BTRY 4830/6830: Quantitative Genomics and Genetics Fall 2012

Project (Version 1)

Posted November 6; Due 11:59PM November 20

1 Introduction and instructions

Introduction: 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 GWAS data by a collaborator who wants to identify positions of causal polymorphisms (loci). You will perform an in-depth analysis of their GWAS data 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 Monica's inbox by 11:59PM, November 20 - 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.

(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. For your Figures and Tables, note that clarity and clear labels is a strategy for maximizing your grade.

(8) For your R code, to 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.

(9) 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 GWAS experiment and the GWAS data

The GWAS experiment: The GWAS data provided to you were collected from an experiment involving *Drosophila melanogaster* (the fruitfly!). The experimental design involved: 1. random selection of females from a wild population, and 2. brother-sister matings among the offspring of individual females for many generations where there was no crossing among the descendants of each original female. The result of this inbreeding experimental design is multiple lines (each line descended from a single female), where every individual in the line is (mostly) homozygous *for the same genotype*! That is, up to an approximation, we can assume that every individual in the line has the same genome! Note that we have not discussed why such lines are a consequence of inbreeding but we will discuss this concept in one of our optional lectures (inbred line designs). The advantage of this type of design is we can measure multiple individuals within the same line, where it is as if we measured the same individual produced under the same condition many times, i.e. you may think of individuals within a line as clones.

The phenotype provided to you is a measurement of the size of a specific area of the wing, where in the data provided to you we have averaged the values within a line, i.e. your n is the number of lines, where each line has one phenotype value. Each line has also been sequenced using next-generation technology such that we have very complete information about all the genetic variation that differentiates one line from another, although for this experiment, we have provided you SNP data for just a few thousand markers (out of several million SNPs!). The total sample for your GWAS is therefore n lines, one phenotype, N genotypes, and no covariates.

The GWAS data: These have been provided to you in three total files: 'project_phenotype.txt', 'project_genotype.ped', 'project_information.map'. Note that these files are PLINK format (google 'PLINK' to read more about it) but you can open them as if they were text files.

The file 'project_phenotype.txt' contains the phenotype data, where each row of the file has information for one of the lines and where the first column is all zero's, the second column indicates the line number, and the third column indicates the average phenotype for the line.

The file 'project_genotype.ped', contains the genotype information for each line, where each row corresponds to an individual, where the first column is all zero's, the second column indicates the line number, the 3rd through 5th columns are all zeros, the sixth column is all -9's, and all of the following columns indicate SNP genotypes (in order) where each genotype is indicated by a pair of columns, i.e. columns 7 and 8 indicate the first genotype, columns 9 and 10 indicate the second genotype, etc. Note that missing genotypes are indicated by zeros!

The file 'project_information.map' contains information on the locations of each of the SNP genotypes in the *Drosophila* genome, where the first row corresponds to the genotype in columns 7 and 8 (in the genotype file), the second row corresponds to the genotype in columns 9 and 10, etc. where the first column indicates the chromosome, the second column contains the 'chromosome name dot SNP position', the 3rd column is all zero's and the 4th column indicates the SNP position.

3 Hints for getting started

A few hints:

- Apply the applicable steps of a 'minimum GWAS' analysis (see lecture 20).
- Apply more than one GWAS statistical testing approach.
- 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 'Flybase' (google it!) to incorporate additional information into your interpretation.