Direct and Indirect Genomic Evaluations in Beef Cattle

D.A.L. Lourenco[1], I. Misztal[1], S. Tsuruta[1], B. Fragomeni[1], I. Aguilar[2], Y. Masuda[1] and D. Moser[3]

[1]Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602

[2]Instituto Nacional de Investigacion Agropecuaria, Canelones, Uruguay 90200

[3]Angus Genetics Inc., St. Joseph, MO 64506

Abstract

We tested two modifications into single-step genomic BLUP (ssGBLUP) that allow it to work with a large amount of genotyped animals. The first method is based on genomic recursions (APY) to construct the inverse of the genomic relationship matrix without directly inverting it; all available genotyped animals are included in ssGBLUP with APY, but they are split into base and non-base, and the method returns direct predictions. The second method is an interim genomic evaluation (IP) for young genotyped animals; only a reference set of animals are used in ssGBLUP with IP, and the method returns indirect predictions for young genotyped animals. A dataset from American Angus with records for growth traits was used. Over 8 million animals were in the pedigree, of which 51,883 were genotyped. The ssGBLUP with APY was as accurate as regular ssGBLUP when the number of genotyped base animals was at least 10,000; this method was also faster and required less memory. The ssGBLUP with IP mimicking the previous official evaluation returned the same accuracy of GEBV for young animals as the regular ssGBLUP. While the first method enables complete genomic evaluations for huge genotyped populations, the second allows for quick genomic predictions on young animals without including their information into a new run of evaluation.

Key words: single-step genomic BLUP, genomic recursions, interim GEBV, big genotyped population

Introduction

Genomic selection in beef and dairy cattle has currently been performed with multistep methods, which use deregressed EBV to estimate SNP effects and then direct genomic value (DGV) for selection candidates based on their genotypes (Meuwissen *et al.*, 2001). In multistep, new animals are easily evaluated if DGV is computed as a sum of marker effects, but not if selection indexes including DGV and parent average (PA) are used.

When pedigrees, phenotypes, and genotypes are jointly available, single-step genomic BLUP (ssGBLUP) (Aguilar *et al*., 2010) is a simple choice for genomic evaluation. However, in its current implementation, ssGBLUP uses direct inversion of genomic matrices (Aguilar *et al*., 2011), which has a cubic cost and a limit of 150,000 animals (Aguilar *et al*., 2013). The increasing number of genotyped animals (e.g. US Holsteins and American Angus) has been a challenge for ssGBLUP. Recently, Misztal *et al*. (2014) presented an algorithm to obtain the inverse of the genomic relationship matrix, which is based on recursions on a fraction of the genotyped population. This makes ssGBLUP suitable and inexpensive for huge populations.

Another way to circumvent the issue on the number of genotyped animals is to enable ssGBLUP to provide quick evaluations for young genotyped animals, without running a complete evaluation that requires several hours to converge. In this way, quick predictions can be calculated indirectly, where genomic EBV (GEBV) for young animals are obtained from SNP effects based on ssGBLUP solutions.

The goals of this study were to 1) present a method for calculating indirect predictions on young genotyped animals using ssGBLUP, and 2) test the feasibility of ssGBLUP with genomic recursions for evaluations in American Angus.

Materials and Methods

Data

Datasets from American Angus Association (AAA) were available that included birth weight (BW), weaning weight (WW), and post-weaning gain (PWG). Table 1 shows number of animals with records and heritability for all traits. A total of 51,883 animals were genotyped for 54,609 segregating SNP from the BovineSNP50k v2 BeadChip (Illumina Inc., San Diego, CA).

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

Trait	h^2	Number of records	Genotyped animals with records
BW	0.41	6,189,661	50,784
WW	0.20	6,890,625	51,830
PWG	0.20	3,387,252	36,196

ssGBLUP with indirect predictions (IP) for young animals

For ssGBLUP with indirect predictions, SNP effects can be calculated using the current run of ssGBLUP with all but young animals, and genomic predictions for young animals are obtained by multiplying the SNP content by the SNP effect to obtain direct genomic value (DGV); a more complete GEBV can also be available through a selection index that combines DGV and parent average (PA). In order to explain how it works, consider the equation for the GEBV of a single individual in ssGBLUP as a combination of equations in Aguilar *et al*. (2010) and VanRaden and Wright (2013):

$$GEBV=w_1PA+w_2YD+w_3PC+w_{4.1}DGV-w_{4.2}PP$$

where YD is yield deviation, PC is progeny contribution, and PP is pedigree prediction for the subset of genotyped animals.

The flow for ssGBLUP with IP was:

1) Run ssGBLUP with a reference population to obtain GEBV. In this study, 3 reference populations were tested:

ref_2k: reference population with top bulls and top cows (n=1,896);

ref_8k: reference population with all parents that were genotyped (n=8,285), this includes ref_2k;

ref_33k: reference population with all genotyped animals born up to 2012 (n=33,162), this includes ref_8k;

2) Split GEBV into all the components shown before. DGV for an animal i in the reference population was calculated as in Aguilar *et al*. (2010);

3) Calculate SNP effects using DGV from the reference population:

$$\hat{\mathbf{u}}= \mathbf{D}\mathbf{Z}'\mathbf{G}^{-1}(\mathbf{DGV})$$

where $\hat{\mathbf{u}}$ is a vector of estimated SNP effects, **D** is a diagonal matrix of weights (standardized variances) for SNP (identity matrix in this case), **Z** is a matrix of centered genotypes for each animal, and **G** is the genomic relationship matrix;

4) Calculate DGV for young genotyped animals (DGV_y):

$$\mathbf{DGV}_y=\mathbf{Z}_y\hat{\mathbf{u}}$$

where DGV_y is a vector of direct genomic values and $\mathbf{Z}_y$ is a matrix of centered genotypes for young animals not included in ssGBLUP evaluation, respectively.

5) Combine DGV_y with PA for young genotyped animals:

$$GEBV_y \approx w_1PA + w_{4.1}DGV_y$$

where $GEBV_y$ is GEBV obtained via IP for young animals, w_1 and $w_{4.1}$ are weights identical for all animals and calculated based on selection index.

The ability to predict future phenotypes on 18,721 young genotyped animals was the validation method chosen in this study.

ssGBLUP with G inverted by a recursive algorithm

When the number of genotyped animals is large and there is a need for using all of them in ssGBLUP evaluations to get direct predictions for all, including young animals, an algorithm that splits genotypes into base (b) and non-base (c) animals and uses recursion to obtain the inverse of the **G** matrix was proposed by Misztal *et al*. (2014). This algorithm is known as APY, and $\mathbf{G}^{-1}$ containing all genotyped animals can be calculated as:

$$\mathbf{G}^{-1} = \begin{bmatrix} \mathbf{G}_{bb}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{bb}^{-1}\mathbf{G}_{bc} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_g^{-1} \begin{bmatrix} -\mathbf{G}_{cb}\mathbf{G}_{bb}^{-1} & \mathbf{I} \end{bmatrix}$$

where the subscript b stands for base and c for non-base animals; each element of $\mathbf{M_g}$ for the i^{th} animal is $m_{g,i} = g_{ii} - \mathbf{G}_{ib}\mathbf{G}_{bb}^{-1}\mathbf{G}_{bi}$.

For this analysis, the definitions of base animals were according to the amount of information the animals had; this included the definitions used for IP (ref_2k, ref_8k, and ref_33k), a fourth definition had 3,872 genotyped parents of genotyped animals as base (ref_4k), and the last definition had 10,000 animals with the highest EBV accuracy as base (ref_10k). We also randomly sampled 5,000 (rand_5k), 10,000 (rand_10k), 15,000 (rand_15k), and 20,000 (rand_20k) animals to be in the base population.

Feasibility of APY for American Angus was tested as correlation between regular ssGBLUP and ssGBLUP with APY.

Results & Discussion

Predictive ability for indirect prediction via conversion of DGV into SNP effects is shown in Figure 1. When the reference population included top bulls and top cows (ref_2k), the predictivity of indirect DGV_y was smaller than predictivity for PA for the three traits (0.23 vs. 0.29 for BW; 0.28 vs. 0.34 for WW; 0.19 vs. 0.23 for PWG). Predictivity for GEBVy calculated as an index of indirect DGV_y with PA was greater than those for EBV for the three traits (0.31 vs. 0.29 for BW; 0.36 vs. 0.34 for WW; 0.24 vs. 0.23 for PWG), however, this predictivity was smaller than the ones from full ssGBLUP (except for WW).

With a larger reference population (ref_8k), all indirect DGV_y were similar or more accurate than PA, and $GEBV_y$ had similar predictivity as the full ssGBLUP. With the largest reference population (ref_33k), all indirect DGV_y were almost as accurate as GEBV from full ssGBLUP, with the index marginally improving predictivity for WW.

The marginal improvement observed for WW may be caused by the use of less than optimal genetic parameters, e.g., zero covariance between direct and maternal effects (to reduce computing costs). The DGV_y obtained with ref_33k reference population were more accurate than GEBV from full ssGBLUP obtained with ref_8k reference population.

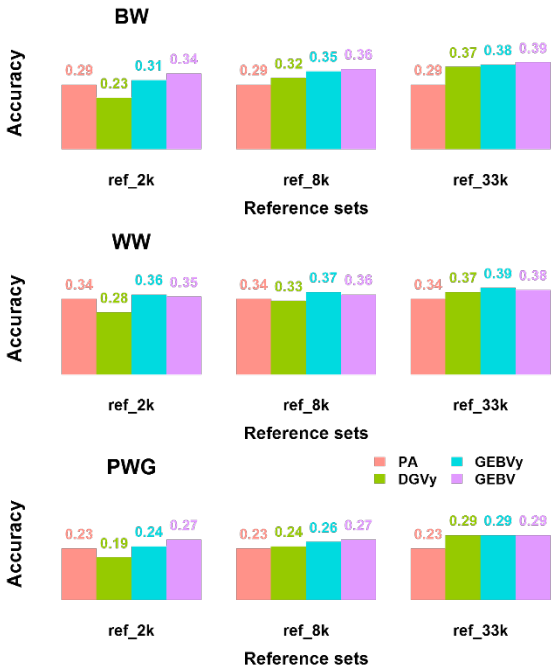

Figure 1. Predictive ability of indirect predictions on 18,721 young genotyped animals when using reference populations ref_2k, ref_8k, and ref_33k animals to run single-step genomic BLUP (ssGBLUP) and derivate SNP effects.

Overall, we observed that when the number of genotyped animals in the reference population is small, there is a need to combine DGV_y with PA; however, when the reference population is large (e.g. previous official evaluation), DGV_y becomes the most important component and there is no need for an index.

For young animals, indirect predictions via SNP effects from ssGBLUP seem a viable alternative as it can be done separately from the full evaluation. As SNP effects are calculated based on trait DGV, indirect predictions are easily obtained for multi-trait models, as done in this study. However, if young animals and particularly full-sibs are intensively selected, selection on the Mendelian sampling will not be accounted for, leading to pre-selection bias (Patry and Ducrocq, 2011). Analyses by ssGBLUP with all genotypes subject to selection are expected to account for pre-selection (VanRaden and Wright, 2013).

Correlation between GEBV from regular ssGBLUP and ssGBLUP when the inverse of **G** is computed with APY is shown in Figure 2.

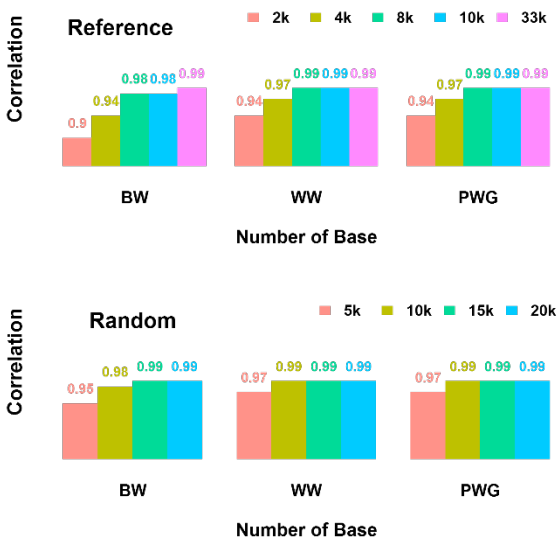

Figure 2. Correlation between GEBV from regular ssGBLUP and ssGBLUP with APY, when base animals were from reference group of animals (top: Reference) or from random samples (bottom: Random).

When base animals were from reference sets (ref_2k, ref_4k, ref_8k, and ref_33k), correlations with regular ssGBLUP increased with the number of base animals and reached 0.99 when at least 8,000 animals were in the base population, especially for WW and PWG. When the animals in the base population were randomly sampled, 10,000 was enough to reach correlation of 0.99 with regular ssGBLUP. Therefore, in ssGBLUP, using genomic recursion to invert **G** while conditioning on enough number of base animals has the same prediction power as **G** using regular inversion. The memory required for APY $\mathbf{G}^{-1}$ using ref_2k, ref_4k, ref_8k, ref_10k, and ref_33k was approximately 0.8, 1.6, 3.3, 4.1, and 13.7 Gbytes, respectively, whereas the amount of memory for the regular $\mathbf{G}^{-1}$ was 21.6 Gbytes. The computing time for constructing APY $\mathbf{G}^{-1}$ in ref_10k was about 8 times smaller than for regular $\mathbf{G}^{-1}$. Therefore, APY $\mathbf{G}^{-1}$ makes computations less costly and faster.

With APY conditioning on 10,000 animals, for example, the only inverse required is for a block of **G** for 10,000 animals, and additional genotypes require only linear storage and computations. Subsequently, computations with a much larger number of genotyped animals may be feasible with predictivity similar to the regular inversion. APY would be the algorithm of choice for ssGBLUP evaluations with a very large number of genotyped animals.

Conclusions

Both methods presented here are applicable to the American Angus population. With a sufficient number of animals in the reference population, indirect prediction for young animals via SNP effects provides similar predictivity to full single-step genomic BLUP, allowing for quick interim evaluations without running a complete evaluation. Use of the algorithm for base and non-base animals in single-step genomic BLUP, with 10,000 base animals, allows for incorporation of a large number of genotyped animals into American Angus evaluations without losing prediction power.

Acknowledgements

This study was partially funded by the American Angus Association (St. Joseph, MO), Zoetis (Kalamazoo, MI), and by Agriculture and Food Research Initiative Competitive Grants no. 2015-67015-22936 NIFA-USDA. The authors thank Tonya Amen, Lizhen Wang, Andres Legarra, and Ivan Pocrnic for helpful suggestions.

References

Aguilar, I., Misztal, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2010. Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci. 93*, 743–752.

Aguilar, I., Legarra, A., Tsuruta, S. & Misztal, I. 2013. Genetic evaluation using unsymmetric single step genomic methodology with large number of genotypes. *Interbull Bulletin 47,* 222–225.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 157*, 1819–1829.

Misztal, I., Legarra, A. & Aguilar, I. 2014. Using recursion to compute the inverse of the genomic relationship matrix. *J. Dairy Sci. 97,* 3943–3952.

Patry, C. & Ducrocq, V. 2011. Evidence of biases in genetic evaluations due to genomic preselection in dairy cattle. *J. Dairy Sci. 94*, 1011–1020.

VanRaden, P.M. & Wright, J.R. 2013. Measuring genomic pre-selection in theory and in practice. *Interbull Bulletin 47,* 147–150.