Quantitative Genomics and Genetics - Spring 2016
BTRY 4830/6830; PBSB 5201.01

Final Exam - available online 11:59PM, Mon., May 16

Final Exam is due by 11:59PM, Thurs., May 19 - Key posted May 24

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 Mahya, 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 DR. MEZEY's EMAIL INBOX before 11:59PM Thurs., May 19. It is your responsibility to make sure that it is in Dr. Mezey's 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 (10 total, multiple parts per question) - make sure you answer all parts of all questions (!!):

Written answers below, see html file for the code and graphs.

Part 1: Performing a GWAS

You have been provided GWAS data in the following files: a first phenotype file "QG16_final_continuous.csv", a second phenotype file "QG16_final_categorical.csv", and a genotype file "QG16_final_genotypes.csv". Each of these three files is in ".csv" format. Note the following:

- The first phenotype file contains data on a "continuous" phenotype measured for each of the n individuals in the sample. The first row contains the name of the phenotype and each following rows contain the name of an individual in the first column and the value of the phenotype measured for the individual in the second column.
- The second phenotype file contains data on a "categorical" (case = 0 / control = 1) phenotype measured for each of the n individuals in the sample. The first row contains the name of the phenotype and each following rows contain the name of an individual in the first column and the value of the phenotype measured for the individual in the second column.
- The "genotypes" file contains data on N genotypes measured for each of the n individuals in the sample. The first row contains the names of the N genotypes. Each of the following rows contains all of the genotypes measured for a specific individual. For each of these rows, the FIRST column contains the name of the individual, EACH OF THE FOLLOWING COLUMNS presents data for a specific genotype, where the state of each genotype for an individual is coded as follows: '0' = homozygote, '1' = heterozygote, '2' = homozygote.

1. **(a)** Import the continuous phenotypes in the file "QG16_final_continuous.csv" [hint: use read.csv("QG16_final_continuous", row.names = 1)] and plot a histogram of the phenotypes. **(b)** For the phenotypes in part 'a', using no more than one sentence, explain why a linear regression is appropriate for analyzing these phenotypes in a GWAS. **(c)** Import the categorical phenotypes in the file "QG16_final_categorical.csv" [see previous hint] and plot a histogram of the phenotypes. **(d)** For the phenotypes in part 'c', using no more than one sentence, explain why a logistic regression is appropriate for analyzing these phenotypes in a GWAS.

   (b) These phenotypes could be well approximated by a normal distribution, such that a linear regression with an normally distributed error term is an appropriate model for analyzing these data in a GWAS.

   (d) These phenotypes could be well approximated by a Bernoulli distribution, such that a logistic regression with a Bernoulli distributed error term is an appropriate model for analyzing these data in a GWAS.

2. Import the genotype data [hint: use read.csv("QG16_final_genotypes.csv") and note that N=32114 so this might take a few minutes to run]. **(a)** Calculate the minor allele frequency (MAF) for each SNP and plot a histogram of these MAFs (provide your code!). **(b)** Perform

a PCA on the genotypes and create a plot that projects the samples onto PC1 and PC2 (label your axes and provide your code!).

3. **(a)** For the phenotypes in the file "QG16_final_continuous.csv", calculate p-values for tests of associations of the phenotypes with the genotypes in "QG16_final_genotypes.csv" 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 a linear regression model for each genotype with NO covariates (provide your code!). 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, and the F-statistic, although you may use the function pf() to calculate the p-value from your F-statistic (provide your code!). **(b)** Provide a Manhattan plot for these p-values. **(c)** Provide a QQ plot for these same p-values (provide your code!). **(d)** Given this QQ plot, using no more than two sentences, explain whether you think the analysis you have applied resulted in appropriate model fit to the data and explain the reasoning behind your answer based on the shape of the QQ plot.

   (d) The QQ plot leaves the 45 degree line early, indicating that unaccounted for covariates are producing significant p-values for large numbers of genotypes. The statistical model applied in the GWAS is therefore not an appropriate model for assessing whether individual genotypes are indicating the position of causal genotypes.

4. **(a)** For the phenotypes in the file "QG16_final_continuous.csv", calculate p-values for tests of associations of the phenotypes with the genotypes in "QG16_final_genotypes.csv" 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 a linear regression model for each genotype WITH THE FIRST TWO PCs calculated in question '2b' as covariates (provide your code!). NOTE (!!): again, 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, and the F-statistic, although you may use the function pf() to calculate the p-value from your F-statistic (provide your code!). **(b)** Provide a Manhattan plot for these p-values. **(c)** Provide a QQ plot for these same p-values (provide your code!). **(d)** Given this QQ plot, using no more than three sentences, explain whether you think the analysis you have applied resulted in appropriate model fit to the data, explain the reasoning behind your answer based on the shape of the QQ plot, and the biological reasoning as to why adding covariates produced this result.

   (d) The QQ plot is along the 45 degree line for the bulk of the p-values and then leaves the line for the more extreme -log p values producing a tail. This indicates that there is no evidence most of the genotypes are in linkage disequilibrium (LD) with causal genotypes, which matches our expectations for the impact of causal genotypes on phenotypes in general, while just a few markers are in LD with causal genotypes, indicating that the statistical model applied in the GWAS is an appropriate model. Given that the first few PCs seem to indicate population structure among the individuals in the sample, population structure seems to be a reasonable explanation as to why adding PCA covariates produced an acceptable QQ plot.

5. For this question, consider the analysis results in question '4'. **(a)** When controlling the study-wide type 1 error of 0.05, what is the appropriate p-value cutoff for assessing which genetic markers are significant when using a Bonferroni correction (provide the formula you

used to calculate this cutoff as part of your answer)? **(b)** How many separate peaks did you observe that were greater than the Bonferroni correction level? Using no more than two sentences, provide a description of how you decided on the number of peaks. **(c)** For each of these separate peaks, list the p-value of the most significant marker in the peak and the 'rsID' of this marker. **(d)** Calculate the correlation between the most significant markers for each peak identified in part 'c' and the markers on either side of each of these most significant markers. **(e)** Is the most significant marker in each peak necessarily closer to the causal polymorphism (assuming the peak indicates a causal polymorphism) than either of the markers on each side of the most significant marker? Use no more than two sentences to explain your answer.

(d) Note that calculations of the correlation with the markers that are physically on either side of the most significant marker in each peak or answers that calculated the correlations with the next most significant markers on either side of the most significant marker within each peak will be accepted.

(e) The most significant marker in a peak is not necessarily physically closer to the causal polymorphism than a less significant marker in the peak, because p-values are impacted by many factors, including sampling variation (i.e., the particulars of the sample), by minor allele frequency of a marker, correlations with covariates, etc. As a consequence, depending on these many factors, the measured markers that are physically closest to the causal polymorphism need not be the most significant in a given sample (not necessary for the answer: although they would be expected to on average across many independent GWAS assuming an appropriate statistical model!).

6. **(a)** For the phenotypes in the file "QG16_final_categorical.csv", calculate p-values for tests of associations of the phenotypes with the genotypes in "QG16_final_genotypes.csv" 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 a logistic regression model for each genotype with NO covariates (provide your code!). 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!). **(b)** Provide a Manhattan plot for these p-values. **(c)** Provide a QQ plot for these same p-values (provide your code!). **(d)** Given that you are analyzing the same genotypes as in question '4', was this QQ result guaranteed to be similar to the QQ result produced in question '4c'? Explain your answer using no more than two sentences.

(d) Note that this question was corrected to refer to question 3, where appropriate answers when comparing to question 3 or question 4 are accepted.

When comparing to question 3: Given that there population structure in this sample that had an effect on the continuous phenotype to produce an inflated QQ plot and this covariate was not corrected for in the question 6 analysis, assuming population structure also had an effect on the case / control phenotype, we would expect the QQ plot to be similarly inflated (note that other answers are acceptable if well justified).

When comparing to question 4: Given that there population structure in this sample that had

an effect on the continuous phenotype that was corrected for to produce an appropriate QQ plot and this covariate was not corrected for in the question 6 analysis, assuming population structure also had an effect on the case / control phenotype, we would expect the QQ plot to be inflated and therefore not similar (note that other answers are acceptable if well justified).

7. **(a)** For the phenotypes in the file "QG16_final_categorical.csv", calculate p-values for tests of associations of the phenotypes with the genotypes in "QG16_final_genotypes.csv" 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 a logistic regression model for each genotype with ONLY THE FIRST PC calculated in question '2b' as a covariate (provide your code!). 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!). **(b)** Provide a Manhattan plot for these p-values. **(c)** Provide a QQ plot for these same p-values (provide your code!). **(d)** Given this QQ plot, using no more than three sentences, explain whether you think the analysis you have applied resulted in appropriate model fit to the data, explain the reasoning behind your answer based on the shape of the QQ plot, and explain why using a different number of covariates than in question '4' can result in a perfectly legitimate and interpretable result.

(d) After including a covariate for population structure, the QQ plot is along the 45 degree line for the bulk of the p-values and then leaves the line for the more extreme -log p values producing a tail. This indicates that after correcting for the influence of population structure on variation in hte phenotype, there is no evidence most of the genotypes are in linkage disequilibrium (LD) with causal genotypes, which matches our expectations for the impact of causal genotypes on phenotypes in general, while just a few markers are in LD with causal genotypes, indicating that the statistical model applied in the GWAS is an appropriate model. An appropriate covariate correction for any factor, such as population structure, only requires that the impact of the factor on the phenotype be accounted for with the covariate(s) such that the impact on the same factor on different phenotypes could be modeled in different ways to produce an appropriate statistical model for the GWAS.

8. For this question, consider the analysis results in question '7'. **(a)** When controlling the study-wide type 1 error of 0.05, what is the appropriate p-value cutoff for assessing which genetic markers are significant when using a Bonferroni correction (provide the formula you used to calculate this cutoff as part of your answer)? **(b)** How many separate peaks did you observe that were greater than the Bonferroni correction level? Are any of these peaks the same as identified in question '5'? If so, do these necessarily indicate the same causal polymorphism as identified in question '5' (use no more than two sentences to explain your answer if so)? **(c)** For each of these separate peaks, list the p-value of the most significant marker in the peak and the 'rsID' of this marker. **(d)** Calculate MAF of the most significant marker in each peak and the MAF of the marker on each side of the most significant marker. **(e)** Using no more than two sentences, explain how these MAFs are expected to comparatively impact the power of the statistical test applied to the most significant marker and the markers on either side of the most significant marker for each peak.

(b) Two peaks were identified, one of which appears to be in the same position as one of the peaks in question 5, possibly indicating the same causal polymorphism has an impact on both phenotypes but also possibly indicating that there are two causal polymorphisms in the region, each impacting a different phenotype.

(d) Note that calculations of the MAF with the markers that are physically on either side of the most significant marker in each peak or answers that calculated the MAF with the next most significant markers on either side of the most significant marker within each peak will be accepted.

(e) MAF of a marker impacts the power of the statistical test, such that the markers on either side of the most significant marker in each peak with a lower MAF are expected to have lower power and therefore expected to have a lower (higher -log) p value (not necessary for the answer: and vice versa, where a higher p-value of the significant marker with a lower MAF occurred due to sampling variation or some other factors).

9. **(a)** Using no more than one sentence, provide at least one reason as to why it is essential that the sample be iid for the data analyzed in question '4' and question '7' given the models you applied. **(b)** Assuming the sampling distribution of phenotypes in both cases are iid, describe the correlation matrix of these distributions. **(c)** Using no more than two sentences, provide one scenario that could result in these sampling distributions NOT being iid and explain why this scenario results in a non-iid sample.

(a) Several possible answers, an example of an acceptable answer: The statistical analysis in question 4 and 7 makes use of a likelihood equation to calculate the MLE (subsequently used to calculate p-values) that involves the multiplication of the likelihood equation for each sample, which is correct only if the sample is iid.

(b) An nxn identity matrix.

(c) Many possible answers, an example of an acceptable answer: If individuals in the sample are closely related (e.g., they are within the same nuclear family) one would expect an iid assumption to not be appropriate for the sampling of the phenotypes of these individuals, since they would tend to be similar to one another based on shared factors due to being in the same family (both environment and genetics).

10. **(a)** Define causal polymorphism. **(b)** Using no more than two sentences, provide an explanation as to why a GWAS is unlikely to identify a causal polymorphism. **(c)** Given your answer in part 'b', using no more than two sentences, describe how might you use a GWAS and other data to identify a candidate causal polymorphism. **(d)** Using no more than two sentences, describe the best possible experiment THAT IS REALISTIC FOR HUMANS, which would allow you to determine whether a candidate polymorphism identified in a GWAS is causal.

(a) A causal polymorphism for a given phenotype is a polymorphic site in the genome where directly swapping one allele for another produces a change in value of the phenotype under some condition (or symbolically $A_1 \rightarrow A_2 \Rightarrow \Delta Y$).

(b) A GWAS is unlikely to identify a causal polymorphism, since measurements (genotyping) of causal polymorphisms are often not included within GWAS data and even if causal polymorphisms are measured these will be in linkage disequilibrium with many other non-causal polymorphisms such that it is (usually) not possible to precisely distinguish which of the polymorphisms are causal among the correlated p-values of the correlated group of markers plus the causal polymorphism.

(c) Many possible answers, examples of acceptable answers: candidate loci could be identified by considering the function of known genes in the indicated region, by considering the functional annotation of specific polymorphisms (e.g., coding, non-coding), by considering expression Quantitative Trait Loci (eQTL) data, by considering other association data (e.g., other GWAS, linkage analyses), etc.

(d) Many possible answers, an example of an acceptable answer: A CRISPR experiment performed in a lab maintained human cell population relevant to the phenotype of interest where the ONLY alteration is a swap of an allele of the candidate causal polymorphism for the other allele segregating in humans, where the impact on molecular pathways relevant for the phenotype of interest is then assessed.