R Notebook

Data and Functions

The GAPIT demo datasets were used for this assignment. They are available from "http://www.zzlab.net/GAPIT/index.html" All credit for these data belong to the Zhiwu Zhang Laboratory.

Question 1 and 2

(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). (2) The package should perform PCA and incorporate PCs as cofactors for GWAS. Your package should also automatically exclude the PCs that are in linear dependent to the covariates provided by users. (25 points).

JKGWAS Summary

The JKGWAS Package contains four functions that are summarized briefly as follows, more information is located in the JKGWAS Package documentation:

- JKPCA takes genotype (X) data and covariate data (CV), computes the PCA on X, then automatically removes PCs that are linearly dependent to the CVs by method of comparing matrix rank. PCs are removed from the matrix in succesion and those that do not change the rank by removal are determined to be linearly independent because they do not provide additional information.

- JKGLM takes phenotype (y), genotype (X), covariate (CV), and principal component (PC) inputs (ideally provided from JKPCA) and returns p-values calculated for the association tests between the phenotype and SNPs

- JKQQ takes the pvalues from JKGLM and visualizes them by QQ plot. Expected p-values of length m are simulated from the continuous distribution.

- JKManhattan visualizes the pvalues from JKGLM by Manhattan plot. User input QTNs can also be visualized. The significance threshold can be set, or it will default to Bonferoni correction for alpha = 0.05

Question 3

(3) Develop a user manual and tutorials. Name your package and create a logo. (20 points).

The JKGWAS package is named for Pabitra Joshi and Lindsey Kornowske, the label is displayed in Figure 1.

The JKGWAS package documentation is provided in a separate file, "JKGWAS$_0.1.0_{P}ackageDocumentation.pdf"andfurther

Question 4

(4) Perform GWAS on the data provided or your own data which must contain cofactors (15 points).

Figure 1: JKGWAS Package Logo

First, GWAS was performed with the phenotype data provided. In the Manhattan plot below, we can see that 4 SNP were detected, but because we do not have information about the QTNs, we do not know whether these significant observations represent true positives or not. Next, we use a simulated phenotype to better assess the performance of the JKGWAS approach.

(5) Demonstrate that your method is superior to the competing method (GWASbyCor) through simulation with at least 30 replicates (25 points).

See file HW4_JKGWAS_functions.R for the function source code.

In order to compare GWASbyCor and GWASbyGLM, we created a function called compareGWASnTimes. The function arguments are:

- n, the number of times to run the simulation
- X, the numeric genomic data
- qtn, the number of qtns to be simulated
- CV, the covariate matrix to be passed to the JKGLM function. The default value is NULL
- PC, the principal component matrix to be passed to the JKGLM function. The default value is NULL

For each iteration, the G2P function simulates the phenotype for X with a heritability of 0.75. Then, the output phenotype is used to compute the GWAS by cor and the GWAS by GLM. The number of True Positives, as well as the True Positive Rate, which is calculated as the number of QTNs that is correctly identified out of all significant SNPs (p-value is smaller than 0.05/total pvalues, Bonferroni correction is automatic). These two dataframes are output as a list, where the first item is the count of true positive QTNs and the second item is the true positive rate.

Statistical Inference

H_o; the mean rates are equal H_1; the mean rates are not equal

significance threshold: 0.05