Use of causative variants and SNP weighting in a single-step GBLUP context

B.O. Fragomeni[1], D.A.L. Lourenco[1], A. Legarra,[2,] M.E. Tooker[3], P.M. VanRaden[3], I. Misztal[1].

[1] University of Georgia, Edgar L. Rhodes Center for Animal and Dairy Science, 30602, Athens, USA
fragomen@uga.edu (Corresponding Author)
[2]GenPhySE, INRA, INPT, INP-ENVT, Université de Toulouse, 31326, Castanet-Tolosan, France.
[3] Animal Genomics and Improvement Laboratory, Agricultural Research
Service, USDA, 20705, Beltsville, USA

Summary

Much effort has been recently put into identifying causative quantitative trait nucleotides (QTN) in animal breeding, aiming more accurate genomic prediction. Among the genomic methods available, single-step GBLUP (ssGBLUP) became the choice because of its simplicity and potentially higher accuracy. When QTN are known, they need to be properly weighted, so the accuracy can be maximized. The weighted ssGBLUP is still under development, and a proper weighing algorithm is needed. The objectives of this study were to investigate whether ssGBLUP is useful for genomic prediction when causative variants are known and to verify the impact of different SNP weighting in ssGBLUP compared to GBLUP. Analyses involved about 4M records for stature of 3M cows. Genotypes were available for 27k sires for a regular 54k chip (BovineSNP50; Illumina), and imputed with extra 17k sequence variants having largest effects, including causative variants affecting stature and 32 other traits. Direct genomic value (DGV) and genomic EBV (GEBV) were calculated using GBLUP and ssGBLUP with regular genomic relationship matrices (**G**). Later, **G** was weighted based on the squared value of SNP effects or with nonlinear A equations, which limits the changes in SNP weights. In GBLUP, the residuals were either homogeneous or heterogeneous. Reliability (R^2) was assessed from forward prediction of DGV or GEBV for young sires with at least 10 daughters. The lowest to highest R^2 for DGV were by GBLUP with homogeneous residuals, GBLUP with heterogeneous residuals, and extracted from ssGBLUP. SNP weighting by nonlinear A increased R^2 by up to 1.7% with homogeneous residuals and by up to 0.2% with heterogeneous residuals, compared to unweighted GBLUP. Linear weighting reduced accuracy in GBLUP and had no effect in ssGBLUP. Overall, adding 17k causative variants increased accuracy up to 0.6% in GBLUP, but had no impact in ssGBLUP. Reliability for DGV extracted from ssGBLUP was at least 0.6% more accurate than any DGV from GBLUP. Linear weighting is not helpful with causative variants with small effect. Gains with SNP weighting in multistep (GBLUP or SNP BLUP) may be partly due to corrections in modeling issues associated with pseudo-observations.

Keywords: Genomic relationship matrix, reliability, genome wide association

Introduction

The interest in finding and using causative variants for prediction has been recently revived due to advances in genotyping technologies and decrease in costs. Data from the 1000 bull genomes project (Daetwyler et al., 2014; Hayes et al., 2014) has been used by different authors with the objective of incrementing accuracy of genomic prediction and searching for causative variants. The majority of the research performed in that context used some Bayesian hierarchical models to estimate breeding value based on multilocus association models as in Meuwissen et al. (2001). Those methods provide potential advantage for oligogenic traits due to SNP weighting or selection when compared to marker estimated relationship matrices methods such as GBLUP and single step GBLUP (ssGBLUP) (Aguilar et al., 2010; Christensen & Lund, 2010; VanRaden, 2008). In commercial evaluations, single step genomic BLUP (ssGBLUP) is becoming the method of choice due to inclusion of non-genotyped animals without the necessity of de-regressing breeding values to be used as pseudo-observations, which can bring potential problems such as double counting of phenotypic and pedigree information, and preselection bias (Legarra et al., 2014).

To successfully account for causative variants in polygenic traits it is necessary to better understand the SNP weighting process in the ssGBLUP context. In this way, the objectives of the present study were to investigate the impact of adding causative variants into GBLUP and ssGBLUP and to test different SNP weighting methods.

Material and methods

Data was provided by Holstein Association USA Inc. (Brattleboro, VT) and included 3,999,631 stature records measured on 3,027,304 cows, from 1990 to 2016. The total number of animals in the pedigree was 4,661,872, and genotypes on 54,087 SNP markers were available for 26,877 bulls. In a separate analysis, a total of 16,648 causative variants, reported in VanRaden et al. (2017) were included in the SNP array, totaling 70,735 markers.

Predictions were made by GBLUP and ssGBLUP. In the first method, the assumption of the distribution of the additive genetic effect () was , where is the genetic additive variance and is the genomic relationship matrix (**G**), as described by VanRaden (2008). The residual (**e**) was distributed as , where is the residual covariance matrix. Residual variances were considered homogeneous with or heterogeneous with , where is daughter equivalent contributions (VanRaden, 2008). Pseudo observations for this method were daughter deviations (DD) (VanRaden & Wiggans, 1991). Daughter deviations were calculated with phenotypic data up to 2011, with outputs from the software BLUPF90 and ACCF90 (Misztal et al., 2002).

In ssGBLUP, the additive genetic effects were distributed as , where **H** is the relationship matrix combining pedigree and genomic information, as defined in Legarra et al. (2009). Single-step GBLUP evaluations were done using a repeatability single trait model as described in Tsuruta et al. (2002).

The **G** was weighted () to account for unequal SNP variances in both methods, as:

(1)

where **M** is a centered SNP content matrix, k is number of markers, is the minor allele frequency of SNP , and is a diagonal matrix of weights; when **G** is unweighted, **,** where **I** is the identity matrix**.** In a first approach, weights for the **G** were calculated as SNP variances: , where is the SNP effect and is the SNP weight.

Additionally, weights were calculated using the nonlinear A method described in VanRaden (2008):

, (2)

where is the absolute estimated SNP effect for the marker *i* and is the standard deviation of the vector with SNP effects. The maximum change in weights was limited to 10. The nonlinear A method differs mainly from the linear method by limiting the maximum and minimum changes in the weights every iteration. Meanwhile, the linear method allows extreme SNP variance.

Additionally, we used the same weights as in VanRaden et al. (2017); this weighting approach will be defined as PVRW hereinafter. All weighting methods are iterative and, based on preliminary results, we chose 10 as the ideal number of iterations for this study.

Raw reliabilities were calculated as the coefficient of determination from regressing DD from 2016 data on GEBV (ssGBLUP) or DGV (GBLUP) from 2011 data. Reliabilities were then adjusted (R^2) following the procedure suggested by VanRaden et al. (2009):

(3)

where is the reliability of DD and PA adjustment was the difference between the published reliability of parent average and the result of R^2 of PA divided by mean reliability of daughter deviations, as in (VanRaden et al., 2009). The validation group included bulls with no daughters in 2011 and at least 10 daughters with records in 2016.

Results and Discussion

Reliabilities of genomic predictions from GBLUP and ssGBLUP are in Table 1. Adding 17k causative variants obtained from a sequence GWA (VanRaden et al., 2017) did not increase reliabilities in ssGBLUP for stature. However, a small increase of 0.6% was observed for GBLUP in a scenario with linear weighting and homogeneous residual variance. Although this was the highest increase, the reliabilities were still less than a scenario with no weights (55.5 vs. 58.7). When no weights were used, the increase in reliability was 0.5 points. Reliabilities were not improved by adding causative variants when weights were from nonlinear A and homogeneous residual variance was considered. The greatest reliability from GBLUP (58.9) was obtained in a scenario with heterogeneous residual variance and weights calculated by VanRaden et al. (2017).

Although adding causative variants with a proper weight in GBLUP increased reliabilities, the values from ssGBLUP were 2% greater than the best GBLUP scenario (PVRW). One reasonable explanation for the discrepancy in results between methods is that ssGBLUP deals with more information than multistep procedures, especially when de-regressions are used in the later. Another reason is the explicit contribution of parent average in ssGBLUP. In addition, some reliability can be lost in multistep because of approximations.

Figure 1 shows reliabilities for all 10 iterations when the causative variants were present in the data. Iteration 1 has equal weights for SNP. If a plateau is reached, the convergence is obtained and SNP variances do not change after that, keeping the reliability steady. It is clear that the convergence was not reached for linear weights. This reveals a problem in using even when causative variants are in the SNP data. Adding a boundary for changes in SNP variance may solve this issue. Drops in reliability were not observed for the nonlinear A weights. Instead, there was a small increase from iteration 1 to 2, meaning using weights was beneficial when residual variances were considered homogeneous. This increase possibly indicates that some weighting approaches can help to correct modeling issues. When we improved the model by considering heterogeneous residual variance and nonlinear A weights, there was no change in reliability by adding weights for SNP. In all 10 iterations, reliabilities of GEBV from ssGBLUP were the same.

When large causative variants are included in the model, we expect an increase in reliability by properly accounting for them. Among the weights we tested, it was clear the linear approach was unable to detect causative variants, because of the extreme changes in weights in every iteration. Nonlinear A method avoids extreme shrinkage and inflation and is more suitable for polygenic traits; however, its implementation for weight updates in ssGBLUP possibly did not converge at maximum reliability. Overall, the lack of improvement in reliability for ssGBLUP can be because none of the weights tested in our study were ideal and because the causative variants may have had only small effects.

This study raises a number of questions. Was any improvement in VanRaden et al. (2017) due to problems with pseudo-observations? However, similar improvement was obtained by VanRaden et al. (2017) using simulated data, where these problems were absent. Is the implementation of GBLUP using nonlinear A equivalent to that as described by VanRaden (2008)? Is ssGBLUP less sensitive to SNP weighting? But using "perfect" weights with causative SNP in ssGBLUP was effective (Fragomeni et al., 2017).

Using trait specific **G** for many traits in ssGBLUP may be expensive and may prohibit multiple-trait models, unless an appropriate extension to ssGBLUP is developed. Once SNP effects are known, SNP-based prediction is simpler. A genomic evaluation in Angus includes multiple-trait ssGBLUP with unweighted **G** for a regular evaluation, and SNP predictions derived from GEBV of ssGBLUP for rapid interim predictions (Lourenco et al., 2015).

Conclusion

Gains in SNP weighting in multistep (GBLUP or the equivalent SNP BLUP) may be partly due to corrections in modeling issues associated with deregressions and improper way to account for all sources of information. Linear weighting of SNP is not beneficial when causative variants have small effect. Additional approaches that approximate true weights in ssGBLUP need to be tested in the future.

List of 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*(2), 743-752.

Christensen, O. F., & Lund, M. S. (2010). Genomic prediction when some animals are not genotyped. *Genet Sel Evol, 42*, 2.

Daetwyler, H. D., Capitan, A., Pausch, H., Stothard, P., van Binsbergen, R., Brondum, R. F., . . . Hayes, B. J. (2014). Whole-genome sequencing of 234 bulls facilitates mapping of monogenic and complex traits in cattle. *Nat Genet, 46*(8), 858-865.

Fragomeni, B. O., Lourenco, D., Masuda, Y., Legarra, A., & Misztal, I. (2017). Incorporation of Causative Quantitative Trait Nucleotides in Single-step GBLUP. *Genet Sel Evol, (accepted)*

Hayes, B. J., MacLeod, I. M., Daetwyler, H. D., Phil, B. J., Chamberlain, A. J., Vander Jagt, C., . . . Liao, X. (2014). *Genomic prediction from whole genome sequence in livestock: the 1000 bull genomes project.* Paper presented at the 10th World Congress on Genetics Applied to Livestock Production (WCGALP).

Legarra, A., Aguilar, I., & Misztal, I. (2009). A relationship matrix including full pedigree and genomic information. *J Dairy Sci, 92*(9), 4656-4663.

Legarra, A., Christensen, O. F., Aguilar, I., & Misztal, I. (2014). Single Step, a general approach for genomic selection. *Livestock Science, 166*, 54-65.

Lourenco, D. A., Tsuruta, S., Fragomeni, B. O., Masuda, Y., Aguilar, I., Legarra, A., . . . Misztal, I. (2015).

Genetic evaluation using single-step genomic best linear unbiased predictor in American Angus. *J Anim Sci, 93*(6), 2653-2662.

Meuwissen, T. H., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics, 157*(4), 1819-1829.

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., & Lee, D. H. (2002). BLUPF90 AND RELATED PROGRAMS (BGF90). *7th World Congress on Genetics Applied to Livestock Production, August 19-23, 2002, Montpellier, France, Commun. No. 28-07*.

Tsuruta, S., Misztal, I., Klei, L., & Lawlor, T. J. (2002). Analysis of age-specific predicted transmitting abilities for final scores in Holsteins with a random regression model. *J Dairy Sci, 85*(5), 1324-1330.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *J Dairy Sci, 91*(11), 4414-4423.

VanRaden, P. M., Tooker, M. E., O'Connell, J. R., Cole, J. B., & Bickhart, D. M. (2017). Selecting sequence variants to improve genomic predictions for dairy cattle. *Genet Sel Evol, 49*(1), 32.

VanRaden, P. M., Van Tassell, C. P., Wiggans, G. R., Sonstegard, T. S., Schnabel, R. D., Taylor, J. F., & Schenkel, F. S. (2009). Invited review: reliability of genomic predictions for North American Holstein bulls. *J Dairy Sci, 92*(1), 16-24.

VanRaden, P. M., & Wiggans, G. R. (1991). Derivation, calculation, and use of national animal model information. *J Dairy Sci, 74*(8), 2737-2746.

Figure 1 – Reliabilities from direct genomic values obtained by GBLUP (GBLUP) and ssGBLUP (ssGBLUP-IND) and GEBV obtained by ssGBLUP (ssGBLUP). Genomic relationship matrix was unweighted in the first iteration for all cases. Weights were calculated for 10 iterations using non-linear A equations, or were calculated by linear SNP variances (LIN). Assumption from residual variance in GBLUP was homogeneous (HOM) or heterogeneous (HET).

Table 1. Raw and adjusted genomic reliabilities and differences from reliability of parent average (PA), for predictions using GBLUP and ssGBLUP using unweighted **G**, 3rd iteration of linear weights (linear W) or 10th iteration of nonlinear A weights (Nonlinear A), using a 54k chip or 70k chip with causative variants. GBLUP was by default using heterogeneous residual variances, but homogeneous residuals were also tested (HOM).

Item	**Raw Reliability**	**Adjusted Reliability**	**Gain from Parent Average (Adjusted)**
Parent Average	31.7	38.6	0
ssGBLUP	60.9	76.1	37.5
ssGBLUP PVR Weights	60.8	76	37.4
ssGBLUP - DGV only	59.5	74.3	36.1
GBLUP 54k	58.2	72.8	34.2
GBLUP 70k	58.7	73.4	34.8
GBLUP Nonlinear A 54k	58.4	73.1	34.5
GBLUP Nonlinear A 70k	58.6	73.4	34.8
GBLUP Nonlinear A 70k PVR	58.9	73.7	35.1
GBLUP Linear Weights 54k	56.4	70.6	32
GBLUP Linear Weights 70k	56.3	70.5	31.9

GBLUP Homogeneous Linear Weights 54k	54.9	68.9	30.3
GBLUP Homogeneous Linear Weights 70k	55.5	69.6	31
GBLUP Homogeneous NonlinearA 54k	56.6	70.8	32.6
GBLUP Homogeneous NonlinearA 70k	56.6	70.8	33