Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Yanlong Yu Date

Genotype prediction based on the gene expression data using Random Forest

By

Yanlong Yu
Degree to be awarded: Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Zhaohui "Steve" Qin, PhD
Committee Chair

_____

Yuan Liu, PhD
Committee Member

Genotype prediction based on the gene expression data using Random Forest

By

Yanlong Yu

Bachelor of management
Southwestern University of Finance and Economics
2018

Thesis Committee Chair: Zhaohui "Steve" Qin, PhD

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Abstract

Genotype prediction based on the gene expression data using Random Forest

By Yanlong Yu

Background: With rapid development of high throughput technologies, thousands of single nucleotide polymorphisms (SNPs) have been identified to be associated with human diseases. It's known that SNPs located in regulatory regions are often eQTLs thatcan modulate gene expression. Generally, gene expression can beaffected by SNP mutations. But since gene expression data is more easily to access than genotype data. We want to explore the relationship between genotype and gene expression and make prediction on SNP genotype based on the gene expression data.

Method: We used random forests as our model to test the classification and prediction problems. First, we first generated a simulated dataset based on the real data to test the strategy. We used out-of-bag (OOB) error rate as our metric to test the simulated data. We next tested hundreds of SNPs and got their AUC values for comparison. For SNPs achieve the highest AUC scores, we conducted a feature importance test.

Result: For the simulation data, the OOB estimate of error rate is 21%. For the real data, the mean AUC scores for the 917 SNPs is 0.559 (std=0.108) and the mean OOB scores is 0. 658 (std=0.056). The max AUC score is 0.933 and OOB score is 0.860. Most of the AUC scores are between 0.5 and 0.7, the OOB scores are between 0.6 to 0.7. We also located important features in SNPs with the highest AUC.

Conclusion: Through this study, we can see that for some SNPs, it is possible to use gene expression data to infer its genotype. However, the majority of the SNPs can not be predicted accurately. Also, we find some features that significantly influence the SNP prediction. Further study is needed.

Classification and prediction of SNPs based on the gene expression data using Random Forests

By

Yanlong Yu

Bachelor of management
Southwestern University of Finance and Economics
2018

Thesis Committee Chair: Zhaohui "Steve" Qin

A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2020

Introduction

Over the past few years, reseach on geneticvariability has increased substantially due to its potential relevance to the differential disease risk among people. One of the researches is genome wide association studies (GWASs). GWASs have identified thousands of single nucleotide polymorphisms (SNPs) are associated with human diseases and demonstrated that most of the variants are found in non-coding regions of the gene and thus are likely to be involved in gene regulations [1]. However, there're still many disease-associated SNPs remained to be found. This analysis of such vairants in the context of gene expression measured in tissues have led to a big field in human genetics studying expression quantitative trait loci (eQTLs). An eQTLs is a locus that can explain a portion of the genetic variance of a gene expression phenotype [2].

It's known that SNPs located in regulatory regions are often eQTLs since they can modulate gene expression [3, 4]. Some studies report an association between eQTLs and GWAS detected SNPs [5, 6]. The identification of a cis association of a SNP with gene expression level has been used to validate candidate genes for complex traits mapped to the same chromosomal locations [7]. Generally, gene expression can be determined by the genotype in SNPs. But due to the technical issue, gene expression data is more easily to access than the genotypes in SNPs. Due to the potential association between gene expression and genotypes, our study is conducted to test if we can predict the

genotypes based on the gene expression data. Limited resources of gene data was a major obstacle to conduct this hypothesis. With recent years increase development on large-scale gene expression analysis, public avaible resources like the Genotype-Tissue Expression (GTEx) project, Gene Expression Omnibus (GEO) [8] provide an effective way to overcome this limitation. And we may assume that not all the SNPs can be predicted precisely by the gene expression since some of the SNPs are uncorrelated with eQTLs.

In the beginning, researchers assumed the gene expression is one-to-one correspond to the SNP. However, there’re thousands of gene expression data and SNPs, conducting large-scale study to identify this correlation may not be feasible. Therefore, combining gene expression data from the same chromosome is a useful approach for researchers to conduct a hypothesis. Researchers have identified that some gene expression data are highly correlated while some gene expression are running indenpendently [9]. Interestingly, some SNPs are not just correlated with one gene expression [10] while some correlations are not that significant. The specific aims of this study are twofolds. First, we explore the association between one SNP and a package of gene expression data which are both from one chromosome by using random forest model. Second, we identify the prediction accuracy on genotype based on gene expression data. Since we know there may exist potential correlation between gene expression data and genotypes, we use machine learning method to test the score correlation for further understanding.

The data we use in this study includes: real data which can be downloaded directly from GTEx and simulation data which is similar to the real data but with a small size. The sample we use is whole blood sample from chromosome 22.

Method

The prediction of the SNP genotype can be as simple as a binary classification problem. The input is the transcriptome of a sample in the form of gene counts matrix, and the output is the probability of SNP genotype of the sample. The method we applied in this dataset is the random forest, which can handle large datasets and can provide a good measure of feature importance. We first test this method in simulated datasets. Subsequently, we conducted real data analysis using GTEx RNA-sequence SNP genotype data.

A random forest is a collection of trees with variations in structure generated using tow modifications to the deterministic tree-growing algorithm [1]. Decision trees are a popular method for various machine learning tasks. Tree learning is invariant under scaling and various other transformations of features values. Besides, it is robust to irrelevant features included. However, a single decision tree is not accurate and easy to overfit the model. The random forest can average thousands of decision trees to significantly lower the variance to prevent the over-fitting problem by building thousands of decision trees. We choose a random forest since it can substantially improve performance in term of prediction and accurace.

In a random forest algorithm, given a training set with n samples $\{X, Y\}$, for $b = 1,2, \ldots B$:

1. Take n samples from X with replacement, make up a collection $\{X_b, Y_b\}$;
2. Train decision tree in sample $\{X_b, Y_b\}$;

After training, taking the average of all models as output:

$$\hat{f} = \frac{1}{B}\sum_{b=1}^{B} \hat{f}_b\,(x')$$

Where B is an adjustable parameter.

Generally, we will use hundreds or thousands of trees or use a cross-validation method to choose the best B.

As we know, a single decision tree can be easily influenced by noise data, which could be highly variant. Implementing a random forest model can successfully solve this problem. However, if we keep training decision trees with the same data, we could also get a highly correlated tree, which may significantly impact our accuracy in the test set. So we apply the bootstrap sampling method to select different data to train decision trees.

When using the bagging method, there's a straightforward way to estimate the test error without using cross-validation. Recall that in bagging, trees are repeatedly fit to bootstrapped subsets. We can prove that, on average, two-thirds of the observations are used to bag the tree while rest one-third observations that not be used to fit the trees are referred to as out-of-bag (OOB) observations. Therefore, we can use these OOB observations to make

predictions. It will yield $B/3$ predictions for the ith observations since we have $B/3$ observations as a test set. We can obtain a single prediction for the ith observation by averaging these responses. Therefore, we can get the overall OOB MSE or classification error. The OOB error is an estimate for the test error. OOB score is another metric in OOB method to evaluate the classification rate. OOB score is calculated as the number of correctly predicted observations from the out of bag samples which were not necessarily used during the model analysis, so with OOB we are not using the full samples. In this way, OOB sample is a little more random than validation set. Therefore, OOB score may on average have a less good accuracy compared to using validation set as prediction. But it's more helpful when we have limited samples since we can't subset a validation set to test the model.

However, while bagging can improve model accuracy in prediction by using a single tree. Unfortunately, it can also lead to difficulty in interpretation. Therefore, random forests provide an improvement over bagging. In bagging, we train a number of decision trees on bootstrapped samples. For every split of a tree, a random sample of m predictors is chosen from the full predictors p . In each split, a new sample of m predictors is taken as a split candidate. This can successfully prevent highly correlated in every tree. For instance, if there's a strong predictor in the first split, then basically all trees will be similar to each other. Therefore, on average $(p-m)/p$ of the splits will not consider the strong predictor. This improvement can be seen as a decorrelating process for

the trees.

This method is very useful in our genetic data since we have a large number of correlated predictors. Our data is downloaded directly from GTEx portal and we use chromosome 22 as our sample. After deleting some unbalanced data and null values, we have around 600 observations with 454 gene expression as features. And we also have hundreds gene SNPs for which each SNP represents a difference in single DNA building block, called a nucleotide. SNP can be used to track the inheritance of disease genes with families, or the association with complex diseases. Our goal is to use random forests to make a classification of SNP and figure out which SNP has the most powerful influence on genetic expression. Besides, for the SNP which achieves the highest AUC score we would like to test the feature importance.

In the real data, our main goal is to make a classification, and we may decide to predict the class values directly. A common way to compare that predicted probabilities for two-class problems is to use the Receiver Operating Characteristic curve (ROC curve). It's a plot of the false positive rate (x-axis) versus the true positive rate (y-axis) for a number of different candidate threshold values between 0 and 1. In other words, it plots the false alarm rate versus the hit rate.

The true positive rate is calculated as the number of true positives divided by the sum of the number of true positive and the number of false negatives. It describes how good the model predicts the positive class when the actual

outcome is positive, which is also referred to the sensitivity.

$$True\ Positive\ Rate = Sensitivity = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The false-positive rate is calculated as the number of false positives divided by the sum of the number of false positives and the number of true negatives. It's also called the false alarm rate because it summarizes how often a positive class is predicted when the actual outcome is negative.

$$False\ Positive\ Rate = 1 - Specificity = \frac{False\ Positives}{False\ Positives + True\ Negatives}$$

The ROC curves can help in deciding the best threshold value. And the area under the curve (AUC) can give the rate of successful classification by the random forest model. The AUC makes it easy to compare the ROC curve of one model to another.

Both AUC and ROC are important evaluation metrics for calculating the performance of any classification model's performance. Therefore, using these two metrics can help us evaluate our model performance.

Simulation

We first use the simulated data to test the model. In order to make the data close to being realistic, I try to simulate the data based on the original dataset. First, I simulated 300 patients with 100 genes and an SNP genotype in each patient. Since an SNP genotype can be associated with the expression of hundreds of genes and mutations can lead to expression changes. In real data, the genotype is coded with three different values: 0, 1, 2, which represent aa,

Aa, AA genotype, respectively. However, in the simulation, I combined values 1 and 2 as one value since their expression features are the same. So in simulation data, the SNP genotype is a binary variable with values 0 and 1. Among the 100 genes, we don't know which one is connected to the genotype. In fact, some of them have nothing to do with this genotype, while others may be the factors that can impact the genotype. So our goal is to find out if the mutated genes have an impact on the genotype. During the simulation, I would generate just a few mutated genes which are different from the normal one for each patient. And each patient has a genotype that is identical to the GTEx data.

The average OOB estimate of error rate is 21%. The parameters for the model we use is 500 number of trees and 10 variables for each split. I run the model 10 times to get a average OOB estimate of error rate This estimate is calculated by counting how many points in the out of bag data were misclassified, which means that if we give a new patient's gene expression data, we have a 21% chance to misclassify his/her SNP.

Data analysis

First, let's take a look at the summary of all data. The data is combined with two datasets, one contains patients' gene expressions, and the other is the SNP data. They're both from chromosome 22 so that the relationships are easier to interpret. Gene expression data can directly be downloaded from GTEx Portal

[11], while the SNP data needs to be selected with at least one eQTL p-value less than 10^-10 since there're a lot of SNPs and our sample size is limited. So we choose the first 1200 SNPs by sort of their normal p-value. After selecting the SNP, I found out that there're a few missing values in the SNP dataset. So we need to drop those nulls, and if one SNP contains too many missing values (the missing value proportion up to 30%) we would drop this whole SNP. There're 32 SNPs I drop because of too many nulls.

Then we combine the gene expression data with the SNP dataset. After matching, we found there're some unbalanced SNPs and gene expression data, which mean too much 0 or 1 (up to 70%). These unbalanced data should not be considered in the final model, so we delete these data. The final dataset contains a total of around 600 observations; each observation has 454 gene expressions as features. We have 917 SNPs as our outputs, which means we need to run the random forests model at least 917 times. And since we use 10-fold cross-validation method, we will run every single SNP 10 times and calculate the mean AUC score for each SNP. Applying 10-fold cross-validation has one potential advantage that it often gives more accurate estimates of the test error rate than does LOOCV or another method. Then we would have 917 AUC scores and want to know which SNP gives the highest AUC scores. And we would make a ROC curve and Precision-Recall (PR) curve.

We will use two different evaluation metrics to compare our final results: OOB scores and mean of AUC scores from 10-fold cross-validation. OOB score is the accuracy of example x_i using all the trees in the random forest ensemble for which it was omitted during training. Thus, it kind of acts as a semi-testing instance. We can get a sense of how well our classifier does by using this metric.

In the final model, we test around 917 SNPs and get a distribution plot of their AUC scores and OOB scores. As the following:

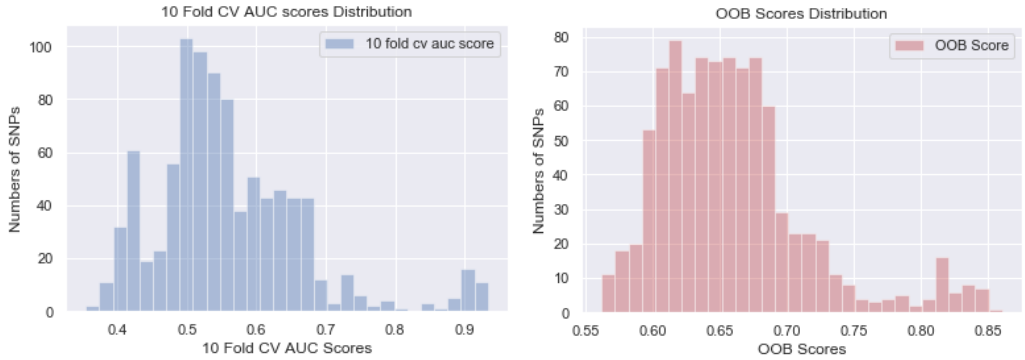

Figure1: Figure1: The distribution of the 10-fold CV AUC scores and OOB Scores. The left plot shows that most of the AUC scores are around 0.5 to 0.7. Just a few of SNPs can achieve a 0.9 AUC score. The right plot shows the distribution is almost the same as the left one. But in the right figure, most of the SNPs' OOB scores are around 0.6 to 0.73. The highest one is about 0.85.

These two plots show that two metrics produce almost the same distribution of SNP scores, although OOB scores have a higher average score than AUC. From figure 2 we can see that the correlation between AUC scores and OOB scores. They follow a linear relationship.

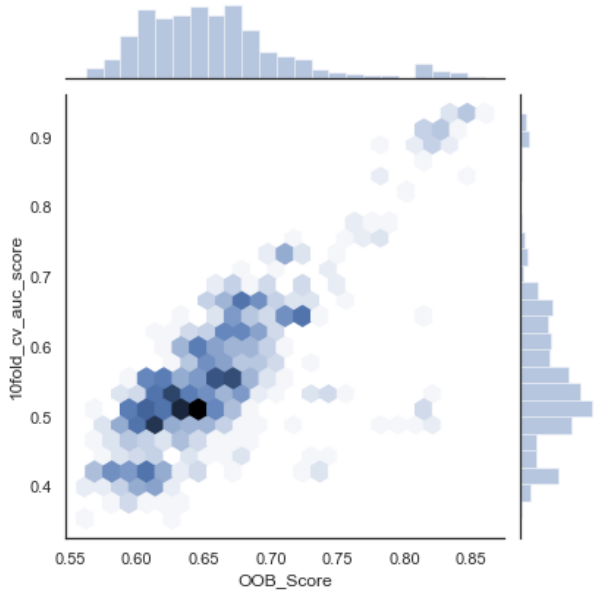

Figure 2: Correlation between AUC scores and OOB scores. From the plot, we can find out their correlations follow a linear line. 10-fold cv AUC scores are calculated as area under the ROC curve while OOB scores are calculated with OOB method.

In the final result, we find out that one SNP: 22_24258777_C_T_b37_C achieves the highest AUC scores with 0.932902 (OOB score 0.845) while the highest OOB score is from 22_24249458_A_C_b37_A with OOB score 0.86 and AUC scores 0.929671. Their ROC curves are as following:

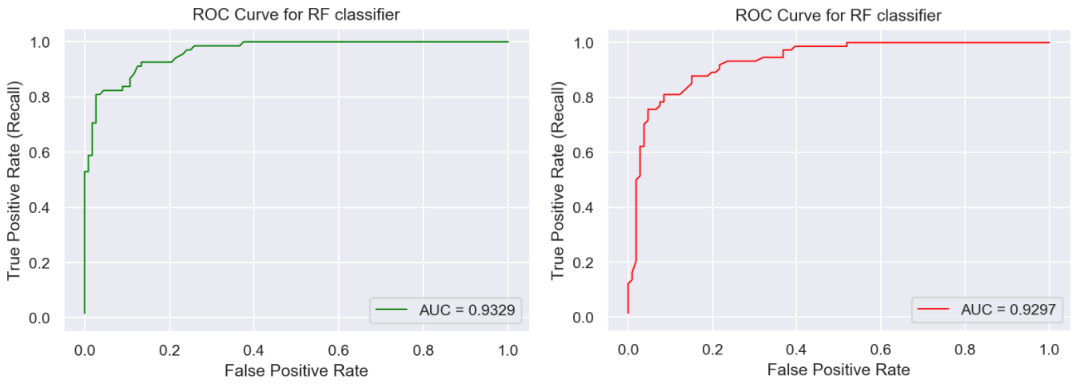

Figure 3: ROC curves for SNP 22_24258777_C_T_b37_C and 22_24249458_A_C_b37_A. From these two curves, we can find out their threads are basically the same. The left one is more steeper than right one.

Then we want to see the importance of the features in the model. Since we have a lot of SNPs, we decided to use the one with highest AUC score. Here is

the plot:

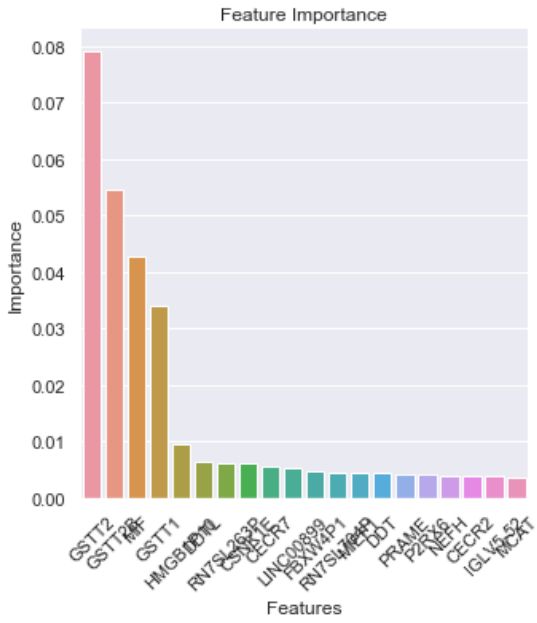

Figure 4: Feature importance. From the plot, we can see only the first 4 features take most of the importance of the model. Other features are basically not that important.

Since we know, the random forest consists of a number of decision trees. Every node in the decision tree is a condition on a single feature., designed to split the dataset into two so that similar response values are assigned in the same set. The measure based on which the optimal condition is chosen is called impurity. The impurity-based feature importance ranks the numerical features to be the most important features. From the table 3 in Appendix above, we can find that the first four features take most of the importance in the model. Although this just one SNP's feature importance, we still can take a look at it since its AUC score is the highest. Below, we make a feature importance summary of the top 6 AUC scores SNPs.

From the table 1 in Appendix, we can see that the important features for 6 SNPs with the highest AUC scores are basically the same, while for some of them,

they may not in the same order. And their values are around the same; other features may not be that important compared to them.

The summary for the AUC scores and OOB scores is in the table2 in Appendix. The summary table tells us that the mean score of AUC is less than OOB, while the highest AUC score is bigger than OOB one. Besides, the standard error for the AUC is larger than OOB, which means that the OOB score is more stable since it has a small variance.

Discussion

From the research, we can see some SNPs genotype can be successfully predicted based on the gene expression data given we have enough data. In this study, we used chromosome 22 as our sample, both SNPs and gene data are from this chromosome. We can conclude that some SNPs can be precisely predicted with AUC scores up to 90% while some other SNPs may not be that precisely predicted. For those with high AUC scores SNPs, we also test their feature importance when fitting the model. The result shows that there are four features playing important roles in random forest classification: GSTT2, GSTT2B, GSTT1, and MIF. Other features may also be important, but in this model, they’re not as important as these four features. The research also shows that most of the SNPs achieve an AUC score around 0.5 to 0.7; only a few can get 0.9. It may imply our model may not be as good as we predicted. They're still a lot of improvement we can do to increase our AUC scores. For now, we

can conclude that the some gene expression is highly associated with the SNP while somo associations are not that significant. Based on the gene expression data, we can precisely predict only a small fraction of SNPs. Due to the highly correlated in some gene expression features and the feature importance test, we may infer that only some features are important on predicting the genotype. However, it still needs further study to determine the specific associations between gene expression and genotypes.

One thing to be considered is that our data is all from GTEx portal and the observations are very limited although we have thousands of SNPs and gene expression When we got access to the dataset, we found out the data is not clear and with lots of null values in the raw dataset. Different datasets are in different formats, and we needed to clear and preprocessed them, which is time consuming. In addition, although there’re thousands of SNPs in the dataset, some data in the SNP are unbalanced, including too much 0 or other values. These unbalanced data would significantly affect the model prediction if we didn’t filter them out.

As for the model, we chose random forests as our model because it has a few advantages, including handling thousands of input variables without variable deletion, providing a reliable feature importance estimate, Maintains accuracy when a large proportion of the data are missing, etc. However, it still has some limitations. When we applied our data in the model, such as easy to be

overfitted, results are less interpretable compared to other models, cost too much time, and computational memory when we have a large dataset. Besides, when fitting the decision trees, if the data contains groups of correlated features of similar relevance to the output, then the smaller groups are favored over large groups. Since there're too many features in the dataset, it's a difficult job to choose the uncorrelated features. Also, our data contains only 600 observations while we have hundreds of features to consider, which makes the model more difficult to fit.

Finally, we would like to note that our work just takes a small part of the gene data, and for some other chromosomes or other tissues, the data may not be the same as our sample, so the model may not be that good when fitting other sample data. This work is just a small piece of genomic work, and we still need more data to validate our hypothesis and make the prediction more accurate.

Reference:

[1]: Nica A C, Dermitzakis E T. Expression quantitative trait loci: present and future[J]. Philosophical Transactions of the Royal Society B: Biological Sciences, 2013, 368(1620): 20120362.

[2]: Brem R B, Storey J D, Whittle J, et al. Genetic interactions between polymorphisms that affect gene expression in yeast[J]. Nature, 2005, 436(7051): 701-703.

[3]: Boggis E M, Milo M, Walters K. eQuIPS: eQTL analysis using informed partitioning of SNPS–a fully Bayesian approach[J]. Genetic epidemiology, 2016, 40(4): 273-283.

[4]: Majewski J, Pastinen T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. Trends Genet. 2011;27(2):72–9.

[5]: Zhu Z, Zhang F, Hu H, et al. Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets[J]. Nature genetics, 2016, 48(5): 481.

[6]: Yang T P, Beazley C, Montgomery S B, et al. Genevar: a database and Java application for the analysis and visualization of SNP-gene associations in eQTL studies[J]. Bioinformatics, 2010, 26(19): 2474-2476.

[7]: Emilsson V, Thorleifsson G, Zhang B, et al. Genetics of gene expression and its effect on disease[J]. Nature, 2008, 452(7186): 423-428.

[8]: Barrett T, Troup D B, Wilhite S E, et al. NCBI GEO: archive for high-throughput functional genomic data[J]. Nucleic acids research, 2009,

37(suppl_1): D885-D890.

[9]: Province MA, Borecki IB. Gathering the gold dust: methods for assessing the aggregate impact of small effect genes in genomic scans. Pac Symp Biocomput. 2008;1:190–200.

[10]: Nicolae D L, Gamazon E, Zhang W, et al. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS[J]. PLoS genetics, 2010, 6(4).

[11]: GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans[J]. Science, 2015, 348(6235): 648-660.

Appendix:

Table 1: Feature importance for the top 6 AUC scores SNPs. We can see the feature importance is almost the same.

22_24292488_G_T_b37_G		22_24249458_A_C_b37_A		22_24266867_G_A_b37_G		22_24250072_AG_A_b37_AG		22_24249616_A_C_b37_A		22_24292178_C_A_b37_C	
Features	Importance	Features	Importance	Features	Importance	Features	Importance	Features	Importance	Features	Importance
GSTT2	0.075119	GSTT2	0.076832	GSTT2	0.072134	GSTT2	0.068589	GSTT2	0.075551	GSTT2	0.074203
GSTT2B	0.055091	GSTT2B	0.043415	GSTT2B	0.049322	GSTT2B	0.062982	GSTT2B	0.041672	GSTT2B	0.056681
GSTT1	0.040015	GSTT1	0.038567	MIF	0.038099	MIF	0.041304	GSTT1	0.040074	MIF	0.039029
MIF	0.034338	MIF	0.033499	GSTT1	0.030196	GSTT1	0.034055	MIF	0.039485	GSTT1	0.035524

Table 2: summary for the AUC scores and OOB scores. The mean of AUC scores are less then the OOB scores while the highest one is larger than OOB scores. Also, the standard error for the OOB scores is less than AUC scores which means OOB scores give a more stable test result

	Mean	Std	Min	25%	Median	75%	Max
AUC Scores	0.559167	0.107518	0.354759	0.494991	0.540776	0.615156	0.932902
OOB Score	0.658146	0.056127	0.561667	0.618333	0.650000	0.681667	0.860000

Table 3: Features Importance. The features are from the highest AUC score SNPs.

	Features	Importance
1	GSTT2	0.078946
2	GSTT2B	0.054679
3	MIF	0.042828
3	GSTT1	0.033928
4	HMGB1P10	0.009399
5	DDTL	0.006537
6	RN7SL263P	0.006151
7	CSNK1E	0.006090