Selection of SNPs with high relevance to diseases

Abbreviation: d1 – Dataset 1, d2 – Dataset 2.

Original files used:

* d1:
  + “d1\_plink1\_om1.5\_rf\_impt\_agre.csv” from the 1st random forest project, “MeanDecreaseGini” values for all the SNPs used in one random forest model
  + “rap\_d1\_go2snp.rd” from the 2nd random forest project, binary R data format of a list mapping from GO term to SNPs
  + “rap\_d1\_go\_oob\_auc.tsv” from the 2nd random forest project, OOB error rate and AUC of ROC curves for each random forest model based on a subset of SNPs associated with a GO term
* d2:
  + “d2\_plink1\_a1e-5\_rf\_impt\_agre.csv” from the 1st random forest project, “MeanDecreaseGini” values for all the SNPs used in one random forest model
  + “rap\_d2\_go2snp.rd” from last project, binary R data format of a list mapping from GO term to SNPs
  + “rap\_d2\_go\_oob\_auc.tsv” from last project, OOB error rate and AUC of ROC curves for each random forest model based on a subset of SNPs associated with a GO term

Code files:

* “045\_script\_4.r”

Result files:

* OOB error rate and AUC of ROC curves, with “1-OOB+AUC”:
  + “rap\_d1\_go\_oob\_auc\_1.tsv”
  + “rap\_d2\_go\_oob\_auc\_1.tsv”
* Selected SNPs:
  + “d1\_top\_snps\_300\_20.txt” (350 SNPs)
  + “d2\_top\_snps\_300\_20.txt” (482 SNPs)

Selection procedure:

1. The top-300 SNPs with the highest “mean\_MeanDecreaseGini” values were selected based on the files “\*\_rf\_impt\_agre.csv”. To recap, “mean\_MeanDecreaseGini” is the mean of “MeanDecreaseGini” values of all the allele variants associated with it (e.g. the mean of SNP\_A-1844768\_\_AA, SNP\_A-1844768\_\_AC, SNP\_A-1844768\_\_CC for SNP\_A-1844768). The higher the “mean\_MeanDecreaseGini” is, the more important the SNP is in predicting genetic predisposition of disease.
2. A metric combining OOB error rate and AUC of ROC curves “1-OOB+AUC” is calculated literally as 1 - OOB + AUC for each random forest model based on a subset of SNPs associated with a GO term. For example, if OOB = 0.37 and AUC = 0.65, then “1-OOB+AUC” = 1 - 0.37 + 0.65 = 1.28. The higher the “1-OOB+AUC” is for a model, the better the model performs.
3. The top-20 GO terms with the highest “1-OOB+AUC” were found, then the SNPs associated with these GO terms were found. These SNPs were ranked in descending order based on the “mean\_MeanDecreaseGini” values from files “\*\_rf\_impt\_agre.csv”, then the top-200 SNPs were selected.
4. The SNPs selected in step 1) and 3) were combined, and duplicates were removed. These selected SNPs are saved in files “d\*\_top\_snps\_300\_20\_200.txt” (replace \* with 1 or 2)