BTRY 4830/6830: Quantitative Genomics and Genetics
Fall 2014

Final - available online Dec. 10

Final exam, due before 11:59PM, Dec. 13

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!). 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 Amanda's or Jin's email inbox (as appropriate) before 11:59PM, Dec. 13. 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!!).

Questions 1-5 ask you to analyze data provided in the files '**QG14_genotypes_final.ped**', '**QG14_genotypes_final.map**', **QG14_phenotypes1_final.txt**' and '**QG14_phenotypes2_final.txt'**, while **Questions 6-10** do not involve analysis of any data (i.e. only written answers required for #6-10).

The genotypes provided in '**QG14_genotypes_final.ped**' are a subset of the genotypes you analyzed for your project. The file contains the genotype information for each individual, where each row corresponds to an individual. The 1st and 2nd columns contain the names of the individuals, the 3rd and 4th columns contain zero's, the 5th column contains the gender of the individual, the 6th column contains the entry '-9', and all of the following columns indicate SNP genotypes (in order) where each genotype is indicated by a pair of columns, i.e. the 7th and 8th column indicate the first SNP genotype, the 9th and 10th columns indicate the second genotype etc. Note that missing genotypes are usually indicated by '-9' in PLINK format (or sometimes '0') and there may or may not be missing data. Also note that the genotypes are provided in order along the chromosome.

The file '**QG14_genotypes_final.map**' contains additional information on the genotypes and has four columns. The 1st column contains the chromosome number of each SNP, the 2nd column contains the 'rsID' = the name of each SNP (where these are presented in order along the chromosome, e.g. the first entry of this column is the name of the SNP where the genotype is in the 7th and 8th column of the first row of the file '**QG14_genotypes_final.ped**'), the 3rd column contains all zeros, and the 4th column contains the physical position of the SNP on the chromosome.

Each individual has also been measured for two phenotypes, a 'continuous' phenotype (=phenotype1) and a 'discrete' phenotype (=phenotype2). Each row of the file '**QG14_phenotypes1_final.txt**' is the value of phenotype1 for an individual in the sample (i.e. the phenotype1 value of the 1st individual is in the 1st row, the phenotype1 value of the 2nd individual is in the 2nd row, ..., the phenotype1 value of the nth individual is in the nth row). Similarly, each row of the file '**QG14_phenotypes2_final.txt**' is the value of phenotype2 for an individual in the sample (i.e. the phenotype2 value of the 1st individual is in the 1st row, the phenotype2 value of the 2nd individual is in the 2nd row, ..., the phenotype2 value of the nth individual is in the nth row)

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

1. **(a)** Calculate the minor allele frequency (MAF) for each SNP in '**QG14_genotypes_final.txt**' after removing all SNPs that have an MAF < 0.05 and plot a histogram of these MAF for the remaining SNPs (provide your code!). **(b)** How many SNPs are left (i.e. what is N after you remove these SNPs)?

   (a) - 8 points: see 'QG14_Final_key' file.

   (b) - 2 points: 1745.

2. **(a)** For the phenotypes in '**QG14_phenotypes1_final.txt**' plot a histogram (provide your code!). **(b)** For each genotype remaining after the filtering in **question #1**, calculate p-values for tests of associations with the phenotypes in '**QG14_phenotypes1_final.txt** when testing the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ versus the alternative hypothesis $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$

when applying the genetic linear regression model for each genotype and provide a Manhattan plot for these p-values. NOTE (!!): use the formulas provided in class, i.e. DO NOT use the function lm() but DO use the formulas for $MLE(\hat{\beta})$, 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 (provide your code!). **(c)** Provide a QQ plot for these same p-values (provide your code!) and using NO MORE than two sentences explain whether you think the analysis you have applied resulted in appropriate model fit to the data (and why)?

(a) - 2 points: see 'QG14_Final_key' file.

(b) - 6 points: see 'QG14_Final_key' file.

(c) - 2 points: see 'QG14_Final_key' and (Many possible answers as long as they are justified!!), e.g. the model fit is not appropriate since the QQ plot leaves the line early, such that perhaps a covariate is needed OR the model fit is appropriate given that the vast bulk of points are quite close to the line, even though they leave the line a little, such that they would be in a reasonable confidence interval for the QQ OR etc.

3. **(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 #2** when using a Bonferroni correction (provide the formula you used to calculate this cutoff as part of your answer)? **(b)** For the p-values obtained in **question #2**, how many separate peaks did you observe that were greater than the Bonferroni correction level (provide a description of how you decided on the number of peaks AND your code as part of your answer!)? **(c)** For each of these separate peaks, list the p-value of the most significant marker and the 'rsID' of this marker.

   (a) - 2 points: see 'QG14_Final_key' file.

   (b) - 4 points: see 'QG14_Final_key' file and (Many possible answers as long as they are justified!!), e.g. there are three peaks since there are three separate positions throughout the genome where sets of markers that are relatively close together are significant at a Bonferroni threshold OR etc.

   (c) - 4 points: see 'QG14_Final_key' file.

4. **(a)** For the phenotypes in '**QG14_phenotypes2_final.txt**' plot a histogram (provide your code!). **(b)** For each genotype remaining after the filtering in **question #1**, calculate p-values for tests of associations with the phenotypes in '**QG14_phenotypes2_final.txt** when testing the null hypothesis $H_0 : \beta_a = 0 \cap \beta_d = 0$ versus the alternative hypothesis $H_A : \beta_a \neq 0 \cup \beta_d \neq 0$ when applying the genetic logistic regression model for each genotype and provide a Manhattan plot for these p-values. NOTE (!!): use the formulas provided in class, i.e. DO NOT use the functions in R that apply a logistic regression but DO use the IRLS algorithm and the appropriate formulas for the MLE and LRT, although you may use the function pchisq() to calculate the p-value from your LRT (provide your code!). **(c)** Provide a QQ plot for these same p-values (provide your code!) and using NO MORE than two sentences explain whether you think the analysis you have applied resulted in appropriate model fit to the data (and why)?

(a) - 2 points: see 'QG14_Final_key' file.

(b) - 6 points: see 'QG14_Final_key' file.

(c) - 2 points: see 'QG14_Final_key' and (Many possible answers as long as they are justified!!), e.g. the model fit is not appropriate since the QQ plot leaves the line early, such that perhaps a covariate is needed OR the model fit is appropriate given that the vast bulk of points are quite close to the line, even though they leave the line a little, such that they would be in a reasonable confidence interval for the QQ OR etc.

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 #4** when using a Bonferroni correction (provide the formula you used to calculate this cutoff as part of your answer)? **(b)** For the p-values obtained in **question #4**, how many separate peaks did you observe that were greater than the Bonferroni correction level (provide a description of how you decided on the number of peaks AND your code as part of your answer!)? **(c)** For each of these separate peaks, list the p-value of the most significant marker and the 'rsID' of this marker.

(a) - 2 points: see 'QG14_Final_key' file.

(b) - 4 points: see 'QG14_Final_key' file and (Many possible answers as long as they are justified!!), e.g. there are three peaks since there are three separate positions throughout the genome where sets of markers that are relatively close together are significant at a Bonferroni threshold OR etc.

(c) - 4 points: see 'QG14_Final_key' file.

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

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

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

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

where for this sample $\mathbf{Y} = [Y_1, Y_2, Y_3, Y_4]$. 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): Which of the sample pairs are positively correlated? Which of the sample pairs are negatively correlated? Which of the sample pairs are uncorrelated?
2 points - Sample pairs that are positively correlated: $(1, 2), (3, 4)$.
6 points - Sample pairs that are negative correlated: none.
2 points - Sample pairs that are uncorrelated: $(1, 3), (1, 4), (2, 3), (2, 4)$.

7. Consider two causal loci 'A' (with alleles '$A1$' and '$A2$') and 'B' (with alleles '$B1$' and '$B2$') **(a)** Write out the 9 possible genotype combinations that could occur at 'A' and 'B'. **(a)** Write out the values of $X_{a,1}, X_{d,1}, X_{a,2}, X_{d,2}$ for the genotype $A_1A_1B_1B_1$. **(c)** Assuming a linear regression model, what is the expected phenotypic value of $A_1A_1B_1B_1$ given the following parameter values: $\beta = [\beta_\mu, \beta_{a,1}, \beta_{d,1}, \beta_{a,2}, \beta_{d,2}, \beta_{a1a2}, \beta_{a1d2}, \beta_{d1a2}, \beta_{d1d2}] = [0.2, 0.1, 0.2, -0.3, 0.17, -0.11, 0.21, 0.08, -0.03]$ (write out the equation you used to do this calculation as part of your answer!).

(a) - 3 points: A1A1B1B1; A1A2B1B1; A2A2B1B1; A1A1B1B2; A1A2B1B2; A2A2B1B2; A1A1B2B2; A1A2B2B2; A2A2B2B2;.

(b) - 3 points: $X_{a,1}(A1A1B1B1) = -1$; $X_{d,1}(A1A1B1B1) = -1$; $X_{a,2}(A1A1B1B1) = -1$; $X_{d,2}(A1A1B1B1) = -1$.

(c) - 4 points:
$\mathrm{E}(Y|g = A1A1B1B1) =$
$\beta_\mu + X_{a,1}\beta_{a,1} + X_{d,1}\beta_{d,1} + X_{a,2}\beta_{a,2} + X_{d,2}\beta_{d,2} +$
$X_{a,1}X_{a,1}\beta_{a1a2} + X_{a,1}X_{d,2}\beta_{a1d2} + X_{d,1}X_{a,2}\beta_{d1a2} + X_{d,1}X_{d,1}\beta_{d1d2}$
$= 0.2+(-1)0.1+(-1)0.2+(-1)(-0.3)+(-1)0.17+(1)(-0.11)+(1)0.21+(1)0.08+(1)(-0.03)$
$= 0.18$

8. Narrow sense heritability (h^2) describes a property of a phenotype in a population impacted by genetics. Interestingly, if $h^2 = 0$ the phenotype does not evolve (i.e. the mean of the phenotype in the population does not change over time), even if the phenotype is under selection. 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:

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

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:

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

Note that there is however a simple relationship between the parameter β_α in equation (6) and β_a, β_d in equation (7):

$$\beta_\alpha = \beta_a\left(1 + \frac{\beta_d}{2}(p_1 - p_2)\right) \tag{8}$$

where $p_1 = MAF$ and $p_2 = 1 - p_1$. That is, for a one locus system with two alleles with a given MAF, there is a true value of β_α for the system. Similarly, for this same system,

there are true values for β_a and β_d. What's more, the true value of β_α for this system can be calculated using the true values of β_a and β_d using the formula in equation (8).

Consider a genetic system with one locus case and two alleles. Assume that the locus is segregating for the two alleles with a given MAF and assume that the true $\beta_a \neq 0$ and true $\beta_d \neq 0$. Provide a formula for p_1 under which $h^2 = 0$ for the system and *show your steps*, i.e. your answer should include an equation of the form $p_1 = equation$ and the steps that you used to derive this equation.

Since $h^2 = \frac{V_A}{V_P}$ only when $V_A = 0$ (and note that we know alleles are segregating with non-zero genetic effects so $V_P = 0$ and this relation exists) we need to solve the equation:

$$0 = 2MAF(1 - MAF)\beta_\alpha^2 \tag{9}$$

where by definition we have

$$0 = 2p1(1 - p1)\beta_\alpha^2 \tag{10}$$

and by substitution

$$0 = 2p1(1 - p1)\left(\beta_a\left(1 + \frac{\beta_d}{2}(p_1 - p_2)\right)\right)^2 \tag{11}$$

we therefore have

$$0 = \left(\beta_a\left(1 + \frac{\beta_d}{2}(p_1 - p_2)\right)\right)^2 \tag{12}$$

and for this to be equal to zero it must be that

$$0 = \beta_a\left(1 + \frac{\beta_d}{2}(p_1 - p_2)\right) \tag{13}$$

we therefore have

$$0 = 1 + \frac{\beta_d}{2}(p_1 - p_2) \tag{14}$$

$$-2 = \beta_d(p_1 - p_2) \tag{15}$$

by subsitution

$$-2 = \beta_d(p_1 - 1 - p_1) \tag{16}$$

$$-2 = \beta_d(2p_1 - 1) \tag{17}$$

$$\frac{-2}{\beta_d} + 1 = 2p_1 \tag{18}$$

$$\frac{-1}{\beta_d} + \frac{1}{2} = p_1 \tag{19}$$

9. Show your understanding of the basic concepts of GWAS by answering the following questions (note: these questions are asking for definitions!!): **(a)** What is a population? **(b)** What is a polymorphism? **(c)** What is a genotype? **(d)** What is a phenotype? **(e)** What is a causal polymorphism / genotype? **(f)** What is a genotypic value? **(g)** What is a genetic marker? **(h)** What is linkage disequilibrium? **(i)** What is a tag polymorphism? **(j)** What is a genetic association?

(a) - 1 point: **Population** - (Many answers acceptable including:) A group of individuals considered to have common ancestry by a specific criteria OR the entirety of individuals considered to be one group by a given criteria OR etc..

(b) - 1 point: **Polymorphism** - the existence of more than one allele at a locus.

(c) - 1 point: **Genotype** - the alleles possessed by an individual at a specified set of loci.

(d) - 1 point: **Phenotype** - any measurable aspect of an individual.

(e) - 1 point: **Causal Polymorphism** - a position in the genome where an experimental manipulation of the DNA produces an effect on the phenotype on average under specified conditions.

(f) - 1 point: **Genotypic Value** - the expected value of an individual given a genotype.

(g) - 1 point: **Genetic Marker** - a measured genotype in a set of individuals in a GWAS.

(h) - 1 point: **Linkage Disequilibrium** - correlation and physical linkage on a chromosome between two or more measured genotypes in a sample (or population).

(i) - 1 point: **Tag Polymorphism** - a genetic marker identified as being in linkage disequilibrium with a causal polymorphism (in a GWAS).

(j) - 1 point: **Genetic Association** - (Many answers acceptable including:) a significant correlation between a genetic marker and a measured phenotype in a sample OR an analysis to identify such significant correlations OR etc. .

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 possible outcomes are possible for this sample? **(g)** What is an example of a statistic that would be a 'terrible' estimator of the parameter p? **(h)** What is the maximum likelihood estimator (MLE) of the parameter p? **(i)** For the MLE($\hat{p}$) statistic, if this equals '0.5' for your observed sample, what is the p-value for $H_0 : p = 0.5$, $H_A : p \neq 0.5$ for a two-sided test? **(j)** Would you reject H_0 in this case for $\alpha = 0.05$?

(a) - 1 point: $\Omega = \{H, T\}$

(b) - 1 point: $\mathcal{F} = \emptyset, \{H\}, \{T\}, \{H, T\}$

(c) - 1 point: $Pr(\emptyset) = 0, Pr(\{H\}) = 0.5, Pr(\{T\}) = 0.5, Pr(\{H, T\}) = 1.0$

(d) - 1 point: $X(\{H\}) = 0, X(\{T\}) = 1$ OR $X(\{H\}) = 1, X(\{T\}) = 0$

(e) - 1 point: $\mathrm{E}X = \sum_{i=0}^{1}(X = i)Pr(X_i) = 0 * 0.5 + 1 * 0.5 = 0.5.$

(f) - 1 point: 2^{10}.

(g) - 1 point: Many possible answers, e.g. a statistic that takes the sample to a constant outside the possible range of the parameter: $T(x_1, ..., x_{10}) = -1$.

(h) - 1 point: $MLE(\hat{p}) = \frac{1}{10}\sum_{i}^{10} x_i$.

(i) - 1 point: p-value = 1.0 (full credit for this answer where the explanation that follows is not necessary: for $T(x_1, ..., x_{10}) = MLE(\hat{p}) - 0.5$, for the observed $T = t = 0.5 - 0.5 = 0$, p-value $= \int_{t=0}^{\infty} Pr(|t||H_0 = true)d(t) = 1.0$.

(j) - 1 point: We would note reject H_0, since p-value $= 1.0 > 0.05$.