CROP 545 HW 4

QN 1. Create an GWAS by GLM package.

Statement: GWAS by Corr provides a good demonstration for the idea of genome-wide assiciate study. However, GWAS by Corr produces relatively high number of false positve, and false negative. Idealy, researchers would preffer a more robust method with higher power. In this section, we implement the PCA and GLM methods for GWAS analysis, and develope the MAPCUE package for automatic process.

Methods: The `GWASbyGLM` is the core function of the MAPCUE package. Based on the genotype and phenotype data provided by the user, the function performs a GWAS analysis by principal analysis and generalized linear regression model method:

$$Y \sim PC_{1:k} + SNP_{1:p} + \epsilon$$

where $PC_{1:k}$ is the first k principal components, and SNP_{1:p} is the p SNP's provided by user. Given a set of genotype and phenotype data, the GWASbyGLM function first perform a PCA on the genotype. Based on the choice of k, the first k principal component will be used to form an predictor matrix with SNP's. Then a linear model is fitted by least square method. For each SNP, we test for $\_H_0$: $\beta_i = 0$ and collect the p value. The function returns the p values of all SNP's.

The rest of the functions in this package focuses on power analysis using the method we used in Homwework 2 and 3.

Results: Source code of the MAPCUE package is enclosed with this report. The package also comes with a user's manual to help users to get started with it.

QN 2. PCA and Covariates

Statement: The structure of the gene can be explored by PCA. PCA may add power to the GWAS analysis. Besides gene effect, there might be other factors that contribute to the phenometric effect. Users might want to include other covariates data in the GWAS analysis.However, if there is collinearity or simply linear dependency between the PCA and covariates, that would casause problem in the least square regression calculation (produces singular matrix). To avoid that, we need to remove any dependent columnes in the data.

Methods: We use the `fixDependence` in the `mgcv` package to detect dependence. And if dependence appears, we remove the dependent columne from the data matrix, then perform the GLM analysis.

Results: The function successfully handled the dependence and perform the analysis. See the result in the example.

QN 3. Develope and user manual and tutorial.

Statement: A user's manual is helpful for beginners to learn how to use a new package. We develope a MAPCUE users manual to explain:

- what is the method we used for analysis;
- how to install and use our functions;
- fundamental trouble shooting;
- what is the expected results in each step;

- examples for each of the function in our package.

QN 4. Perform GWAS analysis on the given data.

Statement In this section we use a standard dataset to test our package. The GWASbyGLM function should identify at least one QTN. However, with this dataset, we cannot identify the false positive and false negative.

Method We set the overall error rate at 5%. The following code performs the GWAS analysis and generates the manhattan plot.

```r
# load package
source("MAPCUE.R")
```

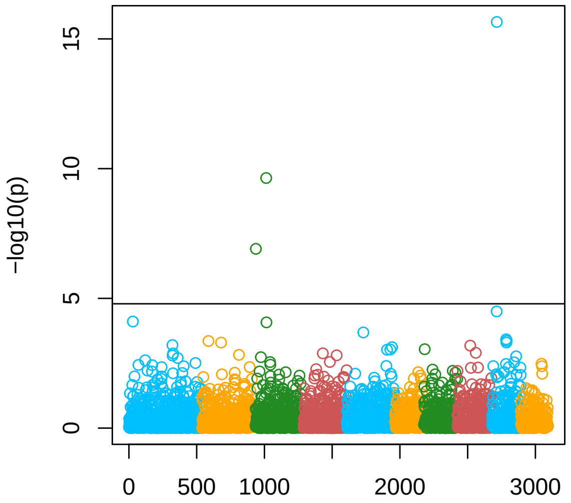
```
## Loading required package: nlme

## This is mgcv 1.8-22. For overview type 'help("mgcv-package")'.
```

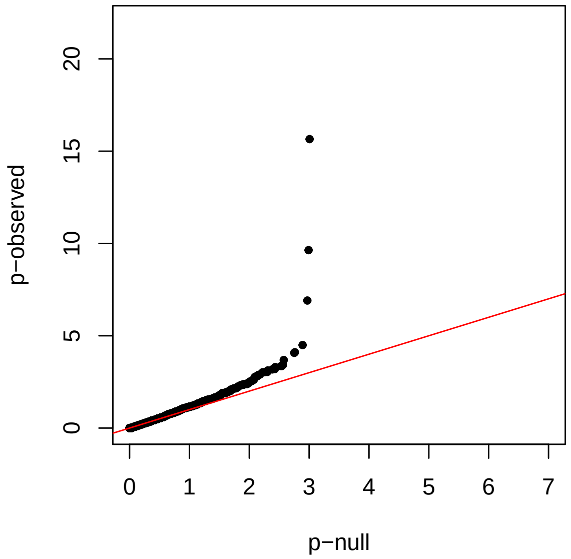
```r
# Import Data
GD=read.table(file="http://zzlab.net/GAPIT/data/mdp_numeric.txt",head=T)
GM=read.table(file="http://zzlab.net/GAPIT/data/mdp_SNP_information.txt",head=T)
y=read.table(file="http://zzlab.net/GAPIT/data/CROP545_Phenotype.txt", head=T)
y=y[,-1]
CV=read.table(file="http://zzlab.net/GAPIT/data/CROP545_Covariates.txt", head=T)
p = GWASbyGLM(geno=GD[,-1], pheno=y, CV=CV, PCA.M = 3)
p[1:5]
```

```
## [1] 0.9427934 0.4319452 0.7815253 0.5851984 0.9611347
```

```r
par(mfrow=c(1,2))
manh.plot(p, GM, 0.05)
qqGLM(p, xlim=7, ylim=22)
```

```
## Warning in plot.window(...): "na.rm" is not a graphical parameter

## Warning in plot.xy(xy, type, ...): "na.rm" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "na.rm" is not
## a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "na.rm" is not
## a graphical parameter

## Warning in box(...): "na.rm" is not a graphical parameter

## Warning in title(...): "na.rm" is not a graphical parameter
```



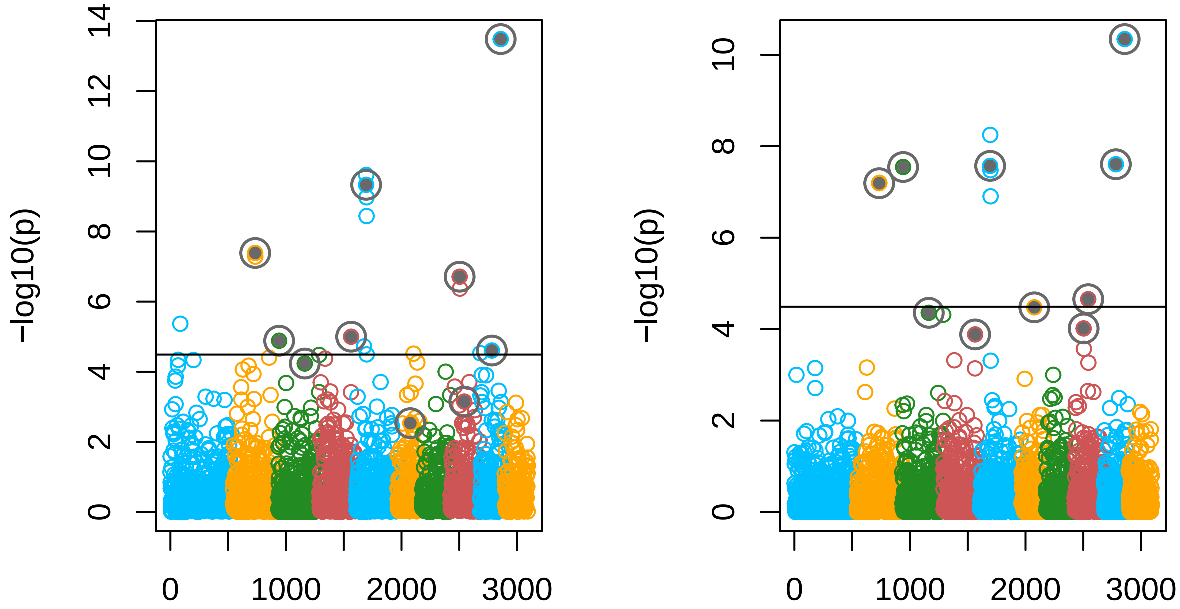


Interpretation Based on the Manhattan plot and the QQ plot, the function identifies three markers as significant QNT's. Since we don't know the number of true QTN's, we cannot make conclusion about the accuracy of the method in this case.

QN 5. Compare GWASbyCor and GWASbyGLM

Statement From previous homework and analysis, we have already known that `GWASbyCor` would produce large number of false positive and false negative. `MAPCUE` not only implements a more sophisticated method, but also allow the chance to include additional data into the model. Therefore, we hypothesis that `GWASbyGLM` would outperform `GWASbyCor`. We can judge it by the number of false positive each method produces.

Method We will use the genome data provided online, simulate an array of phenotype, perform `GWASbyCor` and `GWASbyGLM` respectively, then compare the number of false positive each method produced.



Result and discussion From the Manhattan plot we see that there are more number of false positive in the GWAS by Cor result than the by GLM result.

We run the function 100 times and calculate the mean number of false positives in the two methods. On average, the correlation method produces 13 false positive (exact mean is 12.59), while the GLM method only produces 3 false positive (exact mean is 2.74) which is more than 5 times less than the correlation method.

We performed a paired t-test for H_0: two sample means are equal. The result suggests that the mean false positve number of correlation method and glm method are not equal with a p value of 3E-13. Therefore, in terms of producing less false positive, the GWASbyGLM method outperformed the GWASbyCOR method.

However, the GWASbyGLM method produces high number of false negative, we need to improve the power of this method.