**Data lab 2 - GWAS using PLINK2**

Name:

Date:

Question 1: What does "PC" in the covariate file stands for and why should they be included for GWAS? Any reasonable answers are accepted.

Question 2: How many individuals have available phenotype values? What extension has been added to the GWAS output by PLINK2?

Question 3: Which column do you look at to select significant SNPs? A common significance cutoff is 5e-8. What `awk` command do you use to obtain this information? Which SNP(s), if any, do you find significant?

Question 4: What are the dimensions of the output dataframe? Do you have the same number of rows as the number of SNPs left in your QC-ed dataset?

Question 5: Upload your plot to the report. Do you observe general inflation in the QQ plot?

Question 6: Upload your plot to the report. Do you see the highighted significant SNP?

Question 7: Compare the chromosomal position provided in NCBI with the one in our results, which one is ours based on, "GRCh38" or "GRCh37"?

Question 8: Click "Save PNG" and upload your regional plot to the report.

Question 9: From 1st page of the "Associations" table, list three traits that have been associated with the lead SNP.

Question 10: Which genes are the nearest to the lead SNP? Which one is the closest downstream gene?

Question 11: What protein does the gene encode? Does it sound like a plausible candidate gene for HDL levels?

Question 12: Are the traits for the nearest downstream gene as highlighted in the PheWAS the same for common and rare variants?

Question 13: Check in R or simply from the text-file, is A1 of the investigated SNP the reference or alternative allele? What is it and its frequency?

Extra exercise: Show the command you used to re-run the test. Based on the position of the lead SNP, what is the range of positions ± 500kb around it? Is there any additional association signal in this region? A p-value cutoff of 1e-5 can be used for secondary signal in this example.