**Data lab 1 - Introduction to PLINK and basic QC**

Name:

Date:

Question 1: What is the chromosome number of SNPs in the dataset? In which column do you find this information?

Question 2: What is the marker ID (rs…) of the first SNP? What are the reference and alternative alleles?

Question 3: Which individuals are contained in "interested\_individuals.txt"? Which one has genetic sex of male and which one is female?

What SNPs are contained in "interested\_snps.txt"? What are their reference and alternative alleles, respectively?

Question 4: What genotypes does the two individuals have for the two SNPs, respectively?

Question 5: How many individuals and SNPs are contained in the epihealth dataset?

Question 6: With the filtering criteria for SNP missing call rates>1%, how many SNPs should be excluded?

Question 7: Use `--missing` output files do you get? Among the output files, which reports sample missingness rates? Have a quick view of the file using `head`, and think about how to use `awk` to identify individuals with missing call rate >95%.

Question 8: How many SNPs are excluded if we want those with minor allele frequency >5%?

Question 9: Finally, make the QC-ed PLINK2 dataset "clean\_epihealth" in your own folder that includes individuals with missing call rate<5%, and SNPs with missing call rate<5%, HWE>1e-5 and MAF>2%. You probably already know that `--make-pgen` is the flag to produce new PLINK2 files and `--out` is for specifying the output name. Show the command that you used.

Question 10: In the output log, how many SNPs are left in the QC-ed data?

Extra exercise: Produce another QC dataset "extra1" that constitutes all SNPs in Chr 16 with positions between 250,000 bp and 300,000 bp, recode (export) them as 0,1,2. Use only data from individuals with missing call rate<5%. Load the recoded dataset in R or download them to local computers and open with Excel, calculate the MAF for some SNPs and check the correctness by PLINK2 command `--freq`.