USING RANDOM FORESTS TO IDENTIFY SNP ASSOCIATED WITH LEG DEFECT IN BROILER CHICKEN: IMPACT OF CORRECTING FOR POPULATION STRUCTURES

Y. Li[1], A. George[2], R. Hawken[3], R. Sapp[3], S. Lehnert[1], A. Reverter[1]

[1]CSIRO Agriculture Flagship, St Lucia, QLD 4067, Australia
[2]CSIRO Digital Productivity and Services Flagship, Dutton Park, QLD 4102, Australia
[3]Cobb-Vantress Inc., Siloam Springs, Arkansas 72761-1030, USA

SUMMARY

The machine learning method, Random Forests (RF) has been shown to be effective in genome-wide association studies (GWAS). However, the presence of population structure (PS), e.g. relatedness between individuals, may cause spurious results in a RF analysis. In this study, we examined the impact of correcting for PS on the RF analysis of leg defect data from a commercial poultry population of 826 chickens genotyped for 44,129 SNP (single nucleotide polymorphism) markers. The results show that correcting for PS led to: 1) a significant improvement in the estimates of SNP variable importance values; 2) a significant reduction in false positives identified in the uncorrected data; and 3) a stronger evidence for a set of SNPs associated with the defective phenotype.

INTRODUCTION

One of the challenges of GWAS is that the number of predictors is larger than the number of samples, the so called "large p, small n" problem. During the past decades, a number of parametric statistical approaches have been developed for dealing with this issue, for example: Least Absolute Shrinkage and Selection Operator (LASSO) (Wu *et al.* 2009) and two-step Bayesian variable selection method (Zhang *et al.* 2008). Recently non-parametric machine learning methods have been shown to be efficient in analysing large genomic data (Szymazak *et al*. 2009). One of these methods is Random Forests (RF, Breiman 2001; Chen and Ishwaran 2012), a nonparametric decision tree based ensemble method for classification or regression of multiple predictor variables. Our initial preliminary examination found that this method is a powerful tool in pre-screening candidate genes in GWAS of sheep and cattle datasets (Li *et al*. 2014). Despite the advantage of RF over single marker GWAS methods in accounting for correlations among SNP variables, the existence of population structure (PS) has been shown to cause spurious results in the RF analyses (Zhao *et al.* 2012). In this study, we used a dataset from a commercial poultry population to examine the impact of correcting for PS on the RF analysis of a binary trait – leg defect.

MATERIAL AND METHODS

Data. A total of 826 broiler males from a commercial line of Cobb-Vantress Inc. was genotyped for 51,713 SNPs. The dataset comprised animals from 22 generations with various proportion of animals that had leg related problems, ranging from 29% to 51%. After quality check 7,584 SNPs were removed from the genotype dataset and the remaining 44,129 SNPs were used for the RF analyses. The original recording of an animal's phenotypic leg status was either normal, bowed out, bowed in or rotated. We generated a new binary trait, by merging the latter three categories into a single category "Leg Defect'. Of the 826 animals, 592 were normal (coded "0") and 234 had leg defects (coded "1") (Table 1).

EIGENSTRAT analysis for extracting population structure (PS) information. Unlike a linear model that can accommodate PS by fitting a covariance matrix in the model based on pedigree or genomic relationships, RF as a permutation-based method cannot directly account for such factors. Therefore, prior to a RF analysis, it is necessary to identify and correct any existing

population stratification. In this study we applied a method similar to that used by Zhao *et al*. (2012) to correct for PS. An EIGENSTRAT analysis (Price *et al.* 2006) was initially conducted to extract all eigenvectors from the SNP data. The linear regression models were fitted to regress the first 10 axes of variation (principal components) on: a) individual SNP genotypes, and b) the phenotypic trait values, respectively. The residuals from these analyses were combined for the RF analyses. All analyses were performed using the R program (version 3.1.1, http://www.r-project.org/).

Table 1. Trait distribution of leg related defect attributes in 826 roosters.

Trait	Number	0 (Normal)	1 (Defect)
Bowed Out	826	729	97
Bowed In	826	786	40
Rotated Leg	826	729	97
Leg Defect	826	592	234

Random forests (RF). Details of the RF methodology can be found in Breiman (2001). In brief, six steps are involved: 1) As the training dataset, select a random subsample of 550 individuals (or two thirds) with replacement from the available 826 individuals; 2) Select a random subset of SNPs (parameter mtry; say 420 out of the original 44,129) to form a decision tree; 3) Create a single tree via partitioning of sampled individuals in the subsample (normal *versus* defect) with SNP genotypes (e.g. "AA" *versus* others); and with the order (or arrangement) of SNP in the tree run repeatedly until individuals are perfectly partitioned into normal and defect; 4) Test the tree created in Step 3 with the remaining 276 individuals (i.e. validation) to determine the prediction error rate of the SNP tree; 5) Repeat Steps 1 to 4 to develop a large number of forest trees (parameter Ntree); 6) Compute SNP variable importance value (VIM) by averaging the prediction error values across all forest trees. For a continuous phenotype (e.g. corrected data), Step 3 will build a tree that splits the sampled individuals into subsamples with different data value ranges. Step 4 will calculate the minimized sum of squared error for each SNP. It is worth noting that in a RF analysis, a SNP prediction error value is estimated when the SNP is randomly permuted, i.e. excluded from the forest trees. Therefore, the higher the VIM value, the more important the SNP is.

Two crucial parameters impact the outcome of a RF analysis, i.e., the size of forest trees (Ntree) and the number of markers at each sampling event (mtry). To determine the minimum requirement for these parameters, we examined a range of Ntree and mtry values. These included Ntree = 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, or 2000, and mtry = 1, $\sqrt{p}$, $2 * \sqrt{p}$ or 0.1*p, where p is total number of SNP markers (44,129). Once the minimum parameters were determined, these values were used to run the final RF analyses comprised of 100 RF replicates. To demonstrate the effect of correcting PS on the analysis, we compared the RF results from the corrected data with the uncorrected data. The R program randomForest was used (version 3.1.1).

RESULTS AND DISCUSSION

RF parameter determination. The average SNP VIM values for different parameter combinations of Ntree and mtry are shown in Figure 1. Note that in the context of RF analyses, a high value for VIM is favourable. For both uncorrected and corrected data, the average VIM reached a stable status with Ntree ≥ 1,000. This suggests that the RF analysis with Ntree ≥ 1,000 should produce reasonably accurate VIM values. Among the four parameters tested for mtry, single marker analysis (mtry =1) gave the lowest estimates for VIM, while the other three parameters ($\sqrt{p}$, $2 * \sqrt{p}$ and 0.1*p) produced very similar values.

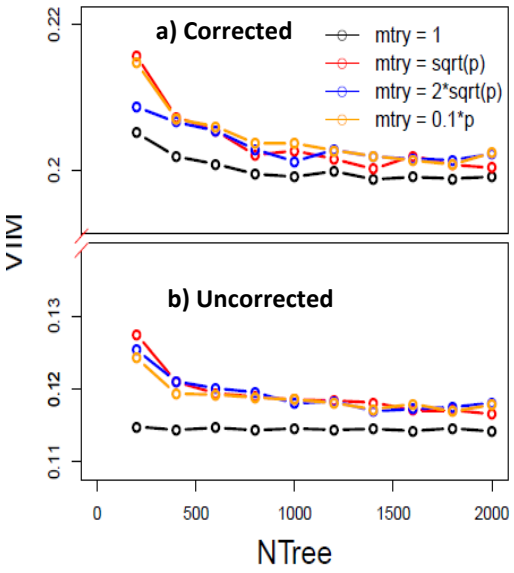

Figure 1. Comparison of the mean VIM values from different combinations of parameters with (top panel) and without (bottom panel) correction for population structure.

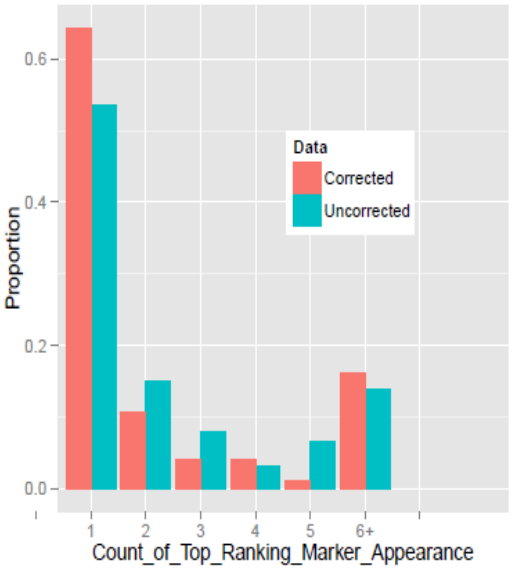

Figure 2. The proportions of top 5 marker appearances in 100 RF replicates with (red bars) and without (blue bars) correction for population structure.

RF analyses in the corrected and uncorrected datasets. Compared to the uncorrected data (Figure 1b), correcting for PS (Figure 1a) clearly resulted in a significant increment in the estimated average VIM values (from 0.120 to 0.205). When investigating the top 5 ranking markers from each of the 100 RF replicates, a total of 166 and 179 markers were found in the uncorrected and corrected data, respectively. The compositions of these marker incidences in both datasets are shown in Figure 2. The top markers appearing only once in 100 replicates had the highest proportion (54% in the uncorrected versus 64% in the corrected data). The uncorrected data tended to have fewer markers (13.85%) with the highest incidence (i.e. captured in 6+ replicates) than the corrected data (16.20%). However, the intriguing results were found when comparing the distributions of the top 5 marker incidences across the whole genome in both datasets (Figure 3). It is very clear that correcting for PS led to a reduction in top ranking SNP incidence in a number of genome regions found to be significant in the leg defect analysis of the uncorrected dataset. The majority SNPs identified in the uncorrected data were no longer in the top ranking markers in the corrected data. Among 166 (uncorrected data) and 179 (corrected data) markers, there were 26 in common (shown by overlapping regions in Figure 3) and 11 of them had a reduced incidence in the corrected data. In contrast, there was a set of 12 common markers closely linked (near the right hand side of the genome), after correcting for PS, the association signal became much stronger.

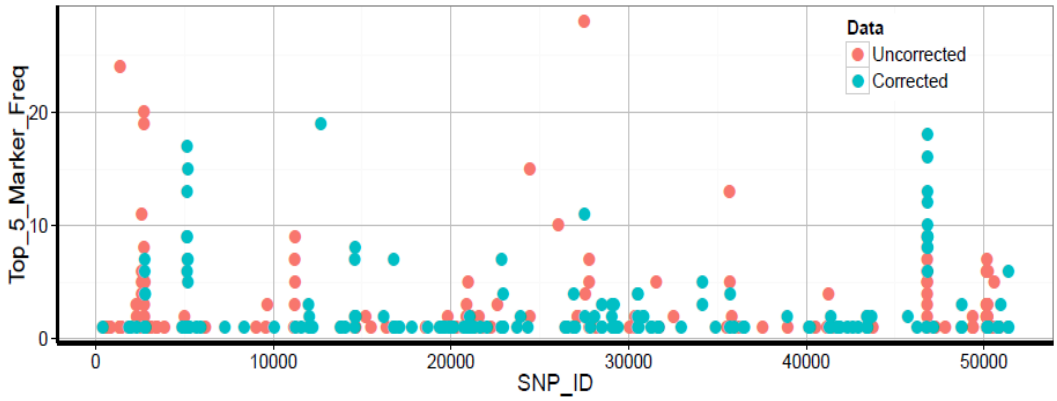

Figure 3. Distribution of top 5 ranking marker incidences across genome for corrected (blue) and uncorrected data (red).

Population stratification or admixture is known to cause different allele or genotype frequencies in subpopulations and that in turn can lead to detection of spurious SNP associations in GWAS (Zhao *et al*. 2012). RF has its advantage over single marker GWAS methods in handling high dimensional genomic data (Chen and Ishwaran 2012), but it has a limited power in dealing with a confounding effect of PS on both genotypes and a phenotype. The results here demonstrate the importance of correcting for population structure prior to RF analysis to minimize false positives. Since the "true" SNPs are unknown, these results are of very limited use for the purpose of method validation. There is a need in future to conduct a systematic evaluation of the method with large simulation datasets.

CONCLUSIONS

Correcting for population structure prior to RF analysis can improve the accuracy of SNP variable importance values and avoid spurious association results. Since RF is a non-parametric permutation based method, a large number of RF replicates is required to get reliable inference of the markers associated with a phenotype.

REFERENECES

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

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

Li Y., Kijas J., Henshall J., Lehnert S.A., McCulloch M. and Reverter A. (2014) 10th WCGALP.

Price A.L, Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A. and Reich D. (2006). *Nature Genet.* **38:** 904.

Szymazak S., Biernacka J.M., Cordell H.J., Gonzalez-Recio O., Konig I.R., Zhang H. and Sun Y. (2009) *Genet. Epidemiol.* **33** (Supplement 1): S51.

Wu T.T., Chen Y.F., Hastie T., Sobel E. and Lange K. (2009) *Bioinformatics*. **25**: 714.

Zhang M., Zhang D. and Wells M.T. (2008) *BMC Bioinformatics*. **9**: 251.

Zhao Y., Chen F., Zhai R., Lin X., Wang Z., Su L. and Charistiani D.C. (2012) *Int. J. Epidemiol*. **41**: 1798.