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_da
   ta/GWAS_project/variant_qc.zip

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

   wget
   https://github.com/WCSCourses/HumanGenEpi/raw/main/course_da
   ta/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.

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

Q3. Filter out SNPs with genotype missingness greater than 0.05

Q4. Filter out samples with individual missingness greater than 0.05

Q5. Filter out SNPs with minor allele frequency less than 0.01. How many samples and SNPs pass the last 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?

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

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

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

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

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?