# BTRY 4830/6830: Quantitative Genomics and Genetics
# Fall 2014

Midterm - available online Oct. 16

**For midterm exam, due before 11:59PM, Oct. 20**

**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 Amanda, Jin, 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 typed 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!). 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 Amanda's or Jin's (as appropriate) email inbox before 11:59PM Mon., October 20. It is your responsibility to make sure that it is in the appropriate 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 human height. They know there are loci scattered throughout the genome that can affect this phenotype, 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 relative measures of height in the file "QG14_phenotypes_midterm.txt" and SNP genotypes in the file "QG14_genotypes_midterm.txt". Note the following:

- Each row of the phenotype file is the relative height value for an individual in the sample (i.e. the height of the 1st sample is in the 1st row, the height of the 2nd sample is in the 2nd row, ..., the height of the nth sample is in the nth row).
- In the "genotypes" file, the first row is an 'index' where each entry is the name of a SNP (e.g. G1, G2, ..., G$N$) and each COLUMN of this file represents a specific SNP (i.e. 1st column (G1) = SNP 1, 2nd column (G2) = SNP 2, ... $N$th column (G$N$) = SNP $N$).
- In the "genotypes" file, after the first row (SNP names) each consecutive PAIR OF ROWS represent all of the states for the $N$ SNP genotypes for an individual (1st and 2nd rows = $N$ SNP genotypes for the 1st individual, 3rd and 4th rows = $N$ SNP genotypes for the 2nd individual, ..., nth-1 and nth rows = $N$ SNP genotypes for the nth individual).
- For each SNP, there are two alleles each represented by a letter (e.g. the 1st SNP has alleles 't' and 'a', the 2nd SNP has alleles 'c' and 'g', etc.). An individual's genotype at a specific SNP is composed of a pair of alleles (e.g. the possible genotypes at the 1st SNP are 'tt', 'ta', 'aa').

QUESTIONS (10 total, multiple parts per question) - make sure you answer all parts of all questions (!!):

1. **(a)** Plot a histogram of the phenotypes (provide your code!). **(b)** What probability distribution could provide a reasonable model for these phenotypes given the histogram? (we are just looking for the name of the distribution - that's it! - you don't need to write any equations or provide any parameter values). **(c)** In no more than one sentence, explain why it is important that the phenotypes be well modeled by this distribution if we are going to use the genetic linear regression to model the relationships between genotypes and this phenotype?

2. **(a)** Calculate the minor allele frequency (MAF) for each SNP and plot a histogram of these MAFs (provide your code!). **(b)** Remove all SNPs that have an MAF $< 0.06$ and plot a new histogram of the MAFs of the SNPs that remain (provide your code!). **(c)** How many SNPs are left (i.e. what is $N$ after you remove these SNPs)?

   NOTE (!!) FOR QUESTIONS #3-8 use only the SNPs that remain after removing those with an MAF $< 0.06$ retaining their original index from the SNP file (e.g. for SNPs G1,G2,G3,G4,G5, ...,G$N$ if you remove SNPs 2 and 4, the SNP indexes you will use will be G1,G3,G5,...,G$N$)!

3. **(a)** For EACH genotype, using the formulas provided in class, calculate the $MLE(\hat{\beta})$ for the three $\beta$ parameters when using the linear regression model $y_i = \beta_\mu + x_{i,a}\beta_a + x_{i,d}\beta_d + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ and plot a histogram for the estimates of each parameter = three histograms

total (provide your code! and make sure you label your plots!). **(b)** For EACH genotype, calculate p-values for testing the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ versus the alternative hypothesis $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$ using the formulas provided in class (i.e. 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() to calculate the p-value from your F-statistic. Plot a Manhattan plot using these p-values (provide your code!).

4. **(a)** Plot a histogram of ALL p-values (not -log(p)! just the p-values!) that you obtained in question #3 (provide your code!). **(b)** Explain why the shape of this histogram makes sense given what you know about p-values, particularly if you had successfully 'tagged' a few causal polymorphisms with the markers in your study?

5. **(a)** For a type 1 error of 0.05, what is the appropriate p-value cutoff for assessing which genetic markers are significant from your analysis in question #3 when using a Bonferroni correction? **(b)** Use the following criteria to identify a significant 'hit' that indicates a causal polymorphism:

   1. Order the markers by their p-values.
   2. Begin with the marker that has the most significant p-value (if there is a tie, start with either!) and determine whether it is significant after a Bonferroni correction:
   - if so, assume the marker indicates a causal polymorphism and continue to step 3
   - if not, assume the marker does not indicate a causal polymorphism and DO NOT continue on to step 3 (=STOP and do not consider this or any more markers significant!).
   3. Assume that the 50 neighboring markers on EACH side (=100 markers total) of the significant marker are tagging the same causal polymorphism, RETURN to step 1 but do not consider the significant SNP or the 100 neighboring SNPs when you start at step 1 (in other words, all 101 SNPS are assumed to represent a single peak = one hit per peak!).

   Using this approach, determine how many causal polymorphisms (='hits') are indicated and list the marker numbers of the most significant marker (=in step 2, using the SNP indexes after removal of those with MAF $<$ 0.06) for EACH separate hit you identified WITH the p-value for these most significant markers (provide your code!).

6. **(a)** Determine the p-value cutoff you would use if you wanted to set your False Discovery Rate (FDR) as close as possible to 0.05 (provide your code! - note that for this question, you may code this up yourself or use any existing function in R that you wish). **(b)** Again, using the criteria of question #5 - part (b) - determine how many hits are indicated when using this new cutoff and list the marker numbers of the most significant marker (=in step 2, using the SNP indexes after removal of those with MAF $<$ 0.06) for EACH separate hit you identified WITH the p-value for these most significant markers (provide your code!). **(c)** Using at most one sentence, state whether this cutoff is more conservative or more liberal than a Bonferroni cutoff and explain the trade-off when using one versus the other in terms of the expected number of false positives?

7. **(a)** For the markers that had the two most significant p-values you obtained overall, calculate the correlation of their $X_a$ random variables using the formula provided in class, i.e. do not use an R function to calculate this correlation! Your script must implement the formula of a correlation to calculate this answer (provide your code!). **(b)** State if you find this correlation value concerning and provide a detailed reason as to why?

8. **(a)** For markers 168 and 1010 produce two x-y plots: $X_a$ vs $Y$ and $X_d$ vs $Y$, such that you will produce 4 plots in total (provide your code! and make sure you label your plots!). **(b)** For each of these two markers, what do you suspect is the true genetic model and provide your parameter estimates to back up your argument?

9. Imagine you are explaining the set-up of the statistical model that you used in this analysis to a statistician that has never heard of a GWAS. List / answer the following when considering the phenotype and a single genotype: **(a)** The sample space $\Omega$. **(b)** The random variables. **(c)** All of the parameters and the possible parameter ranges for the family of probability distributions you are assuming. **(d)** The statistic that you are using to estimate the parameters of interest (provide the formula!). **(e)** Explain why you are interested in a null hypothesis that only includes two of the parameters?

10. 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.)** Why do you observe 'peaks' in your Manhattan plot? **(c.)** Why do the significant peaks in your Manhattan plot (possibly) indicate the genomic position of a causal polymorphism but not the actual causal polymorphism? **(d.)** Why do we use a multiple test correction to assess which of the peaks may indicate the position of a causal polymorphism? **(e.)** Why do we generally assume that that a single peak indicates a single causal polymorphism when considering human data? **(f.)** What is a statistical false positive and what is one reason why a significant peak may be a statistical false positive? **(g.)** What is one reason why a significant peak may produce a biological (but not a statistical!) false positive? **(h.)** What is one reason why you do not expect to identify the positions of all causal polymorphisms affecting height in this study? **(i.)** Why is a larger sample size in a GWAS generally better for identifying the position of causal polymorphisms? **(j.)** What is one reason why you might not be able to identify the causal polymorphisms EVEN if you had correctly genotyped EVERY polymorphic position in the genome and you had an infinite sample?