BMI 704: Data Science 1

Instructors: Chirag J Patel, Arjun K Manrai
Last modified: 1/31/19

Assignment 1:

Connecting inherited variation to disease and phenotypes with genome-wide association studies (GWASs) and polygenic risk scores

Date Due: 2/22/19 by 11:00 pm EST.

Start early. This is a challenging assignment! You may work together, but each student must hand in their code and answers separately. Assignment is a total of 125 points (and 15 points/day will be docked for each day late). The "Execution of GWAS" section of the assignment must be completed in Orchestra (and might take a prohibitively long time to execute if you just rely on your own laptop). Individual-level/Person-level GWAS data can only exist on your own computers for testing your code and debugging; please delete all data after the assignment.

Submit your answers (e.g., figures) and analyses as a .Rmd file that is "knitted" as a .pdf file; your R pipelining script (`gwas.R`), and accompanying output tables as .csv files, into a folder. Zip the directory and name it using the following convention **{eCommonsID}_assignment1.zip**. **Code using the *tidyverse* suite of tools!** Submit via Canvas; we will announce how to submit via Canvas soon.

Background:

Genome-wide association studies (GWASs) are arguably the most significant biomedical advance in the last decade. Utilizing an observational case-control design, populations with disease are analytically compared to those without to discern patterns in differences in genetic variant, or single nucleotide polymorphism (SNP), frequencies in each of the respective populations. If statistically differences in SNP frequencies in cases vs controls are found, one can conclude the SNP locus may have a causal relationship, implicating a gene, genetic region, or gene regulator in the disease of interest.

Analyzing GWASs provide lessons in observational research, including statistical association and multiple testing. They also provide a way to create hypotheses about biological basis of disease for further study. GWASs are also modern-day prerequisites for understanding environmental basis of disease, explored in the next assignments. In this assignment, we will execute a GWAS analysis using data from the Wellcome Trust Case-Control Consortium (WTCCC) on Type 2 Diabetes (and optionally other diseases) versus 1,500 controls. **While there are standard off-the-shelf pipelines to execute GWASs, understanding how to implement one from scratch yourself is a good introductory exercise in biomedical data science.**

Goals: this assignment will get you up to speed in (1) becoming familiar with genetic data and large scale association tests, (2) developing analytics pipelines in R that utilize "vectorized" functions and the tidyverse, and (3) executing jobs on a compute cluster, (4) get a taste of prediction and use of GWAS data in medical decision making with polygenic risk scores.

Description of Data and Readings

Please read the paper by the WTCCC, entitled, “Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls” (link: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2719288/).

Also read Khera et al, entitled “Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations” (link: https://www.ncbi.nlm.nih.gov/pubmed/30104762)

Data for this investigation is located in the folder in Orchestra:
`/n/groups/bmi704_spring19/assignment1/WTCCC`

Each subdirectory in this directory contains a set of GWAS array files, for example in the `58C` subdirectory, there are files called the following: ‘Affx_gt_58C_Chiamo_01.tped.gz’.

The subdirectories are labeled as such:
`58C`: control population 1
`NBS`: control population 2
`RA`: rheumatoid arthritis
`T2D`: type 2 diabetes
`T1D`: type 1 diabetes
`CAD`: coronary artery disease
`HT`: hypertension
`CD`: Crohn’s disease
`BD`: bipolar disorder

All of the GWAS data is broken down per chromosome, 1-22 plus X and is written into the file name, as well as the disease. For example:
Affx_gt_**58C**_Chiamo_**06**.tped.gz

Contains GWAS data for the 6th (**06**) chromosome on *control population 1* (**58C**). The biggest files are chromosome 1 and 2 and smallest is 22 (exercise: why does the data size decrease in chromosome number increases?).

The compressed files (which can be read in to R using `read_delim` *without decompression*) are formatted in ‘tped’ style. Specifically, if we run the following command in R, using the `readr` library in the **tidyverse**:

```
> dat <- read_delim('./Affx_gt_58C_Chiamo_22.tped.gz', delim="\t", col_names=F);
> dim (dat)
[1] 6207 1508
```

Each *row* is a SNP (there are 6207 SNPs measured in Chromosome 22). The first through fourth *columns* denote the chromosome number, the second is the WTCCC snp_id (not to be confused with the rsid), the start coordinate (0), and the end coordinate. Thereafter, each column is a genotype (where each base is separated by a space character) for that SNP for each individual. E.g.,:

```
> dat[1, 1:10]
> # A tibble: 1 x 10
```

```
     X1 X2               X3       X4 X5    X6    X7    X8    X9    X10
  <int> <chr>         <int>    <int> <chr> <chr> <chr> <chr> <chr> <chr>
1    22 SNP_A-4234155     0 32056663 C T   C T   C T   T T   C T   T T
```

In the above, we print out the first 10 columns for the first SNP in the 58C chromosome 22 dataset. Therefore, V1 is the chromosome number, SNP_A-4234155 is the WTCCC snp_id. The fifth column is the first genotype for the first individual, which happens to be "C T". Therefore, this individual is heterozygous for this SNP (CT).

Each subdirectory also contains a '`snps_info.tar.gz`', which contains the snp_id and rsID for each SNP for each chromosome. These files can enable you to map the WTCCC snp_id to the rsid, which is helpful to find the closest genes to a SNP.

Copy the data to your home directory (~) in Orchestra to work on your answers below.

Questions (**125** points total plus **35** Bonus points)

Background questions:

1.) What are the **diseases** of interest ascertained in the WTCCC study? Draw the **E, G, P, and D** diagram and annotate what is investigated in the WTCCC GWASs.(**2 points**)

2.) How are the each of the seven **diseases** assayed phenotypically (**7 points**)?

3.) What is the assay platform for this study? Briefly describe the concept behind the assay technology (**2 points**).

4.) How many SNPs are measured for each individual **for all the chromosomes (you can do a line count of the data files themselves)** (**1 point**)?

5.) Let's start small. The year is 1999, and an investigator has painstakingly genotyped 1 SNP called (rs*GOINGALLIN*) in individuals with bipolar disorder and without. *rsGOINGALLIN* can take on the 3 genotype configurations, CC, CT, TT. He has collected the following data:

	CC	CT	TT
Bipolar	270	957	771
Healthy Controls	436	1398	1170

He asks you how to execute an allele-based test of association "allele frequency test", whereby the reference allele is the **major** allele (the allele, C or T, that is most prevalent in the healthy population).

A.) What is the allele frequency (C and T) in bipolar population? And in Controls? (**4 points**)
B.) Execute a chi-squared test to test the association of the allele frequency in bipolar vs. controls in R and report the pvalue. Is there evidence to conclude there is an association between rsGOINGALLIN and bipolar disorder? (**4 points**)
C.) Execute a chi-squared test amongst healthy controls to test for deviation from Hardy-Weinberg equilibrium (**4 points**).

Execution of GWAS:

1.) We are now going to execute a GWAS on T2D, a moderately heritable disease in which both the genome and exposome are said to play a role. Write R code, from scratch, to execute a GWAS for the WTCCC called **gwas.R** for **a given chromosome (code: 30 total points=10 points for correct setup [e.g., reading, writing] + 20 points for correct case-control analysis code).**

This script will **input** the a) control files, b) the disease file, c) chromosome number for a given disease and chromosome and **output** a .csv file that contains `SNP id (rsid), chromosome number, minor allele, major allele, minor allele frequency in disease, minor allele frequency in controls, odds ratio (major vs. minor allele), p-value, Hardy-Weinberg deviation p-value`

Your code should execute an "allelic test" of association, or test the frequency of "**minor**" alleles in disease versus controls. For this assignment, assume the "major allele" for a SNP **is the allele that occurs most frequently in the control population.**

Second, your code should calculate the **deviation from Hardy-Weinberg equilibrium** chi-squared p-value using the **control population**.

Hint 1: **combine** the 'control 1' (58C) and 'control 2' (NBS) population to a unified control population.
Hint 2: use *vectorized* code in R, such as `table, apply, which, sum, broom, summarize` along with the *tidyverse*
Hint 3: make sure your code works for 1 SNP, then scale it up to all the SNPs, then all the chromosomes. You can execute the GWAS for each chromosome ***separately*** by submitting 22 jobs to Orchestra/O2 via the slurm commands. Please see here:
https://wiki.rc.hms.harvard.edu/display/O2/Using+Slurm+Basic

Using your R code, execute GWAS on T2D for chromosomes 1 through 22 (excluding the sex chromosome). Filter out SNPs that deviate from Hardy-Weinberg (deviation p-value less than 0.05) or have a minor allele frequency less than 1%.

A.) After filtration, how many SNPs are available for analysis? Why filter out SNPs that deviate from Hardy-Weinberg? What is the Bonferroni-threshold of significance (round to 1 significant figure) (**2 points**).

B.) Compile the GWAS results into one .csv table and call it 'T2D.csv' (**15 points=10 points for correct findings and format+5 points for correct filtration]**).

2.) Analysis of GWAS results:

- A.) Produce a **manhattan plot** and **qqplot** of the pvalues (**2 points**).
- B.) How many SNPs were found that exceeded the Bonferroni-level of significance for T2D? (**1 point)**
- C.) What could be biasing the associations and how might one control for this phenomenon? (**1 point**)

D.) What is the p-value of association and odds ratio for rs4506565 for T2D? What is the interpretation of the odds ratio? What gene is this SNP associated with and its putative function in T2D pathogenesis? (**10 points**)

3.) You just executed an allele frequency test of association. How would the genotypic test of association be different? Second, how would you test both of these types of allele configurations using a logistic regression? (**2 points**).

4.) What is the main analytical differences between the test in Background Question 4 (candidate SNP) and GWAS? Why would one want to execute one or the other (**1 point**)?

5.) In question 2B.), you identified sources of biases for GWASs. One source of bias can be controlled for using information from the SNP array. What is this factor? (**2 points**).

6.) The genomic inflation factor is arithmetically defined as the median of the **observed** chi-squared values (a function of the pvalues in GWAS) divided by the median of the **expected** chi-squared values. The chi-squared value can be estimated from a pvalue in the following way in R: `qchisq(pvalue, 1).`

Estimate the genomic inflation factor for your GWAS in T2D. What does this quantity estimate? (**10 points [5 points for estimate, 5 points for answer]**).

7.) BONUS. Modify your code to take into account this bias and re-run your associations for 'Executing GWAS', Question 1 (**10 point bonus**).

GWAS and Medical Decision Making: Introduction to the Polygenic Risk Score (25 points)

1.) What is a "polygenic risk score"? What is the underlying assumption? (**1 point**)

2.) Suppose you have a set of *M* SNPs associated with a disease/phenotype (say identified at p-value threshold of *P*), each with an effect size *x_i* (*i* indexes the *i*th SNP of *M* total SNPs). Specifically, *x_i* is the odds ratio of the **risk allele relative to the non-risk allele for M independent SNPs that are not in LD with one another.** You also have SNP data for one individual with the number of risk alleles *n_i* that correspond to the effect sizes *x_i* assuming the effects are additive. Write down the formula to compute the PRS for this individual. (**2 points**)

3.) Using your GWAS summary statistics for T2D, compute the PRS *for each* individual in the cases and controls for T2D. Plot a histogram of the distribution of the PRS in cases and controls (in different colors) and indicate the quintiles (20, 40, 60, 80, 100 percentiles of the distribution). (**5 points**)

- ***BE careful! Make sure all of your ORs are for the risk allele (reference allele is the non-risk allele)!***
- ***There is an emerging literature on what SNPs to choose for a PRS. For this exercise, use a pvalue threshold of 1e-7, HWE pvalue > 0.05 and minor allele frequency of greater than 1% (in both cases and controls)***

4.) Suppose you have an individual in the 85th percentile of the PRS distribution for T2D and you are the individual's physician--what do you advise your patient? Suppose another individual is in the 43rd percentile--what do you advise this patient? (**5 points**)

5.) T2D is a risk factor for cardiovascular disease (see: https://www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke). Using your GWAS summary statistics for T2D, compute the PRS for each individual for the CVD cases . Add this to the histogram. Are individuals with the highest T2D PRS at genetic risk for CVD? (**5 points**)

6.) Suppose you are given SNP data for a set of *new* patients with and without T2D. How would you expect your PRS to perform on the new patients (ie, more discriminative, less discriminative, or the same)? Why or why not? Further, you did not account for LD when estimating the PRS. How does not accounting for LD influence the PRS? (**4 points**)

7.) Encapsulate your work above into an R function that accepts a two column data frame of SNPs and genotypes for a single individual as input and returns a polygenic risk score for that individual. (**3 points**)

Genetic architecture and disease similarity (25 point bonus):

Execute question 2 for all 7 diseases and answer the following questions.

1.) Produce a qqplot for each disease using ggplot. Rank each disease by the number of significant findings (pvalue threshold of 1x10^{-7}) (**10 point bonus**).

2.) What disease, if any, diseases are associated with the the SNP rs6679677 (**2 point bonus**)?

3.) What are the potential biological meaning of a SNPs found in multiple diseases (**3 point bonus**)?

4.) Devise a method to correlate diseases on the basis of their association statistics (e.g., pvalues and/or odds ratios). What diseases are most correlated with each other? (**10 point bonus**).