Quantitative Genomics and Genetics - Spring 2023
BTRY 4830/6830; PBSB 5201.01

Midterm Exam

Available on CMS by 11AM (ET), Weds., March 29
Due 11:59PM (ET) Fri., March 31

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

1. **YOU ARE TO COMPLETE THIS EXAM ALONE!** The exam is open book, so you are allowed to use any books or information available online, your own notes and your previously constructed code, etc. **HOWEVER YOU ARE NOT ALLOWED TO COMMUNICATE OR IN ANY WAY ASK ANYONE FOR ASSISTANCE WITH THIS EXAM IN ANY FORM e.g., DO NOT POST PUBLIC MESSAGES ON PIAZZA!** (the only exceptions are Mitch, Sam, and Dr. Mezey, e.g., you MAY send us a private message on PIAZZA). As a non-exhaustive list this includes asking classmates or ANYONE else for advice or where to look for answers concerning problems, you are not allowed to ask anyone for access to their notes or to even look at their code whether constructed before the exam or not, etc. You are therefore only allowed to look at your own materials and materials you can access on your own. In short, work on your own! Please note that you will be violating Cornell's honor code if you act otherwise.

2. Please pay attention to instructions and complete ALL requirements for ALL questions, e.g. some questions ask for R code, plots, AND written answers. We will give partial credit so it is to your advantage to attempt every part of every question.

3. A complete answer to this exam will include R code answers, where you will submit your .Rmd script and the results of running your code in an associated .pdf file (plus an additional .pdf files if you have separate files for your written answers and code output). Note there will be penalties for scripts that fail to compile (!!). Also, as always, you do not need to repeat code for each part (i.e., if you write a single block of code that generates the answers for some or all of the parts, that is fine, but do please label your output that answers each question!!).

4. The exam must be uploaded on CMS before 11:59PM (ET) Fri., March 31. It is your responsibility to make sure that it is in uploaded by then and no excuses will be accepted (power outages, computer problems, Cornell's internet slowed to a crawl, etc.). Remember: you are welcome to upload early! We will deduct points for being late for exams received after this deadline (even if it is by minutes!!).

Consider the data in the files ('midterm_phenotypes.txt'; 'midterm_genotypes.txt') of the scaled height phenotypes and SNP genotype data (respectively) collected in a GWAS. Note that in the 'phenotypes' file the column lists the individuals in order (1st entry is the phenotype for individual 1, the nth is for individual n). In the 'genotypes' file, the FIRST row is a header where each entry is the name of a SNP. In this file, each column represents a specific SNP (column 1 = SNP1, column 2 = SNP2) the SNPs in the file are listed in order along the genome such that the first SNPis 'SNP1' and the last is 'SNPN'. and each consecutive pair of rows (past the 1st row) represent all of the genotype states for an individual for the entire set of SNPs (row 1 = header, rows 2 and 3 = all of individual 1's genotypes, rows 4 and 5 = all individual 2's genotypes). Also note that for each of the SNPs, there are two total alleles, i.e. two letters for each SNP and there are three possible states per SNP genotype: two homozygotes and a heterozygote.

1. Import the phenotype data from the file 'midterm_phenotypes.txt' and **(a)** Calculate and report the total sample size n, **(b)** Plot a histogram of the phenotypes. NOTE: do not filter or change these data in any way, just analyze them as given!

2. Import the genotype data from the file 'midterm_genotypes.txt', **(a)** Calculate and report the number of SNPs N, **(b)** Calculate the MAF for every SNP and plot a histogram of the MAFs. NOTE: do not filter or change these data in any way, just analyze them as given!

3. Write code to calculate $MLE(\hat{\beta}) = [\hat{\beta}_\mu, \hat{\beta}_a, \hat{\beta}_d]$ for each SNP and **(a)** plot a histogram of all the $\hat{\beta}_\mu$, **(b)** plot a histogram of all the $\hat{\beta}_a$, **(c)** plot a histogram of all the $\hat{\beta}_d$.

4. For each SNP, calculate p-values for the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ versus the alternative hypothesis $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$ when applying the genetic linear regression model. NOTE (!!): in your linear regressions, DO use the X_a and X_d codings provided in class and DO NOT use the function lm() (or any other R function!) to calculate your p-values but rather use the $MLE(\hat{\beta})$ you calculated in question [3] and use the formula provided in class to calculate the predicted values of the phenotype $\hat{y}_i$ for each individual i, and calculate the F-statistic (although you MAY use the function pf() to calculate the p-value for each F-statistic you calculate!).

5. For the p-values you calculated in question [4], (a) Produce a QQ plot for these p-values (label your plot and your axes using informative names!), (b) USING NO MORE THAN TWO SENTENCES answer the following question: based on this QQ plot, do you think you have achieved an appropriate model fit with your analysis and why do you think this is the case?

6. For the p-values you calculated in question [4], **(a)** Produce a Manhattan plot, **(b)** Report HOW MANY SNPs (not which, just how many!) you find to be significant when controlling the study-wide type 1 error of 0.05 using a Bonferroni correction (note: do NOT use adjusted p-values! Just use the p-values you calculated in question [4]),

7. USING NO MORE THAN TWO SENTENCES answer the following question: based on your answer the question [6], how many distinct 'peaks' do you think you have identified and why?

8. Imagine you are explaining the outcome of your analysis to your biological collaborator who does not have a deep understanding of a GWAS. Answer the following: **(a.)** What is a causal polymorphism? **(b.)** USING NO MORE THAN TWO SENTENCES describe why do you observe 'peaks' in your Manhattan plot? **(c.)** USING NO MORE THAN TWO SENTENCES

describe why the peaks in your Manhattan plot may indicate the genomic position of a causal polymorphism but not (necessarily) the actual causal polymorphism? **(d.)** Provide one reason why a peak may NOT indicate the position of a causal polymorphism.

9. **(a)** Provide a rigorous definition of the 'power' of a hypothesis test. **(b)** List three factors that could impact the power of a hypothesis test in a GWAS. NOTE: don't explain why each factor impacts power, just list them!

10. **(a)** Provide a rigorous definition of a random variable, **(b)** Provide a rigorous definition of a statistic, **(c)** Provide a rigorous definition of a p-value. NOTE: you are welcome to use the exact definitions provided in class!