BST 227 Introduction to Statistical Genetics Final Project-- Fall 2 2017

Overview:
Throughout the course of human medical history, physicians have anecdotally noted co-occurrences of diseases or other remarkable traits within a single individual. In the year 2017, we now have a wealth of genetic and epigenetic data that allow us to systematically examine whether two traits actually share a genetic basis (termed pleiotropy). For this project, each group will be assigned two traits where large GWAS summary statistics data is available. Using techniques and methods discussed in class and the homeworks, the goal will be to measure and identify features of genetic relatedness between traits using summary statistics.

Core questions to be answered:

1) What are some basic baseline facts about each condition (e.g. prevalence for a binary trait; mean / standard deviation with units for quantitative trait)
2) What are the approximate heritability estimates for the two traits assigned? How do the heritability estimates change using different methods (e.g. LDScore, twin-studies), if these results are available?
3) Using LD Score regression, determine the two phenotypes' genetic correlation.
4) Show both global (manhattan) and local (LocusZoom) plots for each trait. Specifically, show a region where there appears to be association for both traits. Are the same variants the most strongly associated for each? Hint: you may need the **.bim** file from here https://data.broadinstitute.org/alkesgroup/LDSCORE/1000G_Phase1_plinkfiles.tgz to get the SNP annotation positions.
5) What are the sample sizes used in each summary statistics GWAS file?
6) Determine the degree of genetic inflation using conventional means (lambda GC) and the LD score regression intercept. Interpret these values.
7) What, if any, genes have been implicated by rare variant / family association studies for these traits? If it's a quantitative trait, try looking for examples of families with extreme physiological or disease levels of the trait.
8) Have any non-European GWAS been conducted? How do their results compare to the European summary statistics used in these analyses?

Additional questions (of interest, but not required for full credit):

9) What epigenomic annotations seem to be enriched in both traits? What annotations are enriched for only one trait? Use LDscore regression to determine these enrichments.
10) What other traits have a strong positive or strong negative correlation with these two traits? (hint: use LD hub; don't worry if the exact summary statistics aren't the same)

Important Dates:

- Presentation date: 13 December 2017 in class
    - The presentation can be carried out by one, a few, or all of you team
    - Email slides to **caleblareau@g.harvard.edu** and **aryee.martin@mgh.harvard.edu** **before class on December 13**

- Final written report due: 14 December 2017 by midnight
    - A group effort should be employed to complete the write-up.
    - Email report to **caleblareau@g.harvard.edu** and **aryee.martin@mgh.harvard.edu** **before midnight on December 14**

Presentation Details:

- Prepare 8-10 minutes worth of slides summarizing your findings. We will then have 3-5 minutes of questions from the TA, Instructor, and other students in the class. Please try to stick to this time limit as we have several groups that must present on the same day.

Writeup Details:

- There are no strict page or word limits. Students should address each question posed above using necessary figures and text communicated in a coherent write-up with an **Abstract, Introduction, Methods, Results,** and brief **Discussion**. A supplement showing relevant code should also be included in the submission.

Support:

- G11 will be available Mondays after class; the TA will be present from 6:15-7:30pm
- G13 will be available Wednesdays after class; the TA will be present from 5:15-7:30pm

Data / Help:

- Summary statistics: https://data.broadinstitute.org/alkesgroup/sumstats_formatted/
- LD Scores: https://data.broadinstitute.org/alkesgroup/LDSCORE/
    - Includes baseline epigenetic data
- Lab 5: http://aryee.mgh.harvard.edu/BST227/Labs/Lab5/BST227_Lab5.html

Group 1: Rheumatoid Arthritis and Systemic Lupus Erythematosus
Chris Cosgriff
Rose Huang
Shuhei Miyasaka
Rafi Small
Ryo Uchimido

Group 2: Schizophrenia and Bipolar Disorder
Aaron Aday
Laura Brenner
Lucas Buyon
Gyeyoon Yim
Tian Zhang

Group 3: Type 1 Diabetes and Type 2 Diabetes
Shimin Bi
Fangyuan Hong
Yuxi Liu
Tianyu Xia
Xinan Wang
Tong Zhao

Group 4: Celiac Disease and Primary Biliary Cirrhosis
Nathan Huey
Jane Liang
Kevin Ma
Derek Shyr
Cathy Wang

Group 5: Height and BMI
Leonce Nshuti
Danielle Rasooly
Siquan Wang
Chen Yuan
Sui Zhang