BTRY 4830/6830: Quantitative Genomics and Genetics
Fall 2012

Midterm, **VERSION 2** - available online Oct. 11

Due before 11:59PM, Oct. 15

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 Monica 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. A complete answer to this exam will include two files: a SINGLE text file including all of your R code, and a SINGLE file including all of your written answers and plots (where the latter may be a scan as long as we can read it). Please note that for your R code, to get full credit for all problems, we must be able to run your code and replicate all of your results (with ease!). The best way to do this is to make your file a script such that we can run all the code from the command line (or using "source") and/or you should provide us instructions on how to run your code. We will attempt to run your code if you do not do this but we will deduct points accordingly (note that no code = no credit!).

3. 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 you advantage to attempt every part of every question.

4. The exam must be in Monica's email inbox before 11:59PM Mon., October 15. It is your responsibility to make sure that it is in her email box before then and no excuses will be accepted (power outages, computer problems, Cornell's internet slowed to a crawl, etc.). Remember: you are welcome to hand this in 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 height in humans. They know there are loci scattered throughout the genome that can affect height, 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 sampled from a population and they have provided you scaled height phenotypes and SNP genotypes in two files ("midterm_phenotypes_fall12.txt" and "midterm_genotypes_fall12.txt"). Note that for the SNP genotypes, each column represents the genotypes for an individual and each pair of rows represents a genotype (rows 1 and 2 = genotype 1, rows 3 and 4 = genotype 2), i.e. two letters indicate each genotype and there are three possible combinations per genotype: two homozygotes and a heterozygote. Also note that the genotypes in the file are listed in order along the genome.

1. What is the sample size (n) and how many genotypes per individual are there (N)?

2. Produce a histogram of the phenotypes (label your plot and your axes using informative names!).

   Are these phenotypes approximately normally distributed? Why is this important if we want to test genotype associations using a linear regression model?

3. Write (and provide!) R code to calculate the $MLE(\hat{\beta})$ for the three β parameters when using the linear regression model $y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2_\epsilon)$ to model the relationship between each of the N genotype markers with the phenotype (note that your code must include the formula for the MLE). Using your MLE estimates, produce two x-y plots (and provide your R code that you used to construct these plots): for the first, plot the value of $\hat{\beta_a}$ for each of the N genotype markers on the y-axis and the order of the genotype markers as they are observed in the genotype data file on the x-axis (as you would in a Manhattan plot) and for the second, plot the value of $\hat{\beta_d}$ on the y-axis and the order of the genotype markers as they are observed in the genotype data file on the x-axis. Make sure you label your plots and axes for each using informative names!

   For this GWAS, what do we expect the true β_a and β_d values to be for the *majority* of the genotypes (explain your answer!)? Also explain why very few (none!) of your $\hat{\beta_a}$ and $\hat{\beta_d}$ estimates are exactly these values (provide a general definition of an estimator and explain the goal of constructing an estimator in your answer).

4. Write (and provide!) R code to calculate p-values testing the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ for each genotype with the phenotype (N total tests!), when using the linear regression model $y_i = \beta_\mu + X_{i,a}\beta_a + X_{i,d}\beta_d + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2_\epsilon)$. NOTE that you cannot use $lm()$ (or related functions in R) and that your code must include formulas for: the MLE of the β parameters, the predicted value of the phenotype $\hat{y}_i$ for an individual i, the SSM, SSE, MSM, MSE, and the F-statistic (although you may use the function $pf()$ or related to calculate the p-value from your F-statistic).

   In this case, what is the alternative hypothesis you are testing? What does the value of each p-value represent in this case (provide the general definition of a p-value in your answer)?

5. Produce a Manhattan plot (and provide your R code that you used to construct this plot). Make sure you label your plot and axes using informative names!

6. Assuming a type 1 error for an individual test of $\alpha = 0.05$, calculate and provide the Bonferroni corrected type 1 error for the entire GWAS. Add a vertical line to your Manhattan plot that corresponds to this Bonferroni threshold.

   What is the definition of Type 1 error? Explain the multiple testing problem and why we want to use a Bonferroni corrected Type 1 error in this case? Provide a definition of Type II error, explain why we cannot control Type II error, and explain the trade-off between Type 1 error and Type II error? Does applying a Bonferroni correction increase or decrease the Type II error of a test? Provide a definition of power, explain why we cannot control power, and explain the trade-off between Type 1 error and power? Does applying a Bonferroni correction increase or decrease the power of a test?

7. Provide a list of ALL genotype markers that have p-values that are considered significant by your Bonferroni corrected Type 1 error. Divide this list into sets where each corresponds to a distinct signal of a causal genotype. List the most significant marker in each of these sets (provide the marker number and the associated p-value) and justify why you divided them into these sets.

8. Calculate the Minor Allele Frequency (MAF) for the most significant genotype markers in each of the sets that you defined in question 7 (provide your R code).

   What is the relationship between MAF and power (provide an intuitive explanation for this relationship in your answer)?

9. List the most significant genotype marker identified in your GWAS overall (provide both the marker number and the p-value). For this most significant marker, calculate the correlation between its X_a and the X_a for EITHER the marker that is in the rows above OR the marker in the row below as listed in the genotype data file (list which marker you choose). Also for the most significant marker, calculate the correlation between its X_a and the X_a of marker number 184 in the genotype file (the 184th marker starting from marker 1 at the top of the file).

   Which of these correlations is higher? Why does this match our expectation for a GWAS (provide a definition of linkage disequilibrium in your answer)?

10. Explain the following to your collaborator: why all of the markers that were significant by a Bonferroni correction do not *individually* indicate causal genotypes affecting height? Why are none of the significant markers in the sets you have defined in question 7 likely to be the causal genotype for height (use the definition of causal genotype and linkage disequilibrium in your explanation)? Why each of the sets you have defined in question 7 indicate the existence of a causal genotype but do not *guarantee* the existence of a causal genotype?