Day 3 Practical: Adjusting for Population structure

In this practical we will analyse a simulated GWAS dataset to check for population stratification and employ two different methods to adjust for structure in association studies.

In this session, we will go through the following steps:

1. Perform a test of association at each SNP using PLINK, and evaluate QQ plots to assess evidence for the presence of population structure.
2. Perform a test of association at each SNP using PLINK and employ Genomic Control to assess if the population structure has been accounted for.
3. Perform a test of association at each SNP using PLINK with PCs as covariates and and re-evaluate to assess if the population structure has been accounted for.

Input Dataset

Please find the input dataset in the Plink format here in your VM

```
cd Day3_Popstructure
```

1. Testing for evidence of population structure from association results.

a. Assuming you have a qc-ed dataset (following the steps you have learnt in Day2), we start with the association analysis

```
plink --bfile demo_data --allow-no-sex --assoc --out assoc_results
```

b. View you result file.

```
head assoc_results.assoc
```

⇒ Can you find the column with p-value ?

c. Check whether you have any genome-wide significant hits (i.e. *p-value* $< 5X10^{-8}$).

```
awk '{if($9<0.00000005) print }'  assoc_results.assoc | wc -l
```

⇒ How many SNPs are significantly associated to the trait?

d. Generate a Manhattan Plot

```
Rscript Manhattan_plot.R  assoc_results.assoc
assoc_results.manhattan.jpeg
```

⇒ How does the Manhattan plot look? Do you think something is not right even though you have qc-ed for your samples and SNPs stringently.

e. Generate a QQ Plot

```
Rscript QQ_plot.R  assoc_results.assoc assoc_results.qqplot.jpeg
```

Early deviation from the expected line, as well as Manhattan plot clearly tells you have a very strong population structure. How to deal with it?

So, let's start with adjusting for Genomic Control

2. GC based correction

a. Run the assoc command once again, but this time use an "adjust flag" which generates another file containing several basic multiple testing corrections for the raw *p-values*, including genomic-controlled p-values

```
plink --bfile demo_data --allow-no-sex --assoc --adjust  --out
assoc_results
```

b. View the new file

```
head assoc_results.assoc.adjusted
```

⇒ Can you find the column with *p-value*? The column with header 'UNADJ' provides the raw *p-value* and the with header 'GC' provides the genomic-controlled *p-value*

c. Check if there are any genome-wide significant hits after GC based correction

```
awk '{if($4<0.00000005) print }' assoc_results.assoc.adjusted | wc -
l
```

d. Generate a QQ Plot to check if we have controlled for population structure successfully

```
sed 's/GC/P/g' assoc_results.assoc.adjusted >
assoc_results.assoc.adjusted1

Rscript QQ_plot.R assoc_results.assoc.adjusted1
assoc_results.GCadjusted.qqplot.jpeg
```

Although, a bulk of the structure is gone, we still see a deviation from the expected line before 4.

Next, we will use a PCA-based correction to control for population structure.

3. Principal component (PC) based correction

a. It’s always a good practice to run PCA on LD pruned dataset, so that the analysis in not biased by disproportionately high LD in some regions.

```
plink --bfile demo_data --indep-pairwise 50 5 0.5

plink --bfile demo_data --extract plink.prune.in --pca 'header' --
out demo_data.pca
```

b. View the files generated

```
head demo_data.pca.eigenvec | cut -d " " -f 1-7

head demo_data.pca.eigenval
```

c. The next step is to visualise the PCA plot.

```
#Create input for PCA plot

echo IID Pheno > demo_data.phe
awk '{print $2,$6}' demo_data.fam  >> demo_data.phe

paste demo_data.pca.eigenvec demo_data.phe  | awk '{if ($2==$23)
print $24,$2,$3,$4,$5,$6,$7,$8,$9,$10,$11,$12}' | sed
's/^1/Control/g' | sed 's/^2/Case/g' > demo_data.pca.input
```

The next step is run the R script that uses the ‘demo_data.pca.input’ as input

```
Rscript plot_PCA.R
```

So, you can see that our dataset is highly structured (hopefully you will not see this in real life scenario ☺). In real GWAS you might also choose to remove outliers based on PCs.

d. We will use the first few PCs from the ‘demo_data.pca.eigenvec’ file as a covariate file to run the logistic regression analysis.

```
plink --bfile demo_data --allow-no-sex --logistic hide-covar --covar
demo_data.pca.eigenvec  --covar-name PC1,PC2,PC3,PC4,PC5,PC6 --out
logistic_results
```

e. Remove the lines with NA to avaoid problems in visualization of the results

```
awk '!/'NA'/' logistic_results.assoc.logistic >
logistic_results.assoc_2.logistic
```

e. Run the Manhattan plot script again, do you see any signals

```
Rscript Manhattan_plot.R logistic_results.assoc_2.logistic
logistic_results.manhattan.jpeg
```

e. Run the QQ plot

```
Rscript QQ_plot.R logistic_results.assoc_2.logistic
logistic_results.qqplot.jpeg
```

⇒ Based, on this practical, which of the approaches would you prefer for correcting population structure in your data?

* Please note there are other approaches such as LMM that might better account for population structure.