Explorations, Statistical Inferencing, and Data Mining for GWAS on Gordon

Paul F. Rodriguez
University of California, San Diego
San Diego Supercomputer Center, MC 0505
10100 Hopkins Drive
La Jolla, CA 92093-0505
1-858-534-8326
p4rodriguez@ucsd.edu

ABSTRACT
Genome Wide Association Studies consists of hundreds of thousands of SNPs (single nucleotide polymorphisms) as independent variables but with typically only few thousands of cases. Given that many diseases are rare, and the high correlation between SNPs, it is difficult to identify significant SNPs that can replicate and have predictive value. Here we explore machine learning/data mining techniques on the Gordon machine at SDSC, using standard packages of R and Matlab™. We suggest that importance measures from machine learning can help identify SNPs involved in complex interactions over and above univariate procedures.

Categories and Subject Descriptors
J.3 [Life and Medical Sciences] Biology and Genetics: Genome Wide Association Studies, G.3 [Probability and Statistics] Correlation and Regression Analysis: Variable Selection

Keywords
Genome Wide Association Studies, Variable Selection

1. INTRODUCTION
Genome Wide Association Studies consists of hundreds of thousands of SNPs (single nucleotide polymorphisms) as independent variables but with typically only few thousands of cases. The main goal is to find significant SNPs that can replicate and have predictive value. Moreover, many studies show that individual SNPs do not account for much of the total phenotypic variance. It is likely that disease states depend on complex interactions between many SNPs. Many machine learning techniques are designed to find non-linear interactions. However, in contrast to univariate procedures that assign significance to individual SNPs, machine learning techniques determine importance or relevance of variables for classification or prediction. Moreover, these approaches are also computationally expensive, which results in myriad of trade-offs between the feasibility of analysis, appropriate analysis, techniques that might be applied, and the kind of question one might like to address.

Here we apply Random Forests and Least Angle Penalized Logistic Regression to genome wide data of Primary Sclerosing Cholangitis (a disease of the bile duct) consisting of about 80,000 SNPs after preprocessing and imputation. We compare and contrast the resulting SNP selections with statistical significance from univariate Logistic Regression.

All processing was done on Gordon compute nodes at San Diego Supercomputer Center. Preprocessing was done using the Plink software package, all analysis was performed in R and Matlab. We found that the R package ususally performs better when datasets are partially sampled to about 3000 variables external to R. This saves memory within an R instance and allows one to run one R instance on each core within a compute node. In this way, we could run 50 sample trees across each of 1000 sample data subsets on 4 compute nodes (64 cores) in about 12 hours. We used Matlab implementation of Logistic Regression and our own implementation of multivariate Least Angle Regression (LARs) to select variables. For sampling interactions we first chose 80 variables within a LARs sequence, then applied LARs to the interaction set of those and chose the top 1000 interactions in a second LARs sequence.

Results show that the importance of SNPs from Random Forest are highly correlated with univariate regression, which means that SNPs that individually have statistically significant estimated effects are also important in the samples of decision trees for classification. However, even at liberal p-value thresholds, there are some SNPs that have high importance, and are often involved in 2-way interactions as selected by penalized regression, but are not among those selected by univariate methods. Thus, using an ensemble of variable selection methods might improve the process of identifying SNPs and SNP-interactions for further investigation and disease predictions.

2. ACKNOWLEDGMENTS
Thanks to A.M. Dale, J. C. Roddey, at the MMIL at UCSD, Dept of Neuroscience for advice and consultation on GWAS methods and sharing thier preprocessed PSC dataset. The author is part of the Predictive Analytics Center of Excellence at SDSC. This research was supported in part by the National Science Foundation through the San Diego Supercomputer Center under [#20091748 OCI-0910847] and utilized the Gordon system.

3. REFERENCES
[1] Liu, JZ, et al..2013 Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis. *Nature Genetics.* 45 (Apr. 2013), 670-675.