

HGEN8341 Module 3 Project (Spring 2021)

You are given a few datasets and you will carry out various calculations, using R or any other software if you prefer. Each dataset is an N by M+2 table, with each row representing an individual, and the 1st column indicating the individual ID, the 2nd column indicating the affection status, and the remaining M columns representing genotypes of M variants. Each variant is bi-allelic, with '0' representing the reference allele, and '1' representing the alternative allele. The genotype is in the form of 0/0, 0/1 or 1/1, which represent homozygous reference allele, heterozygote, and homozygous alternative allele. The following are the results to generate.

1. Allele frequency estimates
 - 1.1. Estimate the allele frequency (AF) of the alternative allele for each variant and store the estimates in a file named "AF.txt".
 - 1.2. Plot the distribution (histogram) of the estimated AF
 - 1.3. Discuss your thoughts of the allele frequency pattern regarding genetic association studies.
2. Hardy-Weinberg Equilibrium
 - 2.1. Calculate the p value of HWE for each variant and store the p values in a file named "HWE.txt". Do this only for variants with MAF>0.05.
 - 2.2. Plot the **distribution** of the p values and **QQ plot** of the p values (log scale). An R code for QQ plot is provided.
 - 2.3. Discuss the HWE patterns observed in the datasets.
3. Linkage Disequilibrium
 - 3.1. Remove variants with MAF<0.05
 - 3.2. Calculate the pairwise LD among the first 100 pairs of the variants (D, D' and R²). Store the pair-wise LD for each of the D, D' and R² in 3 files named "LD_D.txt", "LD_Dprime.txt" and "LD_r2.txt", with the first two columns being the names of the pair of variants and the 3rd row being the corresponding LD measurement. The variants in the output files should be **in the same order** as they appear in the input genotype data. You need to use **EM algorithm to estimate the haplotype frequencies**.
 - 3.3. Investigate how LD (D, D' and R²) measures are related to each other
4. Principal component analysis You can also do association analysis in this one. This question id to help you learn to do PCA
 - 4.1. Code the genotypes using an additive model (i.e. use 0, 1 and 2 to code 0/0, 0/1 and 1/1) and carry out PCA. Calculate the proportion of variance accounted for by the 1st, the 2nd, and the 3rd PC.
 - 4.2. Plot the first two PCs of all individuals on an X-Y plot.
 - 4.3. Repeat 4.2 for PC1 vs. PC3 and PC2 vs. PC3. Which plot gives clear patterns about the pop substructure?