Day 5: GWAS project

Download Datasets

1. Create a folder “GWAS_project”
```
mkdir GWAS_project
```

2. Copy the three exercise datasets from GitHub to this folder
```
wget
https://github.com/WCSCourses/HumanGenEpi/raw/main/course_data/GWAS_project/variant_qc.zip

wget
https://github.com/WCSCourses/HumanGenEpi/raw/main/course_data/GWAS_project/binary_trait.zip

wget
https://github.com/WCSCourses/HumanGenEpi/raw/main/course_data/GWAS_project/continous_trait.zip
```

3. Unzip the three files
```
unzip variant_qc.zip

unzip binary_trait.zip

unzip continous_trait.zip
```

Please check that you have three folders inside the “GWAS_project” folder
Now try to solve the following exercises by yourself.

Exercise 1. Variant and Sample QC

For the dataset in the “**variant_qc**” folder do the following:

Q1. Check how many samples have discrepancy between sex reported in the fam file and sex in this dataset.

PLINK command

```
plink --bfile test_data --check-sex --out test.sex
```

Check how many samples have problematic sex satus

```
grep PROBLEM test.sex.sexcheck | wc -l
```

A1. Three individuals

Q2. Remove these individuals from the dataset and retain only the autosomal chromosomes

PLINK command

```
grep PROBLEM test.sex.sexcheck > sex.drop

plink --bfile test_data --remove sex.drop --chr 1-22 --make-bed --out test_data1
```

A2. 474425 variants and 190 people pass filters and QC

Q3. Filter out SNPs with genotype missingness greater than 0.05

PLINK command

```
plink --bfile test_data1 --geno 0.05  --make-bed --out test_data2
```

A3. 125526 variants removed due to missing genotype data

Q4. Filter out samples with individual missingness greater than 0.05

PLINK command

```
plink --bfile test_data2 --mind 0.05  --make-bed --out test_data3
```

A4. 3 people removed due to missing genotype data

Q5. Filter out SNPs with MAF less than 0.01. How many samples and SNPs pass the last QC?

PLINK command

```
plink --bfile test_data3 --maf 0.01  --make-bed --out test_data4
```

A5. 48834 variants removed due to minor allele threshold ; 300065 variants and 187 people pass filters and QC

Exercise 2. Association analysis for a binary trait

For the dataset in the "**binary_trait"** folder answer the following:

Q1. How many cases and controls do you have?

```
awk '{print $6}' casecontrol.fam | sort | uniq -c
```

A1. 170 cases and 182 controls

Q2. Run the association test for the binary trait and generate the Manhattan and QQ plots. Is there any signal below genome wide significance threshold?

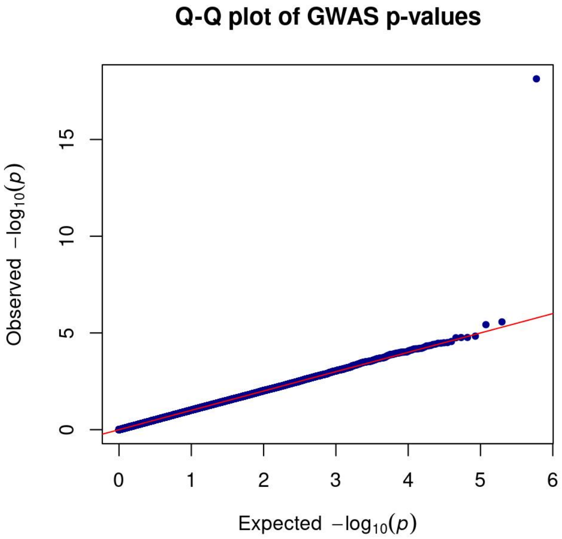
PLINK command

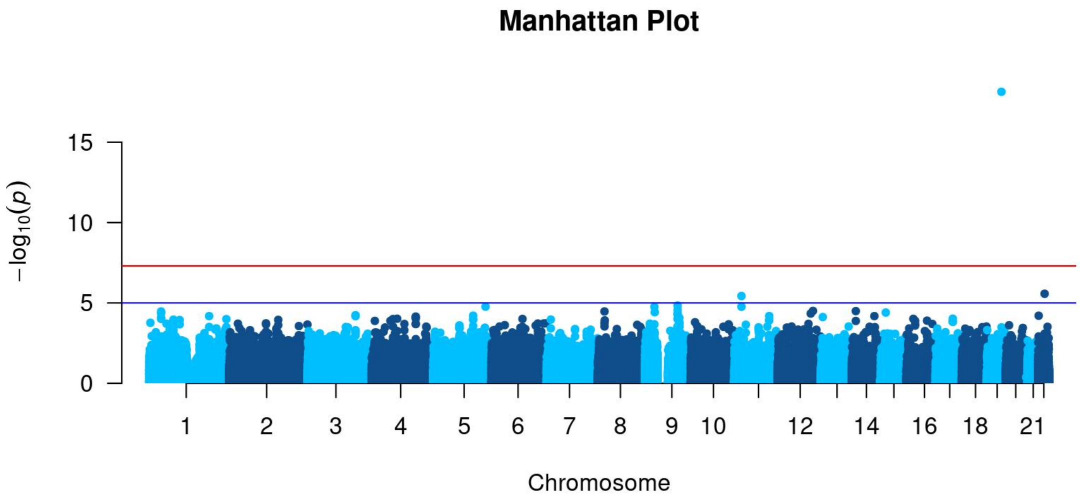
```
 plink --bfile casecontrol --assoc --out casecontrol.assoc
```

Run the R scripts

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

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





A2. There is one SNP p-value below above genome wide significance level at Chr19.

Q3. What is the lambda value. Is there a hint of population structure?

PLINK command

```
plink --bfile casecontrol --assoc --adjust --out
casecontrol.assoc
```

A3. Genomic inflation est. lambda (based on median chisq) = 1.01679. No population structure.

Exercise 3. Association analysis for a continuous trait

For the dataset in the "**continous_trait"** folder answer the following:

Q1. Run a linear regression for the continuous trait including the all the principal components as covariates. Generate the Manhattan and QQ plots for this analysis. Is there any loci below the genome-wide significance threshold, if yes, in which chromosome?

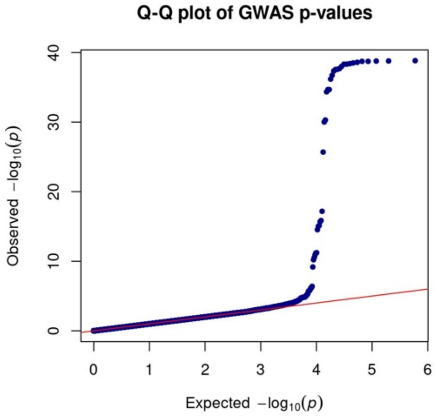
PLINK command

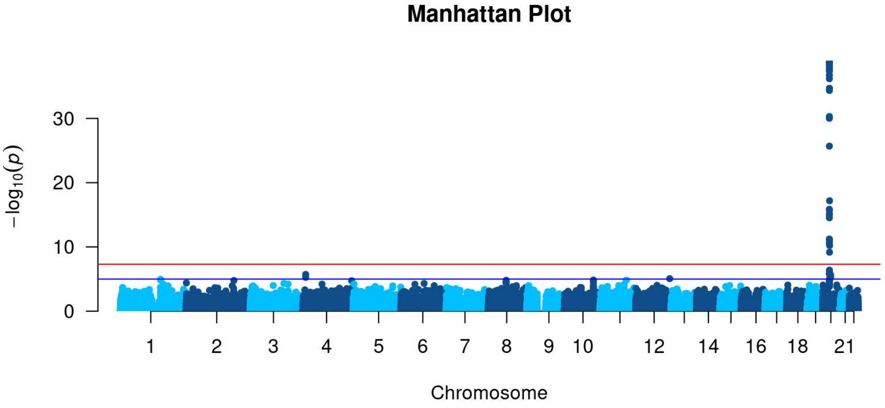
```
plink --bfile continous --linear hide-covar --pheno
continous.phe --covar continous.covar.txt --out
continous.linear
```

Run the R scripts

```
Rscript QQ_plot.R  continous.linear.assoc.linear
continous.linear.assoc.linear.qqplot.jpeg

Rscript Manhattan_plot.R  continous.linear.assoc.linear
continous.linear.assoc.linear.manhattan.jpeg
```





A1. Chr 20

Q2. How many SNPs are below the genome-wide significance level

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

A2. 35

Q3. Identify the SNP with lowest *p-value.*

```
sort -g -k 9,9   continous.linear.assoc.linear | head
```

A3. rs6050598

Q4. Now go to Ensembl and search for this SNP (from Q3). What is its alternate allele frequency in EUR and EAS super-populations? Can you find the gene corresponding to this SNP?

A4. EUR= 0.56 and EAS=0.09; Gene= *GINS1*
http://grch37.ensembl.org/Homo_sapiens/Variation/Population?db=core;r=20:25396757-25397757;v=rs6050598;vdb=variation;vf=332750754
http://grch37.ensembl.org/Homo_sapiens/Variation/Mappings?db=core;r=20:25396757-25397757;v=rs6050598;vdb=variation;vf=332750754