Results of Genome Wide Association Studies Improve the Accuracy of Genomic Selection

***Zhe Zhang*[1], Jinlong He[1], Hao Zhang[1], Ping Gao[1], Malena Erbe[2], Henner Simianer[2], Jiaqi Li[1]**
[1] College of Animal Science, South China Agricultural University, Guangzhou 510642, China, [2] Department for Animal Sciences, Animal Breeding and Genetics Group, Georg-August-Universität Göttingen, 37075 Göttingen, Germany

ABSTRACT: Genomic selection (GS) is widely implemented in livestock breeding due to its potential to accelerate genetic progress. Recently, results of genome wide association study (GWAS) are accumulated for most livestock species. Are these GWAS results useful for GS or not? We validated their usefulness with a dairy cattle population using a BLUP|GA model. Genotypes and phenotypes of 2,000 bulls and a cattle QTL list from animalQTLdb were used. We compared the accuracy of BLUP|GA and GBLUP with five-fold cross validation. Results showed that the public GWAS results can improve the accuracy of GS via BLUP|GA. BLUP|GA outperformed GBLUP for traits with large effect genes. Both the prior knowledge of QTL counts and p value were useful to improve GS accuracy. BLUP|GA deserved further investigations for species for which GWAS results are publically available.

Key words: genomic selection; genome wide association study; dairy cattle; genetic architecture.

INTRODUCTION

Predicting unknown phenotypes or genetic values for complex traits is an interesting and fast developing area in the context of human disease studies as well as in animal and plant breeding. In this context, genomic selection (GS) (Meuwissen et al. (2001)) and genome wide association studies (GWAS) (Klein, et al. (2005)) were widely used approaches. Both use genomic and phenotypic data in a combined analysis. Though hundreds of GWAS were conducted for each common livestock species, which results potentially reveal the genetic architecture of complex traits in a comprehensive manner and should be potent in improving GS, these GWAS results could not be directly used to improve GS with usual genomic selection method (de Los Campos et al. (2012)). Recently, the new model BLUP|GA proposed by Zhang et al (2014) can link abundant GWAS results to GS. The objective of this research was to evaluate the performance of BLUP|GA in different ways of incorporating GWAS results.

MATERIALS AND METHODS

Data. Genotypic data of 5,024 German Holstein bulls were genotyped with the Illumina Bovine SNP50 Beadchip. After quality control (MAF > 0.01, call rate > 0.95), 42,551 SNPs were remaining for further analyses. Conventional estimated breeding values for milk fat percentage (FP), milk yield (MY) and somatic cell score (SCS) with reliabilities greater than 70% were available for all bulls. These three traits represent three different possible genetic architectures of complex traits. We chose 2,000 bulls with the highest reliabilities in the trait MY to decrease the computing time. In order to consider the scenarios with even smaller population size, we randomly selected a subset of 500 and 125 individuals out of these 2,000 individuals.

The list of GWAS and QTL mapping results for dairy cattle was obtained from animalQTLdb (Hu et al. (2007)) (http://www.animalgenome.org/QTLdb, Release 22). The number of SNPs from the genotype data which were located in these QTL regions and the number of QTL reports for these SNPs are counted and employed as GWAS results in further analyse. In this study, the marker weights were calculated as sum of QTL counts or sum of -log(p_value, 10).

Model. The statistical model for the genomic BLUP approach is $\mathbf{y} = \mathbf{X}\mu + \mathbf{Zg} + \mathbf{e}$. The GBLUP approach assume $\mathbf{g} \sim N(0, \sigma_g^2 \mathbf{G})$, and $\mathbf{G} = \mathbf{MIM}^{\mathrm{T}} / c$ (VanRaden (2008)). Hence, GBLUP assumed that all markers in **M** contributed equally. However, this may not be proper for all complex traits. Based on GBLUP, we proposed to use $\mathbf{T} = \omega\mathbf{S} + (1-\omega)\mathbf{G}$, where ω is an overall weight for large effect markers, $\mathbf{S} = \mathbf{M}_1 diag(h_1,...,h_{m_1})\mathbf{M}_1^{\mathrm{T}} / c_1$, and **h** is a vector of marker weights obtained from GWAS results. It should be noted that the **S** matrix was built only for a sub set of large effect markers (N=m_1), and these markers were selected according to the publicly available GWAS results accessed from animalQTLdb. The matrix **S** is supposed to capture the genetic architecture part for the trait under consideration. Further note that **T** equals **G** for $\omega = 0$. We named this method BLUP|GA ("BLUP approach conditional on the Genetic Architecture"). The variance components were estimated via a combined AI-EM restricted maximum likelihood algorithm via the DMU software package (Madsen et al. (2008)).

Model validation. Five-fold cross-validation (CV) procedure and accuracy was used to assess the predictive ability of GBLUP and BLUP|GA. For all scenarios, the five-fold CV was replicated 20 times, resulting in 20 average accuracies. The accuracy was defined as the Pearson correlation coefficient between traditional estimated breeding value and genomic estimated breeding values.

RESULTS AND DISCUSSION

In this study, predictive ability of GS was measured via five-fold cross-validation procedures, applying the BLUP|GA approach with genetic covariance structure given by the trait-specific variance-covariance matrix **T**. The weights **h** for the m_1 markers in **T** were chosen based on counts of how often a marker was reported to be within a significant QTL region during association studies previously carried out in the literatures, a knowledge we retrieved from publicly available QTL

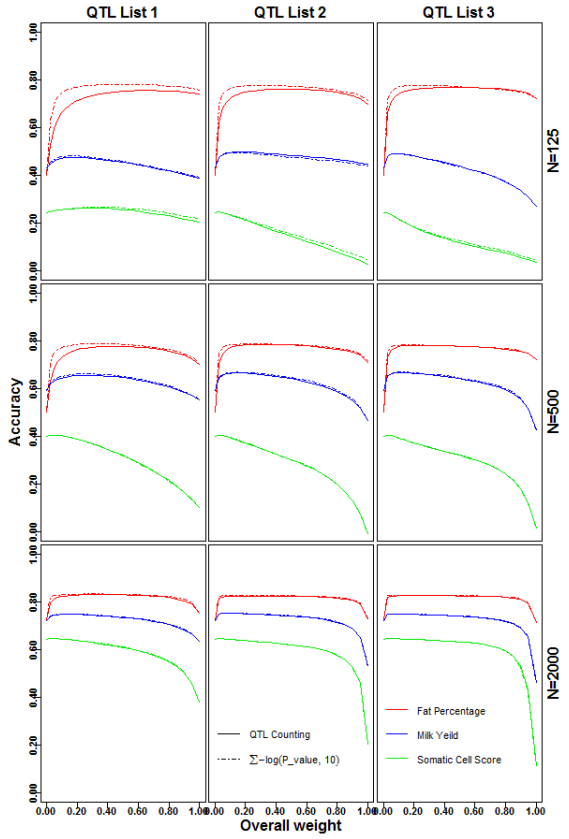

Figure 1: Accuracy of BLUP|GA and GBLUP for milk yield, milk fat percentage, and somatic cell score for scenarios with different population size and different weights.

databases. We set different thresholds to the counts of a marker to obtained three QTL lists, and build **T** matrix with them. We compared the accuracy of BLUP|GA with the standard GBLUP approach.

The result clearly showed that BLUP|GA performed better than GBLUP (the left point of each line with $\omega = 0$, Figure 1) for FP and MY, but not for SCS. This trend is valid for all scenarios with different population sizes and different QTL lists. This suggested that GWAS results can help to improve the accuracy of GS via BLUP|GA. It is also clear that the accuracy for FP is higher when we used p values to build the **S** matrix than that of QTL count. This implied that a strength weighting vector **h** is needed for traits with significant QTLs.

The difference between BLUP|GA and any other GS approach is that BLUP|GA can model any "existing knowledge" about the genetic architecture of complex traits, including publicly available GWAS or QTL mapping results. This can be achieved by building the **S** matrix according to a list of important markers and their corresponding weights which are obtained from "existing knowledge", then build the **T** matrix as a weighted sum of **S** and **G**, finally predict the genetic value of all individuals by solving the mixed model equations, in which the covariance structure is given by the **T** matrix. Hence, an important step for BLUP|GA is the selection of SNPs to build **S**. These SNPs should lie in trait associated chromosomal regions and their corresponding marker weights should represent their relative contributions to the genetic architecture of the trait under consideration. In the present study, the SNPs was chosen according to the times that it was reported be within a QTL region, and the counts or p value was used to be the weights in **S**. However, the ways to select significant markers and weights used to build **S** are not limited to the rules we proposed in this study. BLUP|GA provided a port for all kinds of existing knowledge about the genetic architecture to the prediction model, and should be validated in more situations.

CONCLUSION

The BLUP|GA is a special GS model that can link publicly available GWAS results to GS. Via the BLUP|GA model, GWAS results can help to improve the accuracy of GS. Hence, the existing publicly available GWAS result can well reflect the genetic architecture of a complex trait. The advantage of BLUP|GA depends on the characteristic of the genetic architecture underlying a complex trait and the comprehensiveness of the public knowledge on that trait. BLUP|GA outperformed GBLUP for two out of the three traits and GWAS p values were better than QTL count while building **S**. The BLUP|GA approach deserves further investigations for species where GWAS results are publically available.

LITERATURE CITED

de Los Campos, G., Hickey, J.M., Pong-Wong, R. et al. (2012). Genetics, 193: 327-345.

Hu, Z.L., Fritz, E.R., Reecy, J.M. (2007). Nucleic Acids Res., 35: D604-609.

Klein, R.J., Zeiss, C., Chew, E.Y. et al. (2005) Science, 308: 385-389.

Madsen, P., and Jensen, J.A. (2008). DMU. Version 6, release 4.7.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E. (2001). Genetics, 157: 1819-1829.

VanRaden, P.M. (2008). J. Dairy Sci., 91: 4414-4423

Zhang, Z., Ulrike, O., Erbe M. et al. (2014) PLoS ONE, 9(3):e93017