Haixiao Dong and Ryan Oliveira

Statistical Genomics (CROPS 545), Spring 2017 Homework #4

(1-3) An R software package was developed to perform GWAS using a general linear model using SNPs, user-input covariates, and principal components analysis (PCA) as the fixed effects. The addition of PCA is considered an added feature of the package and features prominently in its name: **A**ssociation **L**everaging **P**rincip**A**l **C**omponents **A**nalysis (ALPACA). However, PCA is not used if any of the columns are found to be linearly dependent on the user-provided covariates. The total inputs include phenotypes (y), genotypes (X), covariates (C), chromosome positions for SNPs (GM), a manual p-value cutoff (cutoff), and positions of known QTNs if running a simulation (QTN.position). The covariates, cutoff value, and QTN positions are all optional. If the covariates or QTN positions are not entered, they are not used in analysis (with SNPs and PCAs alone used as fixed effects). If a manual cutoff value is not used, the subsequent cutoff value is based off a simple Bonferroni correction applied to an alpha value of 0.05. The function returns a list including probability values for each marker to its phenotype and a list of ordered SNPs. If it is used to run a simulation with known QTNs, it outputs power, false discovery rate, and type I error. The program also creates a QQ plot, a Manhattan plot, and a principal component plot. If known QTNs are used, it generates a plot of power versus false discovery rate. In addition, a csv file is generated with a list of all tested SNPs, their chromosome numbers and position, their p-values, and their rank by p-value.

The software package is attached as ALPACA.R. There is also a user manual PDF and an associated R File for the examples used in the manual (ALPACAExamples.R).

(4) GWAS using ALPACA.R is performed on the provided genotypes, phenotypes, and covariates from ZZLab.net. As the principal components generated by ALPACA were found to be linearly dependent on the provided covariates, they were eliminated from the analysis. Four SNPs were found to be significant and are described in the table below:

Table 1. Significant SNPs detected

QTN position	**SNP ID**	**Probability value**
2715	PZA03058.17	0
2717	PZA03058.21	0
1013	PZA02699.1	1.35e-10
937	PZA00615.3	8.69e-7

The Manhattan and QQ plots display this data. As the first two SNPs are very close on the list by location (positions 2715 and 2717), they appear as one marker on the Manhattan plot but are distinct on the QQ plot. A generated Bonferroni cutoff is visible as a green line on the Manhattan plot; the value is 1.62 x 10^{-5}.

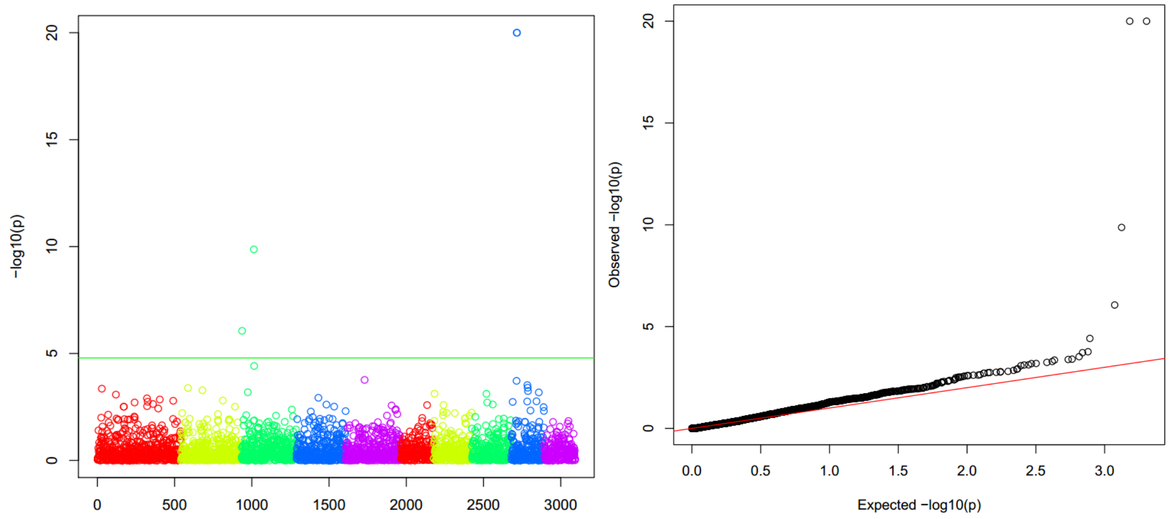

Fig 1. Manhattan plot and QQ plot for Question 4

(5) GWAS using ALPACA.R and GWAS by correlation (using the function GWASbyCor from ZZLab) are run on simulated phenotypes using myG2P from ZZLab.net. In the Manhattan and QQ plots (Fig 2), ALPACA ("GLM Manhattan") resolves the same number of QTNs as the Cor function with fewer false positives. The QQ plot is closer to the line with separation of the highest p-values. Based on 100 times replicates for various number of QTNs (NQTN) and heritability (h2), in Fig 3, the area under the curve for power versus FDR is always higher for GWAS using ALPACA ("GLM") than for GWASbyCor, indicating it is more specific at resolving quantitative trait nucleotides.

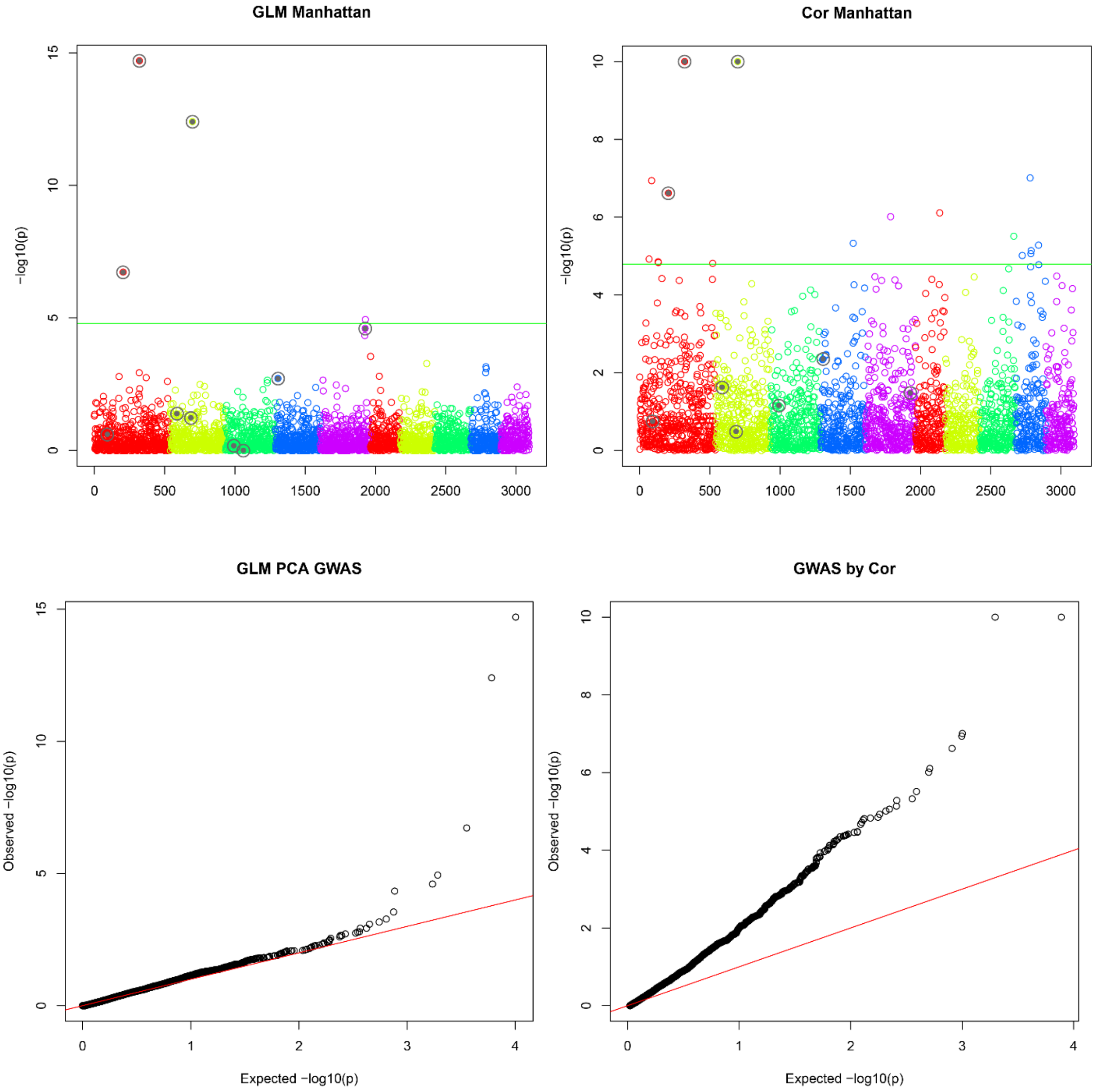

Fig 2. Manhattan plots and QQ plots for GWAS by GLM or Cor

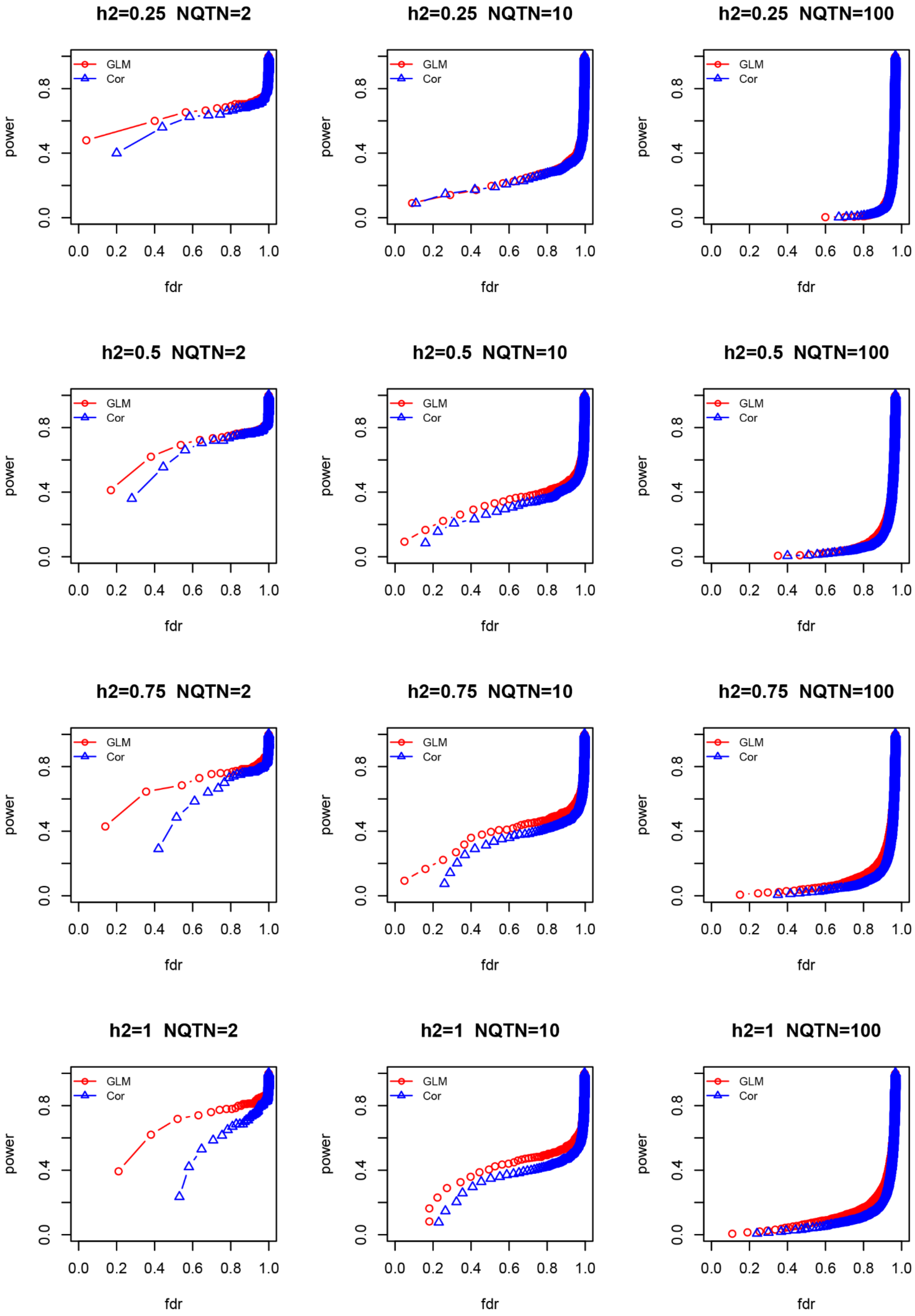

Fig 3. Comparison of Power and FDR for GWAS by GLM or Cor