Quantitative Genomics and Genetics - Spring 2022
BTRY 4830/6830; PBSB 5201.01

Final Exam

Available on CMS on Weds., May 11
Due 11:59PM (ET) Sat., May 21

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 Beulah, Yajas, 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 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 (ET) Sat., May 21. 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 the location of causal polymorphisms that can impact risk for Chronic Kidney Disease and causal polymorphisms that can impact the level of the kidney metabolite Citrulline in humans, so they have performed a GWAS experiment and they would like you to perform the analysis. They have provided you the following data: Citrulline Level measured for each of the n individuals in the file '2022QG_finalexam_citrulline.csv', where each row is the value of the Citrulline Level phenotype measured for an individual; Kidney Disease measured for each of the n individuals in the file '2022QG_finalexam_kidneydisease.csv', where each row is the Kidney Disease phenotype measured for an individual (Healthy = 0, Kidney Disease = 1); and SNP genotype data in the file '2022QG_finalexam_genotypes.csv' which contains data on N SNPs measured for each of the n individuals in the sample, with each row containing all of the genotypes measured for a specific individual (e.g., row 1 = all of the first individual's genotypes, row 2 = all of the second individual's genotypes, etc.) and each column presenting data for a specific SNP (e.g., column 1 = SNP 1, column 2 = SNP 2) and where the genotype of each SNP for an individual is coded as follows: '0' = homozygote, '1' = heterozygote, '2' = homozygote. Also note that the SNPs in the file are listed in order along the genome such that the first SNP (column 1) is 'SNP 1' and the last (column N) is 'SNP N'. PLEASE NOTE: do NOT filter these genotypes (i.e., for these questions, use all N SNPs)!

1. **(a)** Import the Citrulline Level phenotypes. **(b)** Plot a histogram of the Citrulline Level phenotypes (label your plot and your axes using informative names!). **(c)** Import the Kidney Disease phenotypes. **(d)** Plot a histogram of the Kideny Disease phenotypes (label your plot and your axes using informative names!). **(e)** Report the sample size n.

2. **(a)** Import the genotype data. Again please note: do NOT filter these genotypes for this or any of the questions, i.e., use all N! **(b)** Report the number of SNPs N. **(c)** Calculate the MAF for each SNP. **(d)** Plot a histogram of the N MAFs you calculated.

3. **(a)** Using the Citrulline Level phenotype you have imported in question [1a] and the genotype data you have imported in question [2], 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 class and DO NOT use the function lm() (or any other R function!) 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 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!). **(c)** Produce a QQ plot for these p-values (label your plot and your axes using informative names!).

4. **(a)** Using the Kidney Disease phenotype you have imported in question [1c] and the genotype data you have imported in question [2], 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 logistic regression model with NO covariates. NOTE (!!): in your logistic regressions, DO use the X_a and X_d codings provided in class. DO NOT use any functions in R that apply a logistic regression but DO use the IRLS algorithm presented in class and the appropriate formulas for the MLE and LRT presented in class, although you may use the function pchisq() to calculate the p-value for each LRT you calculate. **(b)** Produce a

Manhattan plot for these p-values (label your plot and your axes using informative names!). **(c)** Produce a QQ plot for these p-values (label your plot and your axes using informative names!).

5. Given the QQ plots you produced in questions [3] and [4], using no more than two sentences, explain whether you think the analyses you have applied resulted in appropriate model fit to the data in each case and explain the reasoning behind your answer based on the shapes of the QQ plots.

   The analyses **DID NOT** result in appropriate models, because according to both QQ plots in 3c and 4c, black lines seem to be linear, which means we observe a uniform distribution of p-values (after they were -log transformed). We do not have enough information to detect causal polymorphism positions, and we need to add covariates.

6. **(a)** Perform a PCA on all N genotypes you imported in question [2] using the R code: 'geno_pca <- prcomp(genotypes)' where 'genotypes' is the R data object resulting from you genotype data import. **(b)** Create a plot that projects the n samples onto PC1 and PC2 using the following R code 'plot(geno_pca$x[,1], geno_pca$x[,2], main = "Genotype PC projections", xlab = "PC1", ylab = "PC2")'.

7. **(a)** Using the Citrulline Level phenotype you have imported in question [1a] and the genotype data you have imported in question [2], 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 THE FIRST TWO PCs calculated in question [6] as covariates $X_{z,1}$ and $X_{z,2}$ by setting $X_{z,1}$ = geno_pca$x[,1] and $X_{z,2}$ = geno_pca$x[,2]. 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 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 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!). **(c)** Produce a QQ plot for these p-values (label your plot and your axes using informative names!).

8. **(a)** Using the Kidney Disease phenotype you have imported in question [1c] and the genotype data you have imported in question [2], 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 logistic regression model WITH THE FIRST TWO PCs calculated in question [6] as covariates $X_{z,1}$ and $X_{z,2}$ by setting $X_{z,1}$ = geno_pca$x[,1] and $X_{z,2}$ = geno_pca$x[,2]. NOTE (!!): in your logistic regressions, DO use the X_a and X_d codings provided in class. DO NOT use any functions in R that apply a logistic regression but DO use the IRLS algorithm presented in class and the appropriate formulas for the MLE and LRT presented in class, although you may use the function pchisq() to calculate the p-value for each LRT you calculate. **(b)** Produce a Manhattan plot for these p-values (label your plot and your axes using informative names!). **(c)** Produce a QQ plot for these p-values (label your plot and your axes using informative names!).

9. Given the QQ plots you produced in questions [7] and [8], using no more than two sentences, explain whether you think the analyses you have applied resulted in appropriate model fit to the data in each case and explain the reasoning behind your answer based on the shapes of the QQ plots.

The analyses **DID** result in appropriate models, because QQ plots in 7c and 8c are ideal. This happens because most of the p-values observed follow a uniform distribution (i.e. they are not in LD with a causal polymorphism so the null hypothesis is correct!) but the few that are in LD with a causal polymorphism will produce significant p-values (extremely low = extremely high -log(p-values)) and these are in the "tail".

10. **(a)** For question [7b], your Manhattan plot should show two clear peaks. Report the number of the SNP (i.e., a number between '1' and 'N') that has the most significant p-value for EACH peak (i.e., report the numbers of two SNPs total!) and for each of these two SNPs, answer whether you would reject the null hypothesis when controlling the study-wide Type 1 error to 0.05 using a Bonferroni correction (i.e., for each of the two SNPs, answer "Yes" or "No" as to whether you reject the null hypothesis!) and provide the formula you used to calculate this cutoff as part of your answer. **(b)** For question [8b], your Manhattan plot should show two clear peaks. Report the number of the SNP (i.e., a number between '1' and 'N') that has the most significant p-value for EACH peak (i.e., report the numbers of two SNPs total!) and for each of these two SNPs, answer whether you would reject the null hypothesis when controlling the study-wide Type 1 error to 0.05 using a Bonferroni correction (i.e., for each of the two SNPs, answer "Yes" or "No" as to whether you reject the null hypothesis!) and again provide the formula you used to calculate this cutoff as part of your answer. **(c)** Are the two peaks in your Manhattan plot for question [7b] indicating the positions of the same causal polymorphisms as the two peaks in your Manhattan plot for question [8b]? Explain your reasoning using no more than two sentences. **(d)** For each of the peaks in question [7b] and question [8b], is the most significant SNP in each peak necessarily closer to the causal polymorphism (assuming the peak indicates a causal polymorphism!) than either of the SNPs on each side of the most significant SNP? Use no more than two sentences in your answer.

10A: 9882 20138 20140

For SNPs No. [9882, 20138, 20140], we can use Bonferroni correction to reject the null hypothesis. I used the formula

$$\alpha = \frac{0.05}{N}$$

For each of the SNP, if its p-value is smaller than the new alpha value, we can reject the null hypothesis.

10B: 4210 20138 20140

For SNPs No. [4210, 20138, 20140], we can use Bonferroni correction to reject the null hypothesis. I used the formula

$$\alpha = \frac{0.05}{N}$$

For each of the SNP, if its p-value is smaller than the new alpha value, we can reject the null hypothesis.

10C: Because both 7B and 8B share SNP No. 20138 same, they indicate the positions of the same causal polymorphisms.

10D: Not necessarily. According to linkage disequilibrium, there is a trade-off between resolution and the ability to map causal polymorphisms and that there is a theoretical limit to the resolution of a GWAS experiment. Sometimes p-value at a site can be raised by a nearby site p-value.