Genome Wide Association Study

James O'Reilly
Student Number: r0773125

1 Testing for association under an additive model

1.1 Question One

Give the full Plink command used to perform your logistic regression analysis.

```
./plink --bfile gwa_clean --logistic --ci 0.95 --out gwa_clean
```

1.2 Question Two

Give the full Plink command used to obtain the genomic control inflation factor lambda. Also give the obtained GC inflation factor.

```
./plink --bfile gwa_clean --assoc --adjust
```

The estimate inflation factor is 1.15427.

2 QC for GWAS

2.1 Question Three

Give the full Plink command used to perform the association analysis accounting for age and sex.

```
./plink --bfile gwa_clean --logistic --covar gwa_clean.covar --covar-name SEX,AGE -
out gwa_clean_sexage
```

2.2 Question Four

Give the full Plink command you would use when you want to do a GWAS for age (so to use age as an outcome trait instead of the affection status (case-control))

```
./plink.exe --bfile gwa_clean --linear --pheno gwa_clean.covar --mpheno 2
```

3 Identifying individuals of divergent ancestry

3.1 Question Five

Give the full Plink command used to perform the association analysis accounting for the first 10 PCs.

```
./plink --bfile gwa_exclpop --logistic --hide-covar --ci 0.95 --covar gwa_exclpop.pca
.eigenvec --covar-number 3-12 --make-bed --out gwa_exclpop_pca10
```

4 Exploring results using web based tools

4.1 Question Six

*Which SNP is your most significantly associated SNP, and what is its genomic location according to dbSNP? Does this correspond with the location in your *.map (*.bim) file? Why (not)?*

To get the most significantly associated SNPs, I use the following command:

```
awk '$12 != "NA"'  gwa_exclpop_pca10.assoc.logistic | grep ADD | sort -k12 -g | head
```

The most significant hit is `rs2066844` with a significance of $4.308e^{-13}$ and an odds ratio of 2.462. According to dbSNP its genomic location is chromosome 16 on position 50745926. This corresponds with the genomic location in the map file.

4.2 Question Seven

Check the degree of LD between your most significant SNP and the second most significantly associated SNP. Give both SNP IDs; and the D' and r^2 value between them. Can you speculate on why the D' and r^2 values are so different? What does it mean?

The SNP IDs for the most significant and second most significantly associated SNPs are `rs2066844` and `rs2066847`. According to the tool the SNPs are in linkage equilibrium, with a D' of 1.0 and an r^2 of 0.0024. The fact that the D' is equal to 1 indicates that there is no evidence of recombination. Then why is the r^2 value so small? This is the case because the allele frequencies are not the same. The minor allele frequencies were 0.1161 and 0.06618 for the first and second most signifacnt SNPs, respectively. So when the rarer allele is present it is always inherited, but it is often the case that the common allele is not found with it (which gives the low r^2 value). From this we conclude that the SNPs are actually in linkage disequilibrium.

5 Research Paper

5.1 Methods

The SNPs were tested for association with the phenotype under an additive model in a logistic regression framework. The regression analysis was first performed on the entire dataset with a confidence interval of 0.95. The results were then visualised to get an overview of the results and see if there is any obvious systematic bias. The genomic control inflation factor was calculated to evaluate stratification or other genome-wide confounding factors in the data such as family structure, cryptic relatedness.

The visualisations and genomic control inflation factor from this model indicated that there was some systematic bias. To account for population stratification PCA was used identify population outliers using genotype data available from the HapMap project [2]. Significantly associated loci with large effect sizes were invesigated further.

5.2 Results

Manhattan and QQ plots for the model which does not account for population stratification were used to visualise potential bias (see Figure 1).

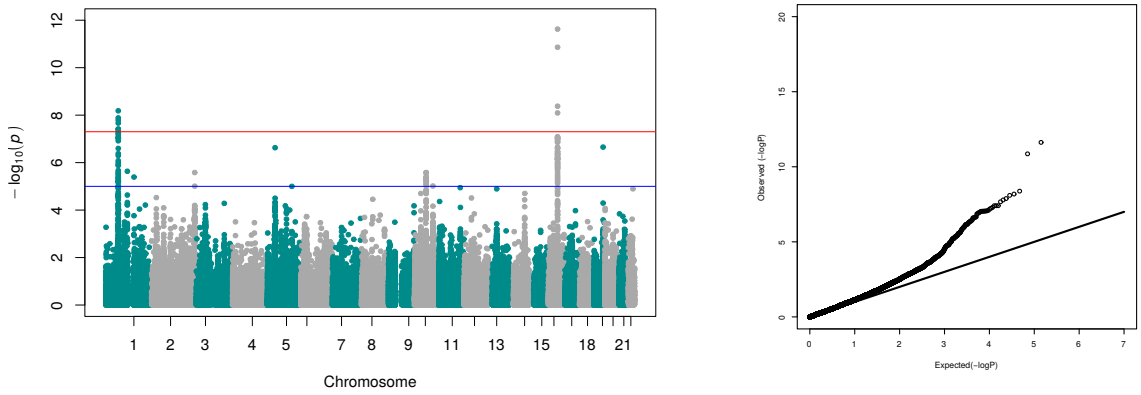


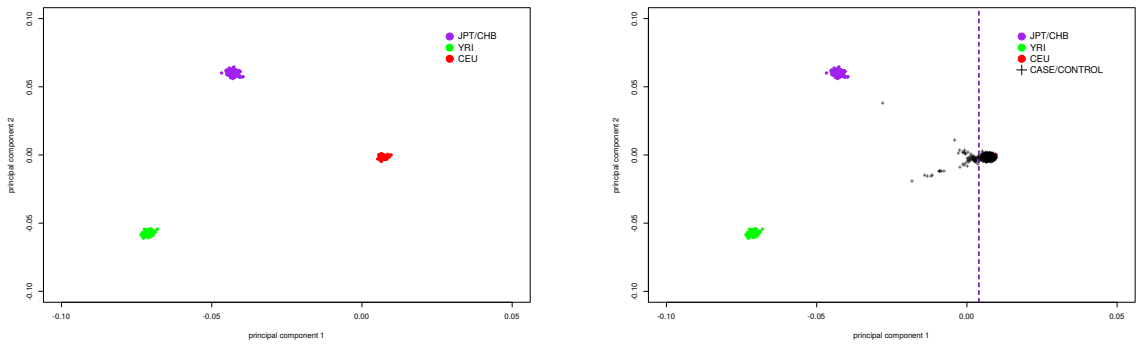
Figure 1: Manhattan and QQ plots for the basic model (not accounting for population stratification).

Both the Manhattan and QQ plots indicate some level of bias which should be corrected for. Although it is worth mentioning that the QQ plot is already quite good, and so any improvements may only be minor. The genomic control inflation factor λ was calculated as 1.15427.

PCA was used identify population outliers using genotype data available from the HapMap project. Scatter plots of the first two principal components for both the HapMap data and also the study data are given in Figure 2.

It is clear that this study population is from central europe but that there are a number of individuals from the case dataset which are outliers. These outliers were removed by specifying a threshold for PCA1 and removing individuals below this threshold (blue line in Figure **??**).

A new model was trained on the pruned data. Manhattan and QQ plots for the pruned data are given in Figures 3. These plots show little improvement over the plots given previously. The genomic control inflation factor λ was recalculated, giving a value of 1.14152 (a slight improvement from before).



(a) (b)

Figure 2: Scatter plots for both the HapMap data and the combined HapMap/study data. The blue line shows the cutof threshold used to remove outliers.

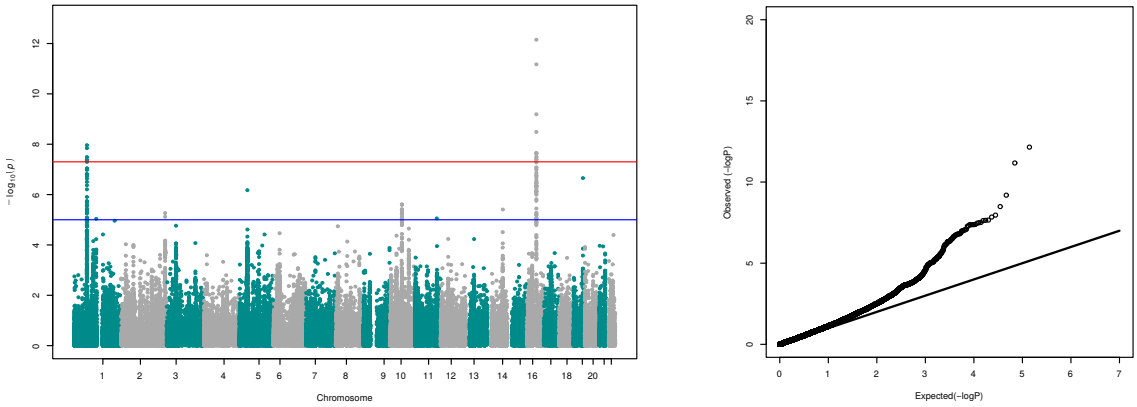


Figure 3: Manhattan and QQ plots for the model accounting for population stratification.

9861 SNPs were found to be significantly associated using a significance level of 0.05. The genomic locations of the most significantly associated SNPs was checked on NCBI dbSNP. The most significantly associated SNP was `rs2066844` which is found on the NOD2 gene. A number of other significantly associated SMPs were found to be located on the NOD2 gene. The significant associated SNP with the largest effect size was `rs111359611` which did not have any disease associations. The NOD2 gene and other significantly associated SNPs for with large effect sizes were searched on the GWAS catalogue to find previous disease associations [1]. Variations in the NOD2 gene have been associated with an increased risk of Crohn disease and other inflammatory disorders.

References

[1] Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research 47*, D1 (2019), D1005–D1012.

[2] Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Ch'ang, L.-Y., Huang, W., Liu, B., Shen, Y., et al. The international hapmap project.