Homework 4
Statistical Genomics
CROPS 545, Spring 2016
Professor: Zhiwu Zhang

Due on March 23, 2016, Wednesday, 3:10PM, PST

Data files: The following files can be download at http://zzlab.net/GAPIT/data/.

a) mdp_numeric.txt from GAPIT demo data. The data file contains 281 individuals (row wise) and 3093 SNPs (column wise) coded as 0/1/2.
b) mdp_SNP_information.txt. The file contains SNP ID, chromosome and position.
c) CROP545_Phenotype.txt. The file contains taxa name and phenotype.
d) CROP545_Covariates.txt. The file contains taxa name and two covariates.

Hand in: Each team (maximum of three people) email your report (PDF, limited to five page), R source code (text file), and user manual/tutorial (PDF, no page limitation) with email subject of "CROPS545 HW4" to Zhiwu.Zhang@WSU.edu. Name your files as following:
Homework4_ PackageName.pdf and Homework4_ PackageName.R
Grade components: 1) Hypothesis or statement; 2) Results; 3) Methods; 4 presentation; 5) R source code (clarity, simplicity and documenting comments)
Objectives: Develop your own R package to perform GLM GWAS.

In order to ease development and deployment, our group opted to build the package using Github and document the functions internally using Roxygen2. The package can be accessed at the following URL:
GWASbyGLM

On the main page, there is a short description of the package, installation instructions, and directions for how to access the help documents for package functions. Additionally, there are examples provided in the function documentation that demonstrate how to use the functions with test data.

(1) The package should contain at least three input: y, X , and C that are R objects of numeric data frame. Their dimensions are n by 1, n by m, and n by t corresponding to phenotype, genotype and covariate data, where n is number of individuals, m is number of markers, and t is number of covariates. The function should return probability values with dimension of 1 by m for the association tests between phenotype and markers. Markers are tested one at a time with covariates in C included as covariates (15 points).

The GWASbyGLM function will perform this function by stipulating the 'cov' option. To manually perform this feature using test data, refer to the Examples section of the help document for the function.

(2) The package should also provide additional co-factors to improve performance of GWAS, such PCA. Name your package with an acronym that describes the features of your co-factors. Your package should also automatically eliminate your own co-factors if they are in linear dependent to the covariates provided by users (25 points).

The GWASbyGLM function will also perform PCA analysis and include the results as cofactors by stipulating the 'pca' option. To manually perform this feature using test data, refer to the Examples section of the help document, but change the option parameter to 'pca' instead of 'cov'.

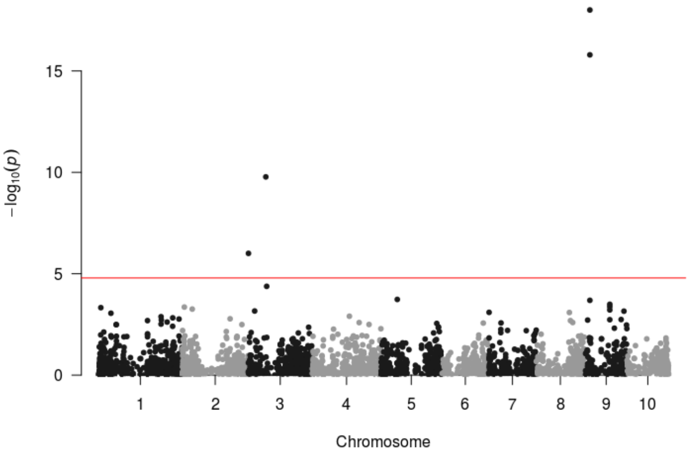
(3) Develop a user manual and tutorials (20 points).

Help documents were embedded in the package for each function. Tutorials are also included in the form of examples. To ensure example compatibility for any machine, permanent sample datasets were added to the Github page.

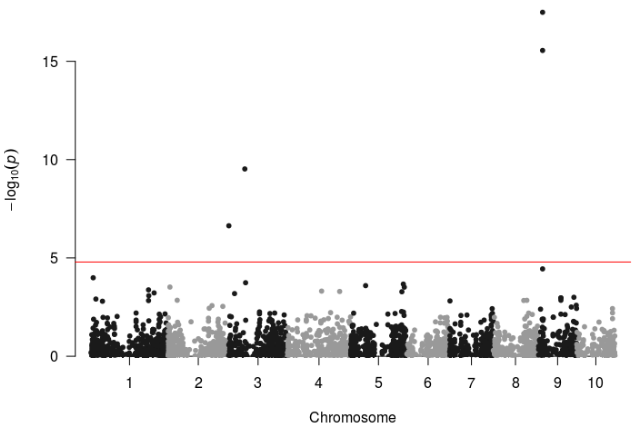
(4) Perform GWAS on the genotypes, phenotypes and covariates provided (15 points).

Using the developed package, GWAS was run using both the 'cov' and 'pca' options. The plot_manhattan function was called for both results and are presented as follows:

GLM with covariates only:

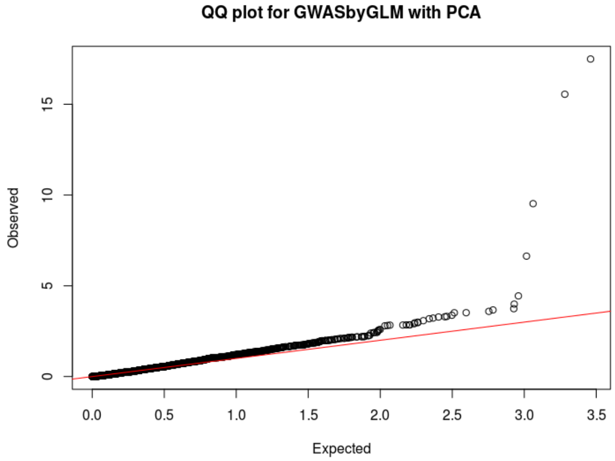


GLM with covariates and PCA analysis:



In terms of significant SNPs, there is little difference between the two manhattan plots. However, there do seem to be some differences between near-significant SNPs close to the Bonferoni cutoff.

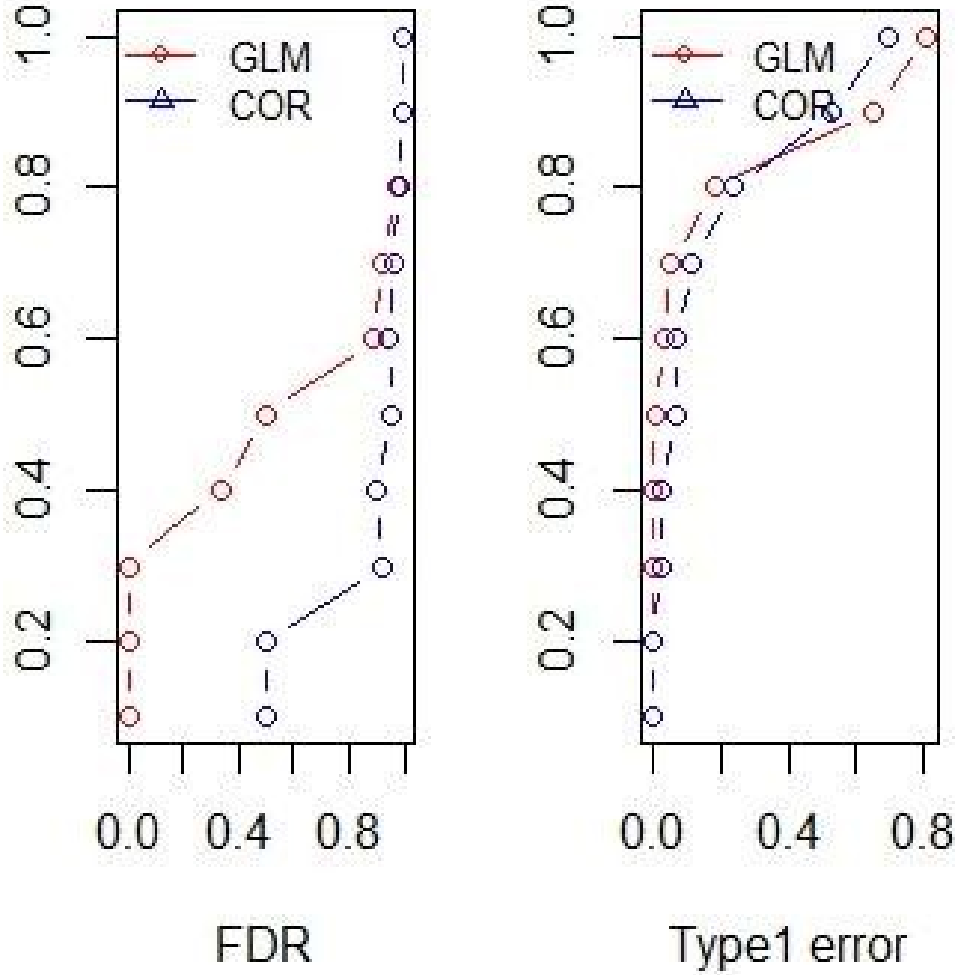
The q-q plot for the PCA GLM analysis looks very promising based on examples presented in class:



(5) Demonstrate that your method is superior to the competing method (GWASbyCor) through simulation (25 points).

Both the GLM package and the GWASbyCor function were performed using 5 different simulated sets of phenotypes using the same method demonstrated in class.

Here are the aggregate Power vs. FDR and Power vs. type 1 error graphs:



There does seem to be a difference between the two algorithms. However, the sample size is relatively small and needs to be run many more times to confirm this further. However, there is criticism whether this accuracy simulation is truly even capable of discerning the true difference between the two algorithms using the simplistic phenotype simulation strategy presented in class. This is primarily due to the fact that the confounding present in the genotype and phenotype data addressed by PCA analysis is 'locked' to the true state of the populations. Simply generating phenotypes from a distribution of artificial QTLs with genetic effects without addressing the underlying confounding population structure that should also be affecting the end phenotype does not seem the best strategy to assess whether including PCA or external covariates in the model performs better than simple GWAS-by-correlation.