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

Final - available, Sun., May 5

Final exam due before 11:59PM, Tues., May 7

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 Tues., May 7. 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 Type I Diabetes (T1D) in humans. They know there are loci scattered throughout the genome that can affect T1D, 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: T1D phenotypes ('final2019_pheno.csv'), and SNP genotypes ('final2019_genotypes.csv'). In the file containing phenotypes, the column indicates the phenotype (health = 0 / T1D = 1) for each individual (i.e., row 1 contains phenotype of the first individual, row 2 contains the phenotype of the second individual, etc.). 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 phenotype data from the file 'final2019_pheno.csv', **(b)** Calculate and report the total sample size n, **(c)** Plot a histogram of the phenotypes (label your plot and your axes using informative names!).

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!).

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 logistic regression model with NO COVARIATES. NOTE (!!): in your logistic regressions, DO use the X_a and X_d codings provided in class and DO NOT use an existing function in R function to calculate your p-values but rather implement the IRLS algorithm to calculate the $MLE(\hat{\beta})$ under the null and alternative hypotheses and use the formula for the Likelihood Ratio Test (LRT) statistic provided in class to calculate the p-value, where 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 Quantile-Quantile (QQ) plot for these p-values (Same!).

4. **(a)** Perform a PCA on the genotypes **(b)** create a plot that projects the samples onto PC1 and PC2 (label your plot and your axes using informative names!)

5. **(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 THE FIRST PC (calculated from question '5') AS A COVARIATE X_Z. NOTE (!!): in your logistic regressions, DO use the X_a and X_d codings provided in class and DO NOT use an existing function in R function to calculate your p-values but rather implement the IRLS algorithm to calculate the $MLE(\hat{\beta})$ under the null and alternative hypotheses and use the formula for the Likelihood Ratio Test (LRT) statistic provided in class

to calculate the p-value, where 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 Quantile-Quantile (QQ) plot for these p-values. (label your plot and your axes using informative names!).

6. **(a)** For the p-values you produced in '5a' 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 '5b', 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.

7. **(a)** Consider a (hypothetical) GWAS with $n = 4$ samples, where we are using a mixed model to analyze each marker:

$$\mathbf{Y} = \beta_\mu + \mathbf{X_a}\beta_a + \mathbf{X_d}\beta_d + \mathbf{a} + \epsilon \tag{1}$$

$$\epsilon \sim multiN(\mathbf{0}, \mathbf{I}\sigma^2_\epsilon) \tag{2}$$

$$\mathbf{a} \sim multiN(\mathbf{0}, \mathbf{A}\sigma^2_a) \tag{3}$$

Say that you are given the following $\mathbf{A}$ matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 0.5 & 0 & 0 \\ 0.5 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0.5 \\ 0 & 0 & 0.5 & 1 \end{bmatrix}$$

For a marker where the true $\beta = [\beta_\mu, \beta_a, \beta_d] = [1.5, 0, 0]$, answer the following questions about the $Pr(\mathbf{Y}|\mathbf{X})$ for the phenotypes of the sample under the model in equations (1-3), i.e., for these questions, list out the sample pairs appropriate for each (e.g., $(1,2), (1,3)$, etc.): **(b)** Which of the sample pairs are positively correlated? **(c)** Which of the sample pairs are negatively correlated? **(d)** Which of the sample pairs are uncorrelated?

8. Narrow sense heritability (h^2) describes a property of a phenotype in a population impacted by genetics. The formula for narrow sense heritability is:

$$h^2 = \frac{V_A}{V_P} \tag{4}$$

where V_P is the variance of the phenotype in the population and V_A accounts for the variance attributable to the (orthogonal) linear impacts of allele substitutions. While additive genetic variance V_A has quite complicated formulas for models that can account for more general genetic systems, for a phenotype that can be modeled with a linear regression (i.e. normal error term) and considering only a single locus (with two alleles), the V_A has the following formula:

$$V_A = 2MAF(1 - MAF)\beta^2_\alpha \tag{5}$$

where the parameter β_α is determined by the following linear regression model:

$$Y = \beta_\mu + X_a\beta_\alpha + \epsilon \tag{6}$$

INSTEAD of the genetic regression model:

$$Y = \beta_\mu + X_a\beta_a + X_d\beta_d + \epsilon \tag{7}$$

(a) What is the relationship between the parameter β_α in equation (6) and β_a, β_d in equation (7)? **(b)** What is the narrow sense heritability when MAF = 0?

9. For a genetic system, assume there are two causal polymorphisms with alleles '$A1$' and '$A2$' and alleles '$B1$' and '$B2$': **(a)** Write out the 9 possible causal genotype combinations that could occur for this system. **(b)** Write out the values of $X_{a,1}, X_{d,1}, X_{a,2}, X_{d,2}$ for the genotype $A_1A_1B_1B_1$ (NOTE: it does not matter which homozygote is '-1' for the X_a codings!) **(c)** Assume that a linear regression model is the correct model for this genetic system and the parameter values are as follows (where there is epistasis between the two causal polymorphisms!):

$$\beta = [\beta_\mu = 0.2, \beta_{a,1} = 0.1, \beta_{d,1} = 0.2, \beta_{a,2} = -0.3,$$
$$\beta_{d,2} = 0.17, \beta_{a1a2} = -0.11, \beta_{a1d2} = 0.32, \beta_{d1a2} = 0.08, \beta_{d1d2} = -0.03]$$

What is the expected phenotypic value of an individual with the $A_1A_1B_1B_1$ genotype? NOTE: write out the equation you used to do this calculation as part of your answer!

10. Show your understanding of the basic concepts of probability and statistics by answering the following questions concerning a coin (system) that you would like to know about, where the question of interest is whether it is a 'fair' coin, where you will attempt to answer this question by making use of individual flips of the coin (i.e. experiment = single coin flip). **(a)** What is the sample space for the experiment? **(b)** What is the sigma algebra for this sample space? **(c)** For a fair coin model, what is the probability function on this sigma algebra? (i.e. for each event in the sigma algebra, you should write out the appropriate probability) **(d)** How would you define a random variable on this sample space such that a Bernouli distribution would be an appropriate probability distribution for the random variable? **(e)** What is the expected value of this random variable given the fair coin probability model (write out the equation and the calculations you used to get to this answer)? **(f)** If you were to generate a sample by performing 10 experimental trials that are i.i.d., how many outcomes are possible for this sample? **(g)** What is an example of a statistic that would be a 'terrible' estimator (i.e., an estimator that is wrong / not even close to the right answer for most or all possible samples) of the parameter p? **(h)** What is the maximum likelihood estimator (MLE) of the parameter p (i.e., for MLE($\hat{p}$) what is the equation)? **(i)** If you were to use the statistic $T(\mathbf{x})$ = 10*MLE($\hat{p}$) as your test statistic to assess the null hypothesis $H_0 : p = 0.5$ versus $H_A : p > 0.5$ for a one-sided test, what is the p-value if the observed value of the statistic is $T(\mathbf{x})$ = 10*MLE($\hat{p}$) = 8? **(j)** Would you reject H_0 in the case of 10i' for $\alpha = 0.05$ (i.e., yes or no answer)?