Integration of single step DGV in conventional genetic evaluations using DGV-PBLUP

H. Eding[1], L. van Kempen[1], M. van Pelt[1] and J. Vandenplas[2]

[1] Animal Evaluation Unit, Cooperative CRV u.a., Rijksweg West 2, 6842 BD Arnhem, The Netherlands

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

Abstract

Routine single-step genomic evaluations can be costly in time and computer resources. Hence, newly genotyped animals initially receive a genomic prediction of their direct genomic values (DGV). If genomic predictions of DGVs of animals become available, it may be convenient to estimate GEBV of such animals using some form of integration into conventional pedigree BLUP evaluations. DGV-PBLUP is a novel method of integration of DGV from genomic predictions, into a conventional pedigree BLUP (PBLUP) evaluation. This is done by setting the prior mean of the animal genetic effect (which usually is zero in linear mixed models) to the DGV to be incorporated. In this paper we report on the application of this methodology to the Dutch-Flemish genetic evaluation. Results showed a high correlation (0.99 or higher) between GEBV of animals associated with a genotype in single step SNP-BLUP and a GEBV in DGV-PBLUP. Run time of DGV-PBLUP evaluations were comparable to conventional pedigree evaluations and much shorter than routine single-step SNP-BLUP evaluations. DGV-PBLUP promises to be a convenient method of integration of genomic information into pedigree BLUP evaluations, without the need for sharing or accessing SNP genotypes.

Key words: Single-step, pedigree BLUP, integration, genomic evaluations

Introduction

Routine single-step evaluations can be costly in time and computer resources. Hence, newly genotyped animals initially receive a genomic prediction of their direct genomic values (DGV). If for some animals genomic predictions of DGV become available it still may be convenient to compute their GEBV using some form of integration into conventional pedigree BLUP evaluations.

There may also be cases where only DGV of animals are available for evaluation, without genotype data, due to legal or legislative considerations. This is the case at CRV, which consists of a commercial half, the corporation, and a cooperative half, with dairy farmer membership. The cooperative publishes national genetic evaluations. However, the single step evaluation is corporately owned. For reasons of IP protection, the corporation cannot share genotypes or allele substitution effects. The cooperative and the corporation have entered in an agreement, where the corporation supplies the cooperative with DGV for inclusion in the national evaluation. If such DGV are to be used in national genetic evaluations, integration is still required.

Integration of genomic data into genetic evaluations has been a long standing subject in the field of animals quantitative genetics and breeding. Methods of integration saw an evolution from linear post-processing after evaluation, via methods using pseudo-records during evaluation, where DGV are fitted as observations on a pseudo-trait added to the evaluation and correlated to the target trait (Stoop *et al*.; 2014) to single-step models, where genotypes are fitted in the statistical model of evaluations. Integration methods of genomic information were successful in achieving their stated goals, but true

equivalency between such methods and single-step evaluations were not achieved.

In this paper we present a model of integration that is mathematically equivalent to single-step SNP-BLUP (ssSNPBLUP) models, but only requires DGV of genotyped animals, in addition to conventional phenotypic and pedigree data.

Materials and Methods

Model

The equations of the model were derived from the ssSNPBLUP linear equations proposed by Liu et al. (2014). If we assume that estimates of SNP effects $\hat{\mathbf{g}}$ are known before performing a single-step genomic prediction, then the vector $\mathbf{d}$ with predicted DGV of genotyped animals can be computed as $\mathbf{d} = \mathbf{Zg}$, where $\mathbf{Z}$ is the genotyped matrix centered with observed allele frequencies, and we can assume the following prior multivariate normal (MVN) distribution for the genetic additive effects $\mathbf{u}$:

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

with

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

and

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

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

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}_{nn} & \mathbf{A}_{ng} \\ \mathbf{A}_{gn} & \mathbf{A}_{gg} \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{A}^{nn} & \mathbf{A}^{ng} \\ \mathbf{A}^{gn} & \mathbf{A}^{gg} \end{bmatrix}$$

is the inverse of the pedigree relationship matrix partitioned between genotyped and ungenotyped animals, w is the proportion of additive genetic variance explained by the residual polygenic effects, σ_u^2 is the genetic variance, $\mathbf{d}$ is the vector with DGV of genotyped animals, and $\mathbf{I}$ is an identity matrix.

The system of equations associated with these assumptions, hereafter called DGV-PBLUP, is written as follows:

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

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

The system of equations of DGV-PBLUP is equivalent to a single-step genomic evaluation, provided that the SNP effects $\hat{\mathbf{g}}$ were estimated using the same phenotypic, genomic and pedigree information (Vandenplas et al, 2021). The system of equations of DGV-PBLUP can also be considered as an application of the Bayesian procedure to integrate external information into genetic evaluations (Legarra et al., 2007; Vandenplas and Gengler, 2012), where, in essence a prior mean is fitted for all animals, based on the (imputed) DGV of (un)genotyped animals.

Data

The DGV-PBLUP method was tested on a dataset and associated variance components of the milk production test day model (TDM), which is a 5 lactation, 4th order random regression with Legendre polynomials (5 regressions per lactation), analyzing milk, fat, protein and lactose yield, as well as somatic cell score and urea content of milk.

Phenotypic data were taken from the April 2025 evaluation of CRV. DGV for each of the 25 regressions were taken from a genomic prediction based on SNP effect estimates from a single-step SNPBLUP evaluation on the same phenotypic data (April ’25). The latter included 828,590 genotyped animals. The data in the DGV-PBLUP evaluation consisted of 16,382,568 pedigreed animals, 13,662,463 of which had phenotypic data. Also included were DGV of 851,704 animals.

GEBV from DGV-PBLUP were compared to results from the corresponding single-step SNPBLUP run. GEBV from the current pseudo-record evaluation (Stoop et al., 2014) were also contrasted to these. GEBV were produced for the following traits or trait groups:

1) Milk production (lac. 1-5 and overall)
2) Fat production (lac. 1-5 and overall)
3) Protein production (lac. 1-5 and overall)
4) Lactose production (lac. 1-5 and overall)
5) Somatic cell score (lac. 1-5 and overall)
6) Urea content (lac. 1-5 and overall)

Presented in this paper are the comparison of GEBV from ssSNPBLUP and DGV-PBLUP for young bulls without progeny, born after 2020, since this group of animals is the most sensitive to changes in genomic information in an evaluation. For the overall traits the Pearson correlation were calculated, as well as the fraction of animals whose GEBV differed less than a quarter genetic standard deviation, as an indication of GEBV stability. For reference the same statistics were produced from the current pseudo-record (PSR) method of integrating genomic information into the national evaluation.

Results & Discussion

Breeding values

A comparison of the GEBVs from DGV-PBLUP and the current PSR system for overall GEBV of traits in the milk production test-day random regression model are presented in Table 1. Correlations with ssSNPBLUP GEBV were clearly improved with DGV-PBLUP, with all correlations > 0.99. Changes in GEBV from ssSNPBLUP to integrated GEBV were also considerably smaller for DGV-PBLUP, with virtually all GEBV with ¼ genetic standard deviation. This also indicates a considerable improvement in GEBV stability compared to the PSR system, where the fraction of animals changing more than ¼ s.d. was considerably larger.

An attractive feature of the DGV-PBLUP method is that no extra correlated traits have to be fitted to incorporate DGV information in a pedigree BLUP evaluation. Neither does it require a post-processing step to integrate DGV.

Table 1. Number of selected bulls, correlations with ssSNPBLUP GEBV and fraction of animals whose GEBV changed less than ¼ genetic standard deviation for the DGV-PBLUP method (*dgv*) and the current pseudorecord method of integration (*psr*).

Trait		*Correlation*		*< 1/4 s.d.*	
	N	*dgv*	*psr*	*dgv*	*psr*
Milk	5,629	0.999	0.928	100.0%	72.2%
Fat	5,629	0.999	0.964	100.0%	80.1%
Protein	5,629	0.998	0.934	100.0%	74.8%
Lactose	5,629	0.995	0.972	99.8%	85.3%
SCS	5,629	0.999	0.994	100.0%	98.3%
Urea	5,629	0.999	0.918	100.0%	70.5%

Selected were young bulls without progeny born after 2020.

Table 2. Run times of genetic evaluations of the milk production test day model. Run times are given in hours:minutes for routine ssSNPBLUP evaluations, DGV-PBLUP and conventional pedigree BLUP evaluations.

Trait	*ssSNPBLUP*	*DGV-PBLUP*	*PBLUP*
Milk	51:28	18:17	17:53
Fat	47:26	18:39	17:37
Protein	49:05	19:20	18:08
Lactose	48:47	19:02	19:30
SCS	56:02	18:22	18:31
Urea	63:00	18:49	19:45

All evaluations were run using 5 threads for parallel computing on a server with Intel(R) Xeon(R) Gold 6448H 64bit chips at 4000MHz.

Run time

The wall clock times of all evaluations are presented in Table 2. All evaluations were run without starting values. The run times of DGV-PBLUP were comparable to the run times of conventional pedigree BLUP evaluations, as expected. The run times of routine single-step SNPBLUP evaluations on average were 2.4 times longer than either conventional or DGV-PBLUP evaluations.

Conclusions

DGV-PBLUP presents itself as a superior method of integrating genomic data into conventional pedigree BLUP evaluations, in the sense that it replicates more closely the results of a routine single-step SNPBLUP run than the PSR method of integration currently implemented at CRV. DGV-PBLUP promises to be a convenient method of integrating genomic information into pedigree BLUP evaluations, without the need for sharing SNP genotypes.

References

Legarra, A., J. K. Bertrand, T. Strabel, R. L. Sanchez, and I. Misztal. 2007. Multi-breed genetic evaluation in a Gelbvieh population. J. Anim. Breed. Genet. 124:286–295. DOI: 10.1111/j.1439-0388.2007.00671.x.

Liu Z., M. E. Goddard, F. Reinhardt, and R. Reents. 2014. A single-step genomic model with direct estimation of marker effects, J. Dairy Sci. 97 :5833–5850. DOI: 10.3168/jds.2014-7924

Stoop W.M., H. Eding, M.L. van Pelt, L.C.M. de Haer and G. de Jong. 2014. Using Pseudo-observations to combine Genomic and Conventional Data in the Dutch National Evaluation. Proc. 10th WCGALP, Vancouver, Canada. DOI: 10.13140/2.1.3694.2409

Vandenplas J., Eding H. and Calus M. 2021. Interim genomic prediction considering newly acquired genotypes and phenotypes, Interbull Bulletin 56. https://journal.interbull.org/index.php/ib/article/view/87

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. DOI: 10.3168/jds.2011-4322