Tuning Indirect Predictions Based on SNP Effects from Single-Step GBLUP

***D.A.L. Lourenco*[1]*, A. Legarra*[2]*, S. Tsuruta*[1]*, D. Moser*[3]*, S. Miller*[3]*, and I. Misztal*[1]**

[1]*Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602*
[2]*Institut National de la Recherche Agronomique, UMR1388 GenPhySE, Castanet Tolosan, France 31326*
[3]*Angus Genetics Inc., St. Joseph, MO 64506*

Abstract

The objectives of this study were to investigate whether SNP effects can be accurately estimated when the algorithm for proven and young (APY) is used in single-step GBLUP (ssGBLUP), and how close indirect predictions, based on SNP effects, are to GEBV from regular ssGBLUP. Tests involved an American Angus dataset with 8 million animals in the pedigree. Among those, 80 993 were genotyped. Validation animals (15 040) were born from 2013 to 2014. The reduced dataset had genotypes and phenotypes up to 2012; the complete dataset had genotypes up to 2014 and was used to obtain the benchmark GEBV. Based on the reduced dataset, GEBV were calculated using regular ssGBLUP with direct inversion of **G** ($\mathbf{G}^{-1}$), and APY ssGBLUP ($\mathbf{G}^{-1}_{\mathrm{APY}}$) with 11 000 core animals. The SNP effects were calculated based on a) $\mathbf{G}^{-1}$, b) $\mathbf{G}^{-1}_{\mathrm{APY}}$, or c) the inverse of the core portion of **G** ($\mathbf{G}^{-1}_{\mathrm{cc}}$). Direct genomic values (DGV) for validation animals were obtained as the sum of SNP effects weighted by the genotype content, and the difference between pedigree and genomic base was added to obtain indirect predictions. Correlations between SNP effects obtained with $\mathbf{G}^{-1}$ and $\mathbf{G}^{-1}_{\mathrm{APY}}$ were > 0.99; the lower correlation (0.93) was observed when using $\mathbf{G}^{-1}_{\mathrm{core}}$. Correlations between the benchmark GEBV and DGV from $\mathbf{G}^{-1}$, $\mathbf{G}^{-1}_{\mathrm{APY}}$, and $\mathbf{G}^{-1}_{\mathrm{core}}$ were all 0.99. The average difference between benchmark GEBV and DGV was 113.95, indicating a large bias. Indirect predictions that include DGV and the difference between pedigree and genomic base were less biased, and therefore, comparable to GEBV. Accurate indirect predictions can be obtained when APY ssGBLUP is used. Backsolving genomic predictions to SNP effects may require only a group of genotyped animals representing the dimensionality of the genomic information.

Key words: algorithm for proven and young, direct genomic value, interim evaluations

Introduction

Genomic BLUP (GBLUP) and SNP-BLUP are equivalent models (VanRaden, 2008), therefore, SNP effects can be derived from GEBV and the inverse of the genomic relationship matrix (**G**) in GBLUP and single-step GBLUP (ssGBLUP) when needed. With the increasing number of genotyped animals, obtaining SNP effects may have a high computational cost. To solve this issue, $\mathbf{G}^{-1}$ from the algorithm for proven and young (APY) can be potentially used.

If SNP effects are available, indirect predictions (IP) can be calculated for young genotyped animals in between official ssGBLUP runs, as an interim evaluation (Lourenco *et al.,* 2015). Indirect predictions may also be useful for genotyped animals that have incomplete pedigree. Such animals can increase bias and reduce reliability of GEBV if included in official ssGBLUP evaluations (Yutaka Masuda, personal communication), given their coefficients in **G** are not compatible to the ones in the pedigree relationship matrix (**A**).

Lourenco *et al.* (2015) showed that the correlation between IP and GEBV was greater than 0.99 for large genotyped populations. However, averages were very different, meaning IP could not be compared to GEBV. The objectives of this study were to fine-tune IP to be compatible with GEBV and to investigate whether SNP effects can be accurately estimated when APY is used in ssGBLUP.

Materials and Methods

Data

The dataset was provided by the American Angus Association (AAA) and contained phenotypes for birth weight (BW), weaning weight (WW), and post-weaning gain (PWG). Table 1 shows the number of animals with records and heritability for all traits. A total of 80 993 animals were genotyped using the BovineSNP50k v2 BeadChip (Illumina Inc., San Diego, CA). A total of 38 421 segregating SNP remained after quality control.

Table1. Heritability (h^2), numbers of records and animals with records and genotypes

Trait	h^2	Number of records	Genotyped animals with records
BW	0.46	6 177 145	51 844
WW	0.28	6 877 731	52 970
PWG	0.27	3 413 415	40 891

A reduced dataset had genotypes and phenotypes up to 2012, whereas a complete dataset had genotypes up to 2014 and was used to obtain the benchmark GEBV. Validation animals (N = 15 040) were born in 2013 and 2014, and had genotypes excluded from the reduced dataset. Based on the reduced dataset, GEBV were calculated and backsolved to SNP effects. Details about the three-trait model used in this study are provided in Lourenco *et al.* (2015).

SNP effects from ssGBLUP with APY

The SNP effects were calculated based on:

a) $\mathbf{G}^{-1}$, using the formula derived by Wang *et al.* (2012): $\boldsymbol{\hat{a}} = \boldsymbol{\lambda D\, Z'}\mathbf{G}^{-1}\boldsymbol{\hat{u}}$; where $\boldsymbol{\hat{a}}$ is a vector of SNP effects, $\boldsymbol{\lambda}$ is the ratio of SNP to additive genetic variance, $\boldsymbol{D}$ is a diagonal matrix of weights for SNP (identity), $\boldsymbol{Z}$ is a matrix of SNP content, $\boldsymbol{\hat{u}}$ is GEBV from regular ssGBLUP.

b) $\mathbf{G}^{-1}$ computed with APY ($\mathbf{G}^{-1}_{\mathrm{APY}}$): $\boldsymbol{\hat{a}} = \boldsymbol{\lambda D\, Z'}\mathbf{G}^{-1}_{\mathrm{APY}}\boldsymbol{\hat{u}}_{\mathrm{APY}}$; where $\boldsymbol{\hat{u}}_{\mathrm{APY}}$ is GEBV obtained with APY ssGBLUP. A total of 11 000 core animals were used. Core definition was either high reliability (H) or randomly chosen (R) animals. In APY, the inverse of **G** is constructed as (Misztal, 2016):

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}^{-1}_{\mathrm{cc}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}^{-1}_{\mathrm{cc}}\mathbf{G}_{\mathrm{cn}} \\ \mathbf{I} \end{bmatrix} \mathbf{M}^{-1}_{\mathrm{nn}} \begin{bmatrix} -\mathbf{G}_{\mathrm{nc}}\mathbf{G}^{-1}_{\mathrm{cc}} & \mathbf{I} \end{bmatrix}$$

where *c* represents the core animals and *n* the non-core.

c) $\mathbf{G}^{-1}$ only for the 11 000 core animals ($\mathbf{G}^{-1}_{\mathrm{cc}}$), using either H or R core: $\boldsymbol{\hat{a}} = \boldsymbol{\lambda D\, Z'}\mathbf{G}^{-1}_{\mathrm{cc}}\boldsymbol{\hat{u}}_{\mathrm{cc}}$; where $\boldsymbol{\hat{u}}_{\mathrm{cc}}$ is GEBV for core animals obtained with APY ssGBLUP.

Indirect predictions for validation animals were obtained as the sum of SNP effects weighted by the genotype content. Correlations between the benchmark scenario ($\mathbf{G}^{-1}$) and the other scenarios were calculated for SNP effects and IP.

Adjustments for indirect predictions

Having IP in the same scale of GEBV enables the use of IP as interim evaluation and, therefore, to compare young or non-qualified (incomplete pedigree) animals with animals that are part of the official evaluation. Although correlation between a benchmark GEBV and IP for young animals can be as high as 0.99 (Lourenco *et al.*, 2015), $E(\boldsymbol{\hat{a}})$ =0 and $E(\boldsymbol{\widehat{IP}})$≈0. Preliminary tests showed $\bar{\boldsymbol{\hat{u}}}$ for PWG for young animals was 99.92, whereas $\overline{\boldsymbol{IP}}$ was almost 0. In this case, GEBV and IP are not comparable.

Once SNP effects are available, the expectation of the conditional distribution of GEBV ($\boldsymbol{\hat{u}}$) and SNP effects ($\boldsymbol{\hat{a}}$) in the GBLUP context is:

$$\boldsymbol{E(\hat{u}|\hat{a}) =}$$
$$\boldsymbol{E(\hat{u}) + Z\frac{1}{2\sum p(1-p)}\left(I\frac{1}{2\sum p(1-p)}\right)^{-1}(\hat{a} - E(\hat{a}))}$$

with $\boldsymbol{E(\hat{u}) = 0}$; then,

$$\boldsymbol{E(\hat{u}|\hat{a}) = Z\frac{1}{2\sum p(1-p)}\left(I\frac{1}{2\sum p(1-p)}\right)^{-1}(\hat{a} - 0)}$$

Therefore,

$$E(\boldsymbol{\hat{u}}|\boldsymbol{\hat{a}}) = \mathbf{Z}\boldsymbol{\hat{a}}$$

In ssGBLUP, ungenotyped and genotyped animals are combined through the realized relationship matrix (H):

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{G} - \mathbf{A}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix}$$

where the subscript 1 represents ungenotyped and 2 genotyped animals.

Usually, **G** is centered based on current allele frequencies, fixing the average breeding value for genotyped animals to 0 (VanRaden, 2008). However, genotyped animals represent the most recent and selected population, and average breeding values should be different from 0. When ssGBLUP is used, there is a difference between pedigree and genomic base, because base animals in **A** are the founders of the pedigree and base animals in **G** are the genotyped animals. This difference can be modeled as a mean for the genotyped animals (Vitezica *et al.*, 2011):

$$\boldsymbol{p(u_2) = N(1\mu, G)}$$

where $\boldsymbol{u_2}$ is breeding value for genotyped animals;

$$\boldsymbol{\mu} : \boldsymbol{G = 11'a + bG}^{*}$$

where $\boldsymbol{a} = \frac{1}{\mathrm{n}^2}(\sum_{\mathrm{i}}\sum_{\mathrm{j}}\mathbf{A}_{22}$ - $\sum_{\mathrm{i}}\sum_{\mathrm{j}}\boldsymbol{G})$, $\boldsymbol{G}^{*}$ is the genomic relationship matrix without adjustments, and $\boldsymbol{b = 1 - \frac{1}{2}a}$; $\boldsymbol{Var(\mu) = a\sigma_u^2}$.

This difference between pedigree and genomic base is taken into account when GEBV are estimated, but not when IP are obtained based on SNP effects as $\boldsymbol{Z\hat{a}}$.

Three approaches were tested to account for the difference between pedigree and genomic base in IP:

a) Formulas derived in Legarra (2017):

$$\boldsymbol{\hat{u}_{ip}} = \boldsymbol{Z_v\hat{a}} + \boldsymbol{\hat{\mu}} + \boldsymbol{\hat{u}_p}, \text{ with}$$

$$\boldsymbol{E(\hat{\mu}|\hat{u})} = \mathrm{a}\ \alpha\ \mathbf{1'}\ \mathbf{G}^{-1}\boldsymbol{\hat{u}}, \text{ and}$$

$$\boldsymbol{\hat{u}_p = 0.5(\hat{u}_{parents,sire} + \hat{u}_{parents,dam})}$$

$$\boldsymbol{E(\hat{u}_{parents}|\hat{u})} = \boldsymbol{A_{22}}\ (1\text{-}\alpha)\ \mathbf{G}^{-1}\boldsymbol{\hat{u}}$$

where $\boldsymbol{\hat{u}_{ip}}$ is the estimated indirect prediction; $\boldsymbol{Z_v}$ is the matrix of SNP content for validation animals; $\boldsymbol{\hat{u}_p}$ is the average of $\boldsymbol{\hat{u}_{parents}}$; $\boldsymbol{\hat{u}_{parents}}$ is the contribution of parents' breeding values and account for the fact that a small portion (1-α) of $\boldsymbol{A_{22}}$ was blended to **G** to avoid singularity problems; 1-α is 0.05 and α is 0.95.

b) Linear regression with double fitting: intercept (b_0) and regression coefficient (b_1) of GEBV on $\boldsymbol{Z\hat{a}}$ were estimated for animals in the reduced data.

$$\mathrm{GEBV}_{reduced} = b_0 + b_1\mathbf{Z}\boldsymbol{\hat{a}}_{reduced}$$

The coefficients were then applied to the validation animals, so the base difference is adjusted by b_0:

$$\boldsymbol{\hat{u}}_{ip} = b_0 + b_1\mathbf{Z}_{\boldsymbol{v}}\boldsymbol{\hat{a}}$$

c) Average of GEBV: a simple average of GEBV for genotyped animals in the reduced dataset was added to $\mathbf{Z}_{\boldsymbol{v}}\boldsymbol{\hat{a}}$:

$$\boldsymbol{\hat{u}}_{ip} = \bar{\hat{u}}_{reduced} + \mathbf{Z}_{\boldsymbol{v}}\boldsymbol{\hat{a}}$$

The calculation of SNP effects in this study accounted for blending and tuning; therefore, the formula from Wang *et al.* (2012) was multiplied by two extra components (α and b).

The difference between GEBV from the complete data and $\boldsymbol{\hat{u}}_{ip}$, for validation animals, was used as a measure of bias. Accuracy was calculated as the correlation between GEBV and $\boldsymbol{\hat{u}}_{ip}$. In addition, the regression coefficient of GEBV on $\boldsymbol{\hat{u}}_{ip}$ was used as a measure of inflation.

Results & Discussion

Comparisons between the benchmark scenario using $\mathbf{G}^{-1}$ and alternative approaches for estimating SNP effects and IP (e.g., $\mathbf{Z}_{\boldsymbol{v}}\boldsymbol{\hat{a}}$) are shown in Table 1. Correlations between SNP effects from $\mathbf{G}^{-1}$ and $\mathbf{G}^{-1}_{\mathrm{APY}}$ were greater than 0.99, indicating SNP effects can be accurately backsolved when APY ssGBLUP is used. Several studies have shown GEBV from APY ssGBLUP are accurate (Fragomeni *et al.*, 2015 and Lourenco *et al.*, 2015); therefore, given the theory of APY is based on the limited dimensionality of genomic information (Misztal , 2016), SNP effects and IP obtained from regular or APY ssGBLUP are also expected to be highly correlated.

Table 1. Correlations between SNP effects and IP calculated based on $\mathbf{G}^{-1}$ and $\mathbf{G}^{-1}_{\mathrm{APY}}$ or $\mathbf{G}^{-1}_{\mathrm{cc}}$ with two definitions of core animals.

Scenario	$\hat{\boldsymbol{a}}$	$\mathbf{Z}_v\hat{\boldsymbol{a}}$[1]
$\mathbf{G}^{-1}_{\mathrm{APY\_H}}$	0.999	0.998
$\mathbf{G}^{-1}_{\mathrm{APY\_R}}$	0.998	0.997
$\mathbf{G}^{-1}_{\mathrm{cc\_H}}$	0.930	0.989
$\mathbf{G}^{-1}_{\mathrm{cc\_R}}$	0.900	0.988

[1] Indirect predictions as $\mathbf{Z}\hat{\boldsymbol{a}}$ for validation animals; H core is composed by high reliability animals; R core is composed by randomly picked animals.

When SNP effects were backsolved based on $\mathbf{G}^{-1}$ and GEBV for the core subset ($\mathbf{G}^{-1}_{\mathrm{cc\_H}}$ and $\mathbf{G}^{-1}_{\mathrm{cc\_R}}$), correlations were lower than expected, especially for the random core. However, differences cancelled out when IP were obtained as linear functions of SNP effects, given the correlations approached 0.99 in all scenarios.

Because $\mathbf{G}$ is a centered matrix, $\hat{\boldsymbol{a}}$ and $\mathbf{Z}_v\hat{\boldsymbol{a}}$ approach 0, making comparisons between official GEBV and IP difficult. In fact, average $\mathbf{Z}_v\hat{\boldsymbol{a}}$ was 2.38 and average GEBV was 116.33 for all validation animals. Differences between the benchmark GEBV and alternative approaches to balance pedigree and genomic base in IP ($\hat{\boldsymbol{u}}_{ip}$) are shown in Figure 1.

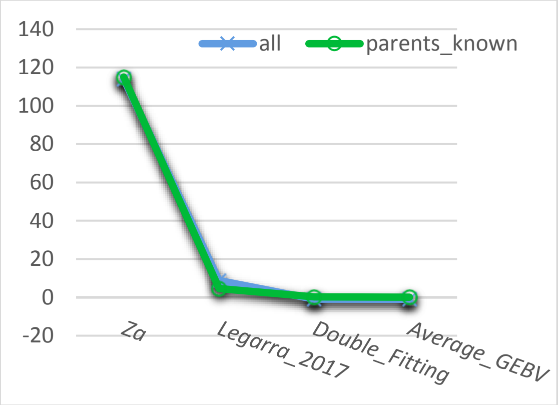

Figure 1. Difference between benchmark GEBV and $\hat{\boldsymbol{u}}_{ip}$ (measure of bias) for all validation animals (blue) or validation animals with parents known (green).

The difference between IP, simply calculated as $\mathbf{Z}_v\hat{\boldsymbol{a}}$, and GEBV was 113.95 for all validation animals and 115.03 for the ones with known parents, confirming a large bias for IP as a linear function of SNP content and effect without any adjustments. When the formulas from Legarra (2017), linear regression with double fitting, or average GEBV were used to obtain $\hat{\boldsymbol{u}}_{ip}$, bias was greatly reduced towards 0 for both groups of animals.

Correlations between GEBV and $\hat{\boldsymbol{u}}_{ip}$ for all scenarios, including $\mathbf{Z}_v\hat{\boldsymbol{a}}$ approached 0.98 for all validation animals and 0.99 for the ones with known parents. In fact, correlations are not expected to change considerably when a constant or small values (e.g., $\hat{\boldsymbol{u}}_p$) are added to $\mathbf{Z}_v\hat{\boldsymbol{a}}$.

Regression coefficients of GEBV on $\hat{\boldsymbol{u}}_{ip}$ are shown in Figure 2. Indirect predictions based on formulas from Legarra (2017) were slightly inflated compared to adding average GEBV to $\mathbf{Z}_v\hat{\boldsymbol{a}}$. Conversely, the double fitting approach was deflated.

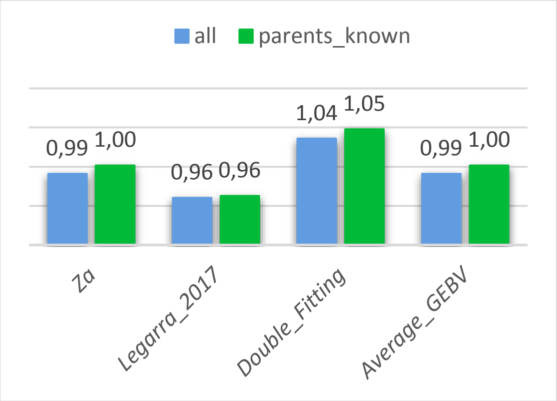

Figure 2. Regression coefficients of GEBV on $\hat{\boldsymbol{u}}_{ip}$ (measure of inflation) for all validation animals (blue) or validation animals with parents known (green).

Although in GBLUP the $E(\hat{\boldsymbol{u}})$ is zero, in ssGBLUP this expectation is modeled as the difference in population base:

$$E(\hat{\mathbf{u}}|\hat{\boldsymbol{a}}) = \hat{\mu} + \mathbf{Z}\frac{1}{2\sum p(1-p)}\left(\mathbf{I}\frac{1}{2\sum p(1-p)}\right)^{-1}(\hat{\boldsymbol{a}} - 0)$$

Therefore,

$$E(\hat{\mathbf{u}}|\hat{\boldsymbol{a}}) = \hat{\mu} + \mathbf{Z}\hat{\boldsymbol{a}}$$

Based on the difference between benchmark GEBV and $\hat{\boldsymbol{u}}_{ip}$, correlations, and regression coefficients, adding the average GEBV is a reasonable approach to adjust IP for the dissimilarity between pedigree and genomic base ($\hat{\mu} = \overline{GEBV}$). After this adjustment, IP becomes compatible to GEBV and animals not included in the official evaluation can be compared to the ones that participated in the evaluation.

Conclusions

Indirect predictions that include the linear function of SNP content and effect ($\mathbf{Z}_v\hat{\boldsymbol{a}}$) summed to the average of GEBV, as the difference between pedigree and genomic base, are less biased, therefore, comparable to GEBV. Accurate indirect predictions can be obtained when APY ssGBLUP is used. Backsolving genomic predictions to SNP effects may require only a group of genotyped animals representing the dimensionality of the genomic information.

Acknowledgements

This study was partially funded by the American Angus Association (St. Joseph, MO), and by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 from the US Department of Agriculture's National Institute of Food and Agriculture (Washington, DC).

References

Fragomeni, B.O., Lourenco, D.A.L., Tsuruta, S., Masuda, Y., Aguilar, I., Legarra, A., Lawlor, T.J. & Misztal, I. 2015. Hot topic: use of genomic recursions in single-step genomic BLUP with a large number of genotypes. *Journal of Dairy Science 98*, 4090-4094. https://doi.org/10.3168/jds.2014-9125

Legarra, A. 2017. Bases for genomic prediction. 81p. Accessed February 26, 2018. (http://genoweb.toulouse.inra.fr/~alegarra/GSIP_git.pdf)

Lourenco, D.A.L., Tsuruta, S., Fragomeni, B.O., Masuda, Y., Aguilar, I., Legarra, A., Bertrand, J.K., Amen, T., Wang, L., Moser, D.W. & Misztal, I. 2015. Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J. Anim. Sci. 93*, 2653-2662. https://doi.org/10.2527/jas.2014-8836

Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics 202*, 401-409. https://doi.org/10.1534/genetics.115.182089

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91*, 4414-4423. https://doi.org/10.3168/jds.2007-0980

Vitezica, Z.G., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. *Genet. Res. 93,* 357-366. https://doi.org/10.1017/S001667231100022X

Wang, H., Misztal, I., Aguilar, I., Legarra, A. & Muir, W.M. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. *Genet. Res. 94:2*, 73-83. https://doi.org/10.1017/S0016672312000274