A deregression method for single-step genomic model using all genotype data

***Z. Liu*[1]*and Y. Masuda*[2]**

[1] *IT solutions for animal production (***vit***), Heinrich-Schroeder-Weg 1, D-27283 Verden, Germany*

[2] *University of Georgia, USA*

Abstract

For various applications in dairy cattle evaluation, pseudo-phenotype data are needed for cows with own records or bulls with daughters. EBV deregression using pedigree is performed routinely to generate deregressed proofs (DRP) for the bull MACE evaluation. In some countries DRP for cows with own data or bulls with daughters are used as pseudo-phenotype in the current multi-step genomic step for dairy cattle genomic evaluations. When more and more countries upgrade their current genomic evaluation to a single-step model, genomic-free EBV must be guaranteed for Interbull's conventional MACE evaluation. Statistical methods were proposed to deregress genomic breeding values of the single-step evaluation using the inverse of the genomic relationship matrix **H** of the single-step GBLUP model. A high number of genotyped female animals in some countries may lead to a **H** matrix too large to be inverted, new GEBV deregression methods are therefore needed that are feasible for using genotypes of millions of animals. The purpose of this paper was to develop a GEBV deregression method for the single-step model using all genotype data. A special single-step SNP BLUP model was applied to the GEBV deregression. All animals with own phenotype data and all genotyped animals including young candidates were included in the GEBV deregression. The same pedigree file as well as the same genotype data were considered in the GEBV deregression process as in the original single-step evaluation. Thanks to the efficient single-step SNP BLUP model, the proposed GEBV deregression should be feasible for processing millions of genotyped animals. Methodological and technical issues were addressed, and a validation procedure was proposed for the GEBV deregression method of the single-step evaluation. Analysis of real data will be required to verify the developed GEBV deregression method.

Key words: deregression, single-step model, genomic evaluation, pseudo-phenotype, validation

Introduction

Pseudo-phenotypes of animals are useful data for diverse genetic or genomic analyses. Pseudo-phenotype data, by definition, should be ideally corrected for all model effects affecting the original phenotype records, such as fixed and non-genetic random effects, so that only additive genetic effects and random errors are contained in the pseudo-phenotype records. The special property of the pseudo-phenotype data makes subsequent genetic or genomic evaluations easier, because only the additive genetic effects need to be estimated. Daughter yield deviations (DYD, VanRaden and Wiggans, 1991; Liu et al. 2004) and deregressed proofs (Jairath et al. 1998) are two commonly used pseudo-phenotype data. Calculation of DYD for complex statistical models like the random regression test-day model (Liu et al. 2004) is technically challenging. In contrast, EBV or proof deregression can be applied to any linear function of estimated breeding values (EBV) to bypass complex evaluation models. A single trait model can be applied in most cases for the proof deregression. By doing so, the proof deregression provides an easier alternative than the DYD calculation for generating pseudo-phenotypic data.

For the conventional MACE evaluation of bulls, national EBV of bulls are first deregressed using a function of effective daughter contribution (EDC) as weights following the iterative procedure (Jairath et al. 1998). A multiple country model is then applied to the deregressed proofs (DRP) of bulls for the international bull MACE evaluation. Bull MACE EBV on a country scale can be deregressed with the same method. To enlarge the genomic reference population, most countries use foreign bulls' DRP of Interbull MACE evaluation as the pseudo-phenotype data. Corresponding to the bulls' MACE EBV are EDC based on daughter information from all countries participating in the bull MACE evaluation, the EDC can be calculated with the procedure by Liu (2011).

Single-step genomic model utilizes all sources of information on phenotype, genotype, and pedigree for an unbiased genomic prediction (Misztal et al. 2009, Aguilar et al. 2010). In comparison to the single-step genomic BLUP model (ssGBLUP), a single-step SNP BLUP model (ssSNPBLUP, Liu et al. 2014) allows a direct estimation of SNP marker effects together with all other model effects. Genomic breeding value estimates (GEBV) of bulls or cows can be deregressed, in analogue to deregressed EBV of the conventional model, to generate the pseudo-phenotype containing only additive genetic and random error effects. The deregressed GEBV of the bulls from each country provide more accurate pseudo-phenotype data for the current bull MACE evaluation than the deregressed EBV from a conventional evaluation because the impact of genomic selection has been properly accounted for in the single-step evaluation.

The main objective of this study was to derive deregressed GEBV of the single-step model for bulls with daughters or cows with own phenotypic records using all available genotype data.

Materials and Methods

Deregression of estimated breeding values

Estimation of breeding values of animals can be understood as a process of regressing phenotype data on additive genetic effects using pedigree information. The deregression of EBV is to reverse the regression process in the genetic evaluation process (Jairath et al. 1998) to obtain the pseudo-phenotype data for the animals with own phenotype records. In literature, many non-iterative deregression methods were proposed. However, the deregression methods on an animal-by-animal basis, maybe labelled as *scalar deregression*, were proven to be sub-optimal (Calus et al. 2016). The *iterative deregression* method (Jairath et al. 1998), also called *matrix deregression*, uses the pedigree data. This optimal deregression process was demonstrated to be reversible (Mark et al. 2002; Calus et al. 2016).

Deregression of bull MACE EBV

On a given country scale MACE EBV of all bulls with daughters in any participating country can be deregressed using the iterative deregression procedure (Jairath et al. 1998). Liu (2011) developed a statistical method to calculate MACE EDC of domestic or foreign bulls using their EDC from all the countries and genetic correlation matrix in the MACE evaluation. The MACE EDC can be then used for the deregression of bull MACE EBV. The deregressed MACE EBV of foreign bulls are commonly used in national genomic evaluation to increase the size of genomic reference population. For German Holsteins, the deregressed MACE EBV of foreign bulls are routinely utilized in the current two-step genomic evaluation. Based on the MACE evaluation of August 2020, a systematic validation of the deregressed MACE EBV of all the bulls was conducted on the German country scale (***vit*** 2020, unpublished data). The deregressed MACE EBV of all the bulls with daughters in any participating country were used as input data together with their MACE EDC for the validation under a single-trait BLUP model. The resulted EBV for those bulls were shown to be identical to their original MACE EBV for all the 38 traits evaluated in MACE. The reversibility of the bull MACE EBV deregression was successfully confirmed and the MACE bull EBV deregression method was, therefore, validated.

Deregression of EBV for domestic cows

To add cows into a bull genomic reference population under the two-step genomic model (Alkhoder et al. 2017), EBV of domestic cows were deregressed with effective record contribution (ERC) as a measure of accuracy, following the same deregression method as for bulls. ERC of cows corresponded to lower reliability than EDC of bulls, particularly for some traits with low heritability. For German Holsteins, genotyped cows have been included in a mixed reference population for all evaluated traits in routine genomic evaluation since 2019 (Liu et al. 2019). Although a validation of the cow EBV deregression had been performed for all test-day yields and somatic cell scores and some conformation traits already in 2016, deregressed cow EBV were systematically investigated again for all evaluated traits via a validation study in 2020 (***vit***, unpublished data). A single-trait BLUP evaluation was performed using DRP of all cows with own records and ERC as weights. The resulted EBV from the special BLUP evaluation gave identical estimates as original national EBV of the cows, even for some low heritability traits like female fertility. Somewhat lower EBV correlations, 0.98, were obtained for traits with binary values like longevity or calf survival. The validation study clearly demonstrated the reversibility of cow EBV deregression across all the evaluated traits.

For bulls with daughters in the bull EBV deregression validation or for cows with own records in the cow EBV deregression validation, we got equal EBV for these animals with own data in the validation studies. In addition, we obtained also equal EBV for ancestors of the two groups of animals with phenotype data. For instance, sires of the cows in the cow deregression validation had equal EBV as their EBV in the original conventional evaluation.

Besides the reversibility of the EBV deregression, we obtained also nearly equal reliability values from the deregression validation as original reliability for the bulls or cows with own data as well as for their ancestors. To ensure the reversibility of the deregression process, it is important that all cows with own records for cow EBV deregression or all bulls with daughters for bull EBV deregression must be included in their respective deregression process.

A GEBV deregression method for the single-step SNP BLUP model

The deregression of conventional EBV (Jairath et al. 1998) makes use of the inverse of numerator relationship matrix **A**. Masuda et al. (2021) suggested using the inverse of the genomic relationship matrix **H** (Misztal et al. 2009) for deregressing GEBV of the ssGBLUP model (Misztal et al. 2009). The idea of using the inverse of matrix **H** for the GEBV deregression had also been independently proposed by Esa Mäntysaari and Zengting Liu of the *Interbull Working Group Genomic-free EBV for MACE* in 2019. A major concern of the GEBV deregression using the $\mathbf{H}^{-1}$ matrix was raised because of a fast-increasing, large number of genotyped cows seen in many countries like the United States and Germany. Therefore, we propose here a GEBV deregression method assuming the ssSNPBLUP model (Liu et al. 2014) that can efficiently utilize all genotype data.

A GEBV deregression model

For deregressing GEBV of the single-step genomic model, we assume a special ssSNPBLUP model:

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{u} + \mathbf{e} \qquad [1]$$

where $\mathbf{y}$ is a vector of deregressed GEBV of animals with own phenotype data, **1** is a vector of 1s, µ is the general mean, $\mathbf{u}$ is a vector of GEBV for the animals with own phenotype data, and $\mathbf{e}$ is a vector of residuals. The deregressed GEBV $\mathbf{y}$ are unknown and will be estimated in the deregression process. Furthermore, it is assumed that

$$[var(\mathbf{e})]^{-1} = \mathbf{D}\sigma_e^{-2} = diag\{n_i\}\sigma_e^{-2} \quad [2]$$

where $\mathbf{D}$ is a diagonal matrix containing EDC of bulls with daughters or ERC of cows with own phenotype records on the animal-model basis, n_i , for animal i, $i = 1, \ldots, n$, and n is the number of animals with phenotype data. σ_e^2 is residual variance. We use genotype data of all animals, including culled animals or young animals, for the GEBV deregression. In addition, all animals with any source of phenotype data are included in the deregression process.

The definition of animals with own phenotype data depends on the applied genomic model. For the single-step evaluation with phenotype data stemming exclusively from one country or population, the animals with own phenotype data are usually domestic cows with own phenotypic records. In the single-step evaluation without foreign phenotype data, bulls with daughters are not considered as animals with own phenotype data here rather as ancestors, because their domestic daughters are treated as animals with own records.

As usual, we define non-genotyped animals as Group 1 and genotyped animals as Group 2 of animals. In contrast to the usual definition of the Group 1, we require here that all animals of Group 1 must have own phenotype records. Ancestors of the animals in Groups 1 and 2 are assigned to Group 0. Under the ssSNPBLUP Liu-Goddard model (Liu et al. 2014), GEBV of the genotyped animals $\mathbf{u}_2$ are:

$$\mathbf{u}_2 = \mathbf{Zg} + \mathbf{a}_2 \quad [3]$$

where $\mathbf{Z}$ is a design matrix of order n_g x m containing all genotypes, n_g is the number of the genotyped animals of Group 2, m is the number of SNP markers fitted, $\mathbf{g}$ is m x 1 vector of additive genetic effects of the SNP markers, and $\mathbf{a}_2$ represents residual polygenic (RPG) effects of the genotyped Group 2 animals. The RPG effects of the genotyped animals follow:

$$\mathbf{a}_2 \sim N(\mathbf{0}, k\sigma_u^2\mathbf{A}_{22}) \quad [4]$$

Where σ_u^2 is additive genetic variance, k is the proportion of additive genetic variance not explained by the SNP markers, and $\mathbf{A}_{22}$ is pedigree relationship for the genotyped animals. Furthermore, we assume that the SNP marker effects have a Normal distribution:

$$\mathbf{g} \sim N(0, (1-k)\sigma_u^2\mathbf{B}) \quad [5]$$

with matrix $\mathbf{B}$:

$$\mathbf{B} = \frac{1}{\sum_{j=1}^{m} 2p_j(1-p_j)}\mathbf{I} \quad [6]$$

where p_j represents allele frequency of SNP marker j.

For the single trait genomic model [1] we define a variance ratio:

$$\lambda = \sigma_e^2 \big/ \sigma_u^2 \quad [7]$$

For the deregression of GEBV from the single-step model [1], it is assumed that SNP effects, $\hat{\mathbf{g}}$, are known without error. The SNP effect estimates may be directly obtained from a genomic evaluation with the ssSNPBLUP model (Liu et al. 2014). For a genomic evaluation using the ssGBLUP model (Misztal et al. 2009), SNP effects can be back solved using GEBV of all reference animals:

$$\hat{\mathbf{g}} = (1-k)\mathbf{BZ}'\mathbf{G}_{22}^{-1}\hat{\mathbf{u}}_2 \quad [8]$$

where the weighted genomic relationship matrix is:

$$\mathbf{G}_{22} = (1-k)\mathbf{ZBZ}' + k\mathbf{A}_{22} \quad [9]$$

Mixed model equations for the deregression

As stated above for the single-step evaluation without foreign phenotype data, all cows with own phenotype data as well as all genotyped animals are treated as animals with data in the GEBV deregression. We denote the number of non-genotyped animals with own phenotype n_{NP}. Let n_T be the total number of genotyped or phenotyped animals. As mentioned before, ancestors of the genotyped (Group 2) or phenotyped and non-genotyped animals (Group 1) are denoted as Group 0. Unknown parent groups (UPG) for the special ssSNPBLUP model [1] are set up via the Quaas-Pollak transformation according to Vandenplas et al. (2021a). We use matrix $\mathbf{Q}_2$ to assign the genotyped animals to their UPG. Let $\mathbf{t}$ denote additive genetic effects of the UPG. Furthermore, we use a diagonal matrix for ERC for the non-genotyped animals with phenotype data (Group 1):

$$\mathbf{W}_1 = diag\{n_{11} \quad n_{21} \quad \cdots \quad n_{n_{NP}1}\} \qquad [10]$$

Likewise, ERC for the genotyped animals of Group 2 are represented with a diagonal matrix as well:

$$\mathbf{W}_2 = diag\{n_{12} \quad n_{22} \quad \cdots \quad n_{n_g2}\} \qquad [11]$$

If a genotyped animal has no phenotype, e.g., a young calf for a regular trait, then its ERC is 0.

Mixed model equations (MME) of the single-step model [1] are given in [12]. Under the assumption of known SNP effects, RHS of GEBV for all the genotyped animals can be modified prior to the deregression process, following the idea of Vandenplas et al. (2021b). MME without the SNP effects are given in [13]. The alternative form of MME [13] can be computationally more efficient because the adjusted RHS for GEBV of all the genotyped animals eliminates the steps of processing genotype data of potentially millions of animals during the solving process of MME.

Input data for the GEBV deregression are GEBV of all animals of Groups 1 and 2. It is important that all the animals with either own phenotype records or genotype data included in the original single-step evaluation must be chosen here for solving MME [12] or [13]. Because no foreign phenotype data are assumed in the original single-step evaluation, the animals with own phenotype data are only domestic cows for all dairy traits, except heifer fertility, calf survival or direct calving traits. Bulls, as sires of the cows, are not considered as the animals providing phenotype data for the deregression. ERC of cows on the animal-model basis, n_i, can be used as weights in $\mathbf{W}_1$ and $\mathbf{W}_2$ of MME [12] or [13].

Right-hand-sides (RHS) for the non-genotyped animals with own phenotypes (Group 1) or the genotyped animals (group 2) are respectively:

$$\mathbf{\Delta}_1 = \mathbf{W}_1\mathbf{y}_1 = \{n_{i1}y_i\} \qquad [14]$$

$$\mathbf{\Delta}_2 = \mathbf{W}_2\mathbf{y}_2 = \{n_{i2}y_i\} \qquad [15]$$

where n_{i1} or n_{i2} represent ERC of the i-th animal of Group 1 or 2, respectively.

Following the idea of the iterative deregression method by Jairath et al. (1998), the input GEBV of the n_T animals with either own phenotype or genotype data must be kept unchanged when solving MME [12] or [13], only GEBV of the ancestors (Group 0 animals, $\mathbf{u}_0$), the unknown parent groups ($\mathbf{t}$), and the general mean (μ) need to be estimated. The SNP effect estimates from the original single-step evaluation are assumed to be known in the deregression and must not be changed during the solving process either. An iterative solving algorithm, like pre-conditioned conjugate gradients, can be applied to solve MME [12] or [13]. Because of the unknown parent groups UPGs in MME [12] or [13] for either the genotyped or the non-genotyped animals derived using the Quaas-Pollak (QP) transformation (Vandenplas et al. 2021a), the GEBV of the ancestor animals automatically contain the effects of UPG (Quaas, 1988), like GEBV of the genotyped or non-genotyped animals.

Computing deregressed GEBV for the animals with phenotype data

When the solutions of MME [12] or [13] are converged, RHS for the non-genotyped animals with own phenotype data (Group 1) are calculated as:

$$\begin{aligned}\Delta_1 = {} & \mathbf{W}_1\mathbf{1}\hat{\mu} + \lambda\mathbf{A}^{10}\hat{\mathbf{u}}_0 \\ & +(\mathbf{W}_1 + \lambda\mathbf{A}^{11})\hat{\mathbf{u}}_1 + \lambda\mathbf{A}^{12}\hat{\mathbf{u}}_2 \qquad [16]\end{aligned}$$

RHS for the genotyped animals (Group 2) are computed with:

$$\begin{aligned}\Delta_2 = {} & \mathbf{W}_2\mathbf{1}\hat{\mu} + \lambda\mathbf{A}^{20}\hat{\mathbf{u}}_0 + \lambda\mathbf{A}^{21}\hat{\mathbf{u}}_1 \\ & + \left(\mathbf{W}_2 + \lambda\left(\mathbf{A}^{22} + \left(\tfrac{1}{k} - 1\right)\mathbf{A}_{22}^{-1}\right)\right)\hat{\mathbf{u}}_2 \\ & -\lambda\tfrac{1}{k}\mathbf{A}_{22}^{-1}\mathbf{Z}\hat{\mathbf{g}} \qquad [17]\end{aligned}$$

If animal i with own phenotype is not genotyped, its deregressed GEBV is:

$$\hat{y}_i = {}^{\Delta_{i1}}/_{n_{i1}} \qquad [18]$$

Otherwise, deregressed GEBV is for the animal *i* with phenotype and genotype data:

$$\hat{y}_i = {}^{\Delta_{i2}}/_{n_{i2}} \qquad [19]$$

where Δ_{i1} or Δ_{i2} are the *i*-th element of RHS [16] or [17] corresponding to animal *i*, respectively. Because animals without own phenotype data have zero EDC or ERC, their deregressed GEBV are not defined. Therefore, only animals with own phenotype data receive deregressed GEBV.

Computing deregressed GEBV for cows with own phenotype records

Depending on the availability of own genotype data, a cow with phenotype records receives its deregressed GEBV with Equation either [18] or [19]. Because ERC of the cow may be rather small for traits with low heritability, Equation [18] or [19] may lead to a DRP having extreme value for the cow. A practical way for avoiding such extreme DRP values is to add a constant to ERC of all cows (or EDC of all bulls) in MME [12] or [13]. This remedy seemed to work well for cow EBV deregression in German Holsteins (unpublished data). By adding a constant to EDC or ERC of all animals with phenotype data also improved the rate of convergence, as the MME for the deregression became more diagonal-dominant.

It is important to make sure that the same genotype and pedigree data are used for the GEBV deregression as in the original single-step evaluation. The SNP effect estimates from the single-step evaluation must be kept unchanged during the deregression process. Also, GEBV of all genotyped animals and GEBV of all cows with own phenotype records must not be changed during the deregression process. As in the original single-step evaluation, the same animals with genotypes, the same cows with phenotype records and the same ancestors plus phantom parent groups must be included for the GEBV deregression.

Computing deregressed GEBV for bulls with daughters

In the conventional bull EBV deregression process (Jairath et al. 1998), bulls with daughters are treated as animals with own phenotype data. Full pedigree of the bulls, based on sires and dams of the bulls, is usually considered. This bull EBV deregression procedure may be called a ***bull-model deregression***. The deregressed bull EBV from a national conventional evaluation are routinely used as trait values for the conventional bull MACE evaluation at Interbull.

The proposed GEBV deregression method above is based on the animal model and it can use genotype and phenotype data of all male or female animals, including also genotype data of young animals. Like the single-step evaluation, the ***animal-model deregression*** of GEBV treats bulls with daughters as ancestors (Group 0), not as animals with direct phenotype information of their own, in case no foreign bulls are considered. RHS of the ancestors (Group 0) in MME [12] or [13] differ with RHS of animals of Group 1 or 2 only in EDC or ERC matrix $\mathbf{W}_1$ for non-genotyped animals or $\mathbf{W}_2$ for genotyped animals, and the ancestors have zero EDC or ERC. However, we can use the RHS of MME [12] or [13] for the bulls, as sires of cows, to derive DRP for the bulls with Equation [18] or [19], as if we conducted a ***bull-model deregression*** of GEBV. Matrix $\mathbf{W}_1$ and $\mathbf{W}_2$ would contain conventional EDC for the non-genotyped or genotyped bulls with daughters. We would apply Equation [16] or [17] to calculate RHS for the non-genotyped or genotyped bulls, respectively. Deregressed GEBV for the bulls with daughters would be computed using Equation [18] or [19].

$$\begin{bmatrix} \mathbf{1}'\mathbf{W}_1\mathbf{1}+\mathbf{1}'\mathbf{W}_2\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1}'\mathbf{W}_1 & \mathbf{1}'\mathbf{W}_2 & \mathbf{0} \\ & \lambda\mathbf{A}^{tt}+\lambda(\frac{1}{k}-1)\mathbf{Q}_2{}'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 & \lambda\mathbf{A}^{t0} & \lambda\mathbf{A}^{t1} & \lambda\mathbf{A}^{t2}-\lambda\left(\frac{1}{k}-1\right)\mathbf{Q}_2{}'\mathbf{A}_{22}^{-1} & \lambda\frac{1}{k}\mathbf{Q}_2{}'\mathbf{A}_{22}^{-1}\mathbf{Z} \\ & & \lambda\mathbf{A}^{00} & \lambda\mathbf{A}^{01} & \lambda\mathbf{A}^{02} & \mathbf{0} \\ & & & \mathbf{W}_1+\lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{12} & \mathbf{0} \\ & & & & \mathbf{W}_2+\lambda(\mathbf{A}^{22}+\left(\frac{1}{k}-1\right)\mathbf{A}_{22}^{-1}) & -\lambda\frac{1}{k}\mathbf{A}_{22}^{-1}\mathbf{Z} \\ & symm. & & & & \lambda(\frac{1}{1-k}\mathbf{B}^{-1}+\frac{1}{k}\mathbf{Z}'\mathbf{A}_{22}^{-1}\mathbf{Z}) \end{bmatrix}$$

$$\mathrm{X}\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{t}} \\ \hat{\mathbf{u}}_0 \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \hat{\mathbf{g}} \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{W}_1\mathbf{y}_1+\mathbf{1}'\mathbf{W}_2\mathbf{y}_2 \\ \mathbf{0} \\ \mathbf{0} \\ \boldsymbol{\Delta}_1 \\ \boldsymbol{\Delta}_2 \\ \mathbf{0} \end{bmatrix} \qquad [12]$$

$$\begin{bmatrix} \mathbf{1}'\mathbf{W}_1\mathbf{1}+\mathbf{1}'\mathbf{W}_2\mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{1}'\mathbf{W}_1 & \mathbf{1}'\mathbf{W}_2 \\ & \lambda\mathbf{A}^{tt}+\lambda(\frac{1}{k}-1)\mathbf{Q}_2{}'\mathbf{A}_{22}^{-1}\mathbf{Q}_2 & \lambda\mathbf{A}^{t0} & \lambda\mathbf{A}^{t1} & \lambda\mathbf{A}^{t2}-\lambda\left(\frac{1}{k}-1\right)\mathbf{Q}_2{}'\mathbf{A}_{22}^{-1} \\ & & \lambda\mathbf{A}^{00} & \lambda\mathbf{A}^{01} & \lambda\mathbf{A}^{02} \\ & & & \mathbf{W}_1+\lambda\mathbf{A}^{11} & \lambda\mathbf{A}^{12} \\ & symm. & & & \mathbf{W}_2+\lambda(\mathbf{A}^{22}+\left(\frac{1}{k}-1\right)\mathbf{A}_{22}^{-1}) \end{bmatrix}$$

$$\mathrm{X}\begin{bmatrix} \hat{\mu} \\ \hat{\mathbf{t}} \\ \hat{\mathbf{u}}_0 \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{1}'\mathbf{W}_1\mathbf{y}_1+\mathbf{1}'\mathbf{W}_2\mathbf{y}_2 \\ -\lambda\frac{1}{k}\mathbf{Q}_2{}'\mathbf{A}_{22}^{-1}\mathbf{Z}\hat{\mathbf{g}} \\ \mathbf{0} \\ \boldsymbol{\Delta}_1 \\ \boldsymbol{\Delta}_2+\lambda\frac{1}{k}\mathbf{A}_{22}^{-1}\mathbf{Z}\hat{\mathbf{g}} \end{bmatrix} \qquad [13]$$

Alternatively, we can derive DRP of bulls with daughters using a ***bull-model GEBV deregression*** method by treating bulls with daughters as animals with own phenotype data, like in the conventional bull EBV deregression (Jairath et al. 1998). Some daughters of the bulls may be genotyped, therefore, those genotyped daughters with own phenotype records must be treated as animals with data as well in this GEBV deregression. To avoid double counting of the contribution by the genotyped daughters, EDC of the bulls must be subtracted by the contribution of their genotyped daughters, following the Interbull document (2018), the bulls would then represent only non-genotyped daughters in the special GEBV deregression procedure.

Genotyped daughters with own phenotype records would receive DRP as the bull himself. Suppose the bull i has n_{ni} non-genotyped daughters with phenotype records and n_{gi} genotyped daughters with phenotype data. Final deregressed GEBV for the *i*-th bull is calculated by combining DRP of the bull ($\hat{y}_i$) representing his non-genotyped daughters and DRPs of all his genotyped daughters ($\hat{y}_{ij}$):

$$\hat{y}_i^{final} = \frac{\sum_j^{n_{ni}} w_j}{\sum_j^{n_{gi}+n_{ni}} w_j}\hat{y}_i + 2\frac{1}{\sum_j^{n_{gi}+n_{ni}} w_j}\sum_j^{n_{gi}} w_j\ (\hat{y}_{ij} - \tfrac{1}{2}\hat{u}_{dam\_j}) \qquad [20]$$

where $\hat{y}_i^{final}$ is the final deregressed GEBV of the bull *i* as weighted average of deregressed GEBV of his individual genotyped daughters and all his non-genotyped daughters, $\hat{y}_i$ is DRP of this bull based on only the non-genotyped daughters, $\hat{y}_{ij}$ is deregressed GEBV of the *j*-th genotyped daughter of this bull, $\hat{u}_{dam\_j}$ is GEBV of dam of the *j*-th genotyped daughter of this bull *i*, w_j represents weight contributed by the genotyped daughter *j*:

$$w_j = \varphi_{ij}/(\varphi_{ij} + \lambda) \qquad [21]$$

where φ_{ij} is ERC contribution by the j-th daughter to her sire, the bull i, adjusted for her dam data contribution (Interbull, 2018), and

$$\lambda = (1 - h^2)/h^2 \qquad [22]$$

with h^2 represents heritability of the analyzed trait.

Single-step evaluation with foreign bull phenotype data included

When all phenotype data originate within a country or population and no foreign phenotype data are included in either the single-step evaluation or the later GEBV deregression process, conventional EDC of the bulls based on only domestic daughters can be used for calculating the bulls DRP via Equation [18] or [19].

If foreign bull phenotype information is included in the single-step like Alkhoder and Liu (2021) and the subsequent GEBV deregression, then for a bull *i* with foreign daughter information his EDC in Equation [18] or [19] should be the sum of EDC contributed by his domestic and foreign daughters, not the difference between MACE and national EDC which was used in the single-step evaluation with the integrated foreign bull phenotype data (Alkhoder and Liu, 2021).

A validation of the GEBV deregression for the single-step evaluation

As a validation for the proposed GEBV deregression method, a designated genomic evaluation based on the special single-step SNP BLUP model [1] needs to be conducted. The input pseudo-phenotype data for the validation study are the deregressed GEBV for all animals that had own phenotype data in the original single-step evaluation. All genotyped animals that were evaluated in the original single-step model must be included in this validation, too. For genotyped animals without own phenotypes, e.g. young candidates, their phenotype data are treated as missing in the validation study, because those genotyped animals without own phenotype records have no deregressed GEBV defined. The SNP effect estimates from the original single-step evaluation must be assumed to known without error. In contrast to MME [12], MME [13] with a modified RHS based on the SNP effects for the genotyped animals can be solved as a pure conventional model. After the solutions of MME [12] or [13] are converged, GEBV of all the animals with phenotype or genotype data are compared to the original single-step genomic evaluation. If the GEBV from the validation evaluation are equal to those from the original single-step evaluation, then the GEBV deregression process is proven to be correct and the GEBV degression process is said to be reversible.

Results & Discussion

Generation of accurate pseudo-phenotype data is required by numerous statistical analyses in animal breeding. For example, the bull MACE evaluation requires for bulls with daughters' pseudo-phenotype data that are free of any environmental and non-genetic random effects. For the two-step genomic evaluation with a mixed reference population of cows and bulls (Liu et al. 2019), pseudo-phenotype data are needed for the reference cows, in addition to the reference bulls. DYD of bulls or yield deviations (YD) of cows (VanRaden and Wiggans, 1991) are a form of pseudo-phenotype data for bulls or cows. However, the calculation of DYD or YD depends on the complexity of genetic evaluation model. For the random-regression test-day model, DYD of bulls or YD of cows are rather difficult to compute and without guarantee of no extreme values (Liu et al., 2004). In contrast, the EBV or GEBV deregression can be done using a single trait model and thus by-pass the technical difficulty associated with the complex multi-trait models. That is why we prefer the EBV or GEBV deregression methods to the DYD or YD calculation for the purpose of generating pseudo-phenotype data.

In last decades conventional EBV of bulls or cows have been accurately deregressed using the iterative, matrix deregression method by Jairath et al. (1998) to obtain unbiased pseudo-phenotype data for the bull MACE evaluation at Interbull or for domestic reference cows of the multi-step genomic model in German Holsteins. With the routine use of the single-step model, either ssGBLUP or ssSNPBLUP, the proposed deregression of GEBV provides a feasible way for generating unbiased pseudo-phenotype data from the single-step evaluation. Using the inverse of the joint genomic relationship matrix **H** of the ssGBLUP model, GEBV deregression can be conducted (Masuda et al. 2021), following the same principle of the deregression of conventional EBV. Due to the fast-increasing number of genotyped animals, the **H** matrix might become too large to be inverted in coming years for Holstein breeds in some large countries. However, the proposed GEBV deregression that is based on the efficient ssSNPBLUP model should be feasible to evaluate millions of genotyped animals.

To obtain accurate DRP for all animals with phenotype data in the single-step evaluation, the same genotype data and pedigree file must be used in the GEBV deregression process as in the original single-step evaluation. The same animals with genotype data or phenotype data must be included in the GEBV deregression as in the single-step evaluation as well. In fact, the GEBV deregression can be seen as an exact reverse engineering process of the original single-step evaluation which may be interpreted as a regression process of phenotype data on breeding values. Based on the idea of using the same phenotype and genotype data as in single-step evaluation, we also proposed here a validation procedure for the GEBV deregression.

One major concern about the deregressed GEBV for the bull MACE evaluation is that DRP of the bulls from the single-step evaluation may lead to an inflation of the conventional MACE evaluation. However, the possibly inflated variance of evaluation can be avoided if we can exactly reverse-engineer the single-step evaluation in the GEBV deregression process.

Obviously, the single-step evaluation and the subsequent GEBV deregression must not include foreign bulls' phenotype data for the purpose of submitting DRP of bulls to the current conventional MACE evaluation.

The current bull MACE evaluation relies on accurate DRP of bulls from national conventional evaluation. Because of the strong genomic selection in Holsteins, we can no longer guarantee unbiased conventional EBV and consequently unbiased conventional DRP of these bulls. Therefore, we developed the GEBV deregression method for the single-step genomic evaluation in order to meet the requirement of the conventional bull MACE evaluation.

Interbull performs bull EBV deregression at Interbull centre routinely for the bull MACE evaluation. However, due to the requirement of direct access to genotype data by our GEBV deregression method, each country must perform the GEBV deregression by itself. To facilitate the GEBV deregression by all the countries, a common software might be jointly developed and shared by all the countries. In this study we made only a theoretical contribution to the topic of GEBV deregression for the single-step model. The proposed deregression method must be verified with real data analysis.

Conclusions

For the bull MACE evaluation, national conventional EBV of bulls with daughters are routinely deregressed with the iterative, matrix deregression method by Jairath et al. (1998). It is also relevant and important to generate similar pseudo-phenotype data for bulls with daughters and cows with records in the era of single-step genomic evaluation. In this study we presented a GEBV deregression method based on a special single-step SNP BLUP model (Liu-Goddard). All animals with phenotype data and all animals with genotypes, including genotyped young animals and culled candidates, must be considered in the GEBV deregression process as in the original single-step evaluation. We use the SNP effect estimates from the original single-step evaluation for the deregression process. Two formulae were presented for computing deregressed GEBV for phenotyped animals, e.g. cows with own phenotype records, with or without genotype data. For bulls with daughters, two alternative GEBV deregression methods were proposed, animal-model and bull-model based deregression. To validate the proposed GEBV deregression methods, we suggested a reversibility test for the GEBV deregression. Thanks to the efficient single-step SNP BLUP model, the proposed GEBV deregression method should be feasible for processing millions of genotyped animals.

Acknowledgements

Members of Interbull working group of Genomic-free EBV for MACE are thanked for their valuable discussion.

References

Calus, M. P. L., Vandenplas, J., ten Napel, J., and Veerkamp, R. F. 2016 *J Dairy Sci.* 99:6403-6419.

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S., and Lawlor, T. J. 2010. *J. Dairy Sci.* 93:743-752.

Alkhoder, H., Z. Liu, F. Reinhardt, and R. Reents. 2017. *Interbull Bulletin* 51:86-90.

Alkhoder, H., and Z. Liu. 2021. Application of a single-step SNP BLUP model to conformation traits of German Holsteins. Interbull annual meeting, June 2021. *Interbull Bulletin* (in press).

Interbull, 2018. A supplementary document to the Interbull genomic reliability method. https://interbull.org/static/web/A_supplementary_document_to_the_Interbull_genomic_reliability_method-1.pdf.

Jairath, L., Dekkers, J.C.M., Schaeffer, L.R., Liu, Z., Burnside, E.B., and Kolstad, B. 1998. *J. Dairy Sci.* 81:550-562.

Liu, Z., F. Reinhardt, A. Bünger, and R. Reents 2004. *J. Dairy Sci.* 87:1896–1907.

Liu, Z. 2011. Use of MACE Results as Input for Genomic Models. *Interbull Bulletin* 43.

Liu, Z., Goddard, M. E., Reinhardt, F., and Reents, R. 2014. *J. Dairy Sci.* 97, 5833-5850.

Liu, Z., H. Alkhoder, Ch. Kubitz, K. Stock, J. Heise, D. Segelke, F. Reinhardt, and R. Reents. 2019. *Interbull Bulletin* 55:35-45.

Mark, T., W. F. Fikse, U. Emanuelson, and J. Philipsson. 2002. *J. Dairy Sci.* 85:2393–2395.

Masuda Y., Liu, Z., and Sullivan, P. 2021. Interbull annual meeting. 2021.

Misztal, I., Legarra, A., and Aguilar, I. 2009. *J. Dairy Sci.* 92, 4648-4655.

Quaas, R. L. 1988 *J Dairy Sci*. 71:91-98.

Vandenplas, J., Eding, H., and Calus, M. P. L. 2021a. *J. Dairy Sci.* 104:3298-3303.

Vandenplas, J., 2021b. Interbull annual meeting. 2021.

VanRaden, P.M., and G.R. Wiggans. 1991. *J. Dairy Sci.* 74:2737-2746.