Using a Random Forest proximity measure for variable importance stratification in genotypic data

Jose A. Seoane[1], Ian N.M. Day[1], Colin Campbell[2], Juan P. Casas[3], Tom R. Gaunt[1,4]

[1]Bristol Genetic Epidemiology Laboratories
[2]Intelligent Systems Laboratories
[4]MRC Integrative Epidemiology Unit.
University of Bristol, Bristol, United Kingdom
[3]Department of Non-communicable Disease Epidemiology
London School of Hygiene and Tropical Medicine, United Kindom
{j.seoane,ian.day,c.campbell,tom.gaunt}@bristol.ac.uk
juan-p.casas@lshtm.ac.uk

Abstract. In this work we study variable-significance in classification using the Random Forest proximity matrix and local Importance matrix. We use the proximity matrix to group the samples across a number of clusters and use these clusters to stratify the importance of a variable. We apply this approach to a cardiovascular genotype dataset for sample classification based on coronary heart disease and we found a number of variations related with cardiovascular disease phenotypes. We also used a set of phenotypes related with this genotype data to match the obtained clusters with coronary heart diseases phenotypes.

Keywords: Random Forest, Proximity Measure, Feature Importance, Genetic Data Analysis, Machine Learning

1 Introduction

According to the World Health Organization (WHO), cardiovascular diseases (CVD) are globally the most significant cause of death [1]. CVD risk can be predicted based in a set of factors (age, sex, obesity, diabetes, blood pressure, etc.) [2]. However, these risk factors do not fully explain these diseases. Using genetic association studies, some genetic biomarkers have been identified as risk factors of cardiovascular diseases [3]. The identification of these genetic risk factors could help to explain the aetiology of CVD and could also be used as targets for new drugs.

In order to find these genetic biomarkers, single association tests are usually performed. However, some machine learning techniques have also been used in the last few years for the same purpose. Some of the most promising results have been obtained using techniques which can deal with high dimensional datasets, such as Support Vector Machines [4] or Random Forest [5, 6].

Random Forest (RF) was proposed by Breiman [7], and has been used in the last years in genotype analysis, because of its good performance with high dimensional datasets. A good review of the use of Random Forest methods in the life sciences can be found in [8]. One of the important features of Random Forest is the possibility of obtaining a set of descriptive measures, in addition to the classification model. These descriptive measures include the proximity matrix and the local importance matrix. Once Random Forest has been trained, the proximity matrix quantifies sample-similarity. The proximity between two samples is calculated by measuring the number of times that these two samples are placed in the same terminal node of the same tree of RF, divided by the number of trees in the forest. On the other hand, the local importance matrix measures the importance of each feature in each sample. This local importance matrix is calculated via two methods. Firstly via the Gini importance which measures the total increase of impurity in a variable, when this variable has been selected for splitting. Secondly, via the permutation importance, which is the mean decrease in accuracy for a predictor variable, calculated as the percentage of correct votes for the correct class in the permuted Out of the Bag (OOB, internal validation dataset) samples subtracted from the percentage of correct votes in the original OOB samples.

In this paper we use these two characteristics of Random Forest (local importance and proximity values) to obtain a descriptive model which indicates which features are most important for classification. We use the proximity matrix in order to obtain a set of clusters, identify the most important features of each cluster, using the local importance matrix, and map these clusters to a phenotype using principal component analysis. Some authors [9] [10] have applied RF for cluster discovery in the life sciences where the Random Forest was trained without independent variables. However, we propose a semi-supervised approach, which detect clusters that arise in a classification of genetic data in Coronary Heart Disease (CHD) and control data.

2 Material and Methods

2.1 Data

We used data from the British Women's Heart and Health Study (BWHHS) cohort study, composed of 4286 health women aged 60-79 years at baseline [11]. A range of baseline data (blood samples, anthropometry, health/medical history, echocardiographic measures, etc.) were collected between 1999 and 2001, and DNA extracted from 3884 participants. Genotyping was performed using the Illumina HumanCVD BeadArray (Illumina Inc, San Diego, USA), which comprises nearly 50,000 SNPs in over 2,000 genes selected on the basis of cardiovascular candidacy [12]. Samples with a genotype call rate <90%, Hardy Weinberg disequilibrium < 1 E-7 and minor allele frequency < 1 % were excluded from the analysis. The different phenotypes used in this study consisted of 11 directed and derived electrocardiogram (ECG) measures, obtained as described in [13], 63 blood measures, 2 blood pressure readings, 3 anthropometric measures and an indicator of whether a patient has suffered coronary heart disease (CHD). These data were measured as described in [11]. Phenotypic data were transformed to a logarithmic scale. Both scaled and unscaled phenotypic data were normalized to zero mean and unit variance.

In order to remove Linkage Disequilibrium effects in the genotypic data, SNPs with correlation higher than 0.75 were removed from the study. We also removed all SNPs with no association with CHD, for which we consider as threshold an optimistic association p-value of 0.1. We used linear regression between the genotype and CHD variable in P-link [14] in order to get these association p-values. A total of 1854 SNPs and 128 phenotypes from 3428 samples were used in this study.

2.2 Study design

A Random Forest implementation in R (http://cran.r-project.org/), "randomForest" was used in this study. The objective is to identify different clusters of cardiovascular genotype in the original data. In order to obtain this, a Random Forest model was built, with genotype data as input and a measure indicating if the patient has suffered coronary heart diseases, as a binary outcome.

This binary classification dataset is highly imbalanced. RF does not achieve best performance with imbalanced datasets when the classification accuracy is biased to the bigger class. In order to remove this bias, some strategies have been proposed [15, 16]. We followed a subsampling approach; where for each built tree, the algorithm selects all samples of lower class, and randomly selects the same number of samples of the bigger class.

Following Breiman, a defined main parameter for RF is the number of features which each new tree selects randomly for classification, given as mtry in "randomForest" package. A cross-validation scheme was used to obtain this parameter. We finally selected an mtry of 45.

We also ran several experiments to select the proper number of trees for the RF. Graphically we observed that the results doesn't improve significantly when the number of trees was bigger than ~2500, so we selected a conservative number set of 3000 trees.

The literature on Random Forest covering the proximity and local importance matrix suggests that there could be some variation in both matrices because of the stochastic nature of the RF model, which randomly selects a set of features to build each tree. These authors [8, 17] recommend running several models and taking some average or proportion of the values from each model. Following these recommendations, we ran 20 RF models and obtained the median of both matrices.

As we mention in the Introduction, Breiman's Random Forest implementation provides two types of importance measures. However, the Gini Variable Importance index showed artificial inflation of feature importance depending on the number of categories for each variable [18], repeated also in genetic data depending on the minor allele frequency of each SNP[19]. So we will use the permutation index both for global and local variable importance. In order to normalize the variable importance, we divide it by the standard error of each variable, obtained from the parameter "importanceSD" from the model.

We built an aggregation of all proximity matrices calculating the median of all the proximity matrices obtained from the 20 RF models. In order to understand this proximity matrix, we built a distance matrix, as 1-proximity measure, for all proximity pairs. Then, we performed a multidimensional scaling [20] on this distance matrix using the R function "cmdscale" from the "stats" R package in order to obtain the three principal coordinate components that was plotted using the cloud plot of R package "lattice".

In order to cluster the samples using these principal coordinate components, we used hierarchical clustering (hclust from R package "stats") using the Euclidean distance between the sample components.

The local importance matrix stores the permutation importance of each feature in each sample classification. In order to obtain the most important features for each cluster, we obtained the mean of the variable importance samples in each cluster.

With the groups and the phenotype data we performed PCA analysis using the R package "FactoMineR" [21] in order to obtain the most important phenotypes for each cluster. Using this package we can obtain a graphical representation of samples grouped by clusters and also a graphical representation of each phenotype. In order to obtain the different distributions of phenotypes between two clusters we use the Kolmogorov-Smirnov test, provided in R package "stats".

3 Results

The Random Forest algorithm used around a third of the samples, named Out of the Bag (OOB) samples, for validation purposes. The algorithm trains with the other two-thirds of the samples and uses OOB samples to calculate error, proximity and local importance values. To ensure the capability of this particular scheme and dataset for classification, we ran a 10-fold cross validation study. In this study, we randomly separated the samples into ten folds, ensuring that all folds have the same proportion of positive samples. Then a RF model was trained with 9 of these 10 folds (using two-thirds of samples of these 9 folds for training and the rest for OOB validation) and the last fold was used for calculate the performance of the classifier. This procedure was repeated 10 times so that each fold was using the validation fold once. This 10-fold cross validation process was repeated 10 times to remove fold sample selection bias. We used a t-test to check the significance of these results. Using this validation scheme, the model reached a test error of 0.3198 ±0.0095 in test folds with a confidence interval of (0.2948-0.3448) at 95% confidence level. This result shows that using only genetic data, the RF model has predictive capability.

However, the main objective of this paper was not to probe the prediction capacity of the RF, but the other important measures of the RF model are useful for model interpretation, namely the proximity matrix and the importance values. Unfortunately, the cross validation scheme cannot be used for this purpose. Although the proximity matrix can be calculated from the final training model using a different dataset (using the prediction function of the "randomForest" package), the local importance matrix needs to be calculated using only the OOB samples. This is the reason why all samples were used to build the final models. As pointed in the Section 2, 20 RF models were built to remove the possible deviation in the importance and proximity matrix. The mean classification error results on OOB samples is 0.273 ± 0.0068, (0.2705-

0.2769) at 95%. This result improves the cross validation error, which shows some overtraining by disregarding the cross validation.

As the next step we studied the importance of the variables for all the 20 RF models. Each model has slight differences in the variable importance, so we take the mean of all of them to get a combined value.

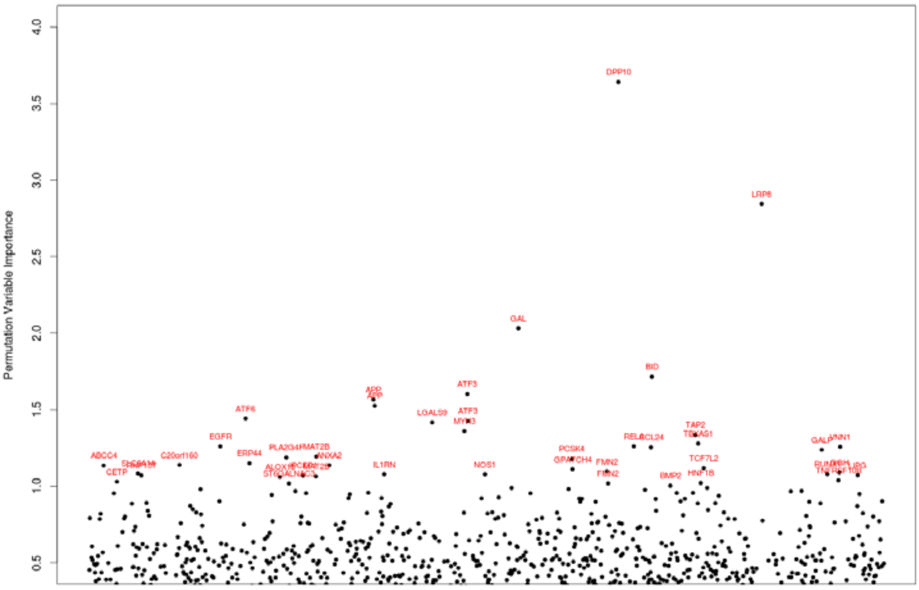

Fig. 1. Permutation based Importance measure for all variations. Names of the genes related with variations with importance greater than one is plotted in red.

In Figure 1 we plot the most important features for classification using the permutation variable importance approach. There are two main important features, SNPs rs1375144 and rs17108202, in genes DPP10 and LRP8 respectively, with an importance around 3 or bigger. The Figure also gives the names of the genes with an importance bigger than 1.

During the next stage of the study we calculated the aggregate proximity matrix of all 20 RF models, taking the median of these values. The result is a 3428 x 3428 symmetric matrix representing the similarity of each pair of two samples. In order to visualize the proximity matrix, the randomForest package uses the function MDSplot. Following the same scheme, we calculated the distance of the proximity matrix and the multidimensional scaling values of this matrix.

The result is shown in Figure 2 as a 3D plot of the distances. Each point represents one sample, and the higher the values in the proximity matrix, the closer are these points in the plot. This plot shows 8 clear clusters, which corresponds to the 8 most important groups in terms of the proximity matrix.

We used a hierarchical clustering algorithm of the components of the multidimensional scaling to obtain the exact classification of each sample.

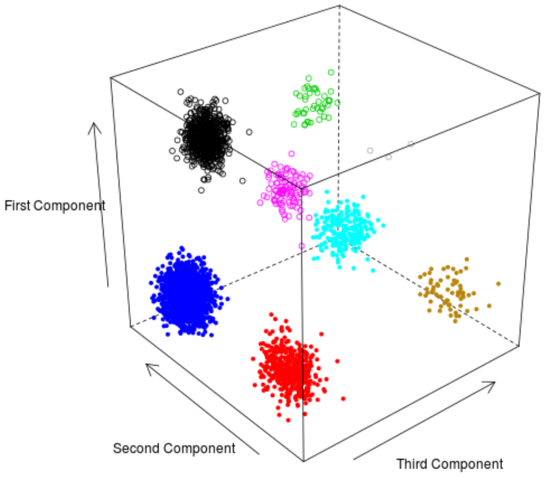

Fig. 2. 3D scatter plot of the coordinate components of the multidimensional scaling of the proximity matrix.

Figure 2 also shows how clusters are correlated with the categories of samples with CHD and controls, upper clusters correspond to CHD samples, while the bottom clusters correspond to controls. That means that the first distance component of the proximity matrix is able to distinguish control from disease. This perfect classification is due to the use of all samples (not only OOB samples) for obtaining the proximity matrix. Using only OOB samples we obtained similar clusters, but some misclassification in terms of CHD/control.

In the next step we used the local importance matrix to discover which features are most important in each of the clusters.

In Table 1 we show the most important features for each cluster, as the mean of the local importance matrix of each feature in the samples belonging to each cluster.

Table 1. 15 most important variants and variant importance score for each cluster

Cluster 1		**Cluster 2**		**Cluster 3**		**Cluster 4**	
DPP10	35.42	LRP8	139.11	LRP8	34.93	ATF6	23.54
LRP8	30.86	ATF6	27.18	ATF6	32.86	GAL	23.43
ATF6	24.91	CCL24	24.77	IL1RN	29.36	CCL24	22.97
ATF3	23.81	MAT2B	22.72	LIPG	28.74	IL1RN	22.92
IL1RN	23.11	IL1RN	22.55	CAV3	27.45	ATF3	22.41
PLA2G4F	22.47	PLA2G4F	21.94	C1QA	26.96	ATF3	22.22
MAT2B	22.23	GCK	21.44	MAT2B	25.99	MAT2B	22.13
CCL24	22.16	APP	21.17	PRKAG2	25.86	BID	22.02
LIPG	20.17	GAL	21.09	ERP44	25.66	APP	21.97
LGR5	19.68	ATF3	20.39	ELP4	24.95	LGR5	21.10

MAT2B	18.76	MYH3	19.78	MYOCD	24.09	PLA2G4F	20.15
Cluster 5		**Cluster 6**		**Cluster 7**		**Cluster 8**	
DPP10	299.96	DPP10	33.57	GPR98	65.16	DPP10	301.66
GAL	26.87	CCL24	26.48	IL1F10	64.50	LRP8	141.30
MAT2B	25.59	ATF6	26.35	LRRFIP1	64.07	ATF6	29.60
ATF6	23.43	MAT2B	25.74	MAT2B	62.35	LGALS9	26.28
TNFRSF11B	23.38	GPR98	23.00	GPR98	61.95	C20orf160	25.89
MYH3	22.77	LGR5	23.00	EGFR	58.74	GAL	24.83
C20orf160	22.20	TAP2	22.99	OSBP2	54.86	PCSK4	23.90
BID	22.19	CAV3	22.56	CDKAL1	54.18	CSF1	23.72
CCL24	21.36	SOD2	21.62	HAT1	52.90	IPCEF1	23.63
ATF3	21.35	ATF3	21.49	RUNX1	49.57	IL1RN	23.58
IL1B	21.10	ACAN	20.97	SORCS1	49.48	MAT2B	22.88

This Table shows that the most important variables rs1375144 and rs17108202 (those related with genes DPP10 and LRP8) represents the 2nd and 3th distance component in the figure 2. Clusters black and gray present higher importance of DPP10 and LRP8, clusters green and red higher importance of LRP8 and clusters cyan and magenta higher importance of DPP10. Analyzing this in terms of genotype values of these two SNPs, clusters black and blue are defined by the wild allele of both SNPs, clusters yellow and grey are defined by a rare variant of both SNPs, cluster red and magenta are defined by a rare allele of rs17108202 and wild allele of rs1375144 and clusters cyan and green are defined by a rare allele of rs1375144 and a wild allele of rs17108202.

In order to find which variants distinguish between the same DPP10 and LRP8 patterns in the two cluster pairs, we analyzed the significant differences between the importances of these variants in each pair of clusters.

For the black and blue cluster pair, genes DPP10 (wild), LRP8(wild), GAL, APP, GCK, ATF3 and SOAT2 are related with CHD and GPR98, MAT2B, ACAN and LIPG with controls.

For the red and magenta cluster pair, genes DPP10, GAL, SOD2, APP and GCK are related with CHD, while LRP8(rare), GPR98, LIPG and S100A9 with controls.

For the green and cyan clusters, genes LRP8, GCK, APP, GAL, BID and ABCG1 are related with CHD and DPP10 (rare), GPR98, PDE1A and NPHS2 with controls.

In the case of clusters grey and yellow, we cannot obtain any reliable information because the yellow-labelled cluster has only 3 samples.

In the final stage of this study we compared the phenotypes of each cluster. In order to do this, we performed a principal component analysis of phenotypes.

In Figure 3 we show the principal components of the phenotypes in the left and the principal components of the samples in the right. We have removed the samples and plot only the centre of the samples of each cluster, representing the centroid of the class.

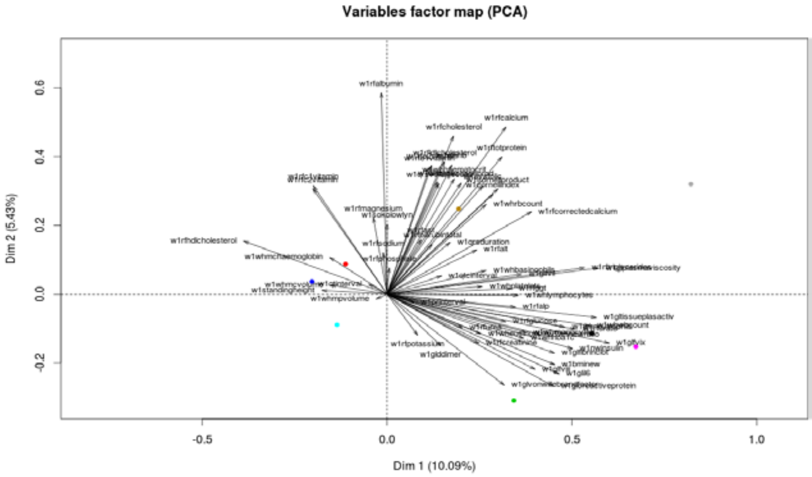

Fig. 3. Variables factor map of the phenotypes and groups.

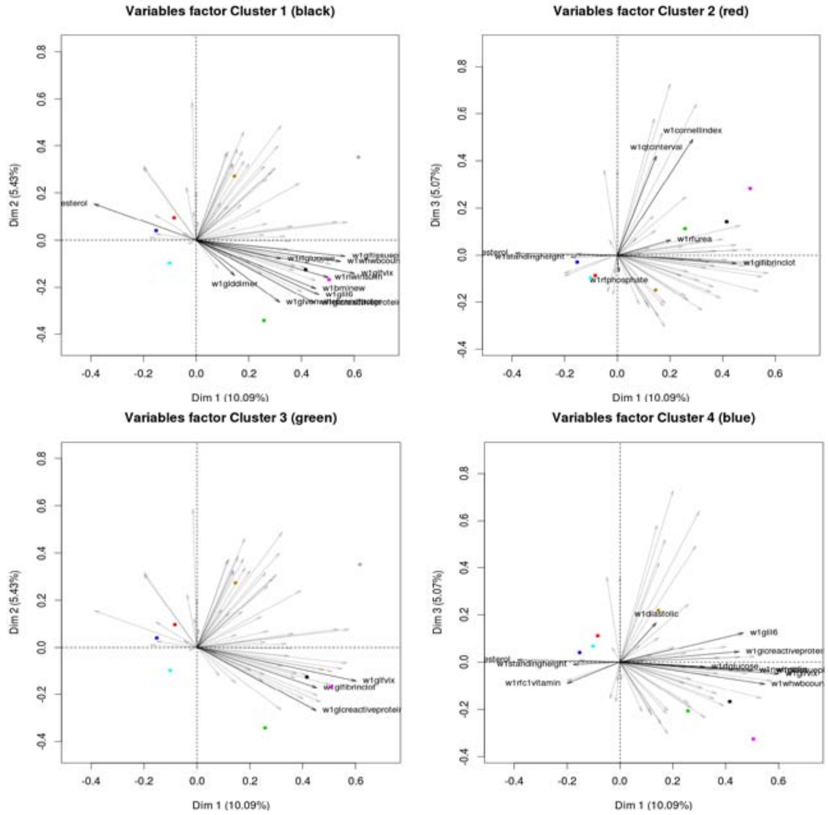

Fig. 4. Variables factor map with significant phenotypes for cluster 1 (panel top left), cluster 2 (top right), cluster 3 (bottom left) and cluster 4 (bottom right)

We can observe how the groups at the bottom right correspond with CHD groups and the groups in the upper left correspond with control. In terms of phenotypes, controls correlates with HDL cholesterol and CHD with phenotypes such as Von Willebrand Factor, Factor IX, white blood cell count, BMI, etc.

In Figures 4 and 5 we show only the phenotypes for each cluster which are different statistically from the others using Kolmogorov-Smirnov test.

Figure 4 shows that cluster 1 (black) is positively correlated with white blood count, tissue plasma activator, Von Willebrand Factor, Factor IX, BMI, glucose, insulin, D-Dimer, C-reactive protein and IL6, and negatively correlated with HDL cholesterol. Cluster 2 (red) is positively correlated with HDL cholesterol and height; and negative correlated with phosphate, urea, fibrin clot, QTC interval and Cornell index. Cluster 3 (green) is correlated with C-reactive protein, Factor IX and fibrin clot and Cluster 4 (blue) is correlated positively with HDL cholesterol, C vitamin and height; and negatively with diastolic blood pressure, white blood count, glucose, insulin, Factor IX, tissue plasma activator IL6 and C-reactive protein.

Figure 5 shows how cluster 5 (cyan) is positively correlated with mean cell volume, phosphate, height; and negatively with LDL cholesterol, QT interval, QTC interval, Cornell index and QRS voltage product. Cluster 6 (magenta) is positively correlated with white blood count, triglycerides, insulin, fibrin clot, IL6, BMI, QTC interval, QRS duration, Cornell Index, Cornell Product, QRS voltage prod; and negatively correlated with HDL cholesterol, C vitamin and height. Cluster 7 (grey) is correlated with lymphocytes and Sokolowin Index. Note that there are only 3 samples, so these results are not significant. Finally, cluster 8 (yellow) is correlated with platelets, neutrophils, potassium, plasma viscosity and magnesium.

3.1 Discussion

The model we have derived for classifying CHD samples in genotypic data shows the importance of a set of SNPs, lead by SNPs rs1375144 and rs17108202 in genes DPP10 and LRP8. Variants in DPP10 have been related with HDL cholesterol and inflammation [22], blood pressure [23], stroke [24] and variants in LRP8 has been associated with premature myocardial infarction related to platelet activation [25, 26]and triglycerides [27]. These variants do not really distinguish between CHD and control, but seems to have some importance in the stratification for further classification, because of the different variable importance pattern in each cluster.

Further replication in other cohorts is needed to ensure the ability of this model to confirm that the most important variations can be reliably related to coronary heart diseases.

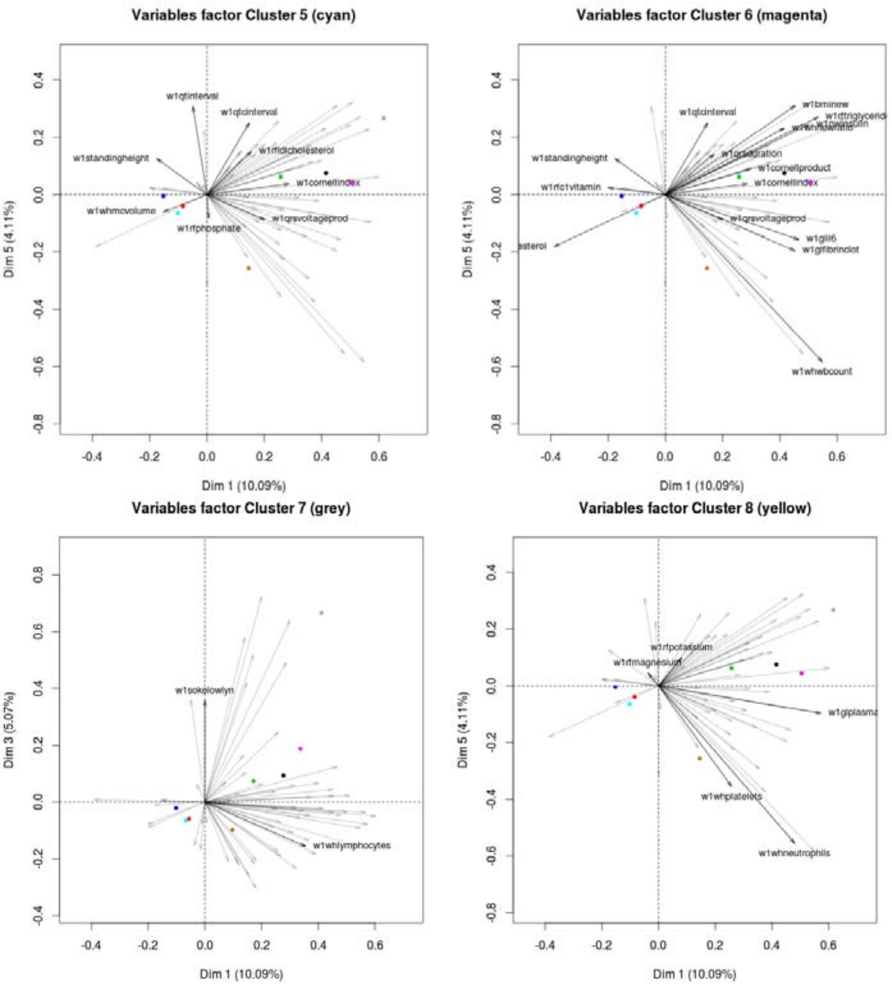

Fig. 5. Variables factor map with significant phenotypes for cluster 5 (panel top left), cluster 6 (top right), cluster 7 (bottom left) and cluster 8 (bottom right)

4 Conclusion

In this work we show that the combination of the Random Forest proximity matrix and local importance matrix could help to understand classification results using groups from variable importance stratification. We used the proximity matrix to cluster the samples into eight groups based in the distances between the samples. We prove, using the local importance matrix, that this clustering is lead by the two most important SNPs in two of the components, and a combination of other SNPs. Using phenotype data and PCA, we found the most important matching phenotypes for each cluster.

In future work, we will explore deeper levels of clustering in order to identify clear phenotype profiles and associate them with a genotype in addition to replicating the variable importance other cohorts.

Acknowledgments

This work was funded by the UK Medical Research Council (grant G1000427). The British Women's Heart and Health Study (BWHHS) has been supported by funding from the British Heart Foundation (BHF, PG/07/131/24254) and the UK Department of Health Policy Research Programme. The BWHHS Illumina HumanCVD BeadChip work is funded by the BHF (PG/07/131/24254, PI Tom Gaunt).

References

1. Mendis, S., Puska, P., Norrving, B.: Global atlas on cardiovascular disease prevention and control. World Health Organization (2011)
2. Folsom, A.R.: Classical and novel biomarkers for cardiovascular risk prediction in the United States. J Epidemiol 23, 158-162 (2013)
3. Kathiresan, S., Srivastava, D.: Genetics of human cardiovascular disease. Cell 148, 1242-1257 (2012)
4. Mittag, F., Büchel, F., Saad, M., Jahn, A., et al.: Use of support vector machines for disease risk prediction in genome-wide association studies: Concerns and opportunities. Human Mutation 33, 1708-1718 (2012)
5. Goldstein, B.A., Hubbard, A.E., Cutler, A., Barcellos, L.F.: An application of Random Forests to a genome-wide association dataset: Methodological considerations & new findings. BMC genetics 11, 49 (2010)
6. Goldstein, B.A., Polley, E.C., Briggs, F.: Random forests for genetic association studies. Stat Appl Genet Mol 10, (2011)
7. Breiman, L.: Random forests. Machine learning 45, 5-32 (2001)
8. Touw, W.G., Bayjanov, J.R., Overmars, L., Backus, L., et al.: Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle? Brief Bioinform 14, 315-326 (2013)
9. Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A., et al.: Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Modern Pathology 18, 547-557 (2004)
10. Seligson, D.B., Horvath, S., Shi, T., Yu, H., et al.: Global histone modification patterns predict risk of prostate cancer recurrence. Nature 435, 1262-1266 (2005)
11. Lawlor, D.A., Bedford, C., Taylor, M., Ebrahim, S.: Geographical variation in cardiovascular disease, risk factors, and their control in older women: British Women's Heart and Health Study. J Epidemiol Community Health 57, 134-140 (2003)
12. Keating, B.J., Tischfield, S., Murray, S.S., Bhangale, T., et al.: Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. PLoS One 3, e3583 (2008)

13. Gaunt, T.R., Shah, S., Nelson, C.P., Drenos, F., et al.: Integration of genetics into a systems model of electrocardiographic traits using HumanCVD BeadChip. Circ Cardiovasc Genet 5, 630-638 (2012)
14. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., et al.: PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81, 559-575 (2007)
15. Xu, B., Ye, Y., Wang, Q., Li, J., et al.: Random forest using tree selection method to classify unbalanced data. In: Conference Random forest using tree selection method to classify unbalanced data, pp. 83344F-83344F-83346. International Society for Optics and Photonics, (Year)
16. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. University of California, Berkeley (2004)
17. Bayjanov, J.R., Molenaar, D., Tzeneva, V., Siezen, R.J., et al.: PhenoLink-a web-tool for linking phenotype to~omics data for bacteria: application to gene-trait matching for Lactobacillus plantarum strains. BMC Genomics 13, 170 (2012)
18. Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T.: Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics 8, 25 (2007)
19. Boulesteix, A.-L., Bender, A., Bermejo, J.L., Strobl, C.: Random forest Gini importance favours SNPs with large minor allele frequency: impact, sources and recommendations. Briefings in Bioinformatics 13, 292-304 (2012)
20. Gower, J.C.: Some distance properties of latent root and vector methods used in multivariate analysis. Biometrika 53, 325-338 (1966)
21. Lê, S., Josse, J., Husson, F.: FactoMineR: An R package for multivariate analysis. Journal of statistical software 25, 1-18 (2008)
22. Sabatti, C., Service, S.K., Hartikainen, A.L., Pouta, A., et al.: Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat Genet 41, 35-46 (2009)
23. Levy, D., Larson, M.G., Benjamin, E.J., Newton-Cheh, C., et al.: Framingham Heart Study 100K Project: genome-wide associations for blood pressure and arterial stiffness. BMC Med Genet 8 Suppl 1, S3 (2007)
24. Ikram, M.A., Seshadri, S., Bis, J.C., Fornage, M., et al.: Genomewide association studies of stroke. N Engl J Med 360, 1718-1728 (2009)
25. Shen, G.Q., Li, L., Girelli, D., Seidelmann, S.B., et al.: An LRP8 variant is associated with familial and premature coronary artery disease and myocardial infarction. Am J Hum Genet 81, 780-791 (2007)
26. Shen, G.Q., Girelli, D., Li, L., Olivieri, O., et al.: Multi-allelic haplotype association identifies novel information different from single-SNP analysis: a new protective haplotype in the LRP8 gene is against familial and early-onset CAD and MI. Gene 521, 78-81 (2013)
27. Shen, G.Q., Li, L., Wang, Q.K.: Genetic variant R952Q in LRP8 is associated with increased plasma triglyceride levels in patients with early-onset CAD and MI. Ann Hum Genet 76, 193-199 (2012)