Lecture 10 Exercises:

ALL of the code and scripts are inside of the file: Lecture_10_code.txt. Because the code for doing some of these tasks is quite involved, please copy and paste from that file, but do make efforts to understand what the code is generally doing.

Exercise 1. Power calculations. Suppose we're writing a grant to do a new project in which we would like to sequence a bunch of people with some continuous trait and then apply SKAT to run the analysis exome wide (i.e. we will be testing 20,000 genes such that the alpha level is 2.5e-6)

We posit the following:

- Average region length will be 5kb
- 20% of the variants with MAF < 5% will be causal
- 20% of the causal variants will decrease the trait value while 80% will increase the trait value
- We assume the magnitude of the effect sizes for the causal variants is equal to $-c \text{Log}_{10} \text{MAF}$ where we set c such that a variant with allele frequency of 1/10,000 is 2.

Let's further assume that we do not have prior sequencing data so we will need to use simulated chromosomes from COSI (let's further assume European ancestry for simplicity).

Calculate the anticipated power if our sample size is 1000.

How many subjects would we need to have 80% power?

Exercise 2. Real data analysis: Please carefully look over the scripts and code for converting the VCF to Plink format and for running ANNOVAR.

Exercise 3. Analyze the 1000 Genomes Chromosome 22 data by testing for association between the variants in each of the genes and the variable Y1 which is inside of the FAM file. Please also adjust for X1 and X2 as potential confounders. Please apply the SKAT method first.

Exercise 4. Now let's run the weighted count based collapsing to test for association between Y1 and the variants in all 400 or so genes.

Exercise 5. Repeat the analysis using the optimal omnibus test.

Exercise 6. SKAT is also used for analysis of common genetic variants (where it is called the SNP-set Kernel Association Test). Just be careful to use kernels that are more appropriate for common variants (e.g. the linear kernel rather than the weighted linear kernel).

Let's re-analyze the transferrin data set.

First, we will map all of the SNPs in the transferrin data set to genes and then we will associate the genes with transferrin levels using SKAT.