STAT 530 Bioinformatics: Homework 2

Due Feb 27, 2017

For problems using R, turn in your answers in the form of a compiled R notebook PDF.

Problem 1 (5 points)

Make a PLINK file set called `qcd` by extracting individuals and SNPs from `hapmap1` using the following QC parameters:

- Exclude samples with missing rates of $> 6\%$.
- Exclude SNPs with missing rates of $> 10\%$.
- Exclude SNPs with MAF ≤ 0.05.
- Exclude SNPs deviating from HWE at $p < 10^{-4}$.

How many samples were removed? How many SNPs are left?

Problem 2 (2 points)

Run one GWAS using logistic regression without controlling for any principal components. Use the flag `--out nopc`. What is the genomic inflation factor? Report the OR of the SNP `rs2222162`. Is having more minor alleles of this SNP associated with higher or lower risk?

Problem 3 (5 points)

Calculate the top 3 PCs using EIGENSTRAT. Use the parameters in `example.perl` script but output 3 PCs instead of 2. Save the top 3 PCs as `qcd.pca`. Report the top 5 lines of `qcd.pca`.

Problem 4 (5 points)

Now run a GWAS controlling for PCs. Create a covariate file `pcs.txt` containing the three principal components calculated using EIGENSTRAT; use the R script `make_pcs.R` provided on the course website. Using this file, run a GWAS controlling for the first principal component. Use the flag `--out pc1`. Have we adequately controlled for population stratification? Report the OR of the SNP `rs2222162`. Is having more minor alleles of this SNP associated with higher or lower risk?

Problem 5 (5 points)

Using the PC-adjusted GWAS results, do any of the SNPs reach genome-wide significance? What is the Bonferroni threshold adjusting for the total number of SNPs tested? Do any SNPs pass this threshold?

Problem 6 (5 points)

Make a Manhattan plot of the results using R and the package `qqman`. Which SNP exceeds the 10^{-5} significance threshold? Using the UCSC Genome Browser, find the RefSeq gene closest to the the SNP.