BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2014

Project (Version 1)

Posted November 5; Due 11:59PM November 26

1 Introduction and instructions

The goal of the class project is for you to demonstrate what you have learned by performing a GWAS analysis on real data. To accomplish this, assume that you have been provided data by a collaborator who wants to identify positions of causal polymorphisms (loci). You will perform an in-depth analysis and write a report for your collaborator that explains your methods and results.

Instructions: While we provide some general guidelines for how to proceed below, the techniques you use to analyze the data and how you construct your report will be up to you. Do however note the following instructions (PLEASE READ THESE CAREFULLY!!):

(1) Your project must be in Amanda's or Jin's inbox (as appropriate) by 11:59PM, November 26 - if it is late for any reason, standard grading policies apply.

(2) You are allowed to work together with other students in the class to analyze these data. However, note that turning in a report that describes exactly the same analyses as a fellow student is not a good strategy for getting a good grade. Also note that you must write your own report.

(3) This is an 'open book' assignment, such that you are allowed to use any resources online, in books, etc. You may also ask third-party (i.e. people not in the class) for suggestions on what analyses to perform but you cannot have a third-party do any of the analyses (or write any code for you!).

(4) You are also allowed to use any software or programming language that you would like as part of your analysis. However, we expect that some of the tasks will be performed in R (also note that you are welcome to use any packages, functions, etc. in R).

(5) Your final project will include a SINGLE report file and a SINGLE text file including all of your R code and / or commands or scripts you used to run other software packages. That is, for your R code, the best way to maximize your grade is to have well commented code that we can run from the command line. If you use other software for some of the tasks, a reasonable approach is to include commented out descriptions in your code that provides details on how you ran the software, e.g. what parameters did you use, etc.

(6) The report file must be no more than 8 pages (single-sided), with NO MORE than 5 pages of text and NO MORE than 3 pages of figures / tables.

(7) For your report, you must describe what you did in detail (a good guide is have you provided enough detail such that someone reading your report could replicate what you have done?). You also need to describe the results you have obtained from your analysis. You may also wish to include some text to describe interpretations and conclusions that may be of interest to your collaborator, including statistical and possibly, biological interpretations. For your Figures and Tables, note that clarity and clear labels is a strategy for maximizing your grade.

(8) We will grade on two broad criteria: 1. the overall quality of the analyses / report, 2. the amount of effort put into your project. Note that 'effort' does not mean run many analyses without thinking carefully about why you are running them or how they fit together to provide a clear picture of results. A guide maximizing your grade on effort is to think carefully about how to produce the best possible report that you can and then put in as many hours as you wish to devote to the project accomplishing this objective (your effort level will be clear to us).

2 The experiment and data

The experiment: Among the very important human genomics resources is HapMap:

`http://hapmap.ncbi.nlm.nih.gov/`

a sample of individuals from population representing distinct ancestry groups from around the world. For each of these individuals, there is a relatively complete (i.e. as about as complete as is currently attainable) and accurate set of SNP genotypes throughout the non-repeat and analyzable regions of the genome with alleles at relatively high MAF in these populations has been measured for each each individual.

From a subset of these HapMap individuals, B lymphocytes (an immune cell type) were collected and 'immortalized' (i.e. they were manipulated so they keep on dividing forever under lab conditions) using the Epstein-Barr virus to generate Lymphoblastoid Cell Lines (LCL). The cell lines for these individuals have been studied extensively and have undergone measurements of many different types, including the collection of data on the expression levels of genes throughout the genome (i.e. the 'transcriptome') using various technologies, including microarray and RNA-Seq. Each of these gene expression measurements may be thought of as a phenotype and one can do a GWAS analysis on each individually, which is called an 'expression Quantitative Trait locus' or eQTL analysis, an unnecessarily fancy name for a GWAS when the phenotype is gene expression.

What you have been provided is a small subset of these data that are publicly available. Specifically, you have been provided most of the SNP genotypes for chromosome 22 for a subset of individuals in the sample from the Yoruba population in Africa (YRI) and a subset of individuals from a Chinese population (CHB). For these same individuals, you have also been provided the expression levels of four genes: MRPL40, GGT5, TTC38, FAM118A. You have also been provided information on the gender of each of these individuals. A description of the broader data set from which these data were extracted can be found in:

http://www.ncbi.nlm.nih.gov/pubmed/22532805

and in other papers relating to analysis of expression data in the HapMap LCL

The data: These have been provided to you in four total files: 'QG14_project_phenotypes.txt', 'QG14_project_covariates.txt', 'QG14_project_genotypes.ped' and 'QG14_project_genotypes.map'. Note that these files are PLINK format (http://pngu.mgh.harvard.edu/~purcell/plink/) but you can open them as if they were text files.

The file 'QG14_project_phenotypes.txt' contains the phenotype data and has five rows. The 1st row contains the name of each individual, for the 2nd-5th rows, each contains the expression value of a gene for all of the individuals, where the gene name is in the first row entry.

The file 'QG14_project_covariates.txt' contains the covariate data and has four columns. The 1st column contains the name of each individual, the 2nd column contains the abbreviation used to refer to the population of the individual, the 3rd column contains the geographic region of the population of the individual, and the 4th column contains the gender of the individual (1=female, 2=male). The first row of the file contains the names of the columns.

The file 'QG14_project_genotypes.ped', contains the genotype information for each individual, where each row corresponds to an individual. The 1st and 2nd columns contain the names of the individuals, the 3rd and 4th columns contain zero's, the 5th column contains the gender of the individual, the 6th column contains the entry '-9', and all of the following columns indicate SNP genotypes (in order) where each genotype is indicated by a pair of columns, i.e. the 7th and 8th column indicate the first SNP genotype, the 9th and 10th columns indicate the second genotype etc. Note that missing genotypes are usually indicated by '-9' in PLINK format (or sometimes '0') and there may or may not be missing data.

The file 'QG14_project_genotypes.map' contains the additional information on the genotypes and has four columns. The 1st column contains the chromosome number of each SNP, the 2nd column contains the abbreviation used to the 'rsID' = the name of each SNP in order (e.g. the first entry of this column is the name of the SNP where the genotype is in the 7th and 8th column of the first row of the file 'QG14_project_genotypes.ped'), the 3rd column contains all zeros, and the 4th column contains the physical position of the SNP on the chromosome.

3 Your assignment and hints for getting started

Your GWAS assignment is to find the position of as many causal polymorphisms as possible for the four expressed genes using the data (note that each 'hit' will potentially indicate an eQTL). You may / should use any and as many analysis approaches as you think that are useful to accomplish this goal. In your report, you will need to describe in detail what you did, why you did it, and describe results in a manner that your 'non-statistical' collaborator will be able to understand, e.g. explain your terms, provide interpretations, etc.

A few hints:

- Apply the applicable steps of a ‘minimum GWAS’ analysis (see lecture 21).
- In your report, justify why you applied each individual step and statistical approach.
- In your report, provide a summary of your results and what they mean.
- You may want to consider going to various resources online (e.g. genecards, UCSC genome browser, dbSNP, many others) to incorporate biological information into your interpretation and hypotheses concerning what you may have found.
- Ask Amanda, Jin, and Jason for thoughts and ideas!

Good luck!