**PRACTICAL 2: TESTING FOR ASSOCIATIONS**

**Introduction**

In Practical 1, you ran sample and SNP quality control procedures on your dataset, ‘genstudy’. In this practical you will again be using the ‘genstudy’ dataset, but to run analyses of association. We will first of all undertake an analysis of association where no adjustment is made for non-genetic variables (univariate analysis of association), and then an analysis of association where adjustment is made for relevant non-genetic variables, (multiple regression analysis).

**Software and datasets**

Following your QC procedures you will now have clean, QC’d PLINK files ready to run the analyses of association.

To run the analyses of association, we use a software package called ‘SNPTEST (Version 2)’ – a freely available package created for the analysis of genome-wide association studies.

The webpage for SNPTEST is available at <https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html#introduction>. The webpage includes a full description of the commands for running SNPTEST, as well as input and output file formats etc, and is a useful point of reference.

In order to use SNPTEST, input files need to be in a specific format – you will require one ‘.gen’ file which includes genotype data and one ‘.sample’ file which includes phenotypic data. It is possible to convert your QC’d PLINK files to ‘.gen’ and ‘.sample’ format using a program called ‘GTOOL’, however for the purpose of this practical the converted files are already provided for you. We have also removed all chromosome X SNPs from these files, since in terms of association we are only interested in SNPs on chromosome 10.

These files can be found in the workshop folder, and are called ‘genstudy\_qc.gen’ and ‘genstudy\_qc.sample’

A ‘.sample’ file has three parts (a) a header line detailing the names of the columns in the file, (b) a line detailing the types of variables stored in each column, and (c) a line for each individual including the information for that individual.

* *Take a look at the first few rows of the ‘genstudy\_qc.sample’ file by typing the following command:*

**head -n 4 genstudy\_qc.sample**

You will notice that the first row, the header line, includes the following column names:

|  |  |
| --- | --- |
| ID\_1 | This column gives a patient identifier for each individual. |
| ID\_2 | This column provides an option to give a second patient identifier for each individual. In our dataset, it is an exact duplicate of ID\_1. |
| missing | This column details the proportion of missing genotype data for the individual. |
| diseased | This column includes the phenotype (also known as outcome) for the individual. In this case, our phenotype is binary and represents whether the patient is diseased or not.  A sample file can include more than one phenotype, and the phenotype can be either binary or quantitative. |
| age | This column includes a covariate representing the individual’s age that may need adjusting for in the analysis of association (see multiple regression analysis below). This covariate is quantitative.  A sample file can include as many covariates as required, and covariates can be binary or quantitative. |
| bmi | This column includes a covariate representing the individual’s BMI that may need adjusting for in the analysis of association (see multiple regression analysis below). This covariate is quantitative. |
| family | This column includes a covariate representing whether the individual has family history of the disease. Again, it may need adjusting for in the analysis of association (see multiple regression analysis below). The covariate is binary, coded 1 (no) and 2 (yes). |
| sex | This column includes a covariate representing the individual’s gender. Again, it may need adjusting for in the analysis of association (see multiple regression analysis below). The covariate is binary, coded 1 and 2. |

The second row includes coding for the type of variable in each column, as follows:

|  |  |
| --- | --- |
| 0 | This is always the entry for the first three columns, since variable type is irrelevant for these columns |
| D | Discrete covariate (i.e. either binary or categorical). All discrete covariates are coded using positive integers (e.g. binary covariates are coded 1 and 2). In our ‘genstudy\_qc.sample’ file, we have ‘family’ and ‘sex’ as binary covariates and so these columns will have an entry of ‘D’ in the second row. Missing values are coded ‘NA’. |
| C | Continuous/quantitative covariate. In our ‘genstudy\_qc.sample’ file we have ‘age’ and ‘bmi’ as quantitative covariates and so these columns will have an entry of ‘C’ in the second row. Missing values are coded ‘NA’. |
| P | Continuous phenotype. Missing values are coded ‘NA’. |
| B | Binary phenotype, with controls coded ‘0’ and cases coded ‘1’. In our ‘genstudy\_qc.sample’ file our phenotype ‘diseased’ is binary and so this column will have an entry of ‘B’ in the second row. Missing values are coded ‘NA’. |

A ‘.gen’ file includes one row per SNP. The first five entries on each line should be as follows:

|  |  |
| --- | --- |
| SNP ID | This entry is usually used to denote the chromosome number. |
| rs number | This entry is a number which uniquely identifies the genotyped SNP. An rs number is an accession number used by researchers and databases to refer to specific SNPs. It stands for ‘Reference SNP cluster ID’. |
| Base pair position of the SNP | A value that described the SNP’s position on the chromosome. |
| Allele coded A (see below for further explanation of this) | The entry here will be A, C, T or G i.e. corresponding to the four possible nucleotides. |
| Allele coded B (see below for further explanation of this) | The entry here will be A, C, T or G i.e. corresponding to the four possible nucleotides. |

The remaining entries on each row represent the genotype of each individual at the SNP. Each individual will have three entries, representing their probability of having the ‘AA’, ‘AB’ and ‘BB’ genotypes respectively. As we do not have any imputed SNPs in the current dataset (we will be looking at this in Practical 3), the probability for each genotype will be either 0 or 1. So, an individual with genotype AA will be coded ‘1 0 0 ‘, with genotype AB will be coded ‘0 1 0’, and with genotype BB will be coded ‘0 0 1’. If genotype is missing for an individual, all three entries are set to ‘0’.

* *Take a look at the first row of the file ‘genstudy\_qc.gen’ by typing the following command:*

**head -n 1 genstudy\_qc.gen**

**Univariate analyses of association**

The command to run univariate analyses of association in SNPTEST includes the following elements:

|  |  |
| --- | --- |
| -data | This specifies the input datasets – typically a .sample file and a .gen file |
| -o | This specifies the name of your output dataset – i.e. a name for the file where your results will be written |
| -pheno | This specifies the name of the phenotype (also known as outcome) you are wishing to test – you may have more than one phenotype to be tested in your .sample file |
| -frequentist | This specifies the genetic model (mode of inheritance) you wish to assume in your analysis of association. It uses numeric coding as follows: 1=Additive; 2=Dominant; 3=Recessive; 4=General (i.e. no mode assumed). |
| -method | This is only of importance when we are dealing with imputed, as opposed to genotyped SNPs, and we will be considering this element of the SNPTEST command in Practical 3. For the purpose of this practical, we always set method to the option ‘threshold’. |

Therefore, if we wish to run univariate analyses of association between all SNPs in our ‘genstudy\_qc’ file and our phenotype ‘diseased’, assuming an additive mode of inheritance, we would use the following SNPtest command:

**snptest \  
-data genstudy\_qc.gen genstudy\_qc.sample \  
-o genstudy\_qc\_univariate\_add.out \**

**-pheno diseased \  
-frequentist 1 \  
-method threshold**

You will note that within the command box above, and in all others used in this practical, we end most lines with a '\' . This is not part of the command; we just use this notation to split each example command over multiple lines to make it easier to read. It is actually equivalent to putting all elements of the command onto a single line, each separated by a space i.e. the above would become:

**snptest -data genstudy\_qc.gen genstudy\_qc.sample -o genstudy\_qc\_univariate\_add.out -pheno diseased -frequentist 1 -method threshold**

You can use either version of the command into Unix and it will work in the same way – for the former it will simply ignore the ‘\’ characters.

* *Run a univariate analysis of association between each SNP and your phenotype ‘diseased’, assuming an additive mode of inheritance by copying the above command.*

Your results will automatically be stored in the ‘genstudy\_qc.out’ file.

* *Take a look at the first 2 lines of your output file by typing the following command:*

**head -n 20 genstudy\_qc\_univariate\_add.out**

You will notice that there are very many columns included in the output file – those of most importance are as follows:

|  |  |
| --- | --- |
| rsid | The unique SNP identifier for each SNP analysed |
| pos | Base pair position of the SNP |
| cases\_AA | Number of cases with the AA genotype |
| cases\_AB | Number of cases with the AB genotype |
| cases\_BB | Number of cases with the BB genotype |
| controls\_AA | Number of controls with the AA genotype |
| controls\_AB | Number of controls with the AB genotype |
| controls\_BB | Number of controls with the BB genotype |
| all\_maf | Minor allele frequency overall |
| cases\_maf | Minor allele frequency in cases |
| controls\_maf | Minor allele frequency in controls |
| missing\_data\_proportion | Proportion of individuals with missing genotypes at the SNP |
| het\_OR | Odds ratio heterozygotes vs wild-type homozygotes |
| het\_OR\_lower | Lower limit 95% confidence interval for above odds ratio |
| het\_OR\_upper | Upper limit 95% confidence interval for above odds ratio |
| hom\_OR | Odds ratio mutant-type homozygotes vs wild-type homozygotes |
| hom\_OR\_lower | Lower limit 95% confidence interval for above odds ratio |
| hom\_OR\_upper | Upper limit 95% confidence interval for above odds ratio |
| all\_OR | Odds ratio per allele |
| all\_OR\_lower | Lower limit 95% confidence interval for above odds ratio |
| all\_OR\_upper | Upper limit 95% confidence interval for above odds ratio |
| diseased\_frequentist\_add\_thresh\_pvalue | p-value for association test |

* *Run an univariate analysis of association between each SNP and your phenotype ‘diseased’, this time assuming a dominant mode of inheritance, saving your results in a new output file ‘genstudy\_qc\_univariate\_dom.out’.*

Due to the large number of output columns and the large number of SNPs investigated, it is useful to look at our results in summary format. A useful way of doing this is by preparing a Manhattan plot to give a graphical representation of our results. Essentially, this is a plot of p-value versus base pair position for each SNP and is a useful way of identifying significantly associated SNPs.

We have a shell script available to run which will prepare a Manhattan plot, however first of all it is necessary to re-format our output file into a format which is appropriate for the shell script. Commands for re-formatting the data are provided below, step-by-step:

STEP 1: Add chromosome number to each row

**awk 'NR>11{$1=10; print}' genstudy\_qc\_univariate\_add.out >genstudy\_qc\_new.out**

STEP 2: Select only columns of interest, and label them

**echo "CHR SNP BP P" > genstudy\_qc\_Manhattan.txt;**

**awk '{print $1,$2,$4,$42}' genstudy\_qc\_new.out >>genstudy\_qc\_Manhattan.txt**

STEP 3: Create a new file (“genstudy\_qc.unimp.snp”) which lists all actually genotyped SNPs (Note: for this practical, this will be a list of all SNPs as they are all genotyped. In Practical 3 it will be a subset of SNPs as some will be imputed)

**awk ' $9==1{print $2}' genstudy\_qc\_univariate\_add.out > genstudy\_qc.unimp.snp**

Now that the output data has been re-formatted as necessary for the Manhattan plot shell script, we can run the script. The command for doing this is:

**sh QuickManhattan.sh**

Once this command is entered, you will be prompted for answers to four questions as follows:

|  |  |
| --- | --- |
| Study Name ? | For example, genstudy\_qc |
| Assoc file ? | Here, you would put the filename for the file you created by re-formatting your output file e.g. genstudy\_qc\_Manhattan.txt |
| Please specify a shorthand for this analysis | This needs to be something to help you remember which analysis your results relate to. For example, it could be ‘Univariate\_Additive’. |
| Finally, please specify a title for your Manhattan plot | This needs to be something to identify which analysis your results relate to. For example, it could be ‘Univariate analyses assuming additive model’ |

* *Create a Manhattan plot for the results of your univariate analysis assuming an additive mode of inheritance, by first of all following steps 1-3 above and then running the Manhattan plot shell script. Please call your analysis ‘Univariate Additive’ and give your plot the following title: ‘Univariate analyses assuming additive model’*
* *Take a look at your Manhattan plot – are there any statistically significant associations ?*

It is also useful to prepare a Q-Q plot of your results to ensure that there is no apparent genomic inflation in your results, which would suggest population substructure. You can do this by using the following script in R. (Note: you will need to type in R on your command line first to take you to the R environment. Type q() to escape from that environment.

**pvals <- read.table("genstudy\_qc\_Manhattan.txt", header=T)**

**observed <- sort(pvals$P)**

**lobs <- -(log10(observed))**

**expected <- c(1:length(observed))**

**lexp <- -(log10(expected / (length(expected)+1)))**

**pdf("qqplot.pdf", width=6, height=6)**

**plot(c(0,7), c(0,7), col="red", lwd=3, type="l", xlab="Expected (-logP)", ylab="Observed (-logP)", xlim=c(0,7), ylim=c(0,7), las=1, xaxs="i", yaxs="i", bty="l")**

**points(lexp, lobs, pch=23, cex=.4, bg="black")**

**dev.off()**

You can also extract a list of the SNPs giving the lowest p-values, so that you can investigate them further. This can be done by filtering based on the p-value column in your “genstudy\_qc\_Manhattan.txt” file using the two lines of code below. (Please note that the threshold on which you are filtering can be amended by changing the value of ‘0.0001’ in the first line of the code. Please also note that the second line of code is only required to filter out SNPs were it was not possible to undertake an analysis of association, in which case the p-values are given as ‘-1’):

**awk ' $4 <0.0001' genstudy\_qc\_Manhattan.txt > lowest\_ps\_univariate**

**awk ‘$4>0’ lowest\_ps\_univariate > lowest\_ps\_univariate\_new**

* *Extract a list of p-values that are less than 0.0001 from your univariate association analysis assuming an additive mode of inheritance. Take a look at the list (using the ‘cat’ unix command as you have done previously), and compare it to your Manhattan plot.*

**Multiple regression analyses**

In the ‘genstudy.sample’ file you will notice four columns to the right of the phenotype column ‘diseased’. These are labelled ‘age’, ‘bmi’, ‘family’ and ‘sex’, and represent four variables believed to be of potential interest in terms of being associated with our disease. ‘age’ represents the participant’s age and is a continuous variable, ‘bmi’ represents their body-mass index and is a continuous variable, ‘family’ represents family history of the disease and is a binary variable (1 meaning no family history; 2 meaning family history ), and ‘sex’ represents gender and is a binary variable (1 meaning male; 2 female).

As these may be associated with disease, it is important that we consider adjusting for them in our analyses of association with each SNP. To decide whether to adjust for each of these variables, we first of all need to test whether they are univariately associated with our phenotype ‘diseased’. As our phenotype is binary, we can use Student’s t-test to test for association with age and bmi, and the chi-square test to test for association with family history and sex. We can do this in the programming language ‘R’, using the commands ‘t.test’ and ‘chisq.test’.

* *Test for association between each of the variables age, bmi, family history and sex by running the R code below(remember you will need to enter the R environment first, as explained above).*

**data=read.table("genstudy\_qc.sample",header=T)**

**data=data[-1,]**

**attach(data)**

**res\_age=t.test(as.numeric(age)~diseased)**

**res\_bmi=t.test(as.numeric(bmi)~diseased)**

**res\_family=chisq.test(family,diseased)**

**res\_sex=chisq.test(sex,diseased)**

**res\_age$p.value**

**res\_bmi$p.value**

**res\_family$p.value**

**res\_sex$p.value**

**sex$p.value**

* *Which variables give a p-value<0.05 ?*

Once we have undertaken analyses of association between each non-genetic variable and outcome, we are ready to run our SNP association analyses using a multiple regression approach. This can again be done in SNPtest, but this time we have to specify which clinical covariates to adjust for.

This can easily be achieved by adding the element ‘-cov\_names’ followed by the clinical variable names to your command. Assuming that we have decided to adjust for all clinical variables giving p<0.05 in our univariate analysis, we will find that in our example we need to adjust for age and family history. So, again assuming an additive mode of inheritance our command will be:

**snptest \  
-data genstudy\_qc.gen genstudy\_qc.sample \  
-o genstudy\_qc\_adjusted\_add.out \**

**-pheno diseased \  
-frequentist 1 \  
-method threshold \**

**-cov\_names age family**

* *Run multiple regression analysis of association between each SNP and your phenotype ‘diseased’, assuming an additive mode of inheritance by running the above command.*
* *Take a look at the first 2 lines of your output file by typing the following command:*

**head -n 20 genstudy\_qc\_adjusted\_add.out**

You will see that the columns in this output file are very similar to those in the output file for the univariate analysis of association.

* *Prepare a Manhattan plot and QQ-plot of the results from your multiple regression analysis by adjusting the code used for plotting the univariate results (remember to change the SNPtest output file name to ‘genstudy\_qc\_adjusted\_add.out’).*
* *Obtain a list of the lowest p-values from this analysis, and compare them with your lowest p-values from the univariate analysis.*