USING RANDOM FOREST TO IDENTIFY SNPS THAT DECREASE ACCURACY OF GENOMIC PREDICTION – BEHAVIOUR OF SNPS WITH NEGATIVE VIM VALUES

Y. Li[1], F.S.S. Raidan[2], M. Naval Sanchez[1], A.W. George[3] and A. Reverter[1]

[1]CSIRO Agriculture & Food, St Lucia, QLD, 4067 Australia
[2]CSIRO Agriculture & Food, Hobart, TAS, 7004 Australia
[3]CSIRO Data61, Dutton Park, QLD, 4102 Australia

SUMMARY

Random Forest (RF) is one of the most popular machine learning methods for large genomics data analysis. It produces the variable important measures (VIMs) for individual features, which can be positive, zero or negative, indicating a positive or negative contribution of the feature. It is easy to interpret single nucleotide polymorphisms (SNPs) with positive or zero VIM values when applying RF for genomic prediction. However, little is known about the interpretation of SNPs with negative VIM values. Most importantly, what impact of these SNPs have on the genomic prediction accuracy of breeding values? In this study, using genotype information from 651,253 SNPs for 2,109 Brahman cattle with yearling weight phenotype, we applied the RF to identify 8,195 SNPs with negative VIM values and investigated their impact on genomic prediction. Specifically, we addressed the questions: 1) How did these SNPs differ from the top SNPs chosen from the RF with positive VIM values or the SNPs randomly selected but evenly spaced along a genome? 2) Did these SNPs have any biological relevance? Our results show that 1) including the SNPs with negative VIM values in the genomic prediction would result in the increase in error variance and decrease in the accuracy of genomic prediction; 2) these SNPs had no biological functions.

INTRODUCTION

Random Forest (RF, Breiman 2001) is one of the most commonly used machine learning methods for large genomics data analysis (Chen and Ishwaran 2012). One of its analysis output parameters is the variable importance measure (VIM). When applied to a continuous phenotype, RF generates the VIM - %IncMSE (percentage increase in Mean Squared Error). It measures an individual feature's contribution to the prediction accuracy of decision trees, via the change of MSE when the data for a feature (here a SNP) is permuted while all others are kept constant, with valid VIM values being positive, zero or negative. The larger the value (i.e., more positive), the more important the feature is. When applying this method to a high-density SNP panel for genomic prediction of a quantitative trait with a moderate heritability, the questions are: 1) how do SNPs with negative VIM values behave? 2) Do they have any biological relevance? In this study, we investigated the impact of SNPS with negative VIM values on the accuracy of genomic prediction and their possible molecular functions.

MATERIALS AND METHODS

Data. A Brahman cattle dataset, consisting of 2,109 genotyped animals with 651,253 SNPs per animal from the CRC for Beef Genetic Technologies (Porto-Neto *et al.* 2014), was used for this study. The animals were measured for yearling weight (YWT), which ranged from 115 to 353 kg with an average of 227.7 kg (±34.32kg). Since RF does not fit fixed effects into the process, prior to the RF analysis, the phenotypic values were adjusted for the fixed effects. These include contemporary group (combination of sex, year and location and 41 levels) and age (302-416 days). The residuals from the linear model of analysis of variance were then combined with the SNP information for the RF analysis.

Identification of SNPs with negative VIM values (SNP$_{negvim}$) using RF. The detailed RF method can be found in Li *et al*. (2018). The algorithm incorporates both training and validation procedures in its process to build decision trees to examine individual SNP contributions to prediction accuracy. We carried out an initial hyper-parameter fine-tuning for tree size (NTree) from 10,000, 12,000, … 20,000 using all SNPs, while the Mtry value was set as two times of the squared root of total number of SNPs. A CSIRO high performance cluster computer with the R program (version 3.4.0) and the library randomForest was used for the analyses.

Genomic prediction accuracy with and without SNP$_{negvim}$. A five-fold cross-validation scheme was applied to the RF and genomic prediction. The population was partitioned into 5 subsets and each time 4 subsets was used for training and the remaining subset was used for validating. In addition to the genomic prediction accuracy comparison between all SNPs with and without SNP$_{negvim}$, we also examined the results from the subsets of the top 1,000, 5,000, 10,000 and 50,000 SNPs with positive VIM values from the RF, and those of the same size but evenly spaced SNPs along the genome (denoted "Even"). A GBLUP model (VanRaden 2008) was used to estimate variance components and genomic breeding values (gEBVs), where the fixed effects in the model included the contemporary group and age. The accuracy of genomic prediction was calculated as the correlation between gEBVs and the adjusted phenotypic values, and then divided by the square root of heritability. The final estimates of genetic parameters were the average values from five validation analyses. The program AIREMLF90 (Misztal *et al*. 2002) was used in the GBLUP analyses.

Gene Ontology (GO) Enrichment Analysis. A locus-based gene ontology enrichment analysis using GREAT v3.00 (McLean *et al*. 2010) was undertaken. SNPs (±10 bp) were translated to human coordinates (GRC37/hg19) using UCSC's liftOver tool (minMatch = 0.1) (Hinrichs *et al*. 2006). A binomial and a hypergeometric test were used to assess the enrichment of molecular function terms and biological process terms.

Functional Enrichment Analysis. Cattle functional annotation was derived from i) histone chromatin marks in liver H3K27ac, and H3K4me3 (Villar *et al*. 2015); ii) ATAC-seq information from CD4+ and CD8+ from the Fr-AgENCODE (Foissac *et al*. 2018); iii) experimental in-house ATAC-seq in liver and muscle tissues; and iv) derived from current UMD3.1 annotation. To assess the significance of overlap between SNP datasets and functional genomic features we performed a Fisher's exact test with false discovery rate correction using the R package LOLA (Sheffield and Bock 2016).

RESULTS AND DISCUSSION

Characteristics of the SNPs with negative %IncMSE values. The distribution of average VIM (%IncMSE) values (from 5-fold training datasets) for ranked SNPs (from the most important to the least important) is shown in Figure 1. Surprisingly, of the 651,253 SNPs, 180,056 (27.7%) were found to have a negative average VIM value. However, when investigated further, we found that only 8,195 of these SNPs had the negative VIM values in all 5-fold datasets, and the remaining 171,861 SNPs varied between the datasets used. This clearly indicates that extreme caution needs to be taken when using the average of the VIM values from a cross-validation scheme as the criteria to identify the SNPs with negative VIM values. An extra step is required to validate the SNPs, because the SNPs with negative VIM values in one population could have positive VIM values in another population.

For these 8,195 SNP$_{negvim}$, the average MAF was 0.21 (with the range 0.01-0.50). We also checked the allele substitution effects from the previous GWAS study on this population (Porto-Neto *et al.* 2014) and found that these SNPs distributed along the whole genome, whereby 4,143 had positive effects and the remaining 4,052 had negative effects. However, the genotypes of these SNP$_{negvim}$ were in fact imputed from an initial low-density panel of cattle 60k. These may reflect the quality of imputation.

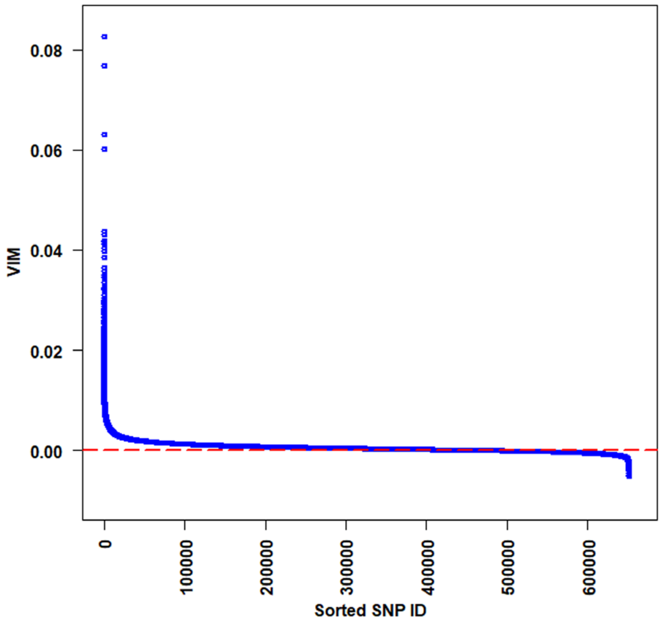

Figure 1. Distribution of average variable importance measures of ranked SNPs

Table 1. Average estimates of variance components and genomic prediction accuracy for different subsets of SNPs

Marker	Additive Model			
	h_a^2	σ_a^2	σ_p^2	†ACC
RF1,000	0.26±0.03	171.6±25.0	658.8±26.1	0.47
RF5,000	0.39±0.04	254.9±32.7	658.2±26.3	0.53
RF10,000	0.42±0.04	278.5±35.2	659.1±26.4	0.55
RF50,000	0.45±0.04	299.0±38.7	669.2±26.7	0.58
Even1,000	0.18±0.03	124.1±22.2	682.8±25.2	0.28
Even5,000	0.30±0.04	218.9±32.2	680.0±26.0	0.47
Even10,000	0.36±0.04	245.4±35.2	681.3±26.3	0.47
Even50,000	0.40±0.04	275.9±38.7	681.4±26.3	0.48
§643,058	0.41±0.05	281.4±39.4	679.5± 26.7	0.59
All SNPs (651,253)	0.41±0.05	281.0±39.6	679.6±26.7	0.55

§ All SNPs without 8,195 VIM negative SNPs; † Accuracy of genomic prediction

Genomic prediction accuracy with and without the negative VIM SNPs. Table 1 presents the estimates of variance components and the genomic prediction accuracies from using different sources of SNPs. In comparison to the accuracy results from using the whole panel (All SNPs, last row in Table 1, ACC = 0.55), the top SNPs from the RF (i.e. RF5,000 and RF10,000) showed very similar or higher (RF50,000) genomic prediction accuracy values. They significantly outperformed the same-size SNPs randomly selected but evenly distributed along the genome (Even-). Interestingly, after removing 8,195 SNP$_{\text{negvim}}$, the genomic prediction with the remaining 643,058 SNPs (Table 1) resulted in an improved accuracy value (0.59) compared to the whole panel (0.55). This value was similar to that of using RF50,000. In addition, we discovered that all the evenly distributed SNP datasets contained about 20% SNP$_{\text{negvim}}$. These results suggest that including SNP$_{\text{negvim}}$ in the whole panel would have caused the reduction in accuracy estimates.

Gene Enrichment Analysis. When comparing the biological functions of the genes near 8,195 SNP$_{\text{negvim}}$ with those of RF5,000 or Even5,000, there was no significant enrichment found for 8,195 SNP$_{\text{negvim,}}$ nor for Even5,000. However, for RF5,000, were enriched for "RNA polymerase II core promoter sequence-specific DNA binding", consisting of several transcription factors such as EGRF1, GATA3, GATA6, NFIL3, PAX6, PAX8 or SOX11. The latter, renowned for its role in embryonic development and determination of cell fate (Jiang *et al*. 2013). Finally, at the functional level, RF 5,000 showed significant enrichment for experimental promoters and muscle regulatory regions.

CONCLUSIONS

In low commodity livestock or aquaculture species, a common practice in applying genomic selection is to genotype parents with a high-density SNP panel, genotype young progeny with a low-density panel and then impute the low-density panel to the high-density panel for genomic prediction. This study demonstrates that it is important to identify and remove the problematic SNPs (with negative VIM values) that increase the error variance and decrease accuracy of genomic prediction. The machine learning method – Random Forest has merit in use as a pre-screening tool for i) identifying problematic SNPs; and ii) identifying subsets of SNPs that have biological functions for low-density panels.

REFERENCES

Breiman L. (2001) *Machine Learning*. **45**: 5.

Chen X. and Ishwaran H. (2012) *Genomics*. **99**: 323.

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerre D, et al. (2018) *bioRxiv* 316091.

HinrichsA .S., Karolchik D., Baertsch R., Barber G.P,. Bejerano G., Clawson H., Diekhans M., Furey T.S., Harte R.A., Hsu F., Hillman-Jackson J., Kuhn R.M., Pedersen J.S., Pohl A., Raney B.J., Rosenbloom K.R., et al. (2006) *Nucleic Acids Res.* **34**: D590.

Jiang Y., Qing Q., Xie X., Libby R.T., Lefebvre V. and Gan L. (2013*). J Biol Chem*. **288**: 18429.

Li B., Zhang N., Wang Y.-G., George A., Reverter A. and Li Y. (2018) *Front. Genet.* ***9****:237.*

Misztal I., Tsuruta S., Strabel T., Auvray B., Druet T., and Lee D.H. (2002) *Proc. 7th World Congr. Genet. Appl. Livest. Prod.* Communication N° 28-07.

McLean C.Y., Bristor D., Hiller, M., Clarke S.L., Schaar B.T., Lowe C.B., Wenger A.M., and Bejerano G. (2016) *Nat. Biotechnol*. **28**:495.

Porto-Neto L.R., Reverter A., Prayaga K.C., Chan E.K.F., Johnston D.J, Hawken R.J., Fordyce G, Garcia J.F., Sonstegard T.S., Bolormaa S., Goddard M.E., Burrow H.M., Henshall J.M., Lehnert S.A. and Barendse W. (2014). *PLOS One*. **9**: e113284.

Sheffield N.C. and Bock, C. (2016) *Bioinformatics* **32:** 587.

VanRaden P.M. (2008) *J Dairy Sci*. **91**: 4414.

Villar D., Berthelot C., Aldridge S., Rayner T.F., Lukk M., Pignatelli M., Park T.J., Deaville R., Erichsen J.T., Jasinska A.J., Turner J.M., Bertelsen M.F., Murchison E.P., Flicek P., Odom D.T. (2015). *Cell* **160**: 554.