**PRACTICAL 4: Population structure in GWAS**

In this practical, we will evaluate the evidence for population structure in the cleaned dataset. We will test for association under an additive model, and generate a QQ plot and evaluate the genomic control inflation factor. We will then perform multi-dimensional scaling of the genotype data to investigate structure within the population. Finally, we will repeat the test for association under an additive model, but this time adjusting for eigenvectors from the multi-dimensional scaling as covariates. We will then re-generate a QQ plot and re-evaluate the genomic control inflation factor to evaluate the evidence of residual population structure.

The data for this practical are stored in the same directory as you have used for previous practicals. If you are not already in this directory, you should begin by moving to it using the command:

cd ~/workshop

1. **Testing for association under an additive model**

Begin by testing for association under an additive model using the PLINK command, but without adjustment for any covariates. We use command:

plink --noweb --bfile EGCUT.clean --logistic --ci 0.95 --out plink.additive

You can investigate the impact of population structure by generating a QQ plot of observed p-values against those that would be expected under the null hypothesis of no association. In the absence of population structure, we would expect most points to lie on the y=x line. You can do this using the command:

R --vanilla --slave --args plink.additive.assoc.logistic ADD qq\_1.pdf < qqPlot.R

The QQ plot also indicates the genomic control inflation factor, which would be expected to be close to 1 in the absence of population structure. What do you conclude?

1. **Performing multi-dimensional scaling**

Multi-dimensional scaling in PLINK is performed in three steps: (i) identifying a subset of “independent” common SNPs (minor allele frequency of at least 5%) that are not in linkage disequilibrium with each other; (ii) calculating the relatedness between each pair of individuals using the set of independent SNPs which is a measure of genetic similarity; and (iii) perform multi-dimensional scaling using the relatedness matrix. The first two of these steps are identical to those used when identifying related individuals in Practical 1.

To perform these three steps, use the following commands (type each command on one continuous line), which might take some time to run, so please be patient:

plink --noweb --bfile EGCUT.clean --maf 0.05 --indep-pairwise 50 5 0.05

plink --noweb --bfile EGCUT.clean --extract plink.prune.in --Z-genome

plink --noweb --bfile EGCUT.clean --read-genome plink.genome.gz --cluster --mds-plot 2

Note that the option --Z-genome produces a compressed genome file and saves on storage. The option --mds-plot specifies how many eigenvectors from the multi-dimensional scaling to summarise in the output file. The eigenvectors are output to the file plink.mds. Have a look at the format of the file using the command:

head plink.mds

The output file has one row per individual, and provides the identifier used in the PLINK family file, together with the first two eigenvectors (columns C1 and C2). You can plot the first two eigenvectors against each other using the command:

R --vanilla --slave --args plink.mds EGCUT.clean.fam mds.pdf < mds.R

In this plot, each point corresponds to an individual. Do you observe any obvious clusters of individuals that might reflect population structure?

1. **Testing for association under an additive model with adjustment for confounding**

To account for population structure in our analysis, we can repeat our test of association under and additive model, but this type adjusting for eigenvectors from the multi-dimensional scaling as covariates. You can do this using the command (type on one continuous line):

plink --noweb --bfile EGCUT.clean --logistic --covar plink.mds --covar-name C1,C2 --ci 0.95 --out plink.additive.mds

You can investigate the impact of population structure that is not accounted for by the first two eigenvectors from multi-dimensional scaling by generating a QQ plot with the command:

R --vanilla --slave --args plink.additive.mds.assoc.logistic ADD qq\_2.pdf < qqPlot.R

What do you conclude?

Recall how to find the lead SNP from the additive model from your association analyses in Practical 2. You can obtain association summary statistics for the SNP using the command:

grep *SNPID* plink.additive.mds.assoc.logistic

In this command, you replace *SNPID* with the identifier of the lead SNP. What impact has adjustment for the two eigenvectors had on the association signal for this SNP? Are the two eigenvectors associated with case-control status?