Proc. Assoc. Advmt. Anim. Breed. Genet. 22:281-284

USING MACHINE LEARNING METHODS TO IDENTIFY SUBSETS OF SNP FOR GENOMIC PREDICTION

B. Li[1,2], A. George[3], A. Reverter[1] and Y. Li[1]

[1] CSIRO Agriculture & Food, 306 Carmody Road, St Lucia, QLD, Australia
[2] School of Computer Science and Technology, Shandong Technology and Business University, Yan Tai, Shandong, P. R. China
[3] CSIRO Data61, Dutton Park, QLD, Australia

SUMMARY

Machine learning methods have gained popularity dealing with high dimensionality, highly correlated structure, or "large P, small N" genomic data problems. The methods have been shown to be efficient in GWAS and candidate gene identification. However, the utility of methods in identifying a subset of single nucleotide polymorphism (SNP) for genomic prediction of breeding values has not been explored before. In this study, using 40,184 SNP genotypes and the live weight phenotypes from 1,097 Brahman cattle, we examined the power of two machine learning methods, Random Forests and Gradient Boosting Machine, in the identification of top 1,000 or 3,000 SNP and using them for building a genomic relationship matrix (GRM) for genomic prediction of breeding values. Our results clearly show that using the subsets of SNP identified by the two methods resulted in the improvement both in the heritability estimate and the genomic prediction accuracy.

INTRODUCTION

Machine learning methods have gained popularity dealing with high dimensionality, highly correlated structure, or "large P, small N" problems arising from large genomic data analyses. Two of these methods, Random Forests (RF; Breiman, 2001) and Gradient Boosting Machine (GBM; Friedman, 2001), have been shown to outperform the conventional GWAS methods in association mapping and genomic-wide prediction of estimated breeding values (GEBV) (Chen and Ishwaran 2012; Lukbe *et al.* 2013; González-Recio *et al.* 2014; Waldmann 2016). However, the utility of these methods in identifying a subset of SNP to estimate GEBV has not been evaluated before. In this study, we examined the efficiency of RF and GBM for the identification of a subset of markers and tested these small panels using a GEBV approach.

MATERIAL AND METHODS

Data. We used a SNP dataset consisting of 40,184 SNP genotypes from 1,097 Brahman cattle from the Legacy Database of the CRC for Beef Genetic Technologies (www.beefcrc.com). The animals varying from 373 to 509 days old came from 57 contemporary groups and were measured for live weight (the average being 308.64 kg (± 38.85) with the range from 180 to 430 kg). A quality check of the marker data resulted in the removal of 2,102 SNP having MAF <0.01 or with missing genotypes due to full genotype requirement by the machine learning methods. A total of 38,082 SNP were used for the final analysis. Since machine learning methods are non-parametric approaches, they cannot directly fit fixed effects in the model to account for environmental effects. Therefore prior to any analysis, the phenotypic values were adjusted for the fixed effects of the contemporary group and age. The residuals from the linear model of analysis of variance were used as phenotype for the evaluation of the machine learning methods.

Machine leaning methods – RF and GBM. Details of the RF method can be found in Breiman (2001). In brief, RF uses a bootstrapping method to randomly select a subset of animals as the

training dataset (default being two third of the total number of animals), and a subset of SNP (default being a squared root of total number of SNP) to form a decision tree that splits the sampled individuals into two subsamples with different weight range values. The remaining individuals (one third) are then used as the validation dataset to determine the prediction error of the SNP tree on the phenotypes. The process repeats until a large number of decision trees are forming a forest (the parameter Ntree determines the size of a forest). Each individual decision tree building excise is independent to each other (with replacement). An individual SNP variable importance value (VIM) is determined by averaging the prediction error values of the SNP across all forest trees. GBM also generates multiple random samples to form trees, but subsequent samples always rely on the outcomes from the previous samples. It builds the trees iteratively by adding all "weak learners" – small trees with only a few SNP splits that predict the phenotypes with high bias but low variance (Lubke *et al.* 2013). Therefore, GBM reduces the prediction error by reducing bias through adding all the outcomes from a large number of models. Each method has its own parameter for measuring a SNP variable importance value (VIM). In RF, it is the %IncMSE (% increasing in mean squared error), while in GBM it is the Relative Influence - maximal cumulated estimated improvement in MSE. In both methods, the higher the VIM value, the more important the SNP is. The R libraries randomForest and gbm (https://www.rstudio.com/) were used for the analysis. The parameter Ntree was set as 2,000, the default values were used for RF and the learning rate of 0.1 for GBM.

Identification of top SNP and Gene Ontology (GO) Enrichment Analysis. Based on the ranked SNP VIM values from RF and GBM, the top 1,000 and 3,000 SNP were selected. The sets of genes near the top SNP or all the SNP with positive VIM values were examined for biological processes using the *Bos taurus* Reference from the PANTHER program (http://www.pantherdb.org/).

Construction of additive genomic relationship matrices using top SNPs for estimating genetic variances and genomic prediction of phenotypes. The additive genomic relationship matrix (GRM) was constructed using either 1,000 or 3,000 top SNPs from all animals, following the same method as in our chicken study (Li *et al.* 2016). An additive genomic model, fitting the GRM as random effect and the contemporary group and age as fixed effects, was then applied to estimate the genetic variance explained by each subset of top SNP (1,000 or 3,000). A random five-fold cross-validation scheme was used, i.e. randomly splitting 1097 animals into 5 equal-size groups and each group (20% of the population) was in turn assigned with missing phenotypic values and used as the validation set. The prediction accuracy was calculated as the correlation between the GEBVs of the animals with no phenotypic values and the true phenotypes of the animals adjusted for fixed effects. The program Qxpak v5.02 (Perez-Enciso and Misztal 2011) was used for the analyses.

RESULTS AND DISCUSSION

Profiles of SNP VIM values from RF and GBM. Figure 1 shows the distribution of the ranked SNP VIM values in RF and GBM. It can be seen that the majority of the SNP had very small or zero VIM values in RF and GBM. Of 38,082 SNP, 18,453 (48.5%) and 16,600 (43.6%) SNP were identified with the positive VIM values in RF and GBM, respectively. Between the two methods, there were 8,797 SNP in common. In RF, we also found a total of 6,660 SNP (17.5%) with negative VIM values, corresponding to the lower end of the distribution (Figure 1, RF graph). These negative values indicate that these SNP were problematic and should not be included in a prediction model. The reason was that the new prediction models using randomly permuted SNP positions on the decision trees had a much smaller mean squared error value (MSE) than the initial prediction model, hence a negative %InMSE value.

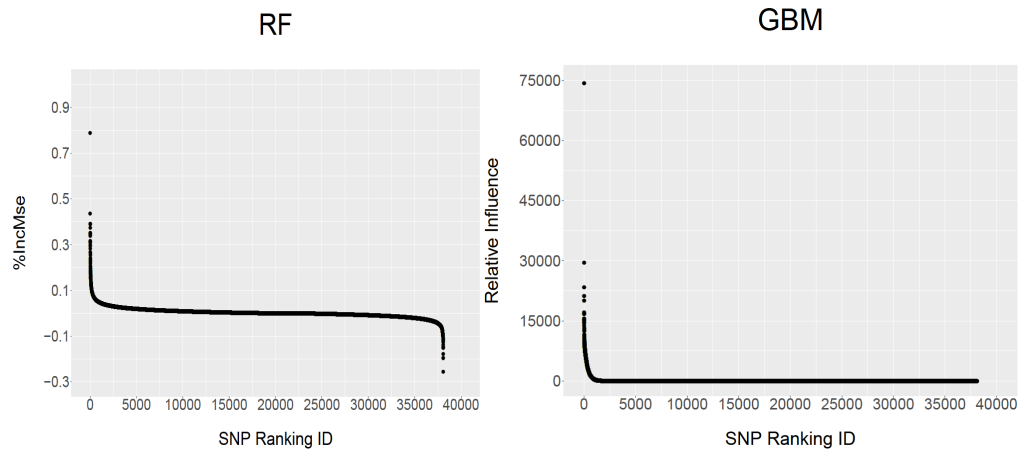

Figure 1. The distribution of ranked SNP variable importance values from RF (%IncMSE) and GBM (Relative Influence)

Gene enrichment analysis for SNP with positive VIM values from RF or GBM. When the sets of genes that were closest to the top 3,000 SNP or all the SNP with the positive VIM values were examined, we found that the top 3,000 SNP were primarily involved in the development, system development, visual perception, nervous system development and cellular activity ($p < 0.0001$). The evidence was much stronger for the genes near all the SNP with positive VIM, involving the growth pathways of development process (RF, *P*=1.54E-07; GBM: *P*= 2.09E-08) and system development (RF: *P* = 5.38E-07*;* GBM: *P* = 2.05E-07).

Both RF and GBM identified the same SNP with highest VIM value. It was ARS-BFGL-NGS-1712 mapped to gene BMPER (BMP binding Endothelial Regulator) on BTA4. A literature search found that BMPER played vital roles in adipocyte differentiation, fat development and energy balance in human and mouse (Zhao et al. 2015). The SNP was a very good candidate for selecting for increased body weight and rump length in cattle breeding (Zhao et al. 2015).

Table 1. Estimates of genetic variance and heritability (h^2) for live weight using different subsets of top ranking SNP identified by RF and GBM with additive genomic model

Method	No of Markers	Genetic Variance	Residual Variance	h^2
RF	1,000	332.60	256.78	0.565
	3,000	373.64	233.56	0.616
GBM	1,000	402.99	204.22	0.664
	3,000	417.05	184.08	0.694
All SNP	38,082	391.29	313.25	0.555

Estimates of genetic variance and heritability (h^2). Table 1 shows the REML estimates of genetic variance and h^2 for a subset of 1,000 or 3,000 top SNP identified by RF or GBM. Equivalent analysis using all 38,082 SNP are also listed in Table 1. It can be seen that there was a significant improvement in the h^2 estimate when the top 3,000 SNP from either RF or GBM were used in an additive genomic model. GBM performed particularly well in both 1,000 or 3,000 SNP cases, where the genetic variance estimates were higher than using all 38,082 SNP. Both RF and GBM captured complex SNP-SNP interactions, hence, resulted in an increased genetic variance.

Table 2. Prediction accuracy of GEBV for live weight using the top 1,000 or 3,000 SNP identified by RF and GBM methods

Methods	R1*	R2	R3	R4	R5	Average
RF1000	0.362	0.449	0.422	0.528	0.477	0.448
RF3000	0.321	0.408	0.421	0.443	0.440	0.407
Average	0.353	0.441	0.404	0.482	0.461	0.428
GBM1000	0.429	0.474	0.546	0.518	0.551	0.504
GBM3000	0.433	0.460	0.469	0.548	0.541	0.490
Average	0.418	0.463	0.476	0.501	0.510	0.474
All SNP	0.134	0.200	0.209	0.275	0.228	0.209

* Randomly selected 20% animals without phenotypic values

Accuracy of GEBV. Table 2 shows the accuracy of GEBV with a subset of SNP markers using an additive genomic model and a random split five-fold cross-validation scheme. In comparison to the additive model with all available SNP, surprisingly, the average prediction accuracy from either top 1,000 or 3,000 SNP outperformed the whole SNP panel, regardless the sources of the SNP chosen from RF or GBM. The prediction accuracy values from RF and GBM were double the amount of those of all SNP, ranged from 0.41–0.45 in RF and 0.43–0.50 in GBM.

Applications of large-scale SNP panels for genomic selection programs have a mixed success in livestock species (Waldmann 2016). While in the dairy cattle industry the genomic prediction of phenotypic values for production traits has achieved high success, the accuracy of GEBVs in beef cattle has been low (Waldmann 2016). We know from large number of GWAS and genomic prediction studies that the majority of SNP had little or no effects on phenotypes at all. This raises the question whether there is a benefit to use only small panel of SNP for genomic prediction? Our results here indicate that the machine learning methods, especially GBM, are efficient methods in identifying a subset of SNP with direct link to the candidate genes affecting the growth trait. It is possible to build a low density SNP panel for a genomic selection program.

In this study, we only examined a phenotype of moderate heritability in beef cattle. Further studies, using systematic approaches, are needed to validate the efficiency of machine learning methods in building low density SNP panels for different species or populations, optimal subset of SNPs and a range of phenotypes with different heritability values.

ACKOWLEDGEMENTS

We would like to acknowledge the financial supports for B. Li from the High Education Science and Technology Planning Program of Shandong Provincial Education Dept. (J16LN14), and Shandong Provincial Science and Technology Development Program (China) (2014GGX101044).

REFERENECES

Breiman L. (2001) *Machine Learning*. **45**: 5.
Chen X. and Ishwaran H. (2012) *Genomics*. **99**: 323.
Conzalez-Ricio O., Rosa G.J.M., Gianola D. (2014) *Livest. Sci.* **166**: 217.
Friedman, J. (2001). *Ann. Stat*. 29: 1189.
Li Y., Hawken R., Sapp R., George A., Lehnert S.A., Henshall J.M. and Reverter A. (2016) *Poult. Sci*. **0**:1.
Lubke G.H., Laurin C., Walters R., Eriksson N., Hysi P., Spector T.D., Montgomery G.W., Martin N.G., Medland S.E. and Boomsma D.I. (2013) *J Data Mining Genomics Proteomics* **4**:143.
Waldmann, P. 2016. *Genet Sel Evol*. **48**:42.
Perez-Enciso M. and Misztal I. 2011. *BMC Bioinformatics*.**12**:202.