MA5106 – Assessment 2

Shane Crinion | 13326096

You have been provided with data from a GWAS study, including a covariate file listing the age of each genotyped individual. You are tasked with carrying out a typical GWAS analysis and producing a report on your findings. This report should include

i) a detailed summary of the data,

ii) details of any tests carried out to determine appropriate thresholds for inclusion in the analysis,

iii) answers to some specific questions as detailed below.

Be sure to include the exact commands used to process the data as well as the first few lines of any intermediate files generated. Your final submission should be a single pdf file which includes your name and student ID on the first page.

Initial Set-Up:

To access the GWAS data for analysis, you must first connect to cluster. To this by making a connection (code 1) with smgate and then connecting to syd (code 2). Connecting to the cluster requires loading of module sge (code 3) and queuing for a root entry (code 4). Enter code 3 and 4 again to be assigned a node in the cluster. Change the directory to the file location (code 5). You then confirm the binary files are in the correct location (code 6) before loading the data.

Results: 3 binary files (gwas.bed, gwas.bim, gwas.fam) and 1 other (gwas.covar) files are found.

> 1. `ssh scrinion@smgate.nuigalway.ie`
> 2. `ssh syd.nuigalway.ie`
> 3. `module load sge`
> 4. `qrsh`
> 5. `cd /home/nextgen2018/data`
> 6. `ls`

Appropriate Thresholds

The appropriate thresholds were selected using the standard thresholds. This was selected due mainly due to the genotyping quality (Output 1). The genotyping quality was very high and the output below indicates that there is low levels of SNP removed. The total number of missing removed is 0.

```
plink --bfile gwas --make-bed --mind 0.1 --maf 0.05 --geno 0.1 --hwe 0.001 --out cleaned
```

QC output:

32193 MB RAM detected; reserving 16096 MB for main workspace. 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. 0 people removed due to missing genotype data (--mind). Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.983323. 0 variants removed due to missing genotype data (--geno). 10010 variants removed due to minor allele threshold(s) (--maf/--max-maf/--mac/--max-mac). 296092 variants and 4000 people pass filters and QC. Among remaining phenotypes, 2000 are cases and 2000 are controls. --make-bed to cleaned1.bed + cleaned1.bim + cleaned1.fam ... done. </code>

Specific Questions

1. QC tests.

i) How many SNPs is individual A2038 missing data for?

You must now load the data to analyse with PLINK software. You must load the data and obtain missing rates (code 7, output 1) statistics. This generates the imiss and lmiss, containing individual data and SNP data respectively. The imiss is inspected for patient A2038 (code 8) using `more` and contains information on missing rates.

Results: Indicate that individual A2038 is missing 5211 SNPs of 306102 genotyped, a rate of 1.7% percent.

> 7. `plink -bfile gwas --missing --out miss_stat`
>
> 8. `more miss_stat.imiss`

Output 1:

PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/ (C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to miss_stat.log. Options in effect: --bfile gwas --missing --out miss_stat

32193 MB RAM detected; reserving 16096 MB for main workspace. 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.983323. --missing: Sample missing data report written to miss_stat.imiss, and variant-based missing data report written to miss_stat.lmiss.

ii) For how many individuals is SNP rs2493272 missing?

This information is used to determine if the SNP can be used for analysis. The missing rate determines the proportion of the sample missing the genotype (code 9, output 2). If the proportion is high, it is inappropriate to use that SNP for further analysis and would not represent the sample accurately. 10 pct is common cut off.

Results: The missingness analysis shows 111 individuals are missing the SNP (table 1) which indicates that it is appropriate for further analysis.

> 9. `plink --bfile gwas rs2493272 --missing -out rs2493272missing`

Output 2

PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/ (C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to rs2493272missing.log. Options in effect: --bfile gwas --missing --out rs2493272missing --snp rs2493272

32193 MB RAM detected; reserving 16096 MB for main workspace. 1 out of 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.97225. --missing: Sample missing data report written to rs2493272missing.imiss, and variant-based missing data report written to rs2493272missing.lmiss.

Table 1

CHR	SNP	N_MISS	N_GENO	F_MISS
1	rs9651273	111	4000	0.02775

iii)Would you consider the missingness rates in general to be high or low? - what might this indicate about the data?

iii) Would you consider the missingness rates in general to be high or low? - What might this indicate about the data?

The missingness level for this sample is very low. The would usually indicate that high quality genotyping was performed. It could also suggest that there is little variability in the sample study.

2. Allele summary calculation.

i) Which is the minor allele for SNP rs4970357 and what is its frequency?

This information can be obtained using allele frequency summary statistics (code 10). This generates a file that states the minor and major allele and their frequency (output 3). The minor allele frequency is used to determine if the SNP occurs in enough of the population to be considered for the study. If the SNP occurs in more than 5% then it is considered a rare variant and will pass QC.

Results: The results of the generated .frq results (table 2) indicate that the minor allele, C, has a frequency of 0.05028 so could be considered for analysis. Considering the value is practically on the normal cut-off, the SNP could be removed from the study in more stringest analysis.

```
1. plink --bfile gwas --snp rs4970357 --freq --out snp2_frq_stat
```

Output 3

PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/ (C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to snp2_frq_stat.log. Options in effect: --bfile gwas --freq --out snp2_frq_stat --snp rs4970357

32193 MB RAM detected; reserving 16096 MB for main workspace. 1 out of 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.982. --freq: Allele frequencies (founders only) written to snp2_frq_stat.frq .

Table 2

SNP	A1	A2	MAF	NCHROBS
rs4970357	C	A	0.05028	7856

3. Basic association (Chi-Squared) test under different genetic models (DOM/REC etc.)

i) Under which genetic model does SNP rs9651273 show the smallest p-value?

Association analysis is performed to test a disease trait with allele frequencies and generates a model for dominant or recessive trait (code 11, table 3). The output generates statistical data including chi-squared statistics, degrees of freedom and asymptotic p-value. The results of the model indicate how likely the SNP would be disease causing if it was dominant and recessively occuring in the individual.

```
11. plink --bfile gwas --snp rs9651273 --model --out snp-info
```

Results: The results (table 3) indicate that there is a significantly low p-value in many of the models. The p-value is lowest for the dominant model which suggests that the SNP may be associated with a dominant disorder.

Output 4

PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/ (C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to snp-info.log. Options in effect: --bfile gwas --model --out snp-info --snp rs9651273

32193 MB RAM detected; reserving 16096 MB for main workspace. 1 out of 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.98575. 1 variant and 4000 people pass filters and QC. Among remaining phenotypes, 2000 are cases and 2000 are controls. Writing --model report to snp-info.model ... done.

Table 3

CHR	SNP	A1	A2	TEST	AFF	UNAFF	CHISQ	DF	P

1	rs9651273	A	G	GENO	322/1013/650	305/939/714	6.085	2	0.04773
1	rs9651273	A	G	TREND	1657/2313	1549/2367	3.995	1	0.04563
1	rs9651273	A	G	ALLELIC	1657/2313	1549/2367	3.892	1	0.04853
1	rs9651273	A	G	DOM	1335/650	1244/714	6.029	1	0.01407
1	rs9651273	A	G	REC	322/1663	305/1653	0.3062	1	0.58

4. Association testing with p-value correction for multiple testing

i) How many SNPs show a significant (<0.05) p-value under all of the multiple testing correction approaches?

The association testing using adjusted p-values is used to correct for multiple testing (code 12). The values generated have multiple properties relevant to association analysis. The values generated are statistical values that can be interpreted to understand if the SNP is actually significant or only appears significant due to chance. The most commonly used value is the Bonferroni correction where the p-value is divided by the number of tests.

Results The top SNPs can be found in Table 4 and were examined using `more` . The SNPs were ordered by p-value when the adjust command was input. The p-value was only to significant for the first 9 SNPs.

> 12. `plink --bfile gwas --assoc --adjust --out p-order`

Output 5

PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/ (C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to p-order.log. Options in effect: --adjust --assoc --bfile gwas --out p-order

32193 MB RAM detected; reserving 16096 MB for main workspace. 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. Using 1 thread (no multithreaded calculations invoked). Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.983323. 306102 variants and 4000 people pass filters and QC. Among remaining phenotypes, 2000 are cases and 2000 are controls. Writing C/C --assoc report to p-order.assoc ... done. --adjust: Genomic inflation est. lambda (based on median chisq) = 1.01019. --adjust values (306102 variants) written to p-order.assoc.adjusted .

> Note the Genomic inflation = 1.01019

Table 4

CHR	SNP	UNADJ	GC	BONF	HOLM	SIDAK_SS	SIDAK_SD	FDR_BH	FDR_BY
3	rs6802898	2.327e-20	3.599e-20	7.123e-15	7.123e-15	7.123e-15	7.123e-15	7.123e-15	9.409e-14
10	rs7901695	6.563e-12	8.366e-12	2.009e-06	2.009e-06	2.009e-06	2.009e-06	1.005e-06	1.327e-05
16	rs8050136	1.006e-08	1.192e-08	0.003078	0.003078	0.003074	0.003074	0.000778	0.01028
16	rs3751812	1.017e-08	1.205e-08	0.003112	0.003112	0.003107	0.003107	0.000778	0.01028
10	rs7904519	2.478e-08	2.913e-08	0.007586	0.007586	0.007558	0.007557	0.001423	0.01879
3	rs7615580	2.789e-08	3.274e-08	0.008537	0.008536	0.0085	0.0085	0.001423	0.01879
10	rs7903146	3.889e-08	4.551e-08	0.0119	0.0119	0.01183	0.01183	0.001435	0.01896
3	rs6768587	3.966e-08	4.639e-08	0.01214	0.01214	0.01207	0.01207	0.001435	0.01896
3	rs2028760	4.22e-08	4.934e-08	0.01292	0.01292	0.01283	0.01283	0.001435	0.01896

i) Is there evidence for population structure which may be confounding the analysis? Explain your answer.

Following a GWAS analysis, you must determine if there is any systematic bias in the results. Population stratification is the systematic differences in alleles found in ethnic groups due to different ancestry. The λgc or the genomic inflation factor will indicate this. The λgc is defined as "the median of the resulting chi-squared test statistics divided by the expected median of the chi-squared distribution" and is calculated from the analysis results. A lambda > 1 may indicate population stratification so our analysis results indicate that population structure could be confounded by systmatic diffferences due to population stratification.

The genomic inflation factor of 1.019 (Output 5) the population stratification is minimal as there is no effect of differences in ethnic groups.

5. Association test using logistic regression which includes sex and age as covariates

By adding covariates to an association test, the test results will account for any effect that is not due to genomic variations. The association test is used to determine if a certain allele is associated with a disease trait. The accounting of covariates can be used to find appropriate subsets in further analysis to avoid covariate bias. Logistic regression models allow for multiple covariates when testing SNP association with a disease.

The association test is performed in this analysis using sex and age as covariates (code 13, output 6). The logisitic regression is used to determine association of each SNP with trait which account for differences due to age and sex. Table 5 indicates the contribution of the age and sex covariates for each allele at the beginning of the file (using head()).

> 13. `--bfile gwas --covar gwas.covar --logistic sex --out logistic`

Output 6

PLINK v1.90b4.4 64-bit (21 May 2017) www.cog-genomics.org/plink/1.9/ (C) 2005-2017 Shaun Purcell, Christopher Chang GNU General Public License v3 Logging to logistic.log. Options in effect: --bfile gwas --covar gwas.covar --logistic sex --out logistic

32193 MB RAM detected; reserving 16096 MB for main workspace. 306102 variants loaded from .bim file. 4000 people (2000 males, 2000 females) loaded from .fam. 4000 phenotype values loaded from .fam. Using 1 thread (no multithreaded calculations invoked). --covar: 1 covariate loaded. Before main variant filters, 4000 founders and 0 nonfounders present. Calculating allele frequencies... done. Total genotyping rate is 0.983323. 306102 variants and 4000 people pass filters and QC. Among remaining phenotypes, 2000 are cases and 2000 are controls. Writing logistic model association results to logistic.assoc.logistic ... done.

Table 5

```
   CHR        SNP      BP A1 TEST NMISS    OR    STAT      P
1    1 rs3934834  995669  T  ADD  3818 1.029 0.38120 0.7031
2    1 rs3934834  995669  T  AGE  3818 1.002 1.11800 0.2635
3    1 rs3934834  995669  T  SEX  3818 1.012 0.19090 0.8486
4    1 rs3737728 1011278  A  ADD  3982 1.019 0.38670 0.6990
5    1 rs3737728 1011278  A  AGE  3982 1.002 1.09800 0.2721
6    1 rs3737728 1011278  A  SEX  3982 1.006 0.09898 0.9212
```

6. Visualise the results of your analysis using a Manhattan Plot

Manhattan plots are used to visualise signal dispersal across each chromosome to determine where most of the signal is coming from.. The Manhattan plot uses negative log P values on the y-axis and genomic location on the x-axis. The Manhattan plot can be visualised using Haploview and R.

The logistic association analysis was performed using code 13. Genomic data is stored on the university cluster. The logistic association test created using PLINK was then extracted from the cluster using code 14-15. This is then exported to local host using code 16-17. The data was first generated using R (Figure 1) and the using HaploView (figure 2) for comparision. The use of Haploview required only installation commands and memory changes.

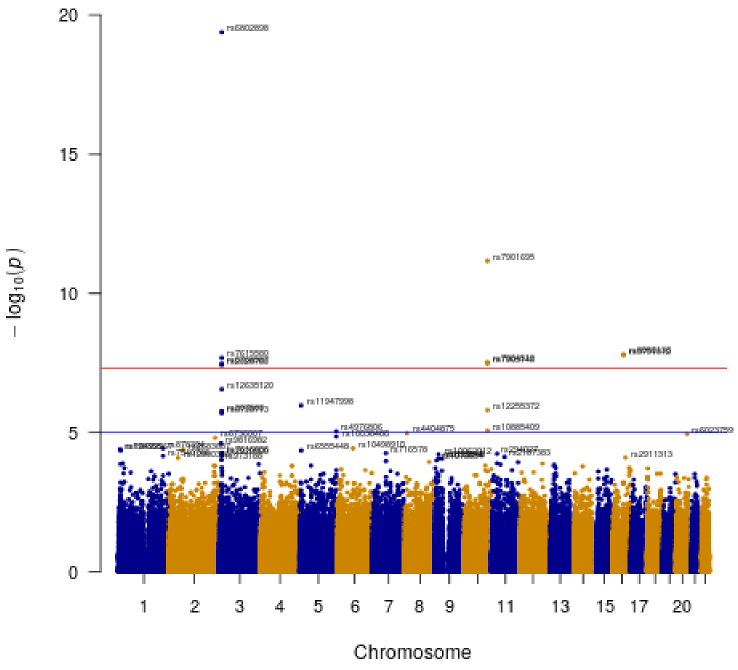
To create a Manhattan plot using R, the qqman package is used. The packaged is installed (code 18) and called from the library (code 19). The data is read in as a table following transfer of association test file to local host (code 20). The tail and head functions are used to ensure data is read in correctly (code 21). The png function is called to write the Manhattan plot to a new file called "manhattanRfinal.png" (code 23).

The generated R plot is more comprehensive than HaploView. Each command from the qqman Manhattan plot function (code 24) is as follows: **annotatePval:** show SNP names for SNPs with p less than 0.0001. **annotateTop = False:** not limited to 1 significant SNP per chromosome. **cex.axis:** reduces font size of axis. **col:** selects colour. **cex:** reduces the point size. The **red line** indicates genome significant SNPs. The **blue line** indicates the suggested signifcant. R stops recording plots using code 25.

Result: The qqplot generated plot shows that each of the most 9 genome wide significant SNPs. This correlates with those outlined in Table 4, indicating that the Manhattan plot is correct and has the same indicated results as through the PLINK analysis.

```
14. scp logistic.assoc.logistic scrinion@smgate:~/

15. scp gwas.bim scrinion@smgate:~/

16. sftp scrinion@smgate.nuigalway.ie

17. get logistic.assoc.logistic
```

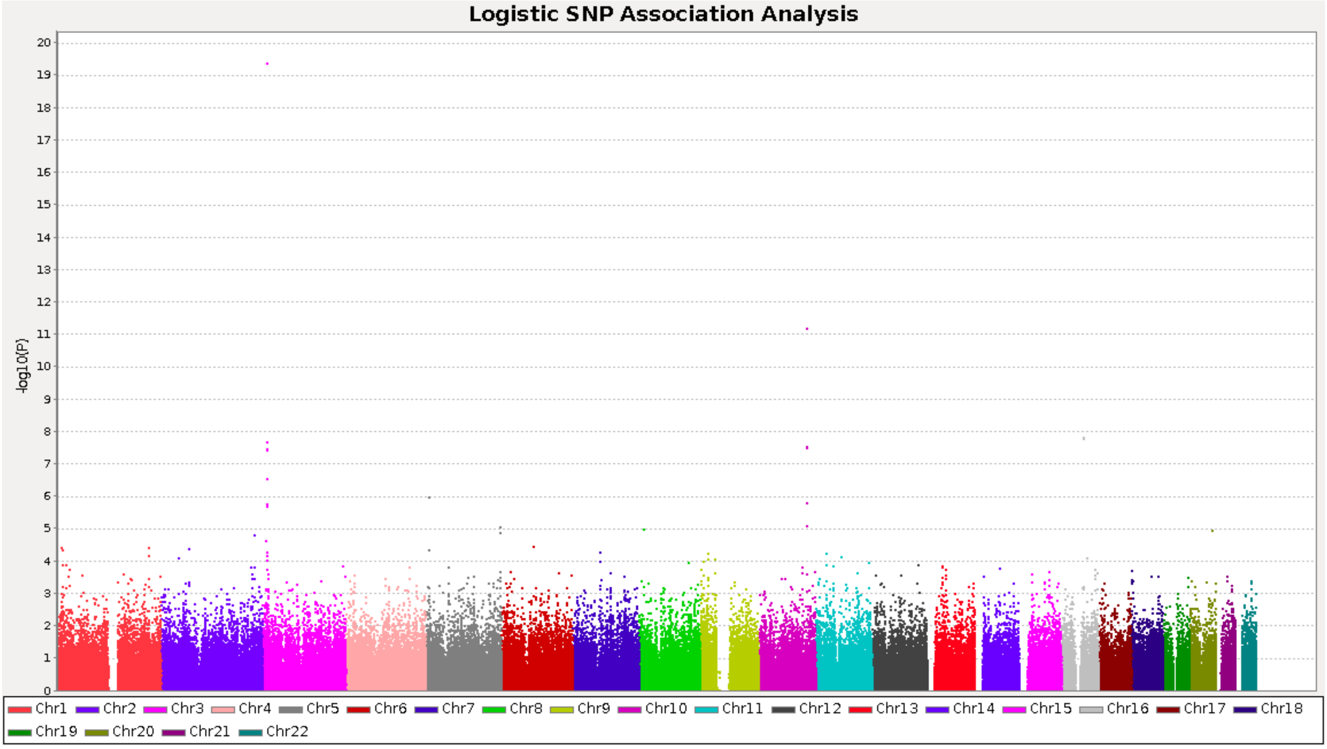
Figure 1: Manhattan Plot of Logistic Analysis Using R



Manhattan plot using R:

```
18. install.packages('qqman')

19. library('qqman')

20. snpdata<-read.table("logistic.assoc.logistic", header=T)

21. tail(snpdata)

22. head(snpdata)

23. png('manhattanRfinal.png')

24. manhattan(snpdata, annotatePval = 0.0001, annotateTop = FALSE, cex.axis = 0.9, col
= c("blue4", "orange3"), cex = 0.6)

25. dev.off()
```

Figure 2: Manhattan Plot of Logistic Analysis using HaploView file:///home/shane/Downloads/manhattanplot.png



7. Are any significant SNPs known to be associated with clinical outcomes?

The following information was obtained on the significantly asssociated SNPs. Most information was obtained using SNPedia and interestingly many SNPs with "no clinical significance" using dsSNP were actually found to be significant in peer-reviewed journals. All SNPs of clinical significance are found to be associated with type 2 diabetes.

rs6802898: Associated with type 2 diabetes. (https://www.ncbi.nlm.nih.gov/pubmed/17554300?dopt=Abstract)

rs7901695: Associated with type 2 diabetes. (https://www.ncbi.nlm.nih.gov/pubmed/17668382?dopt=Abstract)

rs8050136: A allele is associated with a higher risk of type 2 diabetes. (https://www.ncbi.nlm.nih.gov/pubmed/20057365?dopt=Abstract)

rs3751812: No clinical significance.

rs7904519: No clinical significance.

rs7615580: No clinical significance.

rs7903146: Strongly associated with type 2 diabetes. (https://www.ncbi.nlm.nih.gov/pubmed/17671651?dopt=Abstract)

rs6768587: No clinical significance.

rs2028760: No clinical significance.

Summary

Statistics: The results of this analysis indicate 8 SNPs are genome wide significant. Initial QC was done to remove any SNPs that may be of poor quality with `missing`, uncommon with `freq` and unpredictable with `hwe`.

Following initial QC, the number of SNPs missing in a particular individual was identified using the `imiss` file which holds stats for each individual. This also provided the **total genotyping rate** of 98% which indicated high quality genotyping. The `missing` command was used again to determine the SNP missingness rate. This could also be done manually using the lmiss file. The missingness is concluded as very low.

Allele summary calculation The `freq` file was used to determine the frequency of the minor allele of rs4970357 which was allele C and occured in 0.055% of individuals which indicates it is very close to cut off and may have been cut off in more stringent results.

Models Genetic models were created using the dominant and recessive trait models and the `model` command and indicated that the "A" allele of the SNP is most significantly associated with a dominant disorder phenotype.

Association testing with p-value correction for multiple testing This performed correction and then determines that association is true. We find that SNPs are significant following Bonferoni correction.

Association test using logistic regression which includes sex and age as covariates indicates from the p-value that there is minimal association (by no significant p-value) with sex or age.

Manhattan plot confirms the results from previously and visually represents the SNPs of significance.

Significant SNPs all SNPs found to be clinically significant are associated with type II diabetes.