Lecture 4 Exercises:

ALL of the code and scripts are inside of the file: Lecture_4_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. Using PLINK, extract the SNPs from the transferrin data set that are within the 60kb of the *TF* gene and recode them in additive fashion (i.e. as 0, 1, 2 for the number of minor alleles).

Note: From figure 2 of Benyamin et al. (2009), the coordinates for the region of interest is chr 3 from 134840K to 135052K.

Separately, do the same thing for the SNPs on chromosome 12 from 23450K to 25050K. Be careful not to give this the same name as TFsnps!

Exercise 2. Now open up R. Read the SNPs in the TF gene as well as the phenotype information into R. Isolate the SNPs as a matrix called Z and the outcome is a vector y. Let's omit the subjects for which the outcome trait is missing or for which there are any missing SNPs.

Exercise 3. Now let's run a few different tests. The code for these is in the R scripts file.

a. First, let's just run the marginal (individual SNP) association tests and then use the minimum p-value from the SNPs in the TF gene. Try correcting for taking the minimum by Bonferroni correction.

b. Second, let's try collapsing by taking a (weighted) average of the SNPs within the TF gene. We can take a straight average (sum) of the values, or we can use the top principal component.

Exercise 4. Repeat the previous two exercises for the random segment from chromosome 12.

Some more things to think about:

- Returning to the issue of missingness, think about which of the methods we've tried would have worked by just ignoring the SNPs with missing values?
- What sort of hypotheses have we tested and what's the interpretation?
- What happens if we google for the location of the TF gene? Do the positions match what we tested?