Practical: Quality control of genome-wide association study data

In this practical, we will go through the steps in performing quality control (QC) of genotype data from a simulated genome-wide association study of 1000 cases and 1000 controls, typed for 317,503 autosomal and X chromosome SNPs. We will begin by performing sample QC, including calculation of call rates, heterozygosity, and sex discordance. We will then perform SNP QC, including calculation of call rates and deviation from Hardy-Weinberg equilibrium. We will be working mainly with PLINK software (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml) which is widely used in analysis of GWAS data and freely available to download.

The general command structure is:

```
plink --noweb --bfile input.data --options --out output.data
```

This practical is based on "Data quality control in genetic case-control association studies" (Anderson *et al*. 2010, Nature Protocols 5: 1564-73). Handout and data were edited from http://www.well.ox.ac.uk/dtc/.

Before starting the practical, you will need to unpack the genome-wide association binary ped files and additional analysis script files. You can do this by typing the following command in the terminal:

```
tar -xvzf GWAS_practical_1.tar.gz

mv fail-IBD-QC.txt GWAS_practical_1
cd GWAS_practical_1
```

We also need to load the module for plink using:

```
module load plink
```

Part A: Sample QC

Identification of individuals with discordant sex information

Run:

```
plink --noweb --bfile raw-GWA-data --check-sex --out raw-GWA-data
```

This command will calculate the mean homozygosity rate across X-chromosome markers for each individual in the study, outputting raw-GWA-data.sexcheck. We expect males to be homozygous at every locus as they only have one X, but females will have an appreciable number of heterozygous loci. We can produce a list of individuals with discordant sex data by typing:

```
grep PROBLEM raw-GWA-data.sexcheck > raw-GWA-data.sexprobs
```

The 'grep' command is a very useful unix command that searches a file (or directory) for a certain pattern. In this case we are searching for 'PROBLEM'. The '>' sign diverts the output of the grep command to the file 'raw-GWA-data.sexprobs' rather than printing it to the screen.

Use 'less' to look at the file 'raw-GWA-data.sexprobs'. Column 3 denotes ascertained sex and column 4 denotes sex according to genotype data. When the homozygosity rate is more than 0.2 but less than 0.8, the genotype data are inconclusive regarding the sex of an individual and these are marked in column 4 by "0".

In general, any discordances in sex should be reported to study co-ordinators to double check records for errors (e.g. what is the name of the individual? If it is Peter they are likely to be male). In situations in which discrepancy cannot be resolved, the individuals should be excluded from downstream analysis.

Identification of individuals with elevated missing data rates

Run:

```
plink --noweb --bfile raw-GWA-data --missing --out raw-GWA-data
```

This command will create the files "raw-GWA-data.imiss" and "raw-GWA-data.lmiss". The fourth column in the file "raw-GWA-data.imiss" (N_MISS) denotes the number of missing SNPs and the sixth column (F_MISS) denotes the proportion of missing SNPs per individual. A widely used threshold for call-rate is 95% so we will now create a file containing IDs of individuals who do not reach this threshold. Run:

```
awk '$6 > 0.05' raw-GWA-data.imiss > remove.imiss
```

'awk' is a very useful command in unix. Here we are using it to extract all lines with a number > 0.05 in the 6th column. This is all individuals with a call rate of < 95%. These lines are then output to the file 'remove.imiss' which we can use to exclude these samples from downstream analyses.

Identification of duplicated or related individuals

To minimize computational complexity, one would normally create an "independent" set of SNPs to generate the identity by descent (IBS) matrix (i.e. quantify the amount of relatedness between each pair of individuals). This can be done by "pruning" the data so that no pair of SNPs (within a given genomic interval) has an r^2 value (linkage disequilibrium measure) greater than a given threshold, typically chosen to be 0.2. Typically, we would also exclude regions of strong LD, such as the MHC, which are listed in the file "high-LD-regions.txt".

The process of generating an IBS matrix is computationally intensive and so takes a little while to run. It has therefore already been run for you using the command: `plink --noweb --bfile raw-GWA-data --extract raw-GWA-data.prune.in --`

genome --out raw-GWA-data, where raw-GWA-data.prune.in contains only independent SNPs (not in LD). We then ran a script file to identify all pairs of individuals with IBS > 0.185 (a threshold typically used to identify first degree relatives (i.e. siblings, parents, children)) which output the ID of the individual from the pair with lowest call rate (from the previously created file "raw-GWA-data.imiss"). We will then remove the individual with the lowest call rate. This script file can easily be adapted to other data sets and thresholds for "related" individuals. The file "fail-IBD-QC.txt" contains these individuals and can be used to exclude these samples from downstream analyses.

Removal of all individuals failing sample QC

In the terminal, type the following command to concatenate all files listing individuals who have failed previous QC steps:

```
cat raw-GWA-data.sexprobs remove.imiss fail-IBD-QC.txt | grep -v 'FID' | sort | uniq > fail-qc-inds.txt
```

The 'cat' unix/linux command reads in a list of files and prints the content of each in turn. We then pipe '|' (i.e. send the information from the cat command) to a grep command which removes any header lines (containing FID). Finally, we sort and identify only the unique records (using 'uniq') and send the output from all these commands to the file 'fail-qc-inds.txt'.

The file 'fail-qc-inds.txt' should now contain a list of unique individuals failing the previous QC steps. To remove them from the data set, type the following command:

```
plink --noweb --bfile raw-GWA-data --remove fail-qc-inds.txt --make-bed --out clean-inds-GWA-data
```

We will now use the binary ped file 'clean-inds-GWA-data' for subsequent SNP QC analyses.

Part B: SNP QC

Identification of all SNPs with an excessive missing data rate

To calculate the missing genotype rate for each SNP, type the following command at the shell prompt:

```
plink --noweb --bfile clean-inds-GWA-data --missing --out clean-inds-GWA-data
```

The third column in the file 'clean-inds-GWA-data.lmiss' (N_MISS) denotes the number of missing genotypes and the fifth column (F_MISS) denotes the proportion of missing genotypes per SNP.

To remove SNPs with call rate less than 95%, simply add the "--geno 0.05" option to the PLINK command line. We will do this below when creating our final cleaned data set.

Test SNPs for different genotype call rates between cases and controls

To test for differential call rates between cases and controls for each SNP, type the following command at the shell prompt:

```
plink --noweb --bfile clean-inds-GWA-data --test-missing --out clean-inds-GWA-data
```

The output of this test can be found in the file 'clean-inds-GWA-data.missing'. Run the following script to highlight all SNPs with significant differences in case and control call rates ($p<10^{-5}$) from this output file:

```
perl run-diffmiss-qc.pl clean-inds-GWA-data
```

The command creates a file called "fail-diffmiss-qc.txt", which we will use below to exclude these SNPs from downstream association analyses.

Removal of all SNPs failing QC

To remove low-quality SNPs, type the following command at the shell prompt:

```
plink --noweb --bfile clean-inds-GWA-data --exclude fail-diffmiss-qc.txt --geno 0.05 --hwe 0.00001 --make-bed --out clean-GWA-data
```

In addition to removing SNPs identified with differential call rates between cases and controls, this command removes SNPs with call rate less than 95% with --geno option and deviation from HWE ($p<10^{-5}$) with the --hwe option. One additional option is --maf which excludes all SNPs with minor allele frequency less than a specified threshold.

This command will produce cleaned binary ped files for downstream association analyses: "clean-GWA-data.bed", "clean-GWA-data.bim" and "clean-GWA-data.fam".

Part C: Association testing

Test for association with disease using logistic regression

We are now going to test for an association with disease in our final clean data using logistic regression. Run the command:

```
plink --noweb --bfile clean-GWA-data --logistic --out logistic.analysis
```

This will output the file 'logistic.analysis.assoc.logistic' which will be a text file in the format:

```
CHR             Chromosome
SNP             SNP identifier
BP              Position (base-pair)
A1              Minor allele
TEST            Code for the test
NMISS           Number of individuals included in analysis
OR              Odds ratio
STAT            Coefficient t-statistic
P               P-value
```

Which variant is statistically most associated with disease in your file (look for the lowest P-value)? Does it pass the accepted threshold for genome-wide significance ($<5x10^{-8}$)? Finally, what is the size of the effect of this variant on disease risk (the odds ratio?).

To make it easier you might want to run the following command to identify only SNPs with a P-value < 0.0005 and print them to the file logistic.lowpsnps:

```
awk '$9 < 0.00005' logistic.analysis.assoc.logistic > logistic.lowpsnps
```

The command can also incorporate information on sex, age and other variables (i.e. smoking status) if you want to take account of these in your regression model. For example using the option '--covar my_cov_file.txt' where my 'my_cov_file.txt' is a tab delimited file containing a row for each sample and a column for each covariate. The output P-value is then adjusted for each covariate. Alternatively add the --sex option to condition on sex.

If you want to know more, PLINK is very well documented on their web page (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml)

Well done, you have now completed sample QC and a simple association analysis on GWAS data! ☺