Integration of Foreign Estimates of SNP Effects into a Domestic SNPBLUP

J. Vandenplas[1], M.P.L. Calus[1] and G. Gorjanc[2]

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

[2] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Research Centre, Midlothian EH25 9RG, UK

Abstract

The aim of this research was to develop and to test different procedures that integrate estimates of single nucleotide polymorphism (SNP) effects and associated measures of precision from a foreign SNP Best Linear Unbiased Prediction (SNPBLUP), into a domestic SNPBLUP when exchange of genotypes or phenotypes is prohibited for whatever reason. In addition to the foreign estimates of SNP effects, procedures were developed assuming the availability of associated: 1) prediction error (co)variance (PEC) matrix; 2) PEC matrix separately for each chromosome; 3) prediction error variances (PEV) only; 4) PEV, allele frequencies, and linkage disequilibrium (LD) of foreign training set; and 5) as 4) but with LD measured on foreign selection candidates. We tested these approaches with a simulation of two historically related populations for a single trait. We confirmed that integrating foreign estimates of SNP effects and the associated PEC matrix led to the same direct genomic values for selection candidates as the joint SNPBLUP using datasets from both populations. Integrating foreign estimates and PEV only led to biased and inaccurate predictions. Procedures based on partial PEC matrices or on LD information gave almost as accurate and unbiased predictions as the joint SNPBLUP. Therefore, accurate integration of foreign estimates of SNP effects into a domestic SNPBLUP seems possible, even if only PEV and some population statistics are available.

Key words: SNPBLUP, foreign, external, SNP effect, integration

Introduction

Exchange of genetic material among national populations implies comparison of genetic evaluations across populations and ultimately combination of these evaluations for animals of interest. In dairy cattle, these needs were (partly) solved with the implementation of a multiple across-country evaluation (MACE; Schaeffer (1994)) and of a genomic MACE (GMACE; VanRaden and Sullivan (2010)). The MACE and GMACE combine animal-based pseudo-data of sires obtained from national genetic and genomic evaluations, respectively. Pseudo-data is usually derived from (genomic) estimated breeding values ((G)EBV) and associated measures of precision (e.g., reliability).

Increasing size of genotype datasets and exchange of genomic information among national evaluations generate several issues, such as expensive computations of inverted genomic relationship matrices (Fernando *et al.*, 2016) and violation of underlying assumptions of (G)MACE (Liu and Goddard, 2018). Possible solutions for these issues could come from single nucleotide polymorphism (SNP)-based models, instead of conventional animal-based (genomic) models. Therefore, feasibility of single-step SNPBLUP (e.g., Fernando *et al.*, 2016) and of SNP-MACE (Liu and Goddard, 2018) are currently investigated. Their implementations could lead to exchange of estimates of SNP effects, in addition to (or instead of) GEBV, while exchange of genotypes and phenotypes is prohibited for various reasons. Therefore, the aim of this research was to develop and to test different procedures that integrate estimates of SNP effects and associated measures of precision from a foreign SNPBLUP into a domestic SNPBLUP.

Materials and Methods

For deriving procedures that integrate foreign estimates of SNP effects into a domestic

SNPBLUP, we assume that we have a domestic (*d*) and a foreign (*f*) dataset, both with animals phenotyped and genotyped at the same loci. The first part of this section describes (1) domestic and foreign SNPBLUP, (2) a joint SNPBLUP, (3) an exact integration of foreign estimates of SNP effects, and (4) three approximate integrations. The second part describes simulations used to test and validate the different integrations.

Domestic and foreign SNPBLUP

A standard genomic model for the domestic and foreign SNPBLUP is:

$$\mathbf{y}_i = \mathbf{X}_i\ \boldsymbol{\beta}_i^* + \mathbf{Z}_i\ \mathbf{W}_i\ \boldsymbol{\alpha}_i^* + \mathbf{e}_i^*,$$

where $\mathbf{y}_i$ (i = *d*, *f*) is a $n_{obs,i} \times 1$ vector of phenotypes, $\boldsymbol{\beta}_i^*$ is a $n_{f,i} \times 1$ vector of fixed effects, $\boldsymbol{\alpha}_i^*$ is a $n_{SNP} \times 1$ vector of SNP effects, and $\mathbf{e}_i^*$ is the vector $n_{obs,i} \times 1$ of residuals. The matrix $\mathbf{X}_i$ is an incidence matrix linking $\mathbf{y}_i$ with $\boldsymbol{\beta}_i^*$, and the matrix $\mathbf{Z}_i$ is an incidence matrix linking $\mathbf{y}_i$ with $\boldsymbol{\alpha}_i^*$. A $n_{an,i} \times n_{SNP}$ matrix $\mathbf{W}_i$ contains SNP genotypes of $n_{an,i}$ training animals. Without loss of generality, SNP genotypes were coded as 0 for homozygous aa, 1 for heterozygous aA or Aa, and 2 for homozygous AA (Strandén and Christensen, 2011). We assume a multivariate normal prior distribution for SNP effects with mean zero and covariance $\mathbf{I}\sigma_{\alpha_i}^2$, $\boldsymbol{\alpha}_i^* \sim \boldsymbol{MVN}\left(\mathbf{0}, \mathbf{I}\sigma_{\alpha_i}^2\right)$, where $\mathbf{I}$ is an identity matrix, and $\sigma_{\alpha_i}^2$ is the variance of SNP effects. We also assume that residuals are multivariate normally distributed with mean zero and covariance $\mathbf{R}_i\sigma_e^2$, $\mathbf{e}_i^* \sim \boldsymbol{MVN}\left(\mathbf{0}, \mathbf{R}_i\ \sigma_e^2\right)$, where $\mathbf{R}_i$ is a diagonal matrix, and σ_e^2 is the residual variance. For simplicity and without loss of generality, it is assumed in the following development that the residual variances are the same for all analyses. Variance components $\sigma_{\alpha_i}^2$ and σ_e^2 are assumed known, and can be previously estimated from the data.

Domestic and foreign estimates of SNP effects $\widehat{\boldsymbol{\alpha}_i^*}$ are obtained by solving the following system of equations:

$$\begin{bmatrix} \mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\ \sigma_e^{-2} & \mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i\ \mathbf{W}_i\ \sigma_e^{-2} \\ \mathbf{W}_i'\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{X}_i\ \sigma_e^{-2} & \mathbf{W}_i'\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{Z}_i\ \mathbf{W}_i\ \sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_i}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_i^*} \\ \widehat{\boldsymbol{\alpha}_i^*} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_i'\mathbf{R}_i^{-1}\mathbf{y}_i\ \sigma_e^{-2} \\ \mathbf{W}_i'\mathbf{Z}_i'\mathbf{R}_i^{-1}\mathbf{y}_i\ \sigma_e^{-2} \end{bmatrix}. \quad (1)$$

Domestic and foreign direct genomic values (DGV) are obtained by $\widehat{\mathbf{g}_i^*} = \mathbf{W}_i\widehat{\boldsymbol{\alpha}_i^*}$.

Joint SNPBLUP

A standard genomic model for the joint analysis of the domestic and foreign datasets is:

$$\begin{bmatrix} \mathbf{y}_d \\ \mathbf{y}_f \end{bmatrix} = \begin{bmatrix} \mathbf{X}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_f \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_d \\ \boldsymbol{\beta}_f \end{bmatrix} + \begin{bmatrix} \mathbf{Z}_d\ \mathbf{W}_d \\ \mathbf{Z}_f\ \mathbf{W}_f \end{bmatrix} \boldsymbol{\alpha} + \begin{bmatrix} \mathbf{e}_d \\ \mathbf{e}_f \end{bmatrix},$$

where phenotypes from the two datasets are modelled with dataset specific fixed effects $\left(\boldsymbol{\beta}_d, \boldsymbol{\beta}_f\right)$, but a joint set of SNP effects ($\boldsymbol{\alpha}$). We assume a multivariate normal prior distribution for SNP effects with mean zero and covariance $\mathbf{I}\sigma_{\alpha_J}^2$, $\boldsymbol{\alpha} \sim \boldsymbol{MVN}\left(\mathbf{0}, \mathbf{I}\sigma_{\alpha_J}^2\right)$, where $\sigma_{\alpha_J}^2$ is the variance of SNP effects in the joint SNPBLUP. We also assume that residuals are multivariate normally distributed, specifically

$$\begin{bmatrix} \mathbf{e}_d \\ \mathbf{e}_f \end{bmatrix} \sim \boldsymbol{MVN}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{R}_d & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_f \end{bmatrix} \sigma_e^2\right).$$

Joint estimates of SNP effects $\widehat{\boldsymbol{\alpha}}$ are obtained by solving the following system of equations:

$$\begin{bmatrix} \mathbf{X}_d'\mathbf{R}_d^{-1}\mathbf{X}_d\,\sigma_e^{-2} & \mathbf{0} & \mathbf{X}_d'\mathbf{R}_d^{-1}\mathbf{Z}_d\,\mathbf{W}_d\,\sigma_e^{-2} \\ \mathbf{0} & \mathbf{X}_f'\mathbf{R}_f^{-1}\mathbf{X}_f\,\sigma_e^{-2} & \mathbf{X}_f'\mathbf{R}_f^{-1}\mathbf{Z}_f\,\mathbf{W}_f\,\sigma_e^{-2} \\ \mathbf{W}_d'\mathbf{Z}_d'\mathbf{R}_d^{-1}\mathbf{X}_d\,\sigma_e^{-2} & \mathbf{W}_f'\mathbf{Z}_f'\mathbf{R}_f^{-1}\mathbf{X}_f\,\sigma_e^{-2} & \mathbf{W}_d'\mathbf{Z}_d'\mathbf{R}_d^{-1}\mathbf{Z}_d\,\mathbf{W}_d\,\sigma_e^{-2} + \mathbf{W}_f'\mathbf{Z}_f'\mathbf{R}_f^{-1}\mathbf{Z}_f\,\mathbf{W}_f\,\sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_d} \\ \widehat{\boldsymbol{\beta}_f} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_d'\mathbf{R}_d^{-1}\mathbf{y}_d\,\sigma_e^{-2} \\ \mathbf{X}_f'\mathbf{R}_f^{-1}\mathbf{y}_f\,\sigma_e^{-2} \\ \mathbf{W}_d'\mathbf{Z}_d'\mathbf{R}_d^{-1}\mathbf{y}_d\,\sigma_e^{-2} + \mathbf{W}_f'\mathbf{Z}_f'\mathbf{R}_f^{-1}\mathbf{y}_f\,\sigma_e^{-2} \end{bmatrix}. \quad (2)$$

Joint DGV for domestic and foreign animals are obtained by $\widehat{\mathbf{g}}_i = \mathbf{W}_i\widehat{\boldsymbol{\alpha}}$ (i = d, f).

Exact integration

Integration of foreign estimates of SNP effects, $\widehat{\boldsymbol{\alpha}_f^*}$, into a domestic SNPBLUP can be performed by means of absorbing equations corresponding to the foreign dataset in the joint system of equations (2). After some algebra, we get the following system of equations that performs an exact integration of foreign estimates of SNP effects into a domestic SNPBLUP:

$$\begin{bmatrix} \mathbf{X}_d'\mathbf{R}_d^{-1}\mathbf{X}_d\,\sigma_e^{-2} & \mathbf{X}_d'\mathbf{R}_d^{-1}\mathbf{Z}_d\,\mathbf{W}_d\,\sigma_e^{-2} \\ \mathbf{W}_d'\mathbf{Z}_d'\mathbf{R}_d^{-1}\mathbf{X}_d\,\sigma_e^{-2} & \mathbf{W}_d'\mathbf{Z}_d'\mathbf{R}_d^{-1}\mathbf{Z}_d\,\mathbf{W}_d\,\sigma_e^{-2} + \left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1} - \mathbf{I}\sigma_{\alpha_f}^{-2} + \mathbf{I}\sigma_{\alpha_J}^{-2} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{\beta}_d} \\ \widehat{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_d'\mathbf{R}_d^{-1}\mathbf{y}_d\,\sigma_e^{-2} \\ \mathbf{W}_d'\mathbf{Z}_d'\mathbf{R}_d^{-1}\mathbf{y}_d\,\sigma_e^{-2} + \left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}\widehat{\boldsymbol{\alpha}_f^*} \end{bmatrix}, \quad (3)$$

where $\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right) = \left(\mathbf{W}_f'\mathbf{Z}_f'\mathbf{M}_f\,\mathbf{Z}_f\,\mathbf{W}_f\,\sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}\right)^{-1}$ is the prediction error covariance (PEC) matrix associated with the foreign estimates of SNP effects $\widehat{\boldsymbol{\alpha}_f^*}$, with $\mathbf{M}_f = \mathbf{R}_f^{-1} - \mathbf{R}_f^{-1}\mathbf{X}_f\left(\mathbf{X}_f'\mathbf{R}_f^{-1}\mathbf{X}_f\right)^{-1}\mathbf{X}_f'\mathbf{R}_f^{-1}$ being an absorption matrix for foreign fixed effects.

Approximate integrations

Exact integration requires the inverse of PEC matrix from the foreign SNPBLUP, $\left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}$, which might not be available. We propose here three different approximations to overcome this.

First, $\left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}$ can be approximated with the inverse of the PEC matrix associated with each chromosome. This approximation ignores off-diagonal elements among chromosomes, which could be assumed to be close to 0 (Yang *et al.*, 2012). However, similar to $\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)$, obtaining PEC matrix for each chromosome (or any other subset of the genome) separately could be still challenging compared to obtaining only the prediction error variance (PEV) matrix $\boldsymbol{PEV}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)$.

Second, $\left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}$ can be approximated with the inverse of $\boldsymbol{PEV}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)$, that is $\left(\boldsymbol{PEV}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}$. This approximation would be accurate if $\left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}$ has (close to) zero off-diagonal elements, which is dependent on the characteristics of genotypes in the foreign dataset (e.g., allele frequencies, linkage disequilibrium (LD), and population/family structure).

Third, $\left(\boldsymbol{PEC}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)\right)^{-1}$ can be approximated from $\boldsymbol{PEV}\left(\widehat{\boldsymbol{\alpha}_f^*}\right)$, residual and SNP variances, allele frequencies and LD of the foreign training set. Assuming Hardy-Weinberg equilibrium and one record per foreign training animal (i.e.,

$\mathbf{Z}_f = \mathbf{I}$), it can be shown that the product $\mathbf{W}_f'\mathbf{Z}_f'\mathbf{Z}_f\ \mathbf{W}_f$ can be approximated as:

$$\mathbf{W}_f'\mathbf{Z}_f'\mathbf{Z}_f\ \mathbf{W}_f = \mathbf{W}_f'\mathbf{W}_f \approx n_{an,f}(4\mathbf{pp}' + \mathbf{VCV})$$

where $n_{an,f}$ is the number of foreign training animals, $\mathbf{p}$ is a $n_{SNP} \times 1$ vector of allele frequencies in the foreign training set, $\mathbf{V}$ is a $n_{SNP} \times n_{SNP}$ diagonal matrix with the *j*-th diagonal element being equal to $\sqrt{2\mathbf{p}_j(\mathbf{1} - \mathbf{p}_j)}$, and $\mathbf{C}$ is a $n_{SNP} \times n_{SNP}$ matrix of pairwise correlations (LD) between genotypes at SNP loci in the foreign training set. From the approximation of $\mathbf{W}_f'\mathbf{Z}_f'\mathbf{Z}_f\ \mathbf{W}_f$, and by relaxing the assumption of one record per animal, $\left(\boldsymbol{PEC}(\widehat{\boldsymbol{\alpha}_f^*})\right)^{-\mathbf{1}} = \mathbf{W}_f'\mathbf{Z}_f'\mathbf{M}_f\ \mathbf{Z}_f\ \mathbf{W}_f\ \sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}$ can be approximated with $\mathbf{\Lambda}_f(4\mathbf{pp}' + \mathbf{VCV})\mathbf{\Lambda}_f\sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}$ where $\mathbf{\Lambda}_f$ is a $n_{SNP} \times n_{SNP}$ diagonal matrix with the squared *j*-th diagonal element representing the effective number of records for the *j*-th SNP. The diagonal matrix $\mathbf{\Lambda}_f$ can be estimated by solving the nonlinear system of equations $diag\left(\left(\mathbf{\Lambda}_f(4\mathbf{pp}' + \mathbf{VCV})\mathbf{\Lambda}_f\sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}\right)^{-1}\right) = \boldsymbol{PEV}(\widehat{\boldsymbol{\alpha}_f^*})$ through a fixed-point iteration algorithm detailed in Appendix. It is worth noting that this algorithm requires the inversion of a $n_{SNP} \times n_{SNP}$ dense matrix (step 4) at each iteration. This computational cost can be reduced by performing the algorithm with a limited number, let say 2000, of consecutive SNPs.

In practice, the matrix $\mathbf{C}$ for the foreign training set could be unknown, and can be approximated by using a reference panel that includes, for example, available genotypes of selection candidates related with the foreign training population (Yang *et al.*, 2012).

Simulations

We tested the developed methods with simulated data for two populations and a single trait. The data was simulated in 5 replicates with the AlphaSim program, which uses coalescent method for simulation of base population chromosomes and gene drop method for simulation of chromosome inheritance within a pedigreed population (Hickey and Gorjanc, 2012; Faux *et al.*, 2016).

A diploid genome was simulated with 30 chromosomes, each 10^8 base pairs long. Coalescent mutation and recombination rate per base pair were set to 10^{-8}, while effective population size was modelled in line with the values reported by MacLeod *et al*. (2013). Effective population size of the pedigree base was set to 100. For each chromosome, 100 SNP loci per chromosome (3000 per genome) were sampled to serve as causal loci. The allele substitution effect of causal loci was sampled from normal distribution with mean zero and variance 1/3000. The effects were used to simulate a complex trait with additive genetic architecture. In addition, 2000 loci per chromosome (60 000 per genome) were selected to serve as markers with the restriction of having minor allele frequency above 0.05.

The domestic and foreign populations were ancestrally related through the common base population, but otherwise maintained independently, i.e., there was no migration between the populations. Each population was initiated with 10 000 founders (half males and half females) and maintained for 7 generations with constant size. For creating the next generation, 25 males were selected on true breeding value (TBV), and all 5000 females were used as parents.

For every animal in the domestic population, an own phenotype was simulated as the sum of TBV and residual sampled from normal distribution with mean zero and residual variance scaled relative to variance of TBV in the base population such that heritability was 0.3.

For every animal in the foreign population, an own phenotype was simulated as the sum of TBV and the mean of n_{weight} residuals. Each residual was sampled from a normal distribution with mean zero and residual variance scaled relative to variance of TBV in the base population such that heritability was

0.3. The weight n_{weight} was sampled from a geometric distribution with a probability of 0.15, augmented by 1. The average n_{weight} was 6.6. These simulated weighted phenotypes mimic (daughter) yield deviations, deregressed proofs, or repeated records.

For satisfying the assumption of identical residual variance across all evaluations, domestic and foreign phenotype records were divided by the square root of the corresponding residual variance, such that $\sigma_e^2 = 1$.

Analysis

The aim was to validate the exact integration of foreign estimates of SNP effects into a domestic SNPBLUP, and to test the different approximations. For each population, 5000 animals from generation 1 to 6 were randomly sampled to generate a training population. All the 10 000 animals of generation 7 were considered as selection candidates. We assumed that variance components were known and equal to true variances. We also assumed that allele frequencies of the foreign training set were known at the domestic level.

The following SNPBLUP were performed:

a) a separate SNPBLUP for each population (i.e., system of equations (1));
b) a joint SNPBLUP (i.e., system of equations (2));
c) a combined SNPBLUP with an exact integration of foreign estimates of SNP effects (i.e., system of equations (3));
d) the same as c) but approximating the PEC matrix with a partial PEC matrix where PEC between loci on different chromosomes were set to zero;
e) the same as c), but approximating the PEC matrix with a PEV matrix where PEC between loci was set to zero;
f) the same as c), but approximating the PEC matrix with PEV, allele frequencies, and pairwise correlations among all SNPs of the foreign training set. The algorithm for estimating effective number of records per SNP was performed on subsets of 2000 consecutive SNPs;
g) the same as f), but with pairwise correlations among all SNPs computed from foreign selection candidates instead of the training data.

Analysis of the different integrations of foreign estimates was performed by comparing DGV of selection candidates obtained from the different SNPBLUP. The joint SNPBLUP was considered as the reference, because it uses both domestic and foreign information. If integration was fully accurate, no difference should be detected in DGV of selection candidates obtained from the joint SNPBLUP and the SNPBLUP with integration. This can be observed by (a) a Pearson correlation between joint DGV and DGV with integration equal to 1, (b) a regression coefficient of joint DGV on DGV with integration equal to 1, and (c) a mean square error (MSE) of DGV with integration, computed as mean of the squared difference between DGV with integration and joint DGV, equal to 0. These three parameters were used for comparison, and were computed for 10 000 selection candidates of each population separately.

Results & Discussion

Accuracies of DGV without integration and joint DGV

Accuracies, that is Pearson correlations between TBV and DGV, of DGV without integration and joint DGV are in Table 1. Analyzing jointly both domestic and foreign datasets increased accuracy of DGV of domestic selection candidates by 15 absolute points, and accuracy of DGV of foreign selection candidates by 4 absolute points. The lower improvement for foreign candidates can be explained by a higher baseline amount of information in the foreign evaluation.

Table 1. Accuracies of DGV without integration and joint DGV for domestic and foreign selection candidates. Results are averaged across the five replicates (SE between brackets).

SNPBLUP	Domestic	Foreign
Separate	0.47 (0.01)	0.63 (0.01)
Joint	0.62 (0.01)	0.67 (0.01)

Comparison of different integrations

The developed procedure enabled integration of estimates of SNP effects and associated measures of precision into a domestic SNPBLUP. Table 2 compares DGV without and with integration to joint DGV for domestic and foreign selection candidates. The exact integration of foreign estimates, that is by means of the PEC matrix, led to the same DGV for both domestic and foreign selection candidates as with the joint SNPBLUP, as shown by correlations and regression coefficients of 1, and MSE close to 0. Non-zero MSE could be explained by rounding errors (Table 2). For comparison, correlations between joint DGV and DGV without integration were 0.79 and 0.94 for domestic and foreign selection candidates, respectively. Corresponding regression coefficients were 0.92 and >0.99, and corresponding MSE were 0.20 and 0.06 (Table 2).

Approximate integrations based on partial PEC matrices or LD information gave almost as accurate and unbiased DGV as the exact integration for both domestic and foreign selection candidates. This is shown in Table 2, and demonstrates that integration of estimates of SNP effects into a SNPBLUP can be performed accurately without availability of the full PEC matrix, or of genotypes and phenotypes of the foreign training set. These two features are interesting because computing the exact PEC matrix, or even exchanging it between evaluations, could be challenging, and because all genotypes (and phenotypes) of the training set are usually unknown and unavailable for a third party. However, genotypes of animals related with a foreign population might be available locally, and could be used for approximating PEC, as demonstrated by the results of integration using LD information computed from selection candidates (Table 2). For such an approximation, only estimates of SNP effects, associated PEV, and allele frequencies of a foreign training population need to be exchanged.

An approximate integration based on PEV only led to biased DGV with integration. Regression coefficients were lower than 0.90 for both domestic and foreign selection candidates, even if correlation for domestic DGV increased by 18 absolute points (Table 2). Using only PEV, that is ignoring PEC, usually gives satisfying results when integrating conventional EBV (Legarra *et al.*, 2007; VanRaden *et al.*, 2014), mainly because the animal least-squares part of the mixed model equations of a foreign evaluation (i.e., $\mathbf{Z}_f'\mathbf{R}_f^{-1}\mathbf{Z}_f$) can be accurately approximated by a diagonal matrix. For SNPBLUP, the corresponding marker least-squares part (i.e., $\mathbf{W}_f'\mathbf{Z}_f'\mathbf{R}_f^{-1}\mathbf{Z}_f\ \mathbf{W}_f$) is dense. Therefore, off-diagonal elements, or PEC, should not be ignored when integrating estimates of SNP effects.

Potential of the developed procedure

The proposed procedure was developed under simple assumptions, such as datasets from only two populations, SNP genotypes at the same loci, and the same residual variances for all SNPBLUP. These assumptions can be easily removed by extending the developed procedure to multi-trait and multi-population analyses. The SNP-MACE (Liu and Goddard, 2018) is an example of such an extension. It is also worth noting that the integration of SNP effects is similar to the integration of conventional EBV. Therefore, procedures developed for the integration of EBV for traits with different variance components, measurement units/scales, or trait definitions should be easily adapted for the integration of SNP effects, by taking PEC into account if necessary.

Future research is needed to extend the developed procedure to single-step SNPBLUP (e.g., Fernando *et al.*, 2016). Indeed, unlike this study, single-step SNPBLUP considers data of genotyped and non-genotyped animals and, potentially, a residual polygenic effect.

Conclusions

The developed procedure accurately integrated estimates of SNP effects from a foreign SNPBLUP into a domestic SNPBLUP. Therefore, the developed procedure led to DGV as accurate and unbiased as with a joint SNPBLUP that uses all available datasets. We

also showed that accurate integration of estimates of SNP effects was possible when only PEV and some population statistics were available.

Table 2. Comparison of DGV without or with integration to joint DGV for domestic and foreign selection candidates. Results are averaged across the five replicates (SE between brackets)[1].

Type of integration	Domestic population			Foreign population		
	r	b	MSE	r	b	MSE
No integration	0.79	0.92	0.203	0.94	>0.99	0.055
	(0.00)	(0.01)	(0.060)	(0.00)	(0.01)	(0.023)
PEC	1.00	1.00	0.001	1.00	1.00	0.001
	(0.00)	(0.00)	(0.000)	(0.00)	(0.00)	(0.000)
PEC per chromosome	0.99	0.99	0.009	0.98	0.97	0.018
	(0.00)	(0.00)	(0.002)	(0.00)	(0.00)	(0.008)
PEV	0.97	0.90	0.021	0.95	0.86	0.043
	(0.00)	(0.01)	(0.005)	(0.01)	(0.01)	(0.012)
PEV + LD of the foreign	>0.99	0.98	0.026	>0.99	0.98	0.027
training set	(0.00)	(0.00)	(0.013)	(0.00)	(0.00)	(0.015)
PEV + LD from foreign	0.99	0.96	0.035	0.98	0.99	0.031
selection candidates	(0.00)	(0.00)	(0.016)	(0.00)	(0.00)	(0.011)

[1] r = Pearson correlation between joint DGV and DGV without or with integration; b = regression coefficient of joint DGV on DGV without or with integration; MSE = mean squared error of DGV without or with integration.

Acknowledgements

This study was financially supported by the Dutch Ministry of Economic Affairs (TKI Agri & Food project 16022) and the Breed4Food partners Cobb Europe, CRV, Hendrix Genetics and Topigs Norsvin. The use of the HPC cluster has been made possible by CAT-AgroFood (Shared Research Facilities Wageningen UR).

References

Burden, R.L. & Faires, J.D. 2011. *Numerical analysis.* 9th ed. Brooks/Cole; Boston, MA.

Faux, A.-M., Gorjanc, G., Gaynor, R.C., Battagin, M., Edwards, S.M., Wilson, D.L., Hearne, S.J., Gonen, S. & Hickey, J.M. 2016. AlphaSim: Software for Breeding Program Simulation. *Plant Genome 9*.

Fernando, R.L., Cheng, H., Golden, B.L. & Garrick, D.J. 2016. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. *Genet. Sel. Evol. 48:96*.

Hickey, J.M. & Gorjanc, G. 2012. Simulated Data for Genomic Selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods. *G3 GenesGenomesGenetics 2,* 425–427.

Legarra, A., Bertrand, J.K., Strabel, T., Sapp, R.L., Sanchez, J.P. & Misztal, I. 2007. Multi-breed genetic evaluation in a Gelbvieh population. *J. Anim. Breed. Genet. 124,* 286–295.

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

MacLeod, I.M., Larkin, D.M., Lewin, H.A., Hayes, B.J. & Goddard, M.E. 2013. Inferring demography from runs of homozygosity in whole-genome sequence, with correction for sequence errors. *Mol. Biol. Evol. 30,* 2209–2223.

Misztal, I. & Wiggans, G.R. 1988. Approximation of prediction error variance in large-scale animal models. *J. Dairy Sci. 71(Suppl. 2),* 27–32.

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

Strandén, I. & Christensen, O.F. 2011. Allele coding in genomic evaluation. *Genet. Sel. Evol. 43:25.*

Vandenplas, J. & Gengler, N. 2012. Comparison and improvements of different Bayesian procedures to integrate external information into genetic evaluations. *J. Dairy Sci. 95,* 1513–1526.

VanRaden, P.M. & 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. & Hutchison, J.L. 2014. Comparison of single-trait to multi-trait national evaluations for yield, health, and fertility1. *J. Dairy Sci. 97,* 7952–7962.

Yang, J., Ferreira, T., Morris, A.P., Medland, S.E., G.I. of An.T. (GIANT) Consortium, Dia.G.R.A.M. (DIAGRAM) Consortium, Madden, P.A.F., Heath, A.C., Martin, N.G., Montgomery, G.W., Weedon, M.N., Loos, R.J., Frayling, T.M., McCarthy, M.I., Hirschhorn, J.N., Goddard, M.E. & Visscher, P.M. 2012. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat. Genet. 44,* 369–375.

Appendix: Estimation of the effective number of records per SNP

Here we detail the algorithm for computing the effective number of records per SNP from linkage-disequilibrium, allele frequencies, and prediction error variances of estimates of SNP effects $\widehat{\boldsymbol{\alpha}_f^*}$ ($\boldsymbol{PEV}(\widehat{\boldsymbol{\alpha}_f^*})$) of the foreign dataset. We assume that an unknown $n_{SNP} \times n_{SNP}$ diagonal matrix $\boldsymbol{\Lambda}_f$ exists such that:

$$\boldsymbol{PEC}(\widehat{\boldsymbol{\alpha}_f^*}) \approx \left(\boldsymbol{\Lambda}_f \boldsymbol{\Psi} \boldsymbol{\Lambda}_f \sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}\right)^{-\mathbf{1}}$$

where the squared *j*-th diagonal element of $\boldsymbol{\Lambda}_f$ represents the effective number of records for the *j*-th SNP, and $\boldsymbol{\Psi} = 4\mathbf{pp}' + \mathbf{VCV}$ with $\mathbf{p}$ being a $n_{SNP} \times 1$ vector of allele frequencies in the foreign dataset, $\mathbf{V}$ being a $n_{SNP} \times n_{SNP}$ diagonal matrix with the *j*-th diagonal element being equal to $\sqrt{2\mathbf{p}_j(\mathbf{1} - \mathbf{p}_j)}$, and $\mathbf{C}$ being a $n_{SNP} \times n_{SNP}$ matrix of pairwise correlations between genotypes at SNP loci.

The diagonal matrix $\boldsymbol{\Lambda}_f$ can be estimated by solving the nonlinear system of equations:

$$diag\left(\left(\boldsymbol{\Lambda}_f \boldsymbol{\Psi} \boldsymbol{\Lambda}_f \sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}\right)^{-\mathbf{1}}\right) = \boldsymbol{PEV}(\widehat{\boldsymbol{\alpha}_f^*})$$

through a fixed-point iteration algorithm (Burden and Faires, 2011) as follows:

1) $\mathbf{Q}^0 = \left(\mathbf{P}^{0\,-\mathbf{1}} - \mathbf{I}\sigma_{\alpha_f}^{-2}\right) * (diag(\boldsymbol{\Psi})\sigma_e^{-2})^{-\mathbf{1}}$
   where $\mathbf{P}^0$ is a diagonal matrix with the *j*-th diagonal element equal to the PEV of the *j*-th SNP;
2) $\boldsymbol{\Lambda}_f^0 = \sqrt{\mathbf{Q}^0}$
3) $k = 1$
4) $\mathbf{P}^k = diag\left(\left(\boldsymbol{\Lambda}_f^{k-1} \boldsymbol{\Psi} \boldsymbol{\Lambda}_f^{k-1} \sigma_e^{-2} + \mathbf{I}\sigma_{\alpha_f}^{-2}\right)^{-1}\right)$
5) $\mathbf{H} = \left(\mathbf{P}^{k\,-\mathbf{1}} - \mathbf{I}\sigma_{\alpha_f}^{-2}\right) * (diag(\boldsymbol{\Psi})\sigma_e^{-2})^{-1}$
6) $\mathbf{S}^k = \mathbf{Q}^0 - \mathbf{H}$
7) If trace of $\mathbf{S}^k$ is not sufficiently small
   a. $\mathbf{Q}^k = \mathbf{Q}^{k-1} + \mathbf{H}$
   b. If any diagonal element in $\mathbf{Q}^k$ is negative, set it to 0
   c. $\boldsymbol{\Lambda}_f^k = \sqrt{\mathbf{Q}^k}$
   d. $k = k + 1$
   e. Repeat from 4
8) $\boldsymbol{\Lambda}_f^k = \sqrt{\mathbf{Q}^k}$

It is worth noting that this algorithm is similar to algorithms proposed by Misztal and Wiggans (1988) and Vandenplas and Gengler (2012) to estimate effective number of records free of contributions from relatives. The *j*-th diagonal element of $\mathbf{Q}^k$ can therefore be considered as the effective number of records for the *j*-th SNP.