TA-BLUP: A New Genetic Evaluation Method For Genomic Selection

Zhe Zhang[*†], Xiangdong Ding[*], Jianfeng Liu[*], Dirk-Jan de Koning[†] and Qin Zhang[*]

Introduction

Molecular data has become an important information source in selection schemes. Many methods utilizing molecular information have been proposed to estimate genetic value and help breeders to make selection decisions. These methods mainly include marker assisted selection (MAS) (Smith 1967), best linear unbiased prediction with a genomic relationship matrix (G-BLUP) (VanRaden 2008) and genomic selection (GS) (Meuwissen, Hayes and Goddard 2001). MAS employs the information from one or several identified markers and/or pedigree to evaluate a selection candidate. To avoid problems in MAS and take full advantage of high dense marker information, GS first estimates marker effects in a reference population and subsequently sums up SNP effects to obtain estimated genomic breeding values (GEBVs) for a genotyped candidate. The G-BLUP method employs marker information from the whole genome to construct a realized relationship matrix (RRM or G matrix) to replace the numerator relationship matrix (NRM or A matrix) derived from pedigree in the framework of mixed model equations. This matrix was proved to be superior over NRM in variance component estimation and breeding value estimation (Hayes and Goddard 2008; Visscher, Medland, Ferreira *et al.* 2006). However, G-BLUP is equivalent to ridge regression BLUP (RR-BLUP) method in GS (Goddard 2009; Habier, Fernando and Dekkers 2007; VanRaden 2008), which is not as good at predicting ability as BayesB for some scenarios (Meuwissen, Hayes and Goddard 2001).

In this study, a trait specific marker derived relationship matrix (TA) was suggested to replace the G matrix to improve the predicting ability. In the TA matrix, not only marker genotypes but also their effects were utilised to describe the variance-covariance structure between pairs of individuals. The method including a TA matrix into mixed model equations was named TA-BLUP in our study. We compared TA-BLUP to G-BLUP, RR-BLUP and BayesB.

Material and methods

Data simulation: The simulation started from 1,000 generations of historical population with an effective population size of 100 in each generation. After that, 6 generations of with 1,000 individuals each were generated with breeding values, phenotypes of a quantitative

[*] Key Laboratory of Animal Genetics and Breeding of the Ministry of Agriculture, College of Animal Science and Technology, China Agricultural University, Beijing, 100193, China
[†] The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Roslin, EH25 9PS, UK

trait and marker genotypes. The first generation was used as the training population and the subsequent 5 generations were used as validation populations of which only the marker genotypes were available. The genome was simulated with 5 chromosomes and a total length of 5 Morgan. On each chromosome, 2,000 biallelic markers were randomly distributed. Positions of 100 QTL were randomly selected from the whole genome. The effects of QTL were generated assuming a normal distribution. The breeding value was the sum of all QTL effects. The heritability was set to be 0.5.

Estimation of marker effects: RR-BLUP and BayesB were employed to estimate marker effects using phenotype and genotype information of the training population. The statistical model can be written as $y = Xb + \sum_{i=1}^{m} Z_i g_i + e$, where b is a vector of fixed effects (including an overall mean), g_i is the i^{th} marker effect, m is the total number of markers and e is a vector of residual errors. X and Z are design matrices corresponding to b and g. We assumed that residuals e are independent and follow a normal distribution, $e \sim N(0, \sigma_e^2)$. All marker effects g_i were also assumed to be normally distributed, $g_i \sim N(0, \sigma_{gi}^2)$.

In RR-BLUP, the simulated variance components were used as the prior. The i^{th} marker variance can be calculated from $\sigma_{gi}^2 = 2\sigma_a^2/m$, where σ_a^2 is the total additive genetic variance. In BayesB, the prior proportion of loci without effect, π, was estimated from a pre-analysis of the simulated data. The MCMC chain was run for 10,000 cycles with 100 cycles of Metropolis-Hastings sampling in each Gibbs sampling the first 2,000 cycles were discarded as burn-in. All the samples of marker effects after burn-in were averaged to obtain the estimate of marker effects.

Estimation of GEBVs in the validation population: GEBVs of individuals in the validation populations were estimated in two different ways, i.e., by summing up all estimated marker effects obtained using either BayesB or RR-BLUP and by using the TA-BLUP method. In the TA-BLUP method, a mixed linear model, $y = Xb + Zu + e$, was used, where y is the vector of phenotypes of individuals in the training population and u is the vector of breeding values of the individuals in both the training and the validation population. The variance-covariance matrix of u is $A\sigma_a$, where A is a trait specific marker derived relationship matrix (TA matrix). Using A, u is estimated using the mixed model equations in the same manner as conventional BLUP.

The TA matrix was constructed based on the marker genotypes and their estimated effects. Based on the basic scoring rule of Eding and Meuwissen (2001), the allelic relationship at locus k between individual i and j is defined as $S_{ijk} = \sum_{m=1}^{2} \sum_{n=1}^{2} I_{mn} / 4$, where I_{mn} is 1 if allele m in the first individual is identical to allele n in the second individual or 0 otherwise. The score is weighted with the estimated marker effects and the weighted overall relationship between individual i and j is defined as $S_{ij} = \sum_{k=1}^{N} S_{ijk} w_k / \sum_{k=1}^{N} w_k$, where w_k is the weight for marker k and is defined as the absolute value of its effect estimated using either RR-BLUP or BayesB and N is the number of markers used for constructing the TA matrix.

We use TA-B and TA-P to represent the TA-BLUP method with the marker effects being estimated using BayesB and RR-BLUP, respectively. To evaluate the function of weighting in the TA-BLUP method, we also considered the case where w_k was set to be 1 for all markers (i.e., no weighting) and denoted this method as G-BLUP, since in this case the TA-matrix is equivalent to the G-matrix.

Results and discussion

Accuracy of GEBVs was defined as the correlation between GEBVs and TBVs (Meuwissen, Hayes and Goddard 2001). Accuracies after the training population indicate the predicting ability of the methods. As shown in Table 1, the highest accuracy in generation 2 is 0.800 for TA-B, and the lowest is 0.704 for G-BLUP. Accuracy of BayesB is slightly lower than TA-B. Rank correlation for all methods is lower than the corresponding correlation by 1.2 percents on average. The ranks of methods by both correlation coefficients are consistent. This indicates that both can be used to compare methods.

Table 1: Correlation and rank correlation between GEBVs from different methods and true breeding values (TBVs), and regression of TBVs on GEBVs in generation 2

Methods	Correlation	Rank correlation	Regression
BayesB	0.798±0.010	0.786±0.012	1.064±0.031
RR-BLUP	0.748±0.005	0.735±0.006	1.050±0.018
TA-B	0.800±0.009	0.789±0.011	1.034±0.019
TA-P	0.797±0.008	0.786±0.009	1.154±0.025
G-BLUP	0.704±0.006	0.692±0.007	1.597±0.045

Mean ± S.E.. Results based on 10 repeats.

The regression coefficient of TBVs on GEBVs is used to assess the bias of GEBVs. All methods overestimated the breeding values with a regression coefficient greater than 1. The most unbiased method is TA-B and the most biased method is G-BLUP. Results in our simulation show that including marker effects as prior in TA-BLUP when constructing a genomic relationship matrix can increase the predicting ability and reduce bias compared with the G-BLUP method.

Table 2: Correlation coefficients between GEBVs and TBVs in generation 2 to 6

Methods	Generation				
	2	3	4	5	6
BayesB	0.798±0.010	0.755±0.013	0.722±0.010	0.708±0.014	0.685±0.017
RR-BLUP	0.748±0.005	0.681±0.007	0.630±0.011	0.607±0.011	0.584±0.016
TA-B	0.800±0.009	0.756±0.011	0.723±0.009	0.709±0.012	0.686±0.015
TA-P	0.797±0.008	0.749±0.010	0.716±0.009	0.698±0.011	0.677±0.016
G-BLUP	0.704±0.006	0.633±0.009	0.580±0.013	0.566±0.012	0.543±0.016

Mean ± S.E.. Results based on 10 repeats.

The decline of accuracy of GEBVs over generations can be a measure of the persistency of predicting ability. As shown in Table 2, TA-B performs as well as BayesB through the five generations. G-BLUP performs worst in all generations with a decrease of 0.161 in accuracy

from generation 2 to 6. TA-P showed a large advantage in accuracy over RR-BLUP in all generations, and a better persistency than RR-BLUP due to increased difference of accuracy between them, which is 0.049 in generation 2 and 0.093 in generation 6. By taking estimated marker effects as prior, TA-BLUP also showed better predicting ability and persistency than the G-BLUP method.

The proposed TA-BLUP method performed equally to the BayesB method in our simulation (Table 1 and Table 2), although they are based on different models, which is additive marker effect model for BayesB and additive polygenic model for TA-BLUP. However, there are several favourable features of TA-BLUP which BayesB does not possess. First, due to the merit of MME, reliability of GEBVs at individual level for TA-BLUP method can be easily obtained using the rule proposed in traditional BLUP. This can be useful for breeders to make selection decisions. Second, the model of TA-BLUP can be extended to include non-genotyped individuals to utilize more records, which might improve the accuracy of genomic selection and has been discussed for the G-BLUP method (Legarra, Aguilar and Misztal 2009). Third, it is also possible to include the residual polygenic effects in the model of TA-BLUP, which would benefit genomic selection when the available markers are insufficient to explain all genetic variance.

Conclusion

The results of this study show that taking marker effects into account when constructing genomic relationship matrix is a successful trial. The proposed TA matrix is an improvement upon NRM and RRM. The TA-BLUP method performs equally to BayesB in terms of accuracy of genomic selection and holds some favourable features that BayesB lacks.

Acknowledgement

This work was supported by the State High-Tech Development Plan (Grant No. 2008AA101002), the National Natural Science Foundation of China (Grant No. 30800776, 21028089), and the National Key Basic Research Program of China (Grant No. 2006CB102104). DJK acknowledges support from BBSRC through the Institute Strategic Programme Grant.

References

Eding, H., and Meuwissen, T. H. E. (2001). *J. Anim. Breed. Genet.*, 118:141–159.
Goddard, M. E. (2009). *Genetica*, 136:245–257.
Habier, D., Fernando, R. L., and Dekkers, J. C. M. (2007). *Genetics*, 177:2389–2397.
Hayes, B. J., and Goddard, M. E. (2008). *J. Anim. Sci.*, 86:2089–2092.
Legarra, A., Aguilar, I., and Misztal, I. (2009). *J. Dairy Sci.*, 92:4656–4663.
Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). *Genetics*, 157:1819–1829.
Smith, C. (1967). *Anim. Prod.*, 9:349–358.
VanRaden, P. M. (2008). *J. Dairy Sci.*, 91:4414–4423.
Visscher, P. M., Medland, S. E., Ferreira, M. A. *et al.* (2006). *PLoS Genet.*, 2:316–325.