Quantitative Genomics and Genetics - Spring 2019
BTRY 4830/6830; PBSB 5201.01

Midterm - available, Fri., April 12

Midterm exam due before 11:59PM, Sun., April 14

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** (the only exceptions are Olivia, Scott, and Dr. Mezey). 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 in Rmarkdown, where you will submit your .Rmd script and associated .pdf file. 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!!). You should include all of your plots and written answers in this same .Rmd script with your R code.

4. The exam must be uploaded on CMS before 11:59PM Sun., April 14. 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!!).

Your collaborator is interested in mapping genetic loci that can affect Blood Pressure (BP) in humans. They know there are loci scattered throughout the genome that can affect BP, but they do not know the locations of these loci, so they have performed a GWAS experiment and they would like you to perform the analysis. They have collected data for a number of individuals from two populations representing distinct ancestry groups. They have provided you the following data: scaled BP phenotypes and population assignment ('midterm2019_pheno+pop.csv'), and SNP genotypes ('midterm2019_genotypes.csv'). In the file containing phenotypes and population assignments, the first column indicates the sample ID of each individual and the second column the scaled BP measurements for that individual (i.e., row 1 contains the ID for the first individual and the BP for the first individual, row 2 contains the sample ID and BP for the second individual, etc.). Note that the population assignment for each individual is also indicated in the sample ID column of this file, where every individual with a sample ID starting with 'HG' is in the first population and every individual with a sample ID starting with 'NA' is in the second population. In the file containing the SNP genotypes, the genotype state for a homozygote is coded as either '0' or '2' and heterozygote is coded as '1'. In this file each column represents a specific SNP (column 1 = SNP 1, column 2 = SNP 2) and each row represents all of the SNP genotype states for an individual for the entire set of SNPs (row 1 = all of the first individual's genotypes, rows 2 = all of the second individual's genotypes, etc.). Also note that the SNPs in the file are listed in order along the genome such that the first SNP is 'SNP 1' and the last is 'SNP N'.

1. **(a)** Import the sample ID and BP data from the file 'midterm2019_pheno+pop.csv', **(b)** Calculate and report the total sample size n, **(c)** Plot a histogram of the BP phenotypes (label your plot and your axes using informative names!), **(d)** The shape of your histogram will deviate some from a normal distribution. Using no more than two sentences, provide a qualitative explanation for how the shape of this histogram deviates from a normal distribution (i.e., describe in words how it deviates) and also explain from the information provided to you about the GWAS data, what might explain this deviation and why a linear regression may still be appropriate for analyzing these GWAS phenotypes (i.e., again, explain in words / no analysis!), **(e)** Using no more than one sentence, provide ONE reason why a logistic regression would NOT be appropriate for analyzing these GWAS phenotypes.

2. **(a)** Import the genotype data from the file 'midterm2019_genotypes.csv', **(b)** Calculate and report the number of SNPs N, **(c)** Calculate the minor allele frequency (MAF) for each SNP and plot a histogram of these MAF values (NOTE: that the minor allele homozygotes may be encoded with 0 or 2, depending on which SNP you are considering. Also, please label your plot and your axes using informative names!). **(d)** Provide a rigorous definition of the 'power' of a hypothesis test, **(e)** If you were to somehow perform a genetic linear regression hypothesis test on a causal SNP directly, explain to your collaborator how you would expect the MAF of this causal SNP to impact the power of the test (i.e., your answer should be a simple description of how power relates to allele frequency).

3. **(a)** Using the phenotype and genotype data you have imported in '1a' and '2a', for each genotype, 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 a genetic linear regression model with NO covariates. NOTE (!!): in your linear regressions, DO use the X_a and X_d codings provided in classand DO NOT use the function lm() to calculate your p-values but rather calculate the $MLE(\hat{\beta})$ using the formula provided in class, calculate the predicted value of the phenotype $\hat{y}_i$ for each individual i under the null and alternative and use these to calculate SSM and

SSE, and use the formulas for MSM and MSE to calculate the F-statistic, although you may use the function pf() to calculate the p-value for each F-statistic you calculate. **(b)** Produce a Manhattan plot for these p-values (label your plot and your axes using informative names!).

4. **(a)** Produce a Quantile-Quantile (QQ) plot for the p-values you produced in '3a'. **(b)** Using no more than three sentences, explain to your collaborator whether you think the analysis you have applied resulted in appropriate model fit to these GWAS data based on the shape of this QQ plot and provide an explanation for why, if your explanation is correct, you would expect a QQ plot with the observed shape.

5. **(a)** From the sample IDs you imported from the file 'midterm2019_pheno+pop.txt', calculate and report the number of individuals in the first population n_1 and in the second population n_2 (where $n_1 + n_2 = n$). *HINT: Try using the* **substr()** *function to extract a chunk of a string based on the positions of the characters. For example,* substr('a substring' , 3, 5) *will take the 3rd through 5th characters of the string in the first argument, returning the string 'sub' as the output.* **(b)** Using no more than two sentences, provide an intuitive explanation to your collaborator as to why a PCA could have been used to show that the individuals in this sample represent two populations with distinct ancestries if you did not have this population information in the data (NOTE: you do not need to run a PCA!).

6. **(a)** Using the phenotype, population, and genotype data you have imported in '1a' and '2a', for each polymorphism, 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 a genetic linear regression model with WITH A COVARIATE X_Z that codes each of the n_1 individuals as '-1' and the n_2 individuals as '1'. NOTE (!!): in your linear regressions, DO use the X_a and X_d codings provided in class along with your X_Z dummy variable and DO NOT use the function lm() to calculate your p-values but rather calculate the $MLE(\hat{\beta})$ using the formula provided in class, calculate the predicted value of the phenotype $\hat{y}_i$ for each individual i under the null and alternative and use these to calculate SSE($\hat{\theta}_0$) and SSE($\hat{\theta}_1$), and use these to calculate the F-statistic, although you may use the function pf() to calculate the p-value for each F-statistic you calculate. **(b)** Produce a Manhattan plot for these p-values (label your plot and your axes using informative names!).

7. **(a)** Produce a Quantile-Quantile (QQ) plot for the p-values you produced in '6a'. **(b)** Using no more than three sentences, explain to your collaborator whether you think the analysis you have applied resulted in appropriate model fit to these GWAS data based on the shape of this QQ plot and an explanation as to why the QQ plot has the observed shape.

8. **(a)** For the p-values you produced in '6a' when controlling the study-wide type 1 error of 0.05, report the appropriate p-value cutoff for assessing which genetic markers are significant when using a Bonferroni correction and provide the formula you used to calculate this cutoff. **(b)** Given the Manhattan plot in '6a', report how many separate peaks you observe that are greater than the Bonferroni corrected cutoff and, using no more than two sentences, provide a description of the criteria you used to determine the number of separate peaks. **(c)** Using no more than two sentences, explain to your collaborator how many causal polymorphisms you believe each of these separate peaks are indicating and why. **(d)** Using no more than two sentences, explain to your collaborator why these peaks likely ONLY indicate the positions of causal polymorphisms and why it may not be possible to determine the exact causal polymorphisms that are impacting BP from these GWAS data / your analysis.

9. **(a)** For each of these separate peaks you identified in '8b', list the p-value of the most significant SNP in the peak and number of this SNP (i.e., each genotype you list will have a number between 1 and N where remember the SNPs are provided in order). **(b)** Using no more than two sentences, provide ONE explanation as to why the most significant SNP in each peak is not necessarily closest to a causal polymorphism when considering all of the SNPs you have analyzed. **(c)** For each peak, for the most significant SNP in the peak, use the X_a coding of this SNP and calculate the correlation with the (closest) SNP on either side using their respective X_a codings. **(d)** Using no more than two sentences, explain why it makes sense that these correlations are not that close to '0' given what you know about linkage disequilibrium (LD).

10. **(a)** Provide a rigorous definition of a causal polymorphism. **(b)** Using no more than two sentences, describe an ideal experiment (not necessarily realistic!) for demonstrating that a polymorphism is causal. **(c)** Provide a rigorous definition of a p-value. **(d)** Provide three reasons why it could be that in a case where you reject a null hypothesis for a polymorphism in a GWAS there will be no causal polymorphism in the genomic location of the polymorphism you analyzed (i.e., explain why the polymorphism for which you reject the null hypothesis could be a biological false positive).