GENOMIC BEST LINEAR UNBIASED PREDICTION USING DIFFERENTIAL EVOLUTION

H.A. Al-Mamun[1], P. Kwan[2], S. Clark[1], S.H. Lee[3], H.K. Lee[3], K.D. Song[3], S.H. Lee[4] and C. Gondro[1]

School of Environmental and Rural Science, University of New England, NSW, Australia
[2]School of Science and Technology, University of New England, NSW, Australia
[3]The Animal Genomics and Breeding Center, Hankyong National University, Anseong, Korea
[4]Division of Animal and Dairy science, Chung Nam National University, Daejeon, Korea

SUMMARY

In this paper we proposed a method to improve the accuracy of prediction of genomic best linear unbiased prediction (GBLUP). In GBLUP a genomic relationship matrix (GRM) is used to define the variance-covariance relationship between individuals and is calculated from all available genotyped markers. Instead of using all markers to build the GRM, which is then used for trait prediction, we used an evolutionary algorithm (differential evolution – DE) to subset the marker set and identify the markers that best capture the variance-covariance structure between individuals for specific traits. This subset of markers was then used to build a trait relationship matrix (TRM) that replaces the GRM in GBLUP (herein referred to as TBLUP). The predictive ability of TBLUP was compared against GBLUP and a Bayesian method (Bayesian LASSO) using simulated and real data. We found that TBLUP has better predictive ability than GBLUP and Bayesian LASSO in almost all scenarios.

INTRODUCTION

Genomic selection is a method based on marker-assisted selection that is used to determine the genetic value of individuals so that they can be selected as parents in breeding programs. In genomic selection, marker effects are estimated from a *discovery* (or training) dataset that comprises individuals that have both genotypic and phenotypic information. Then genomic estimated breeding values (GEBV) for selection candidates without phenotypic records are estimated based on these marker effects. Within the framework of genomic selection, two different approaches are commonly used to estimate the marker effects in the training data. The first approach assumes all SNP have a non-zero contribution to the variance of the trait of interest and the distribution of the SNP effects follows a normal distribution. Both ridge regression best linear unbiased prediction (RR-BLUP) and genomic best linear unbiased prediction (GBLUP) are based on this assumption. The second approach is based on non-linear methods that emphasize certain genomic regions and allow marker effects to come from different statistical distributions. Bayes A, Bayes B (Meuwissen *et al.* 2001), Bayes C (Habier *et al.* 2011) and Bayesian LASSO (Least Absolute Shrinkage and Selection Operator) (de los Campos *et al.* 2009) are examples of such non-linear methods for genomic selection.

GBLUP was first suggested by VanRaden (VanRaden 2008) and has been used for prediction of breeding values for use in agricultural selection programs (Goddard & Hayes 2009). In GBLUP a genomic relationship matrix (GRM) is used to define the variance-covariance relationship between individuals and is calculated from all available genotyped markers. Most of the proposed (VanRaden 2008; Goddard *et al.* 2011) implementations of the GRM are based on the infinitesimal model which assumes that a very large number of genes are evenly distributed across the genome, each contributing a minute amount to the trait of interest. In GBLUP the same GRM is used for the estimation of GEBV irrespective of the trait. Most traits of interest in animal or plant breeding are in fact polygenic but not necessarily infinitesimal; i.e. different traits are

controlled by (a limited) different sets of genes. The true underlying genetic structure of any trait deviates from the infinitesimal model to a certain extent and most quantitative traits are significantly affected by a finite set of genes (Meuwissen et al. 2001). Therefore, a GRM estimated based on the assumption of the infinitesimal model cannot optimally describe the variance-covariance relationships between individuals for the trait of interest. A model that uses only the SNP that track the relevant regions (QTL) of the traits of interest may be more appropriate to construct the variance-covariance relationship matrix.

Whereas methods that place different weightings on markers have also been proposed (i.e. Bayes A, B, C, and R), studies in which evolutionary algorithms like Differential Evolution (Storn & Price 1995) were applied to solve such a problem are few. Differential Evolution (DE) is a reliable and versatile function optimizer that is easy to implement, fast to converge, and does not require complex initial settings. DE has been successfully used in a wide range of biological optimization problems. The objective of this study was to apply DE in identifying an optimum subset of SNP to construct the variance-covariance relationship matrix for a specific trait, followed by estimation of the GEBV using BLUP based on this matrix. The performance of this method, called *Trait Best Linear Unbiased Prediction* (TBLUP) was assessed by comparing it with GBLUP and a Bayesian method (Bayesian LASSO) on simulated and real data.

MATERIALS AND METHODS

Data. One real dataset and one simulated dataset were used to assess the proposed method. Genotype information on 50K Illumina BeadChip array was available for a total of 1,937 cattle from pure-breed Korean Hanwoo with four phenotypic data: back fat (BF), carcass weight (CW) eye muscle area (EMA), and marbling score (MS). The simulated dataset contains genotype information on 10,000 samples for 40,000 SNP with simulated phenotypes. Genotypes were simulated by random sampling from frequencies under Hardy-Weinberg equilibrium (in effect an unstructured population). Phenotypes were simulated for different numbers (50, 100, 200 and 500) of known QTL. Randomly selected SNP were assigned different effects drawn from a normal distribution. Then the phenotypes were created by summing up the SNP effects plus a random environmental effect component. Both the real and the simulated datasets were divided into discovery and validation populations: 100 samples were randomly selected as validation samples; the remainder of the data were used as the discovery population. The 100 random samples selected for validation were the same for all scenarios.

Evolutionary algorithm. An algorithm based on DE was developed to select the best SNP subset in order to create the genomic relationship matrix (GRM). To select a SNP subset for the GRM, random keys were used. A random key is an evolvable vector of real values (one for each SNP) that are sorted by the objective function. The ranking of the key is then used to rank the SNP. The idea is that SNP that are better for genomic prediction evolve to higher values in the key with the rest to lower values. Once the keys are sorted, they reflect the relative value of a given SNP. An additional parameter to be optimized is the number of SNP in the panel – a *cutoff value*. The DE evolves the cutoff value, sorts the SNP based on their key values and uses the top ranked ones up to the number defined by the cutoff value. More in-depth details on the algorithm are given in (Gondro & Kwan 2012). An objective function was used to calculate the fitness of the selected SNP. In the objective function, the discovery population was further divided into two subsets: i) a subset population with known phenotype, and ii) another subset population with unknown phenotype (phenotypes were set to missing for these samples). A genomic relationship matrix was constructed using only the selected SNP for all discovery samples, which was then used to predict (by using GBLUP) the phenotype for the samples in the unknown subset population. The fitness of a selected SNP subset was defined as the correlation between the actual phenotype and the predicted phenotype. For each phenotype the DE evolved for 1,000 generations.

RESULTS AND DISCUSSION

Figure 1 shows a comparison of true genetic value (TGV) vs predicted breeding value for 50 known QTLs. Table 1 shows the comparison between prediction accuracies for 50, 100, 200 and 500 known QTLs with the simulated dataset for the different methods of genomic prediction (GBLUP, Bayesian LASSO, and our proposed method TBLUP). For the simulated data, the proposed method performed better than GBLUP and Bayesian LASSO. For the real dataset, heritability of the phenotypes were estimated using the GCTA software (Yang *et al.* 2011) which were 0.54, 0.56, 0.53 and 0.43 for BF, CW, EMA and MS respectively. Table 2 shows the genomic prediction accuracies for the validation samples obtained in the real data achieved by the three methods. Once again, the proposed method outperformed the two other methods for all four phenotypes.

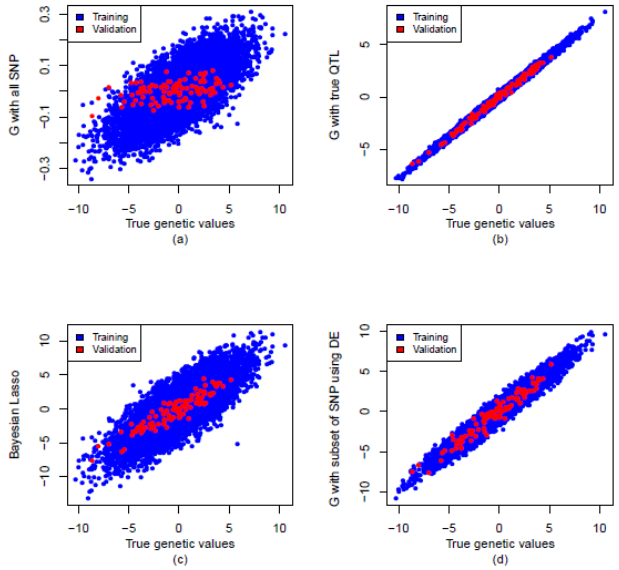

Figure 1. Prediction accuracy comparison (simulated phenotype with 50 known QTL). Blue dots are predicted values for the training data while the red dots are predicted values for the validation data. (a) Accuracy using all SNP, (b) accuracy using only the *true* QTL (QTN), (c) accuracy using Bayesian Lasso (BLR), and (d) accuracy using DE.

Table 1. Prediction accuracy comparison with the simulated data

True QTL	GBLUP	BL*	TBLUP	
			Accuracy	SNP used / QTL found
50	0.40	0.94	0.97	111 / 36
100	0.32	0.89	0.96	172 / 62
200	0.33	0.83	0.98	469 / 107
500	0.20	0.69	0.95	1041 / 186

Table 2. Prediction accuracy comparison with the real data

Trait	GBLUP	TBLUP	BL*
BF	0.370	0.440	0.394
CW	0.350	0.416	0.263
EMA	0.355	0.410	0.325
MS	0.236	0.245	0.233

*Bayesian LASSO

Improved accuracy of genomic prediction has immediate practical and commercial value for agricultural production as it leads to improved accuracy of selection and higher rates of genetic gain. GBLUP and various Bayesian methods for genomic prediction have been successfully employed in a large number of scenarios. The accuracy of these genomic predictions depends on

the genetic architecture of the trait, e.g. number of QTL and their effect sizes (Hayes *et al.* 2010), marker density, linkage disequilibrium (LD) and family relationships (Goddard *et al.* 2011; Clark *et al.* 2012; Wientjes *et al.* 2013), population structure (Moghaddar *et al.* 2014), sample size (Goddard 2009) and also the method used to estimate marker effects (Clark *et al.* 2011). Bayesian methods tend to outperform BLUP approaches when the trait is less polygenic (Clark *et al.* 2011). In practice, differences between methods in prediction accuracy are generally quite small. While these methods have well characterised statistical properties they are constrained by the underlying model assumptions. Given the dimensionality of the solution space, even very small estimates of effects in non-informative markers (noise) will, collectively, reduce prediction accuracy. This is an increasing problem with the increasing number of genetic variants to predict from. In TBLUP, we have attempted to reduce the noise from the system and tried to identify only the SNP that tracked relevant regions. In essence, the approach attempts to create a relationship matrix that tracks relationships between causal regions while excluding spurious associations and even true genetic relatedness that is not relevant to the trait of interest. We suggest that a *model free heuristic optimisation* approach choosing a small subset of best predictors is expected (and shown in the present study) to perform better in the context of genomic prediction.

CONCLUSION

In summary, we have described a novel BLUP method for estimation of breeding values using a trait-based relationship matrix, which we called TBLUP. The only difference between conventional GBLUP and the proposed TBLUP is that TBLUP focuses more closely on those markers that effectively contribute to the variation of the trait of interest and removes some of the noise that reduces accuracy of prediction. The preliminary results with real data were promising but further sudies (with more real data) are required to properly validate the method and better understand its advantages and limitations. In practice, the method can be used to develop smaller panel sets and this should reduce genotyping costs which can lead to a wider adoption by industry.

ACKNOWLEDGEMENTS

This work was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ01134906), Rural Development Administration, Republic of Korea and an Australian Research Council Discovery Project DP130100542.

REFERENCES

Clark S.A., Hickey J.M. & van der Werf J.H.J. (2011) *Genet. Sel. Evol.* **43**: 18.
Clark S.A., Hickey J.M., Daetwyler H.D. & van der Werf J.H.J. (2012). *Genet. Sel. Evol*. **44**: 4.
de los Campos G., Naya H., Gianola D., Crossa J., Legarra A., Manfredi E., Weigel K. & Cotes J.M. (2009) *Genetics* **182**: 375-85.
Goddard M. (2009) Genetica **136**: 245-57.
Goddard M.E. & Hayes B.J. (2009) *Nat. Rev. Genet.* **10**: 381-91.
Goddard M.E., Hayes B.J. & Meuwissen T.H.E. (2011) *J. Anim. Breed. Genet.* **128**: 409-21.
Gondro C. & Kwan P. (2012). IGI Global.
Habier D., Fernando R.L., Kizilkaya K. & Garrick D.J. (2011) *BMC Bioinformatics* **12**: 186.
Hayes B.J., Pryce J., Chamberlain A.J., Bowman P.J. & Goddard M.E. (2010) *PLoS Genet*. **6**.
Meuwissen T.H., Hayes B.J. & Goddard M.E. (2001) *Genetics* **157**: 1819-29.
Moghaddar N., Swan A.A. & van der Werf J.H.J. (2014) Genet. Sel. Evol. **46**: 58.
Storn R. & Price K. (1995) In: *Technical Report, TR-95-012* Int. Comp. Sci. Inst. (ICSI).
VanRaden P.M. (2008) *J. dairy sci.* **91**: 4414-23.
Wientjes Y.C.J., Veerkamp R.F. & Calus M.P.L. (2013) Genetics **193**: 621-31.
Yang J., Lee S.H., Goddard M.E. & Visscher P.M. (2011) *Am. J. Hum. Genet.* **88**: 76-82.