Integration of estimates of SNP effects into a single-step genomic evaluation

J. Vandenplas[1] and R. Bonifazi[1]

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

Corresponding author: jeremie.vandenplas@wur.nl

Abstract

The aim of this research was to develop and validate a method that integrates estimates of single nucleotide polymorphism (SNP) effects and the associated prediction error (co)variance (PECs) matrix from a genomic evaluation into a single-step SNP Best Linear Unbiased Prediction (ssSNPBLUP) evaluation. As the PEC matrix is a dense matrix, the developed method was also tested with two different chromosome-wise matrices (that is, ignoring off-diagonal elements among chromosomes), and with a prediction error variance matrix (that is, ignoring all off-diagonal elements of the PEC matrix). Using simulated data from two dairy cattle populations with a genetic correlation between their traits of 0.80, we compared the genomic enhanced breeding values (GEBVs) predicted by the different integration methods to those of a joint ssSNPBLUP evaluation of both populations. The developed method, using the whole PEC matrix, resulted in GEBVs for selection candidates highly correlated and consistent with those from the joint ssSNPBLUP evaluation. Ignoring off-diagonal elements among chromosomes resulted in similar accurate results, but ignoring all PECs resulted in biased GEBVs in comparison to those of the joint evaluation. Therefore, an accurate integration of estimates of SNP effects and the associated PEC matrix into a single-step genomic evaluation is feasible and accurate when PEC of SNP effects within chromosomes are at least considered. The developed method can be readily implemented in existing software that support ssSNPBLUP models and can be adapted for single-step genomic BLUP models, though further research is needed to address potential computational challenges with these models.

Key words: ssSNPBLUP, SNP effects, integration, Prediction Error Covariance

Introduction

For genomic evaluation in dairy cattle, single-step genomic models have emerged as the models of choice. A major advantage of these genomic prediction approaches is that they simultaneously analyze phenotypic and pedigree information of genotyped and non-genotyped animals with Single Nucleotide Polymorphism (SNP) genomic information of genotyped animals (Legarra *et al.*, 2014). Although the prediction of genomic enhanced breeding values (GEBVs) is the principal goal of the different equivalent single-step genomic evaluations, estimates of SNP effects can also be obtained for all of them, either simultaneously with the GEBV prediction (e.g., Fernando *et al.*, 2014; Liu *et al.*, 2014) or indirectly by back-solving GEBVs (e.g., Lourenco *et al.*, 2015). Models that directly predict GEBVs and SNP effects as random effects will hereafter be referred to as single-step SNP Best Linear Unbiased Prediction (ssSNPBLUP), while models that predict only GEBVs will hereafter be referred to as single-step GBLUP (ssGBLUP).

The exchange of genetic material among populations necessitates the comparison and combination of genetic and genomic evaluations across populations for animals of

interest. In dairy cattle, these needs have been addressed through meta-analysis approaches. These include, among others, the Multiple Across-Country Evaluation (MACE; Schaeffer, 1994), which combines individual-based pseudo-data of sires obtained from national genetic evaluations, the Genomic MACE (GMACE; VanRaden and Sullivan, 2010), which combines individual-based pseudo-data of sires derived from national genomic evaluations, and, more recently, the SNPMACE approaches (e.g., Jighly *et al.*, 2022; Kärkkäinen et al., 2024; Vandenplas *et al*., 2018), which combine SNP-based pseudo-data obtained from genomic evaluations. These meta-analyses facilitate the combination of genetic and genomic evaluations across multiple populations and an optimal use of all available across-country information in each population.

The increasing adoption of single-step genomic models for routine evaluations, along with the implementation of meta-analyses such as SNPMACE, may result in an increased exchange of estimates of SNP effects and their associated measures of precision, potentially replacing the exchange of individual-level pseudo-data (e.g., computed from (G)EBVs). This potential increased exchange of estimates of SNP effects and their associated measures of precision creates a need for methods that can accurately integrate them into national single-step genomic evaluations.

The objective of this research was to develop and validate a method that integrates external estimates of SNP effects and their associated measures of precision into a ssSNPBLUP evaluation. Our method was validated using simulated data from two dairy cattle populations. Results demonstrate that the developed method enables accurate integration of estimates of SNP effects into a single-step genomic evaluation.

Materials and Methods

To develop a method that integrates estimates of SNP effects and their associated measures of precision into a ssSNPBLUP evaluation, we consider two populations, respectively *A* and *B*, both associated with animals phenotyped and/or genotyped at identical SNP loci. We first describe in this section a population-specific ssSNPBLUP evaluation based on mixed model equations (MME) proposed by Liu *et al.* (2014). Second, we describe a joint ssSNPBLUP evaluation that simultaneously analyzes phenotypes and genotypes from both populations. Third, we describe a method to integrate estimates of SNP effects of population *B* into a ssSNPBLUP evaluation of population *A*, assuming an exact prediction error covariance (PEC) matrix is available for population *B*. Finally, we outline four approximations of the PEC matrix. In the second part of this section, we present the simulations used to validate these different methods.

Population-specific ssSNPBLUP

A standard univariate mixed model for a single-step genomic evaluation for population i ($i = A, B$) can be written as:

$$\mathbf{y}_i = \mathbf{X}_i\mathbf{b}_i^* + \mathbf{W}_i\mathbf{u}_i^* + \mathbf{e}_i^*, \qquad (1)$$

where $\mathbf{y}_i$ is the vector of records for population *i*, $\mathbf{b}_i^*$ is the vector of fixed effects, $\mathbf{u}_i^* = \left[\mathbf{u}_{n,i}^{*\prime} \quad \mathbf{u}_{g,i}^{*\prime}\right]'$ is the vector of additive genetic effects for non-genotyped (n) and genotyped (g) animals, respectively, and $\mathbf{e}_i^*$ is the vector of residuals. The matrices $\mathbf{X}_i$ and $\mathbf{W}_i$ are incidence matrices relating records to the corresponding effects.

Additive genetic effects of the genotyped animals, $\mathbf{u}_{g,i}^*$, for population *i* can be decomposed as $\mathbf{u}_{g,i}^* = \mathbf{a}_{g,i}^* + \mathbf{Z}_i\mathbf{g}_i^*$, where $\mathbf{a}_{g,i}^*$ is the vector of the residual polygenic (RPG) effects, $\mathbf{g}_i^*$ is the vector of SNP effects, and $\mathbf{Z}_i$ is the centered matrix of SNP genotypes (coded as 0 for one homozygous genotype, 1 for the heterozygous genotype, or 2 for the alternate homozygous genotype). We assume a

multivariate normal (MVN) distribution for the additive genetic effects $\mathbf{u}_i^*$ and the SNP effects $\mathbf{g}_i^*$, with a mean equal to zero and a covariance matrix $\mathbf{H}_i^*\sigma_{u,i}^2$ with $\mathbf{H}_i^*$ being the covariance structure matrix and $\sigma_{u,i}^2$ being the additive genetic variance for population *i*. Finally, we assume that $var(\mathbf{e}_i) = \mathbf{I}\sigma_{e,i}^2$ where $\mathbf{I}$ is an identity matrix, and $\sigma_{e,i}^2$ is the residual variance for population *i*.

The inverse of $\mathbf{H}_i^*$ for population *i*, $\mathbf{H}_i^{*-1}$, is equal to (Liu *et al.*, 2014):

$$\mathbf{H}_i^{*-1} = \begin{bmatrix} \mathbf{A}_i^{nn} & \mathbf{A}_i^{ng} & \mathbf{0} \\ \mathbf{A}_i^{gn} & \mathbf{A}_i^{gg} + \frac{1-w}{w}\mathbf{A}_{gg,i}^{-1} & -\frac{1}{w}\mathbf{A}_{gg,i}^{-1}\mathbf{Z}_i \\ \mathbf{0} & -\frac{1}{w}\mathbf{Z}_i'\mathbf{A}_{gg,i}^{-1} & \mathbf{K}_i^* \end{bmatrix}$$

where $\mathbf{A}_i^{-1} = \begin{bmatrix} \mathbf{A}_i^{nn} & \mathbf{A}_i^{ng} \\ \mathbf{A}_i^{gn} & \mathbf{A}_i^{gg} \end{bmatrix}$ is the inverse of the pedigree relationship matrix partitioned between non-genotyped and genotyped animals, $\mathbf{A}_{gg,i}$ is the pedigree relationship matrix among genotyped animals, w is the proportion of variance (due to additive genetic effects) considered as RPG effects, and $\mathbf{K}_i^* = \frac{1}{w}\mathbf{Z}_i'\mathbf{A}_{gg,i}^{-1}\mathbf{Z}_i + \frac{1}{1-w}\mathbf{B}^{-1}$ with $\mathbf{B}^{-1} = \mathbf{I}2\sum p_j(1-p_j)$ and p_j being the allele frequency of the *j*-th SNP.

It is worth noting that these assumptions lead to the following MVN distribution for the SNP effects for population *i*, $\mathbf{g}_i$:

$$\mathbf{g}_i \sim MVN\left(\mathbf{0}, \mathbf{B}\sigma_{g,i}^2\right)$$

with $\sigma_{g,i}^2 = (1-w)\sigma_{u,i}^2$.

Joint single-step genomic evaluation

A standard bivariate mixed model for the joint analysis of the phenotypic, genomic, and pedigree datasets of both populations *A* and *B* can be written as:

$$\begin{bmatrix} \mathbf{y}_A \\ \mathbf{y}_B \end{bmatrix} = \begin{bmatrix} \mathbf{X}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_B \end{bmatrix}\begin{bmatrix} \mathbf{b}_A \\ \mathbf{b}_B \end{bmatrix} + \begin{bmatrix} \mathbf{W}_A & \mathbf{0} \\ \mathbf{0} & \mathbf{W}_B \end{bmatrix}\begin{bmatrix} \mathbf{u}_A \\ \mathbf{u}_B \end{bmatrix} + \begin{bmatrix} \mathbf{e}_A \\ \mathbf{e}_B \end{bmatrix} \quad (2)$$

where $\mathbf{b}_i$ $(i = A, B)$ are the vectors of population-specific fixed effects, $\mathbf{u}_i = \begin{bmatrix} \mathbf{u}_{n,i}' & \mathbf{u}_{g,i}' \end{bmatrix}'$ are the vectors of population-specific additive genetic effects for non-genotyped and genotyped animals, and $\mathbf{e}_i$ are the vectors of population-specific residuals.

Similarly to the population-specific model (1), we assume a MNV distribution for the additive genetic effects with mean zero and a covariance matrix equal to $\mathbf{H}_J \otimes \mathbf{G}_J$, where the additive genetic covariance matrix $\mathbf{G}_J$ is equal to $\mathbf{G}_J = \begin{bmatrix} \sigma_{u,A}^2 & \sigma_{u,AB} \\ \sigma_{u,AB} & \sigma_{u,B}^2 \end{bmatrix}$, with $\sigma_{u,AB}$ being the additive genetic covariance between populations *A* and *B*. The inverse of $\mathbf{H}_J$ is computed as for the population-specific $\mathbf{H}_i^{*-1}$ using pedigree and genotype datasets of both populations. Similarly, we also assume a MVN distribution for the residuals, that is $\begin{bmatrix} \mathbf{e}_A \\ \mathbf{e}_B \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{I}\sigma_{e,A}^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_{e,B}^2 \end{bmatrix}\right)$.

Integration of estimates of SNP effects

To develop a method that integrates estimates of SNP effects into ssSNPBLUP, we assume that the estimates of SNP effects and the associated PEC matrix of population *B* are known without approximation and are expressed on the same scale as the trait of population *A*. Furthermore, we assume that all SNP genotype matrices, i.e., $\mathbf{Z}_A$, $\mathbf{Z}_B$, and $\mathbf{Z}_J$, were centered with the same allele frequencies.

The integration of estimates of SNP effects from population *B* into the single-step genomic evaluation of population *A* can be achieved through a method analogous to that proposed by Gianola and Fernando (1986) for integrating external estimated breeding values (EBVs) and the associated PEC into internal genetic evaluations. Therefore, our method relies on the alteration of the mean and covariance matrix of the MVN distribution for SNP effects of population *A*, using the estimates of SNP effects $\hat{\mathbf{g}}_{B,A}^*$ and the associated PEC matrix, $\mathbf{\Delta}_{B,A}^*\sigma_{g,A}^2$,

obtained from a genomic evaluation of population *B* and expressed on the scale of the trait of population *A*, that is $\left[\mathbf{g}_A \middle| \hat{\mathbf{g}}^*_{B,A}, \boldsymbol{\Delta}^*_{B,A}\sigma^2_{g,A}\right] \sim MVN\left(\hat{\mathbf{g}}^*_{B,A}, \boldsymbol{\Delta}^*_{B,A}\sigma^2_{g,A}\right)$.

After some algebra, and ignoring fixed effects for readability, the ssSNPBLUP MME with an integration of estimates of SNP effects of population *B* can be written as follows:

$$\begin{bmatrix} \mathbf{W}'_{n,A}\mathbf{W}_{n,A}\sigma^{-2}_{e,A} + \mathbf{H}^{11}\sigma^{-2}_{u,A} & \mathbf{H}^{12}\sigma^{-2}_{u,A} & \mathbf{0} \\ \mathbf{H}^{21}\sigma^{-2}_{u,A} & \mathbf{W}'_{g,A}\mathbf{W}_{g,A}\sigma^{-2}_{e,A} + \mathbf{H}^{22}\sigma^{-2}_{u,A} & \mathbf{H}^{23}\sigma^{-2}_{u,A} \\ \mathbf{0} & \mathbf{H}^{32}\sigma^{-2}_{u,A} & \mathbf{H}^{33}\sigma^{-2}_{u,A} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{n,A} \\ \mathbf{u}_{g,A} \\ \mathbf{g}_A \end{bmatrix} = \begin{bmatrix} \mathbf{W}'_{n,A}\mathbf{y}_{n,A} \\ \mathbf{W}'_{g,A}\mathbf{y}_{g,A} \\ \mathbf{0} \end{bmatrix} \sigma^{-2}_{e,A} +$$

$$\mathbf{H}^{*-1}_{A,B}\sigma^{-2}_{u,A} \begin{bmatrix} -(\mathbf{A}^{nn}_A)^{-1}\mathbf{A}^{ng}_A\mathbf{Z}'_A \\ \mathbf{Z}'_A \\ \mathbf{I} \end{bmatrix} \hat{\mathbf{g}}^*_{B,A},$$

$$\text{with } \mathbf{H}^{*-1}_{A,B} = \begin{bmatrix} \mathbf{H}^{11} & \mathbf{H}^{12} & \mathbf{H}^{13} \\ \mathbf{H}^{21} & \mathbf{H}^{22} & \mathbf{H}^{23} \\ \mathbf{H}^{31} & \mathbf{H}^{32} & \mathbf{H}^{33} \end{bmatrix} = \begin{bmatrix} \mathbf{A}^{nn}_A & \mathbf{A}^{ng}_A & \mathbf{0} \\ \mathbf{A}^{gn}_A & \mathbf{A}^{gg}_A + \frac{1-w}{w}\mathbf{A}^{-1}_{gg,A} & -\frac{1}{w}\mathbf{A}^{-1}_{gg,A}\mathbf{Z}_A \\ \mathbf{0} & -\frac{1}{w}\mathbf{Z}'_A\mathbf{A}^{-1}_{gg,A} & \mathbf{K}^*_{A,B} \end{bmatrix},$$

and with $\mathbf{K}^*_{A,B} = \frac{1}{w}\mathbf{Z}'_A\mathbf{A}^{-1}_{gg,A}\mathbf{Z}_A + \frac{1}{1-w}\boldsymbol{\Delta}^{*-1}_{B,A}$.

It is worth noting that the only difference between $\mathbf{K}^*_A$ in the MME without integration and $\mathbf{K}^*_{A,B}$ in the MME with integration is the replacement of the diagonal matrix $\mathbf{B}^{-1}$ by the dense matrix $\boldsymbol{\Delta}^{*-1}_{B,A}$.

Approximation of the PEC matrix

In practice, computing the PEC matrix of $\hat{\mathbf{g}}^*_{B,A}$, $\boldsymbol{\Delta}^*_{B,A}\sigma^2_{g,A}$, can be computationally challenging as it requires the inversion of the coefficient matrix of the MME. Therefore, the PEC matrix must be approximated. In this study, without loss of generality, we assume that the genomic evaluation of population *B* is a single-step genomic evaluation, and we approximate the PEC matrix $\boldsymbol{\Delta}^*_{B,A}\sigma^2_{g,A}$ by applying steps 1-3 of Gao *et al.* (2023) to a bivariate single-step genomic evaluation for both populations *A* and *B*, while considering only the phenotypic and SNP genotype datasets of population *B*. Considering the parameters of population *A* in this bivariate approach allow us to approximate the PEC matrix of population *B* expressed on the scale of population *A*.

Briefly, as a first step and in the context of this study, the approach consists of computing reliabilities for a bivariate pedigree-based BLUP for all animals in population *B* for the traits of both populations *A* and *B*. Second, deregressed equivalent record contributions (ERCs) for the genotyped animals in population *B* and for trait *A* are computed by reversing the method of Tier and Meyer (2004). Third, a coefficient matrix of a univariate SNPBLUP is constructed using all genotypes of population *B*, the residual and additive genetic variances of population *A*, and the deregressed ERCs of the genotyped animals in population B. Finally, an approximation of the PEC matrix $\boldsymbol{\Delta}^*_{B,A}\sigma^2_{g,A}$ associated with SNP effects in population *B* for the trait *A*, denoted by $\widetilde{\boldsymbol{\Delta}}^*_{B,A}\sigma^2_{g,A}$, is obtained by inverting the SNPBLUP coefficient matrix. This approach for approximating the PEC matrix is computationally feasible, even for large-scale genomic evaluations, as shown by Gao *et al.* (2023).

Additional approximations could be needed as the properties of the PEC matrix could lead to additional computational challenges when solving the single-step genomic evaluations. Indeed, it is a dense square matrix of size equal the number of SNPs multiplied by the number of traits, and these characteristics can make handling of the inverse of the PEC matrix in an iterative solver prohibitively demanding. To address these issues, we propose below three

approximations, assuming that $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}\sigma^{2}_{g,A}$ is available.

The first approximation of $\boldsymbol{\Delta}^{*-1}_{B,A}\sigma^{-2}_{g,A}$ involves ignoring off-diagonal elements among chromosomes of $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}$, which corresponds to inverting each block of $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}$ associated with a chromosome separately. This approximation results in $\boldsymbol{\Delta}^{*-1}_{B,A}\sigma^{-2}_{g,A} \approx \left(block\_diag\left(\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}\right)\right)^{-1}\sigma^{-2}_{g,A}$ with $block\_diag(.)$ denoting a chromosome-wise block diagonal matrix.

The second approximation of $\boldsymbol{\Delta}^{*-1}_{B,A}\sigma^{-2}_{g,A}$ involves ignoring off-diagonal elements among chromosomes after the inversion of $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}$, which results in $\boldsymbol{\Delta}^{*-1}_{B,A}\sigma^{-2}_{g,A} \approx block\_diag\left(\widetilde{\boldsymbol{\Delta}}^{*-1}_{B,A}\right)\sigma^{-2}_{g,A}$. This block matrix $block\_diag\left(\widetilde{\boldsymbol{\Delta}}^{*-1}_{B,A}\right)\sigma^{-2}_{g,A}$ corresponds to the SNPBLUP coefficient matrix used to compute $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}$ after absorbing the fixed effects and ignoring its off-diagonal elements.

The third approximation of $\boldsymbol{\Delta}^{*-1}_{B,A}\sigma^{-2}_{g,A}$ involves ignoring all off-diagonal elements of $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}$, which corresponds to inverting only the prediction error variances (PEV) of $\hat{\mathbf{g}}^{*}_{B,A}$, and results in $\boldsymbol{\Delta}^{*-1}_{B,A}\sigma^{-2}_{g,A} \approx \left(diag\left(\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}\right)\right)^{-1}\sigma^{-2}_{g,A}$.

Simulations

Two dairy cattle populations originating from the same breed were simulated following the procedure of Bonifazi *et al.* (2023a). Each population had simulated data on one trait with a heritability assumed to be equal to 0.30 in both populations. The genetic correlation between populations was assumed to be equal to 0.80. Briefly, about 2,000 QTLs were simulated to be randomly distributed across 30 chromosomes of 1 Morgan length each, and QTL effects were sampled from a Gaussian distribution. Each population was independently selected for 20 generations. In each population, 15,000 individuals were simulated per generation. Within each population and generation, 40 sires and 3,000 dams were selected to produce offspring for the next generation. Selection was first at random from generation 1 to generation 9, followed by a truncated selection based on within-population pedigree-based genetic evaluation. Pedigree and phenotypic information were assumed to be recorded from generation 7 and generation 10, respectively. The SNP genotypes included about 45,000 SNPs after quality control, and were assumed to be available for animals from generation 15 to generation 20 for both populations. Connectedness between the two populations was simulated by exchanging each generation the eight sires with the highest EBVs in each population throughout the last five generations.

Each scenario was replicated 10 times. The simulation was performed using the R-package MoBPS (Pook *et al.*, 2021), and pedigree-based genetic evaluations were performed with the software MiXBLUP (Vandenplas *et al.*, 2022).

Analysis

Using the simulated datasets, the aim was to validate the integration of estimates of SNP effects in population *B* into a ssSNPBLUP evaluation in population *A*, and to test the different approximations of the PEC matrix $\boldsymbol{\Delta}^{*}_{B,A}\sigma^{2}_{g,A}$.

For each replicate, datasets analyzed with ssSNPBLUP were built as follows. For population *A*, 60,000 phenotypes were randomly sampled for animals from generation 10 to 19. In addition, SNP genotypes for 7,000 animals were randomly sampled from generation 17 to 20. All the genotyped animals of generation 20 in population *A* were considered as selection candidates. For population B, all the 165,000 animals from generation 10 to 20 were associated with a phenotype, and all the 75,000 animals from generation 16 to 20 were associated also with a SNP genotype. Finally, the SNP genotypes of the exchanged sires were added to the genotype dataset of each population.

Using the datasets of both populations *A* and *B*, the following analyses were performed:

a) a joint ssSNPBLUP evaluation based on model (2) and using all datasets of both populations *A* and *B*;
b) a joint ssSNPBLUP evaluation based on model (2) but using genotype and phenotypic datasets of population *B* only. This evaluation is equivalent to a population *B* ssSNPBLUP evaluation based on the model (1), except that it provides also estimates of SNP effects of population *B* expressed on the scale of the trait of population *A*, $\hat{\mathbf{g}}^*_{B,A}$, and the associated approximated PEC matrix $\widetilde{\mathbf{\Delta}}^*_{B,A}\sigma^2_{g,A}$ computed as detailed in the section "Approximation of the PEC matrix", which are used in analyses d) to g) below;
c) a population *A* ssSNPBLUP evaluation based on the model (1) and using genotypes and phenotypes of population *A* only;
d) same as in c), but also integrating the population *B* information summarized by $\hat{\mathbf{g}}^*_{B,A}$, and $\widetilde{\mathbf{\Delta}}^*_{B,A}\sigma^2_{g,A}$;
e) the same as d) but by using $\left(block\_diag\left(\widetilde{\mathbf{\Delta}}^*_{B,A}\right)\right)^{-1}$ instead of $\widetilde{\mathbf{\Delta}}^*_{B,A}$;
f) the same as d) but by using $block\_diag\left(\widetilde{\mathbf{\Delta}}^{*-1}_{B,A}\right)$ instead of $\widetilde{\mathbf{\Delta}}^*_{B,A}$;
g) the same as d) but by using $\left(diag\left(\widetilde{\mathbf{\Delta}}^*_{B,A}\right)\right)^{-1}$ instead of $\widetilde{\mathbf{\Delta}}^*_{B,A}$.

All evaluations were performed with the software MiXBLUP (Vandenplas *et al.*, 2022). Without loss of generality, the pedigree of both populations was used in all evaluations. Furthermore, we assumed that the variance components were known and equal to the simulated variance components, and that the proportion w for RPG effects was assumed to be equal to 0.30. Finally, all genotypes in both populations were centered with the same allele frequencies. Therefore, a regression effect (often called J-factor; e.g., Strandén *et al.*, 2022) that makes the GEBVs independent of the allele frequencies used for centering was fitted for each evaluation.

To evaluate the accuracy of the integration of estimates of SNP effects in ssSNPBLUP, we compared the GEBVs of all population *A* selection candidates obtained with the different population *A* ssSNPBLUP evaluations (i.e., analyses c) to g) above). The joint ssSNPBLUP evaluation was used as reference, because it analyses simultaneously all data from both populations *A* and *B*.

The metrics computed for comparing the joint evaluation with the population *A* evaluations were: (i) Pearson correlations (r) between joint GEBVs and GEBVs without or with integration, (ii) regression coefficients (b_1) of joint GEBVs on GEBVs without or with integration, and (iii) root mean square errors (RMSE) of GEBVs without or with integration, defined as the square root of the mean of the squared differences between joint GEBVs and GEBVs without or with integration, and expressed in genetic standard deviation (SD) units. An accurate and consistent integration will result in Pearson correlation and regression coefficient equal to 1 and in RMSE equal to 0.

Results & Discussion

Integration with the complete PEC matrix

Based on our results, the developed method enables integration of estimates of SNP effects and the associated PEC matrix from a genomic evaluation into a single-step SNPBLUP. Table 1 compares joint GEBVs to GEBVs without or with integration for selection candidates in population *A*. The integration of estimates SNP effects with the approximated PEC matrix $\widetilde{\mathbf{\Delta}}^*_{B,A}\sigma^2_{g,A}$ resulted to almost the same GEBVs for the selection candidates as with the joint ssSNPBLUP, as shown by average correlations and regression coefficients close to 1 (that is, 0.98 and 0.97, respectively), and RMSE close to 0 (that is, 0.10 genetic SDs). For comparison, the average Pearson correlation between joint

GEBVs and GEBVs without integration was 0.74, the average regression coefficient was 0.78, and the average RMSE was 0.40 genetic SDs (Table 1).

Table 1. Comparison of joint GEBVs to GEBVs without or with integration for selection candidates in population *A*. Results are averaged across the 10 replicates (SE between brackets)[1].

Evaluation	R	b_1	RMSE
Pop. *A*	0.739	0.781	0.404
	(0.019)	(0.034)	(0.017)
PEC	0.982	0.973	0.104
	(0.001)	(0.004)	(0. 004)
Chromosome-wise PEC (v1)[2]	0.989	0.951	0.086
	(0.001)	(0.005)	(0.004)
Chromosome-wise PEC (v2)[3]	0.989	0.977	0.080
	(0.001)	(0.004)	(0. 004)
PEV	0.981	0.904	0.123
	(0.002)	(0.005)	(0.004)

[1] r = Pearson correlation between joint GEBVs and GEBVs without or with integration; b_1 = regression coefficient of joint GEBVs on GEBVs without or with integration; RMSE = root mean squared error of GEBVs without or with integration (in genetic standard deviation units).
[2] Off-diagonal elements among chromosomes ignored before inversion.
[3] Off-diagonal elements among chromosomes ignored after inversion.

Non-unity Pearson correlations and regression coefficients, as well as non-zero RMSE, for the integration of SNP effects using the approximated PEC matrix $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}\sigma^{2}_{g,A}$ could be explained by two approximations. First, differences between joint GEBVs and GEBVs with integration can be explained by the fact that the PEC matrices were approximated. Although our results show that our approach based on Gao *et al.* (2023) still results in an accurate integration of SNP effects, other approaches have been proposed in the literature (e.g., Jighly *et al.*, 2022; Vandenplas *et al.*, 2018), and should be also investigated in the context of single-step evaluations. Second, differences between joint GEBVs and GEBVs with integration can be explained by the fact that the contributions of the RPG effects to the additive genetic effects in the genomic evaluation of the population *B* are not integrated in the ssSNPBLUP evaluation of the population *A*. Future research is needed to explore the impact of ignoring the RPG effects in the developed procedure and to extend it for integrating RPG effects if needed.

Integration with chromosome-wise PEC and PEV matrices

Integrations based on chromosome-wise PEC matrices (that is, $\left(block\_diag\left(\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}\right)\right)^{-1}\sigma^{-2}_{g,A}$ and $block\_diag\left(\widetilde{\boldsymbol{\Delta}}^{*-1}_{B,A}\right)\sigma^{-2}_{g,A}$) resulted in accurate and consistent GEBVs, similarly to the integration based on the approximated PEC matrix $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}\sigma^{2}_{g,A}$, as shown in Table 1. Both versions of chromosome-wise PEC matrices resulted in metrics similar to those using $\widetilde{\boldsymbol{\Delta}}^{*}_{B,A}$ (that is, average Pearson correlations of 0.99, average regression coefficients higher than 0.95 and RMSE between 0.8 and 0.9 genetic SDs). These results suggest that the integration of estimates of SNP effects into a ssSNPBLUP evaluation can be performed without the whole PEC matrix. This is an appealing result because considering the whole PEC matrix in a multi-trait context could be challenging as it is a dense square matrix, and ignoring off-diagonal elements among chromosomes results in a relatively sparse block-diagonal matrix that can be easily handled with current computers.

Finally, the integration based on PEV only resulted in highly accurate, but biased, GEBVs, as shown by an average Pearson correlation of 0.98 and an average regression coefficient of 0.90. These results agree with those obtained by Vandenplas *et al.* (2018) in the context of SNPBLUP evaluations.

Implementation of the developed method

Implementing our developed method in existing software should be straightforward for those that already support a ssSNPBLUP model. First, the inverse of the (co)variance matrix of SNP effects must be replaced by the inverse of the (chromosome-wise) PEC matrix in the

coefficient matrix of the ssSNPBLUP MME. Second, the right-hand-side of the ssSNPBLUP MME requires the addition of a vector equal to the multiplication of $\mathbf{H}_{A,B}^{*-1}$ with a vector that includes imputed DGVs for non-genotyped animals ($-(\mathbf{A}_A^{nn})^{-1}\mathbf{A}_A^{ng}\mathbf{Z}_A'\hat{\mathbf{g}}_{B,A}^{*}$), direct genomic values (DGVs) for genotyped animals ($\mathbf{Z}_A'\hat{\mathbf{g}}_{B,A}^{*}$), and the estimates of SNP effects of population *B* ($\hat{\mathbf{g}}_{B,A}^{*}$). By implementing these changes, existing software can efficiently

$$\begin{bmatrix} \mathbf{W}_{n,A}'\mathbf{W}_{n,A}\sigma_{e,A}^{-2} + \mathbf{H}_g^{11}\sigma_{u,A}^{-2} & \mathbf{H}_g^{12}\sigma_{u,A}^{-2} \\ \mathbf{H}_g^{21}\sigma_{u,A}^{-2} & \mathbf{W}_{g,A}'\mathbf{W}_{g,A}\sigma_{e,A}^{-2} + \mathbf{H}_g^{22}\sigma_{u,A}^{-2} \end{bmatrix} \begin{bmatrix} \mathbf{u}_{n,A} \\ \mathbf{u}_{g,A} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_{n,A}'\mathbf{y}_{n,A} \\ \mathbf{W}_{g,A}'\mathbf{y}_{g,A} \end{bmatrix}\sigma_{e,A}^{-2} +$$

$$\mathbf{H}_g^{*-1}\sigma_{u,A}^{-2}\begin{bmatrix} -(\mathbf{A}_A^{nn})^{-1}\mathbf{A}_A^{ng}\mathbf{Z}_A' \\ \mathbf{Z}_A' \end{bmatrix}\hat{\mathbf{g}}_{B,A}^{*},$$

with $\mathbf{H}_g^{*-1} = \begin{bmatrix} \mathbf{H}_g^{11} & \mathbf{H}_g^{12} \\ \mathbf{H}_g^{21} & \mathbf{H}_g^{22} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_A^{nn} & \mathbf{A}_A^{ng} \\ \mathbf{A}_A^{gn} & \mathbf{A}_A^{gg} - \mathbf{A}_{gg,A}^{-1} + \mathbf{G}_{A,B}^{*-1} \end{bmatrix}$

where the inverse of the genomic relationship matrix $\mathbf{G}_{A,B}^{*}$ is equal to $\mathbf{G}_{A,B}^{*-1} = \left((1-w)\mathbf{Z}_A\boldsymbol{\Delta}_{B,A}^{*}\mathbf{Z}_A' + w\mathbf{A}_{gg,A}\right)^{-1}$.

It is worth noting that the form of $\mathbf{G}_{A,B}^{*}$ has the same form as the genomic relationship matrix of ssGBLUP with residual polygenic effects (Christensen and Lund, 2010), except that the diagonal matrix $\mathbf{B}$ is replaced by $\boldsymbol{\Delta}_{B,A}^{*}$. However, replacing $\mathbf{B}$ by $\boldsymbol{\Delta}_{B,A}^{*}$ for computing $\mathbf{G}_{A,B}^{*-1}$ might lead to computational challenges as $\boldsymbol{\Delta}_{B,A}^{*}$ is a dense square matrix of size equal to the number of SNPs multiplied by the number of traits in multi-trait evaluations. Further research to efficiently implement our method in ssGBLUP is therefore needed.

Potential uses of the developed method

Our analyses demonstrate that the integration of estimates of SNP effects and the associated PEC into a single-step genomic evaluation can be performed accurately. Because our developed method does not depend on the form of the genomic evaluation that provides the estimates of SNP effects, it is expected that similar results will be obtained with estimates of SNP effects integrate estimates of SNP effects obtained from a foreign genomic evaluation.

Our developed method can be also extended to ssGBLUP. As explained by Vandenplas *et al.* (2023), the absorption of the equations of SNP effects of the ssSNPBLUP MME result in the ssGBLUP MME based on the Woodbury matrix identity applied to the inverse of the genomic relationship matrix. Applying the same strategy to MME (3) results in the following MME:

computed, e.g., with a SNPMACE approach (Kärkkäinen et al., 2024; Liu and Goddard, 2018). Therefore, our developed method can be used by national organizations to integrate estimates of SNP effects computed by an international genomic evaluation into their national single-step genomic evaluation. As such, our method is an alternative to procedures that integrate pseudo-data computed from (G)EBVs into genetic evaluations (e.g., VanRaden *et al.*, 2014; Bonifazi *et al.*, 2023b)

Our developed method was tested under simple assumptions, such as datasets from only two populations, genotypes at the same SNP loci, same allele frequencies in all evaluations, and PEC matrices of population *B* available on the scale the trait of population *A*. These assumptions can be easily ignored by using or extending procedures developed in the context of SNPMACE (e.g., Jighly *et al.*, 2022; Kärkkäinen et al., 2024; Vandenplas *et al.*, 2018).

Conclusions

In this study, we developed a method that accurately integrates estimates of SNP effects and the associated PEC matrix into a single-step genomic evaluation. Our results demonstrates that the developed method yields GEBVs highly consistent with those of a joint single-step genomic evaluations when the whole PEC

matrix was used. Using chromosome-wise PEC matrices provided similarly accurate results, allowing for computationally efficient implementations in large-scale multi-trait single-step genomic evaluations.

Acknowledgments

The use of the HPC cluster has been made possible by the Regio Deal Foodvalley (through Shared Research Facilities Wageningen UR). Jeremie Vandenplas thanks Zengting Liu and Esa Mäntysaari for fruitful discussions.

References

Bonifazi, R., Neufeld, G.M., Pook, T., Vandenplas, J., and Calus, M.P.L. 2023a. Using genomic data to estimate genetic correlations between countries with different levels of connectedness. *Interbull Bull.*, 59: 1–10.

Bonifazi, R., Calus, M.P.L., ten Napel, J., Veerkamp, R.F., Biffani, S., Cassandro, M., Savoia, S., and Vandenplas, J. 2023b. Integration of beef cattle international pedigree and genomic estimated breeding values into national evaluations, with an application to the Italian Limousin population. *Genet. Sel. Evol.* 55:41.

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

Fernando, R.L., Dekkers, J.C. and Garrick D.J. 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.

Gao, H., Kudinov, A.A., Taskinen, M., Pitkänen, T.J., Lidauer, M.H., Mäntysaari, E.A., and Strandén I. 2023. A computationally efficient method for approximating reliabilities in large-scale single-step genomic prediction. *Genet. Sel. Evol.* 55:1.

Gianola, D., and Fernando R.L. 1986. Bayesian methods in animal breeding theory. *J. Anim. Sci.* 63:217–244.

Jighly, A., Benhajali, H., Liu, Z., and Goddard M.E. 2022. MetaGS: an accurate method to impute and combine SNP effects across populations using summary statistics. *Genet. Sel. Evol.* 54:37.

Kärkkäinen, H., Vargas, N., Lidauer, M., Mäntysaari, E.A., and EG SNP MACE working group. 2024. Multitrait across country genomic evaluations for EuroGenomics countries. Presentation at Interbull Open Meeting 2024, Bled, Slovenia. https://interbull.org/static/web/5_7_Karkkainen_Interbull2024.pdf

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

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

Liu, Z., and Goddard, M.E. 2018. A SNP MACE model for international genomic evaluation: technical challenges and possible solutions. Page 11.393 in *Proceedings of the 11th World Congress on Genetics Applied to Livestock Production*, Auckland, New Zealand.

Lourenco, D.A.L., Misztal, I., Tsuruta, S., Fragomeni, B., Aguilar, I., Masuda, Y., and Moser, D. 2015. Direct and indirect genomic evaluations in beef cattle. *Interbull Bull.* 49:80-84.

Pook, T., Büttgen, L., Ganesan, A., Ha, N.-T., and Simianer, H. 2021. MoBPSweb: A web-based framework to simulate and compare breeding programs. *G3 Bethesda Md.* 11:jkab023.

Schaeffer, L.R. 1994. Multiple-country comparison of dairy sires. *J. Dairy Sci.* 77:2671–2678.

Strandén, I., Aamand, G.P., and Mäntysaari, E.A. 2022. Single-step genomic BLUP with genetic groups and automatic adjustment for allele coding. *Genet. Sel. Evol.* 54:38.

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

Vandenplas, J., Calus, M.P.L., and Gorjanc, G. 2018. Genomic Prediction Using Individual-Level Data and Summary Statistics from Multiple Populations. *Genetics* 210:53–69.

Vandenplas, J., Veerkamp, R. F., Calus, M. P. L., Lidauer, M. H., Strandén, I., Taskinen, M., Schrauf, M., and ten Napel, J. 2022. MiXBLUP 3.0 - Software for large genomic evaluations in animal breeding programs. Pages 1498-1501 in *Proceedings of the 12th World Congress on Genetics Applied to Livestock Production*, Rotterdam, Auckland, The Netherlands.

Vandenplas, J., ten Napel, J., Darbaghshahi, S.N. , Evans, R., Calus, M.P.L., Veerkamp, R., Cromie, A., Mäntysaari, E.A., and Strandén, I. 2023. Efficient large-scale single-step evaluations and indirect genomic prediction of genotyped selection candidates. *Genet. Sel. Evol.* 55:37.

VanRaden, P.M., and Sullivan, P.G.. 2010. International genomic evaluation methods for dairy cattle. *Genet. Sel. Evol.* 42:7.

VanRaden, P.M., Tooker, M.E., Wright, J.R., Sun, C., and Hutchison, J.L. 2014. Comparison of single-trait to multi-trait national evaluations for yield, health, and fertility. *J. Dairy Sci.* 97:7952-7962.