Weighted Single-step Genomic BLUP: an Iterative Approach for Accurate Calculation of GEBV and GWAS

X. Zhang,[1] D. A. L.Lourenco,[1] I. Misztal,[1] I. Aguilar,[2] and A. Legarra[3]
[1]University of Georgia, Athens, Georgia, USA, [2]Instituto Nacional de Investigación Agropecuaria, Las Brujas, Uruguay, [3]Institut national de la recherche agronomique, Castanet-Tolosan, France

ABSTRACT: Three different procedures were implemented to calculate weights for a genomic relationship matrix to restrict the shrinkage along iterations of weighted single-step genomic BLUP (WssGBLUP). The procedures as well as BayesC were tested with 3 simulated data sets. Prediction accuracy for WssGBLUP improved at 2nd or 3rd iteration by updating only the top number of SNP equal to 1 × or 3 × the number of QTL; accuracy increased after 3rd iteration and remained stable by using weights proportional to $2p_i(1- p_i)u_i^2$ + constant. Except in the 5 QTL scenario, accuracies with all WssGBLUP procedures were higher than with BayesC. Noise in Manhattan plots was small with 5 and 100 QTL but large with 500 QTL.

Key words: GWAS; WssGBLUP; BayesC.

INTRODUCTION

Genomic BLUP (GBLUP) is usually associated with equal weights on all SNP while the Bayesian methods are associated with different weights on SNP. If those weights are known, weighted GBLUP provides similar Genomic EBV (GEBV) to a Bayesian procedure using the same weights (Legarra et al., (2010)). Methods were developed that allow for estimation of weights within the GBLUP (Sun et al. (2011)) or single-step GBLUP (Misztal et al. (2009); Aguilar et al. (2010); Wang et al. (2012)). They can be called as WGBLUP and WssGBLUP, respectively. Sun et al. developed two procedures for calculating weights in WGBLUP. In the first one, the weights were calculated as $w_j^{(i)} = \hat{a}_j^{(i)2}$, where $w_j^{(i)}$ is the weight of j-th SNP at i-th iteration and $\hat{a}_j^{(i)}$ is the effect of j-th SNP at i-th iteration. Such a procedure was good for identification of top QTL but shrank small SNP too much, thus reduced accuracy of GEBV. The highest accuracy of GEBV was achieved by modifying the formula for weights to $w_j^{(i)} = \hat{a}_j^{(i)2} + t$, where $t = \frac{\sigma_g^2}{2\sum_{j=1}^{m} p_j q_j}$, σ_g^2 is the genetic variance; p and q are the minor and major allele frequencies at j-th locus, respectively; and m is the number of SNPs. This procedure brought the accuracy of GEBV close to that by BayesC but yielded "noisy" Manhattan plots. Wang et al., (2012) evaluated WssGBLUP with simulation data. They iterated either on SNP alone or on GEBV and SNP. The first option gave a good identification of top QTL, and the second option provided a higher accuracy of GEBV than BayesB, but at the second iteration only.

The objectives of our study were to present new procedures to calculate weights for SNP in WssGBLUP and compare the accuracy and SNP effects with those computed by BayesC (Kizilkaya et al. (2010)) using simulated data.

MATERIALS AND METHODS

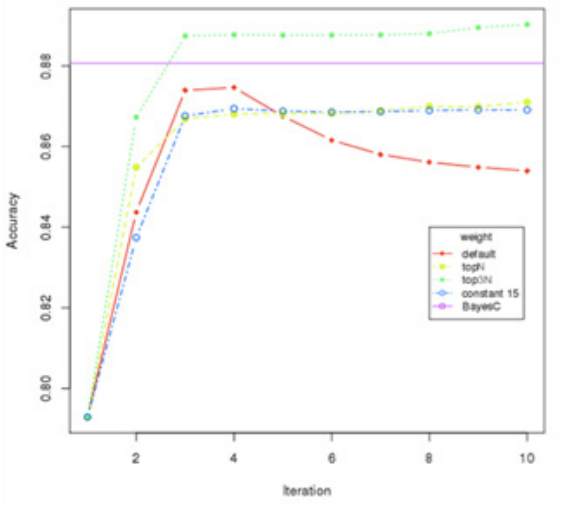
Data simulation. One additive trait with a mean of 1.0, phenotypic variance 1.0 and heritability 0.5 was simulated using QMSim (Sargolzaei and Schenkel, (2009)). A total of 20 chromosomes with average length 82 cM containing 45K evenly distributed SNP were created. Three scenarios were considered involving different numbers of randomly placed QTL: 5, 100, and 500. For the first scenario, QTL effects were sampled from the normal distribution with a minimum absolute value of 0.2. For the later two scenarios, QTL sampling was by the gamma distribution with a shape factor 0.4. Both SNP and QTL were bi-allelic, with no overlapping between their positions. The simulated population was randomly selected for 205 generations and preceded by a historical population with 1000 generations of random mating. 200 males and 2600 females were selected to mate in each generation with a litter size of 1. Generations 200-204 were treated as a training population and 205 as a validation population, with 1240 and 300 genotyped animals, respectively. The complete datasets contained 18,400 individuals in the pedigree, of which 13,000 were phenotyped and 1540 were genotyped. Average LD (r^2) at last generation was about 0.29.

Model. The model for the simulation analysis was included a population mean, a random SNP effect and a random residual error term. Comparisons included WssGBLUP and BayesC. Both GEBV and SNP effects were obtained by BLUPF90 (Misztal et al. (2002)) modified for genomic analyses (Aguilar et al. (2010)), and GenSel (Fernando and Garrick (2009)).

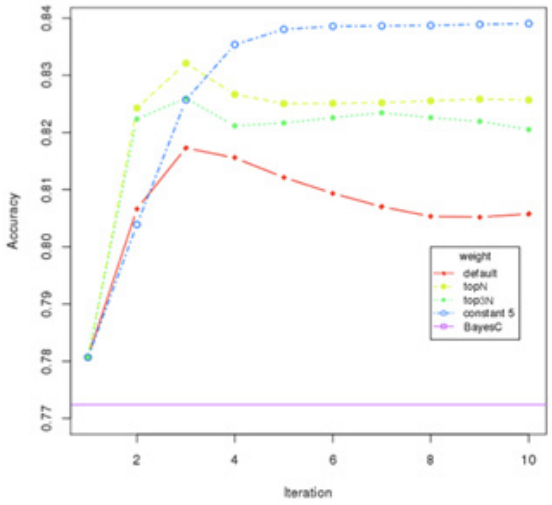
Statistical analysis. The weights were derived from SNPsolutions. Improvements on the SNP weights can be obtained iteratively either by recomputing the SNP effects only or by also recomputing the GEBV (Wang et al., (2012)). The latter was chosen for this study. Four options were used to calculate the SNP weights in ssGBLUP: 1) default: proportional to $2p_i(1- p_i)u_i^2$, where p_i and u_i are frequency and effect of the i-th SNP; 2) top N: weights as in 1, but updating only N SNP with largest effects using the number of SNP is equal to 1 × the number of simulated QTL; 3) top 3N: updating only the top 3N SNP where 3N is equal to 3 × the number of simulated QTL; 4) constant: proportional to $2p_i(1- p_i)u_i^2$ + constant, where the constant was chosen as the weight of top 1 SNP in the first iteration.

Accuracy was defined as the correlation between true breeding value (TBV) and GEBV in the validation population. Comparisons were made among the 4 options and also with BayesC using π equal to the proportion of ignored SNP in WssGBLUP using Option 2.

5 QTL



100 QTL



500 QTL

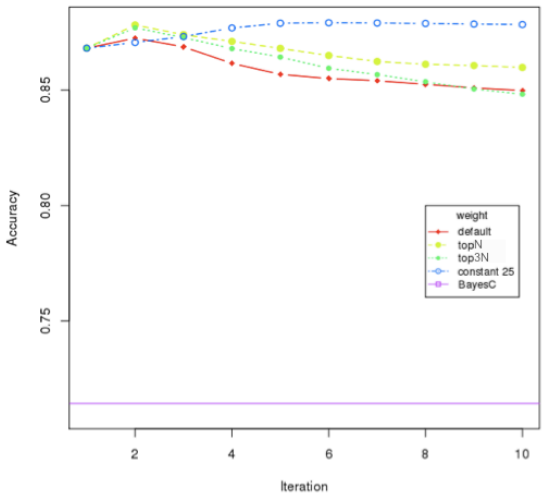
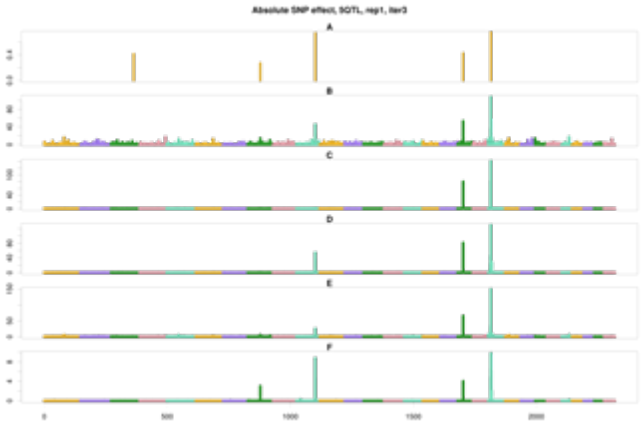
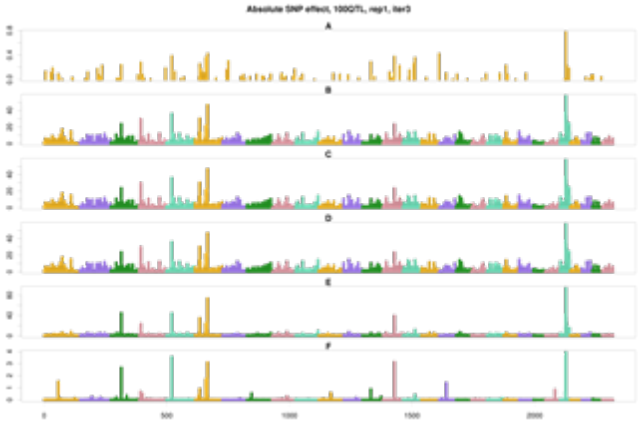


Figure 1. Accuracies with data containing different number of simulated QTL.

5 QTL



100 QTL



500 QTL

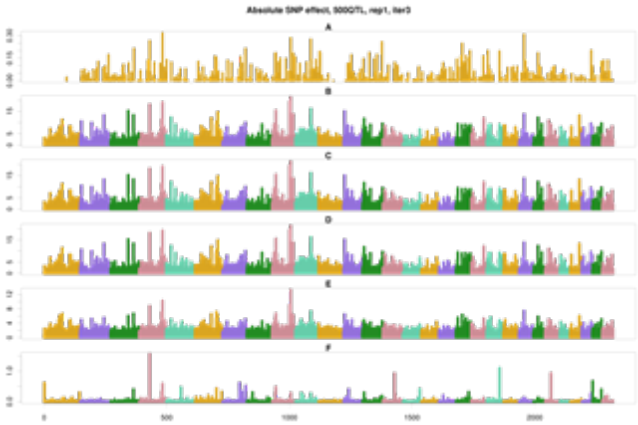


Figure 2. QTL effects and absolute SNP effects of 5 methods using different weight with data containing different number of simulated QTL. A: true QTL effects; B: default: proportional to $2p_i(1- p_i)u_i^2$; C: top N: updating only the top N SNP where N is equal to 1 × QTL number; D: top 3N: updating only the top 3N SNP where 3N is equal to 3 × QTL number; E: constant: proportional to $2p_i(1- p_i)u_i$+ constant, where constant = 15, 5 and 25 for 5, 100 and 500 QTL; F: BayesC with π equal to the proportion of ignored SNP in WssGBLUP using Option 2.

RESULTS AND DISCUSSION

Accuracy. Figure 1 shows accuracies of GEBV for 5 different methods under three scenarios. With Option 1 the accuracy increased initially but declined later. As the number of QTLs increased, the inflection point came earlier (0.875, 0.817 and 0.873 on 4th, 3rd, and 2nd iterations for 5, 100, and 500 QTL scenarios, respectively).

At early (≤3) iterations, options 2 and 3 could reach the best accuracy at 2nd (500 QTL scenario) or 3rd (5 and 100 QTL scenarios) iteration. Option 3 provided the highest accuracy compared to other options by improving the peak accuracy of Option 1 in 2% (0.832 vs. 0.817) and 1% (0.878 vs. 0.873) under 100 and 500 QTL scenarios, respectively. For the 5 QTL scenario, the best accuracy of Option 2 did not outperform Option 1 (0.871 vs. 0.875), but kept increasing throughout the iterations. However, except for the 5 QTL scenario, the accuracy of Options 2 and 3 still dropped after the peak points.

Although Option 4 did not reach a high accuracy at early iteration as Options 2 and 3, the accuracies remained stable once reaching the peak point (0.869 and 0.879, for 5 and 500 QTL, respectively), and kept increasing for the 100 QTL scenario (0.839 at 10th iteration). For 100 and 500 QTL scenarios, the accuracy by Option 4 exceeded those by Option 2 but only by 1%.

Except for the 5 QTL scenario, all WssGBLUP options under all scenarios surpassed BayesC in accuracy. BayesC was 6% lower than the peak accuracy of Option 1 under the 100 QTL scenario (0.772 vs. 0.817), and 18% lower under the 500 QTL scenario (0.714 vs. 0.873). This implies that BayesC performs well when the number of QTL is small, whereas WssGBLUP performs better when the number of QTL is large (>50, results not shown). However, Option 3 in the 5 QTL scenario outperformed BayesC, indicating that the choice of the weight in WssGBLUP affects the accuracy. Options 2 to 4 enhanced the accuracy because over-shrinkage of SNP effects was avoided.

SNP identification. Figure 2 shows the Manhattan plots of SNP effects (graph A) of all methods and scenarios at the iteration corresponding to the best accuracy (graph B-F; iteration 3, 3 and 2, for 5, 100 and 500 QTL, respectively). Under all scenarios, Options 2 to 4 reduced the noise. Although up to 20% of the QTLs did not create large peaks, most of QTL with large effects were identified, and few peaks were due to false positives.

Option 4 distinguished QTL effects more clearly than Options 2 and 3 as adding a constant did not change the scale of each SNP relative to others like in Options 2 and 3, whereas small effects were not shrunk as much as in the other scenarios. However, under the 500 QTL scenario, all Manhattan plots were noisy.

CONCLUSION

Presented procedures to calculate weights of SNP in WssGBLUP can be effective in improving both the accuracy of GEBV and GWAS. By the procedures, GEBVs were more accurate than by BayesC, although different parameters in the latter could change ranking of methods. Option 4 maybe the best choice given that in real data we may not know the true number of QTL. The WssGBLUP method is especially useful for GWAS when the population contains many ungenotyped animals and complex models preclude accurate deregression.

LITERATURE CITED

Aguilar, I., Misztal, I., Johnson, D.L. et al. (2010). *J. Dairy Sci.* 93:743-752.

Fernando, R., and Garrick, D. (2009). *GenSel-User manual for a portfolio of genomic selection related analyses.* 3rd ed., version 2.14. Accessed Mar. 31, 2013.

Garrick, D., Taylor, J. F., and Fernando, R. L. (2009).*Genet. Sel. Evol.* 41:55-62

Groenen, M. A. M., Wahlberg, P., Foglio, M. et al. (2009). *Genome Res.* 19:510-519.

Kizilkaya, K., Fernando, R. L., and Garrick, D. J. (2010). *J. Anim. Sci.* 88:544-551

Legarra A., Robert-Granie, C., Croiseau, P. et al. (2010). Proc. 9th WCGALP, Leipzig, Germany. ID118

Misztal, I., Tsuruta, S., Strabel, T. et al.(2002). Proc. 7th WCGALP, Montpellier, France. ID28.

Misztal, I., Legarra, A.; Aguilar, I. (2009). *J. Dairy Sci.* 92:4648–4655

Sargolzaei, M. and Schenkel, F. S. (2009). *Bioinformatics.*25:680-681.

Sun X., Garrick, D. J., and Dekkers, J. C. M. (2011). 44th ADSA-ASAS Midwest meeting, Iowa, USA. Poster 43393.

Wang, H., Misztal, I., Aguilar, I. et al. (2012). *Genet. Res.(Camb.)*,94:73-83