Lab 5 – GWAS

In this lab we will be performing a genome wide association study. We will be following the lab designed by the Foulkes Lab at Mount Holyoke College. More details can be found here: http://www.stat-gen.org/

The data we are working with is a subset of the data collected by the PennCath study. This was a large cardiac health study. The complete data set has genotype and clinical data for 3850 individuals of European ancestry. We will be using a randomly selected subset of 1401 individuals. We will be trying to identify SNPs that are associated with HDL concentrations.

1. Download the GWAS.zip file from our google drive.

2. Unzip and place in your U drive

3. Start R-studio and make the GWAS folder your working directory.

4. Open the following R programs for inspection
   a. Packages.R – downloads and sets up the R programs we'll be using
   b. DataGlobals.R – sets up the data input ant output directories
   c. ReadData.R – reads data into R

5. Source these programs in the order listed above.

6. Meanwhile, let's look at the data. In the GWAS_TutorialsFiles folder there are a series of data files.

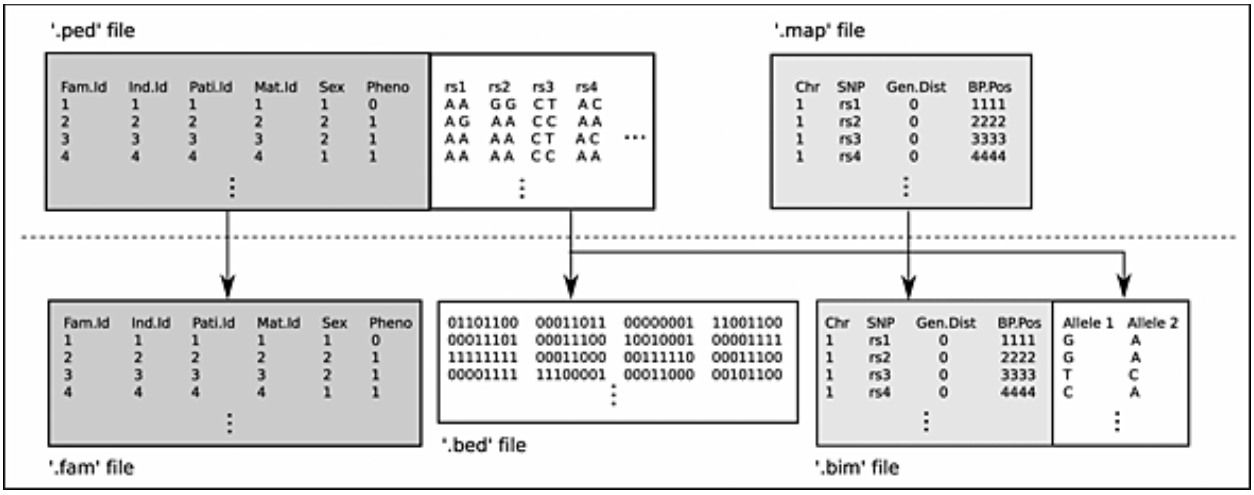

Figure 1: Genome-wide association data files. GWA data files are organized into either .ped and .map files or .bim, .bed, and .fam files. The later set is substantially smaller because the .bed file contains a binary version of the genotype data. R can read in either set of files although the later is preferable. Figure from (Reed et al., 2015).

7. Let's look at the data a little. The first thing this ReadData.R program will print to your screen is information on the data frame SnpMatrix that we have named genotype. From

this, information, how many individuals do we have data for? How many SNPs do we have?

The next thing that is printed to your screen is the first few rows of the genoBim data that you have loaded into R. What does this data look like? What does it tell you? A1 is the minor allele. For the first five SNPs listed in this file, what chromosome are they on? What is the base of the minor allele?

The next set of information printed to the screen is the first few rows of the matrix of clinical data, called clinical. What clinical information do we have about these people? The column CAD indicates if they have been diagnosed with a coronary artery disease or not. If they are a control, the CAD will be 0. From this, are people that have genotypes for which we do not have clinical data of interest? To determine the number of controls in the data and the number of individauls with hdl values, use the following commands in you input window.

```
>sum(clinical$CAD==0)
> length(clinical$hdl[clinical$hdl=='NA'])
```

Filtering

8. Now that we have read in the data, we're going to filter on things we care about. We're going to start by removing genotypes for which we have no clinical data.
   a. Source Step1_SubsetData.R

   Do you lose any genotypes in this step? Why do you think this is?

9. Now we're going to start filtering the data. Open Step2_Filter1.R. We are going to filter the SNPs for two things in this step.
   a. We are filtering for only SNPs for which at least 95% of the individuals we have calls. This will eliminate any SNPs that are low quality across the data.
   b. We are filtering for SNPs whose minor allele is found in at least 1% of the data. We do this because alleles that are less common than that are hard to associate with a common characteristic, like high HDL concentrations. Removing them will save us processing time without sacrificing associations.

   How many SNPs? Do we remove at this point? How many SNPs remain?

10. Now we are going to screen for inbreeding, sample contamination and poor sample quality. We are going to use dramatic change from the predicted Hardy-Weinberg Equilibrium to identify these. As we discussed in class, heterozygotes are predicted in the population with a rate of $2p(1-p)$ where p is the probability of the dominant allele. If there are significant deviations from this, it often implies that there is something wonky about the data, including and of the three conditions above.
    Open Step3_Filter2.R
    We are going to calculate the heterozygosity expected and observed for each locus and remove any that deviate more that 10% from the expected.

We are also going to remove any individuals that have poor data. We are going to define poor data as having calls for less than 95% of the SNPs. This is similar to the cut off we used for SNPs, but in the reverse. We're eliminating individuals instead of SNPs.

11. Source Step3_Filter2.R
    How many SNPs do we loose in this step? Why do you think this is?
    Next we are going to filter out individuals that are too closely related. IF individuals are closely related, their relatedness will cause spurious associations to be identified, as many SNPs unrelated to the HDL characteristics will be correlated. To do this we will be determining the relatedness of each sample/ We're going to be using the same technique, identity-by-descent (IBD) that we saw in the North American Ancestry paper from Ancestry.com (Han et al., 2017). We remove any pair of individuals that have a linkage or kinship greater than 20% or 10% respectively.

12. Open and source Step4_Filter3R
    This step takes a while as it is building a large matrix to compare all the individuals. Eventually a plot will appear in the plot window. This shows the distribution of ancestries of the individuals. This shows the diversity within the population in the sample. As these individuals were all of European ancestry, what does this plot tell you?

13. Now we're going to filter based on HWE. As we know that the HWE can be disrupted by selection, we are only going to look at the control (not case) samples in the data. Basically, we are looking for alleles that are out of HWE. So we are looking for alleles where the numbers of each genotype don't look like the following:
    $$P_{AA} = p_A^2 \ ; \ P_{Aa} = 2p_a p_A \ ; \ P_{aa} = p_a^2$$
    We can calculate how far from the expected we are, this is known as the z-test. It is similar to a chi-squared test. We are going to exclude any SNPs that appear extremely far from HWE using a cut off of 10^{-6} for the z-score.

14. Open and source Step5_HWEfilter.R

GWAS Calculations

We are finally finished prepping the data! WOOT! Therefore, we can now start using the data to generate the files we will use for association analysis.

15. To perform the analysis we want to do some basic feature selection. To do that we are going to calculate the first 10 principal components of the data. This will allow us to determine the latent population substructure that goes beyond "European". We are going to once again prune our data for linkage greater than 20%, then calculate the PCAs.

16. Open and source Step6_PopStrat.R

17. Now we are going to impute the genotypes of SNPs that we didn't genotype. We can do this due to known linkage effects. Since this will significantly increase the size of the

data, we are going to focus on one chromosome where we know there is a gene of interest, chromosome 16.

18. Open and source Step7_Imputation.R
    How many SNPs did we impute to chromosome 16?

19. Now we're going to prep the data for GWAS analysis. To do this we need to combine the genotype data with the clinical data of interest, in this case HDL levels. We also want to make the HDL data fit the type of analysis we are performing. In the next step we'll be using a simple additive model and linear fitting to determine the association, so we want the HDL data to be Gaussian, so we transform the data into a Gaussian using a rank-based inverse normal transformation.

20. Open and source Step8A_PrepareGWAS.R
    Be sure to note the histograms of the HDL. How many SNPs will we be analyzing? Across how many individuals? In addition to the SNP PCs, what other data will we be entering into the GWAS program?

21. Now we are going to run the GWAS. WE are first going to subset the data and only look at those on chromosome 15-17. This will make this run reasonably quickly. Also, we will be running this in parallel on two cores in the computer.

22. Open and source Step8B_RunGWAS.R
    How long did it take to run?

23. Let's add the imputed SNPs. Open and source Step9_ImputeP.R
    How many SNPS did we calculate a p-value for? What is the one with the lowest p-value? How many SNPs are in the CETP gene? How many of those are imputed versus typed?

Analysis

24. Now that we have determined the p-values, let's look at the data visually. We'll start with a Manhattan Plot. Open and source Step11_Manhattan.R
    What does the data show? Are there any SNPs that reach above the Bonferroni Corrected? Do any reach the less stringent threshold? What do you think that mean for these data? What conclusions can you reach?

25. Let's look at these in the context of the chromosome more closely. Open and source Step12_LD.R
    In this you can see the gene and the SNPs in the context of the linkage disequilibrium.

    (This might be broken. It worked on Wed., but not on Thurs. We'll see how it goes)

26. Let's explore these more in context. Open a web browser and go to locuszoom.org. There you can chose to look at published-Interactive.

27. Select the HDL data and chromosome 16. What do you see? How would you interpret this? What region/gene seems to be most involved with HDL regulation?

References:

Han, E., Carbonetto, P., Curtis, R. E., Wang, Y., Granka, J. M., Byrnes, J., … Ball, C. A. (2017). Clustering of 770,000 genomes reveals post-colonial population structure of North America. *Nature Communications*, *8*, 14238. https://doi.org/10.1038/ncomms14238

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015). A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, *34*(28), 3769–3792. https://doi.org/10.1002/sim.6605

Lab Report Instructions

100 points

Due 6/2/17 11:59:00 PM

There should be 4 sections in this report. As always, you should include information and figures that help explain your answers to these questions, and it should be written in paragraph, not bullet format. (10 points for writing style and citations)

I. Data (25 points)

A. Describe the data we used as input. Be sure to include answers to the following questions.

1. How many individuals do we have data for? How many SNPs do we have?

2. For the genoBim data structure, what does the data look like? What does it tell you? A1 is the minor allele. For the first five SNPs listed in this file, what chromosome are they on? What is the base of the minor allele?

3. What clinical information do we have about these people? Are people that have genotypes for which we do not have clinical data of interest? How many control subjects? How many case subjects? How many of the individuals do not have HDL data in their clinical reports?

II. Filtering (15 points)

A. Discuss how we filtered and why. Be sure to indicate the number of data points removed at each filter step.

III. GWAS calculation and results (20)

A. Discuss the imputation. Why did we do it? How many SNPs did we impute? How did we determine p-values for these imputed SNPs?

B. Discuss any additional steps we used to narrow the data and why we chose to use those.

C. How many SNPs did we identify that are significant?

D. Be sure to answer questions posed in the lab description.

IV. Analysis (30)

A. Discuss what these data tell you. Are there any SNPs that are significant? How are you defining significance? What gene are these in?

B. Given the linkage information, do you think the imputation of SNPs is reasonable? Why or why not?

C. If you were a researcher, what would be the next step after these results?