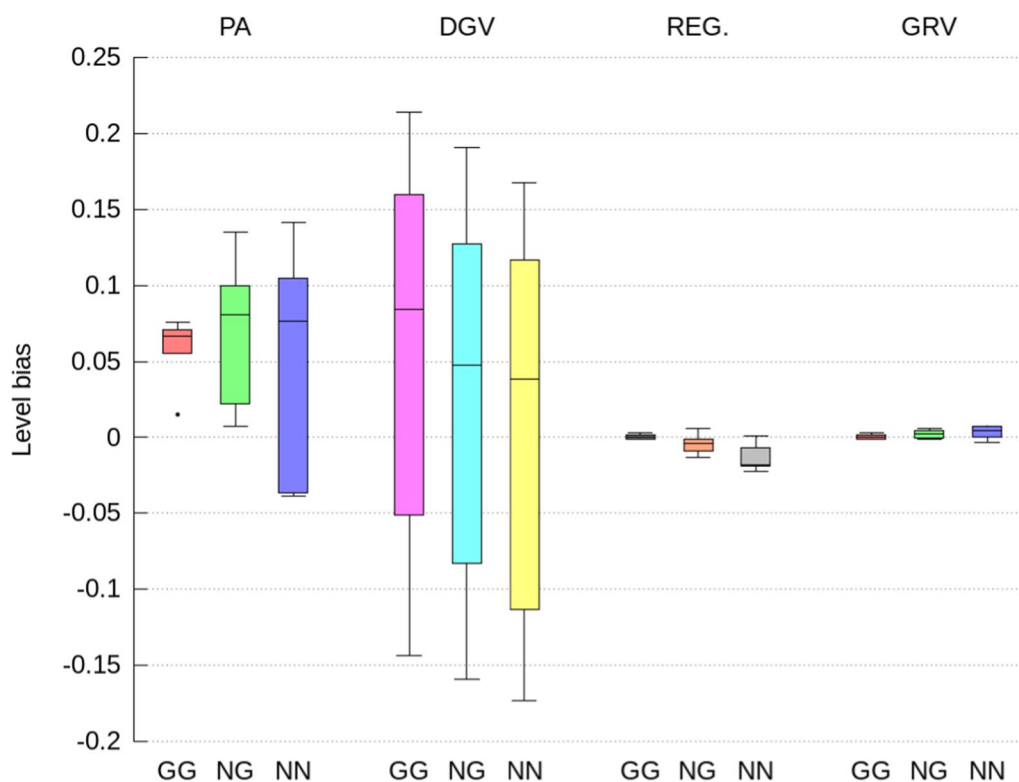
### Discussion

In this study, first, we presented a unified model for ssGTBLUP and ssSNPBLUP, and we introduced different approaches to predict GEBV of genotyped selection candidates based on solutions of a single-step

evaluation that does not consider the pedigree and genotypes of these selection candidates indirectly. The performance of the different models and methods was investigated using a dataset with a total of 2.61 million genotypes. In this section, we will discuss the following three points: (1) the computational similarities and differences between ssGTBLUP and ssSNPBLUP; (2) the indirect prediction of GEBV; and (3) the performance gains by ignoring genotypes of selection candidates in the single-step evaluation models.

### Computational similarities and differences between ssGTBLUP and ssSNPBLUP

Within a PCG iteration, the computations needed in MME (2) and (3) are theoretically similar. The main computational task in each iteration of the PCG method is the MME coefficient matrix times a vector product. Any differences in the necessary MME coefficient matrix times a vector product in MME (2) and (3) are due to differences in  $\mathbf{H}^{-1}$  and  $\mathbf{H}_L^{-1}$ . However, these products can be arranged so that they perform with similar efficiency, as shown with the component-wise approach of ssGTBLUP. The use of  $\mathbf{T}_a$  in the current implementation



**Fig. 4** Level bias of direct GEBV computed as the difference between the average of the indirect predictions and ssSNPBLUP solutions expressed in genetic standard deviation units, for genotyped selection candidates with both parents genotyped (GG), with only one parent genotyped (NG), and no parents genotyped (NN). Indirect prediction approaches are: (1) PA: mean of parent GEBV; (2) DGV: direct genomic values; (3) REG: GEBV with approximated residual polygenic effects; and (4) GRV: exact computation of GEBV

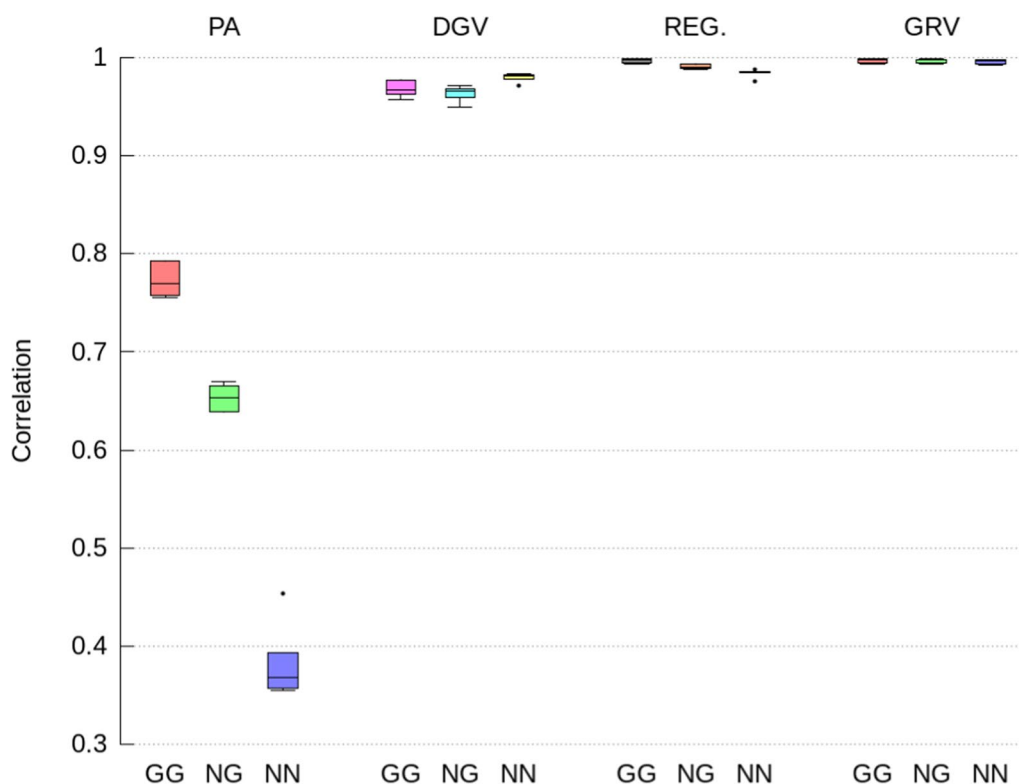
of MiXBLUP explains the RAM and time differences observed between the original ssGTABLUP and the component-wise ssGTABLUP and ssSNPBLUP in MiXBLUP. As expected, the component-wise ssGTABLUP approach and ssSNPBLUP require similar times per PCG iteration as well as similar RAM. The differences in RAM between these two approaches can be explained by the storage of  $L$  in RAM as a double precision dense matrix that requires  $8m^2$  bytes. Consequently, our results illustrate that the component-wise ssGTABLUP (MME (2)) approach and ssSNPBLUP (MME (3)) result in performing a similar number of floating-point operations.

Vandenplas et al. [26] showed that the multiplication of  $Z$  by an array with the use of  $M$  stored using the Plink 1 binary form was more than twice as fast compared to the same multiplication using the Intel MKL DGEMM subroutine, thanks to the efficient use of parallelization, vectorization, and CPU cache with the packed  $M$  matrix. The efficiency of the packed matrix operations is also indicated by the fact that ssGTABLUP and ssSNPBLUP approaches showed a similar reduction in computing time per iteration in the solver step when analysing the reduced data instead of the full data. This illustrates the

fact that both approaches behave numerically similarly when the number of genotyped animals changes.

Although the computations needed for ssGTABLUP and ssSNPBLUP within a PCG iteration are theoretically similar, both ssGTABLUP approaches needed a longer total computing time than for ssSNPBLUP. This longer time is mainly due to the heavy preprocessing step to compute  $T_a$  for the original ssGTABLUP approach and  $L$  for the component-wise ssGTABLUP approach. Both matrices are computed with the same procedure of the program calc\_grm, except that the last steps for computing  $T_a$  from  $L$  and writing it to a file, are skipped for the component-wise ssGTABLUP approach. This current implementation explains the reduced times and similar amounts of RAM for the preprocessing step of the component-wise ssGTABLUP approach in comparison to the original ssGTABLUP approach, as both computations use a genotype matrix stored in a 8-byte array. These preprocessing costs could be reduced by computing a single  $T_a$  or  $L$  for multiple genomic evaluations that share a common set of genotyped animals. Since ssSNPBLUP has no such preprocessing step, for this model it is more attractive to define minimal sets of required genotypes





**Fig. 5** Pearson correlations for maternal GEBV computed from the full ssSNPBLUP and from the indirect prediction approaches for genotyped selection candidates with both parents genotyped (GG), with only one parent genotyped (NG), and no parents genotyped (NN). Indirect prediction approaches are: (1) PA: mean of parent GEBV; (2) DGV: direct genomic values; (3) REG: GEBV with approximated residual polygenic effects; and (4) GRV: exact computation of GEBV

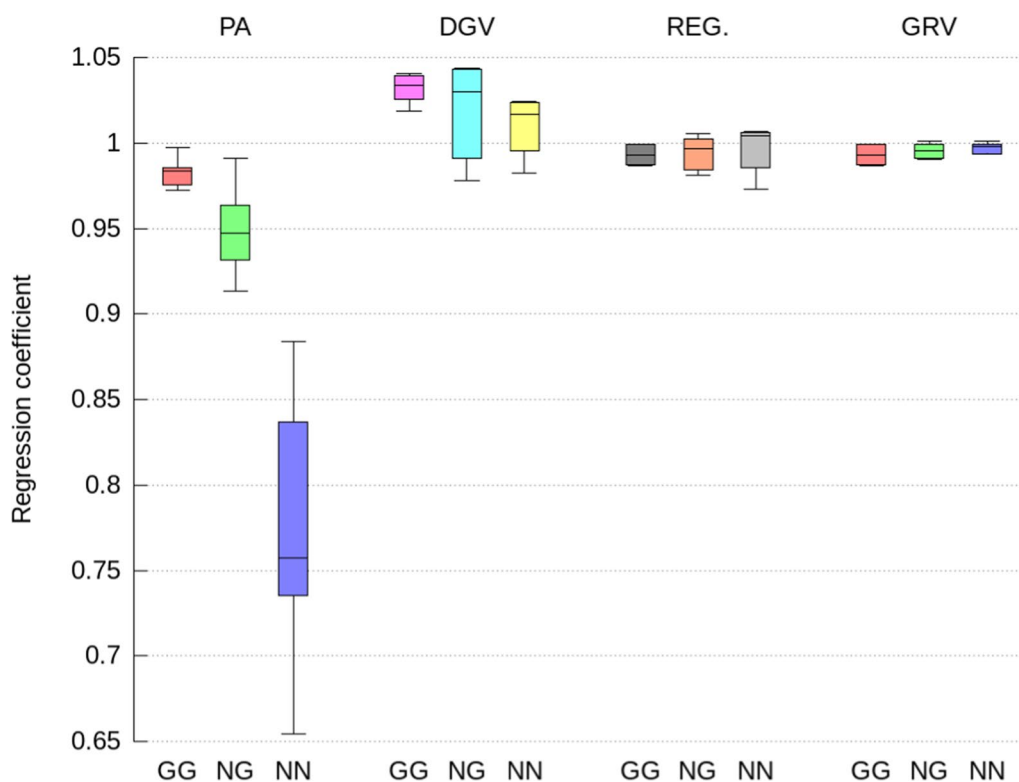
for each genomic evaluation separately, by removing genotypes of animals that had neither own nor progeny phenotypes, as done here.

Finally, both  $\mathbf{H}^{-1}$  and  $\mathbf{H}_L^{-1}$  contain  $\mathbf{K}$ . Because the size of this matrix is limited by the number of markers  $m$  when the SNP effects are assumed to have the same weights across all traits, this dense matrix can be stored in computer RAM. In MME (2), its inverse,  $\mathbf{K}^{-1}$ , or its Cholesky decomposition  $\mathbf{L}_C$ , can be precomputed to gain efficiency in the PCG iterations. It is worth noting that MME (3) can be more easily applicable for a model where  $\mathbf{K}$  is not the same across the traits because  $\mathbf{K}$  does not need to be inverted as in MME (2). Then, instead of precomputing  $\mathbf{K}$  for each trait, a computationally more efficient approach for MME (3) can be the computation of the needed matrix times vector product during the PCG iteration. Consider the product  $\mathbf{K}\mathbf{g} = \mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z}\mathbf{g} + \mathbf{B}^{-1}\mathbf{g}$ . Because the  $\mathbf{s} = \mathbf{Z}\mathbf{g}$  product needs to be computed at each iteration, and because the number of iterations is typically much smaller than the number of SNPs, the multiplication of  $\mathbf{Z}'(\mathbf{C}^{-1}\mathbf{s})$  during each iteration is less demanding than precomputing once the product  $\mathbf{Z}'\mathbf{C}^{-1}\mathbf{Z}$  [26].

The ssGTABLUP approach showed a better convergence than the ssSNPBLUP approach. This can be attributed to a better preconditioner in ssGTABLUP. In the ssGTABLUP, the diagonal of the MME matrix is easier to compute than in the ssSNPBLUP approach where the diagonal for the marker effects is approximated [26], because  $\mathbf{K}$  is not computed explicitly as done for ssGTABLUP. The computation of  $\mathbf{K}$  for ssSNPBLUP would lead to a similar preprocessing time as for ssGTABLUP and could allow faster convergence. However, the total computing time for ssSNPBLUP would be increased. This illustrates that the preconditioner is a compromise that is achieved within the tolerated preprocessing computing time.

**Indirect prediction of GEBV**

As shown by our results, the GRV and REG approaches proposed in this study for predicting indirectly and efficiently GEBV of genotyped selection candidates are accurate and (almost) unbiased. The accuracy and unbiasedness of the GRV and REG approaches can be



**Fig. 6** Regression coefficients of maternal GEBV computed from the full ssSNPBLUP on GEBV computed from the indirect prediction approaches for genotyped selection candidates with both parents genotyped (GG), with only one parent genotyped (NG), and no parents genotyped (NN). Indirect prediction approaches are: (1) PA: mean of parent GEBV; (2) DGV: direct genomic values; (3) REG: GEBV with approximated residual polygenic effects; and (4) GRV: exact computation of GEBV

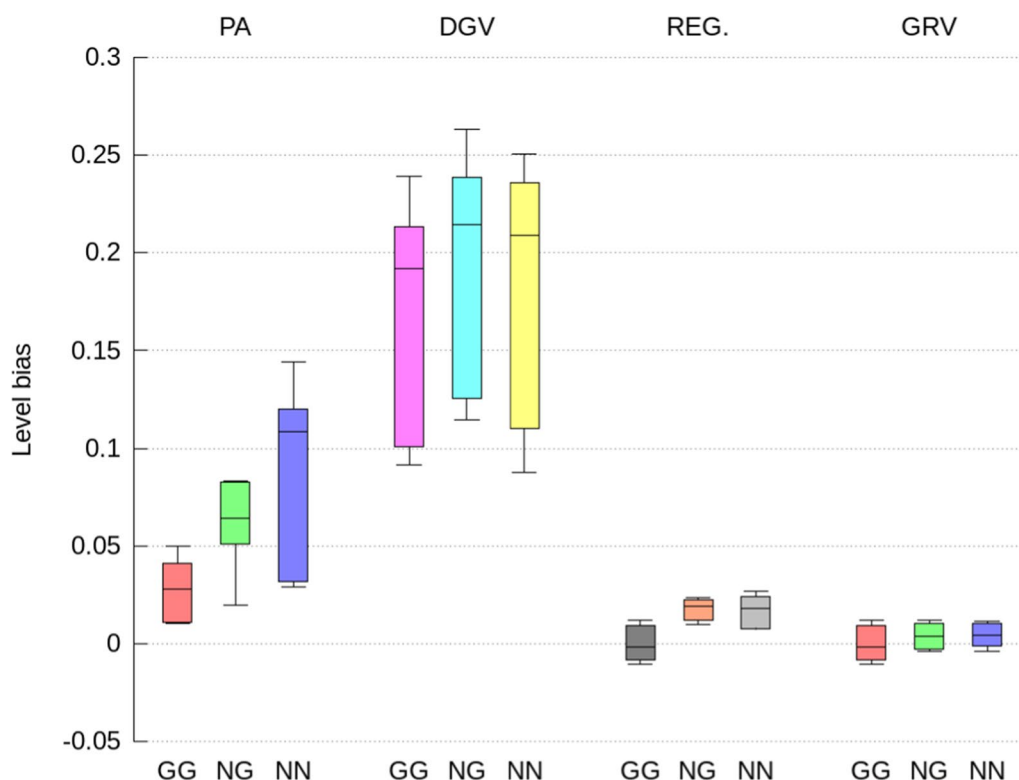
explained by the fact that all components of a GEBV are adequately computed by the proposed indirect prediction approaches. Approximating some components of the GEBV, such as the RPG and SNP effects, or ignoring some of them, such as the contribution of the J-factor, may result in less accurate and more biased indirect GEBV. Hereinafter we discuss alternatives illustrated by the indirect approaches that approximate the GEBV.

The SNP effects can be computed from GEBV estimated with ssGTBLUP without any approximations using Eq. (5), that is  $\hat{\mathbf{g}} = \mathbf{K}^{-1}\mathbf{Z}'\mathbf{C}^{-1}\hat{\mathbf{u}}_g$ . This Eq. (5) yields exact solutions for  $\hat{\mathbf{g}}$  and can be efficiently implemented for all versions of ssGBLUP, in contrast to the equation  $\hat{\mathbf{g}} = \mathbf{BZ}'\mathbf{G}_C^{-1}\hat{\mathbf{u}}_g$  requiring  $\mathbf{G}_C^{-1}$  and often used in the literature [12, 16, 23]. While it can be shown that the equation  $\hat{\mathbf{g}} = \mathbf{BZ}'\mathbf{G}_C^{-1}\hat{\mathbf{u}}_g$  is equivalent to Eq. (5), its implementation for large genotyped datasets requires an approximated  $\mathbf{G}_C^{-1}$  [16, 23], resulting in an approximated  $\hat{\mathbf{g}}$ . An alternative approach to compute  $\hat{\mathbf{g}}$  without computing  $\mathbf{G}_C^{-1}$  has been proposed by Pimentel et al. [12]. This approach consists of solving a SNPBLUP model with GEBV as phenotypes. However, in the presence of an

RPG effect, this approach still leads to an approximated  $\hat{\mathbf{g}}$  because it assumes implicitly that  $\mathbf{C} = \epsilon\mathbf{I}$  instead of  $\mathbf{C} = w\mathbf{A}_{gg}$ .

The computation of the RPG effects for the selection candidates can be efficiently implemented with our exact alternative approach to Eq. (6), or with the approximated REG approach that uses a mean parent RPG value when the genotyped parent's RPG value is known but a regression approach is used for a non-genotyped parent. Ignoring the RPG effects leads to dispersion bias of the indirect GEBV, as shown by the results for the DGV approach. The proposed exact approach gains efficiency because it requires neither  $\mathbf{G}_C^{-1}$  [16, 23] nor  $\mathbf{A}_{gg}^{-1}$  [12, 13], as previously proposed in the literature.

Although our REG approach resulted in accurate indirect GEBV for the selection candidates, other approaches for approximating the RPG effects for the selection candidates have been proposed or could be tested. For example, Lourenco et al. [23] proposed to use the linear regression of individual GEBV on direct genomic values, i.e., estimate  $a$  and  $b$  in the linear regression  $\hat{\mathbf{u}}_g = \hat{a} + \hat{b}\mathbf{Z}\hat{\mathbf{g}}$ , to compute directly GEBV of selection



**Fig. 7** Level bias of maternal GEBV computed as the difference between the average of the indirect predictions and ssSNPBLUP solutions expressed in genetic standard deviation units, for genotyped selection candidates with both parents genotyped (GG), with only one parent genotyped (NG), and no parents genotyped (NN). Indirect prediction approaches are: (1) PA: mean of parent GEBV; (2) DGV: direct genomic values; (3) REG.: GEBV with approximated residual polygenic effects; and (4) GRV: exact computation of GEBV

candidates as  $\hat{\mathbf{u}}_c = \hat{a} + \hat{b}\mathbf{Z}_c\hat{\mathbf{g}}$ . An alternative linear regression to that proposed in the "Methods" section here, would be to use the linear regression of individual on parent average GEBV, i.e., estimate  $a$  and  $b$  in the linear regression  $\hat{\mathbf{d}}_g = \hat{a} + \hat{b}\hat{\mathbf{u}}_{g,PA}$  using genotyped animal values for  $\hat{\mathbf{d}}_g$  and  $\hat{\mathbf{u}}_{g,PA}$ . Thus, the estimated coefficients  $\hat{a}$  and  $\hat{b}$  are used in the prediction equation for the candidate animals to estimate the RPG effects  $\hat{\mathbf{d}}_c = \hat{a} + \hat{b}\hat{\mathbf{u}}_{c,PA}$  where the  $\hat{\mathbf{u}}_{c,PA}$  vector has parent average GEBV for the genotyped selection candidates. This approach has the advantage that the mean parent GEBV is used directly without having to know the genotyping status of either of the parents. Another alternative is to compute the RPG effect of the selection candidates as the three sire parent averages when the dam is missing [12]. Finally, it is worth noting that we used an RPG proportion of 0.20 in this study. The use of an RPG proportion close to 0 will improve the accuracy and the dispersion bias of the indirect GEBV by decreasing the importance of  $\hat{\mathbf{d}}_c$  and increasing the importance of DGV.

No level bias was observed with the GRV method, and the bias was negligible for the REG approach. In the presence of level bias, animals with indirect GEBV cannot be compared to animals with GEBV computed by a single-step evaluation. Level bias can be attributed to unaccounted differences between the pedigree and genomic bases [12, 16, 23], as well as to approximated RPG effects (but with a smaller contribution than the first one), as illustrated by the DGV and REG approaches. The issue of level bias was solved in the literature with different approaches, such as by adding the mean GEBV of the genotyped animals of the previous single-step evaluation [23], or by adding a general mean that is estimated simultaneously with the estimates of SNP effects using GEBV of genotyped animals [12]. The mean computed by these two approaches can be considered as an approximation of the J-factor covariate fitted explicitly in this study. For ssGBLUP approaches in which the J-factor effect is fitted as a random effect and absorbed in the additive genetic effect [20], its value can be easily computed from the GEBV of the genotyped animals as:

$$\begin{aligned}\hat{\mu} &= k\mathbf{1}'\mathbf{G}_C^{-1}\hat{\mathbf{u}}_g = k\mathbf{1}'\left(\mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{Z}\mathbf{K}^{-1}\mathbf{Z}'\mathbf{C}^{-1}\right)\hat{\mathbf{u}}_g \\ &= k\mathbf{1}'\mathbf{C}^{-1}\left(\hat{\mathbf{u}}_g - \mathbf{Z}\hat{\mathbf{g}}\right) = k\mathbf{1}'\mathbf{C}^{-1}\hat{\mathbf{d}}_g,\end{aligned}$$

where  $k$  is a function of the proportion of RPG and the variance of the J-factor effect [20].

### Performance gains by ignoring the genotypes of selection candidates

Ignoring the genotypes of the selection candidates, that is genotyped animals without own or progeny records, can be an effective method to reduce the computational costs of the single-step genomic evaluations in some single-step genomic evaluations. In our study, ignoring 17% of the whole genotype set decreased the computing time for  $\mathbf{T}_a$ , and the computing time per iteration of ssGTABLUP and ssSNPBLUP by between 17 and 23%, while leading to almost the same GEBV for the animals included in both the reduced and full evaluations. The observed reductions in computing time per iteration matches our expectations. The main computational costs within each iteration are due to the multiplication of  $\mathbf{T}_a$  (or  $\mathbf{Z}$ ) and its transpose by an array, for which the computational costs depend linearly on the number of genotyped animals. Therefore, it will be easy to estimate the potential performance gains by ignoring the genotypes of selection candidates in routine single-step genomic evaluations.

It is worth noting that small GEBV differences can be expected between the reduced and the full single-step evaluations for non-genotyped parents of selection candidates, as the genotypes of their non-included progeny will not contribute to the imputation of their genotype within the single-step evaluations. The impact of this effect could be larger and should be further investigated, e.g., for species with large litters because the accuracy of the imputed genotype of a non-genotyped parent increases with the number of genotyped offspring [32].

### Conclusions

In this study, first, we presented a unified model for two single-step approaches, ssGTBLUP and ssSNPBLUP. Second, we presented different approaches to predict indirectly GEBV of genotyped selection candidates based on solutions of a single-step evaluation that does not consider the genotypes of these selection candidates. Based on our results, ignoring genotypes of selection candidates resulted in faster single-step evaluations, and the proposed indirect approaches resulted in accurate indirect GEBV for selection candidates, with almost no dispersion

and level bias. The proposed indirect approaches are also more memory efficient and computationally fast, compared to solving the single-step evaluations. Therefore, they can be computed even on a weekly basis to estimate GEBV for newly genotyped animals while the full single-step evaluation is done only a few times within a year.

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12711-023-00808-z>.

**Additional file 1: Table S1.** Pearson correlations, regression coefficients, and level bias for direct and maternal GEBV computed from the full ssSNPBLUP versus from the indirect prediction approaches for 251,332 genotyped selection candidates with both parents genotyped. **Table S2.** Pearson correlations, regression coefficients, and level bias for direct and maternal GEBV computed from the full ssSNPBLUP versus from the indirect prediction approaches for 155,675 genotyped selection candidates with only one genotyped parent. **Table S3.** Pearson correlations, regression coefficients, and level bias for direct and maternal GEBV computed from the full ssSNPBLUP versus from the indirect prediction approaches for 50,164 genotyped selection candidates with no genotyped parents. **Table S4.** Pearson correlations, regression coefficients, and level bias for direct and maternal GEBV computed from the full ssGTABLUP versus from the indirect prediction approaches for 251,332 genotyped selection candidates with both parents genotyped. **Table S5.** Pearson correlations, regression coefficients, and level bias for direct and maternal GEBV computed from the full ssGTABLUP versus from the indirect prediction approaches for 155,675 genotyped selection candidates with only one genotyped parent. **Table S6.** Pearson correlations, regression coefficients, and level bias for direct and maternal GEBV computed from the full ssGTABLUP versus from the indirect prediction approaches for 50,164 genotyped selection candidates with no genotyped parents.

### Acknowledgements

The use of the high-performance cluster was made possible by Irish Cattle Breeding Federation (P31D452, Link road, Ballincollig, Cork, Ireland).

### Author contributions

IS and JV conceived the study design. JV ran the tests, and wrote the programs. SND and RE prepared and provided the datasets. IS and JV wrote the first draft. IS, JV, and MC discussed and developed the theory. All authors provided valuable insights throughout the writing process. All authors read and approved the final manuscript.

### Funding

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food Project 16022 and LWV20054) 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).

### Declarations

#### Ethics approval and consent to participate

The data used for this study were collected as part of routine data recording for a commercial breeding program. Samples collected for DNA extraction were only used for the breeding program.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

**Author details**

<sup>1</sup>Wageningen University and Research, P.O. Box 338, 6700 AH Wageningen, The Netherlands. <sup>2</sup>Irish Cattle Breeding Federation, Highfield House, Newcestown Road, Bandon, Cork, Ireland. <sup>3</sup>Natural Resources Institute Finland (Luke), Jokioinen, Finland.

Received: 14 October 2022 Accepted: 28 April 2023

Published online: 08 June 2023

**References**

- Aguilar I, Misztal I, Johnson DL, Legarra A, Tsuruta S, Lawlor TJ. Hot topic: a unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J Dairy Sci.* 2010;93:743–52.
- Christensen OF, Lund MS. Genomic prediction when some animals are not genotyped. *Genet Sel Evol.* 2010;42:2.
- Mäntysaari EA, Koivula M, Strandén I. Symposium review: single-step genomic evaluations in dairy cattle. *J Dairy Sci.* 2020;103:5314–26.
- Misztal I, Lourenco D, Legarra A. Current status of genomic evaluation. *J Anim Sci.* 2020;98:skaa101.
- Misztal I, Legarra A, Aguilar I. Using recursion to compute the inverse of the genomic relationship matrix. *J Dairy Sci.* 2014;97:3943–52.
- Fernando RL, Dekkers JC, Garrick DJ. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet Sel Evol.* 2014;46:50.
- Liu Z, Goddard ME, Reinhardt F, Reents R. A single-step genomic model with direct estimation of marker effects. *J Dairy Sci.* 2014;97:5833–50.
- Mäntysaari EA, Evans RD, Strandén I. Efficient single-step genomic evaluation for a multibreed beef cattle population having many genotyped animals. *J Anim Sci.* 2017;95:4728–37.
- Legarra A, Ducrocq V. Computational strategies for national integration of phenotypic, genomic, and pedigree data in a single-step best linear unbiased prediction. *J Dairy Sci.* 2012;95:4629–45.
- Taskinen M, Mäntysaari EA, Strandén I. Single-step SNP-BLUP with on-the-fly imputed genotypes and residual polygenic effects. *Genet Sel Evol.* 2017;49:36.
- Lourenco DAL, Tsuruta S, Fragomeni BO, Masuda Y, Aguilar I, Legarra A, et al. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J Anim Sci.* 2015;93:2653–62.
- Pimentel ECG, Edel C, Emmerling R, Götz K-U. Technical note: methods for interim prediction of single-step breeding values for young animals. *J Dairy Sci.* 2019;102:3266–73.
- Liu Z, Goddard ME, Hayes BJ, Reinhardt F, Reents R. Technical note: equivalent genomic models with a residual polygenic effect. *J Dairy Sci.* 2016;99:2016–25.
- VanRaden PM. Efficient methods to compute genomic predictions. *J Dairy Sci.* 2008;91:4414–23.
- Legarra A, Aguilar I, Misztal I. A relationship matrix including full pedigree and genomic information. *J Dairy Sci.* 2009;92:4656–63.
- Tsuruta S, Lourenco DAL, Masuda Y, Lawlor TJ, Misztal I. Reducing computational cost of large-scale genomic evaluation by using indirect genomic prediction. *JDS Commun.* 2021;2:356–60.
- Legarra A, Lourenco DAL, Vitezica ZG. Bases for genomic prediction. 2022. <http://genoweb.toulouse.inra.fr/~alegarr/GSIP.pdf>. Accessed 10 Jan 2023.
- Strandén I, Matilainen K, Aamand GP, Mäntysaari EA. Solving efficiently large single-step genomic best linear unbiased prediction models. *J Anim Breed Genet.* 2017;134:264–74.
- Colleau J-J. An indirect approach to the extensive calculation of relationship coefficients. *Genet Sel Evol.* 2002;34:409–21.
- Vitezica ZG, Aguilar I, Misztal I, Legarra A. Bias in genomic predictions for populations under selection. *Genet Res (Camb).* 2011;93:357–66.
- Hsu W-L, Garrick DJ, Fernando RL. The accuracy and bias of single-step genomic prediction for populations under selection. *G3 (Bethesda).* 2017;7:2685–94.
- Strandén I, Aamand GP, Mäntysaari EA. Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding. *Genet Sel Evol.* 2022;54:38.
- Lourenco DAL, Legarra A, Tsuruta S, Moser D, Miller S, Misztal I. Tuning indirect predictions based on SNP effects from single-step GBLUP. *Interbull Bull.* 2018;53:48–53.
- Evans RD, Cromie AR, Pabiou T. Genetic evaluations for dam-type specific calving performance traits in a multi-breed population. In: Proceedings of the 70th Annual Meeting of the European Association for Animal Production: 26–30 August 2019. Ghent; 2019.
- Sargolzaei M, Chesnais JP, Schenkel FS. A new approach for efficient genotype imputation using information from relatives. *BMC Genomics.* 2014;15:478.
- Vandenplas J, Eding H, Bosmans M, Calus MPL. Computational strategies for the preconditioned conjugate gradient method applied to ssSNPBLUP, with an application to a multivariate maternal model. *Genet Sel Evol.* 2020;52:24.
- Aldridge M, Vandenplas J, Duenk P, Henshall J, Hawken R, Calus M. Validation with single-step SNPBLUP shows that evaluations can continue using a single mean of genotyped individuals, even with multiple breeds. *Genet Sel Evol.* 2023;55:19.
- ten Napel J, Vandenplas J, Lidauer MH, Strandén I, Taskinen M, Mäntysaari EA, et al. MIXBLUP 3.0.1 manual. V3.0. Wageningen: Wageningen University; 2021. <https://www.mixblup.eu/download.html>. Accessed 15 Jan 2023.
- Vandenplas J, Calus MPL, Eding H, Vuik C. A second-level diagonal preconditioner for single-step SNPBLUP. *Genet Sel Evol.* 2019;51:30.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience.* 2015;4:7.
- Calus MPL, Vandenplas J. Calc\_grm—a program to compute pedigree, genomic, and combined relationship matrices. Wageningen: Wageningen University; 2016.
- Shabalina T, Pimentel ECG, Edel C, Plieschke L, Emmerling R, Götz K-U. Short communication: the role of genotypes from animals without phenotypes in single-step genomic evaluations. *J Dairy Sci.* 2017;100:8277–81.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

