BTRY 4830/6830: Quantitative Genomics and Genetics
Fall 2012

Final Exam, **VERSION 1** - available online Dec. 5

Due before 11:59PM on Dec. 12

PLEASE NOTE THE FOLLOWING INSTRUCTIONS:

1. You are to complete this final 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. For all questions involving coding, all coding must be done in R (no exceptions)! For ANY tasks you perform in R, make sure you include the R code in your answer (i.e. no R 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. 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 to get full credit, 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 R code from the command line (or using "source") and/or you should provide us instructions on how to run your R code. We will attempt to run your R code if you do not do this but we will deduct points accordingly.

5. The exam must be in Monica's email inbox before 11:59PM Weds., December 12. 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 type 2 diabetes risk. They know that there are loci scattered throughout the genome that can increase risk for type 2 diabetes, 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 phenotypes and SNP genotypes in two files ("final_phenotype_fall12.txt" and "final_genotype_fall12.txt"). Note that for the phenotypes, "one" is a case (i.e. an individual with diabetes) and "zero" is a control (a healthy individual). Note that for the SNP genotypes, every column is a marker and every pair of rows is a genotype of an individual (note that two letters indicate each genotype and there are three possible combinations per marker: two homozygotes and a heterozygote). Also note that the markers in the file are listed in order along the genome.

1. Produce a histogram of the phenotypes (label your both plot and your axes using informative names!). What is the sample size (n)? How many genotypes per individual are there (N)? Is there any missing genotype data (explain how you determined this to be the case)?

2. Filter these genotypes by removing all genotypes with a MAF < 0.05. How many genotypes are left after applying this filter?

3. Create the "X_a" and "X_d" dummy variables for the remaining genotypes after applying the filter in question 2 and use the code below to plot the n individuals on the first two principal components of the sample correlation matrix:

   ```
   > W <- (Xa - rowMeans(Xa)) / sqrt(diag((cov(Xa))))
   > geno.pc <- princomp(W)
   > plot(geno.pc$loadings[,c(1,2)], main="Plot of n individuals on the loadings of the first two
   PCs", xlab="PC1", ylab="PC2")
   ```

   Based on this plot, why would we not consider the n individuals to be sampled from the same population (explain your reasoning!)? Assuming that it is appropriate to assign these individuals into two separate populations, define third dummy variable "X_z" that assigns each individual into one or two populations using the value "-0.03" on PC1 as the split point. How many individuals are in each population? Why this is a reasonable split point to separate these individuals into two separate populations (explain your answer)?

4. Implement the IRLS algorithm we considered in class to calculate the $MLE(\hat{\beta})$ for the four β parameters (β_μ, β_a, β_d, β_z) when using a logistic regression to analyze each of the N genotypes. Calculate the mean of the absolute value of $\hat{\beta}_z$ for the N markers and similarly calculate the mean of the absolute value of $\hat{\beta}_a$ and $\hat{\beta}_d$. Is the mean of the absolute values of your $\hat{\beta}_z$ values greater or less than the mean values of the absolute values of the $\hat{\beta}_a$ and $\hat{\beta}_d$? What does this imply about the impact of population structure on the phenotype?

5. Implement the IRLS algorithm we considered in class to calculate the $MLE(\hat{\beta})$ for a model with ONLY the three genetic model β parameters (β_μ, β_a, β_d) when using a logistic regression to analyze each of the N genotypes, i.e. do not include β_z. Note that we will refer to these as ($\beta_{\mu-}$, β_{a-}, β_{d-}) to differentiate them from the parameters estimated for the model in question 4. Calculate the mean of the absolute values of the $\hat{\beta}_{a-}$ for the N markers and compare these to the mean of the absolute values obtained for $\hat{\beta}_a$ in question 4. Which of these are higher? Also do the same for the mean of the absolute values of the $\hat{\beta}_{d-}$ and answer

the same question when comparing these to the absolute values obtained for $\hat{\beta}_d$ in question 4. What could explain this result (hint: think about the impact of population structure)?

6. Calculate p-values for each of the N genotypes when considering the null hypotheses $\beta_a = 0 \cap \beta_d = 0$ and $\beta_{a-} = 0 \cap \beta_{d-} = 0$ for the logistic regression models in questions 4 and 5, i.e. you should obtain two lists with N p-values in each (note: use the IRLS algorithm to obtain the necessary parameter estimates and not the glm() or related functions!). Plot histograms for these two lists of p-values. Why does one of these histograms more closely resemble the uniform distribution than the other (explain your answer)?

7. Produce quantile-quantile (QQ) plots for both of the lists of p-values from question 6. Explain why the QQ plot for the analysis considering the covariate (i.e. question 4) looks appropriate and why the QQ plot for the analysis without the covariate (i.e. question 5) indicates something is wrong.

8. Produce a Manhattan plot for ONLY the p-values obtained when considering a model with a covariate (i.e. question 4). Add a horizontal line above which markers are significant at a Bonferroni corrected threshold for a (single test) $\alpha = 0.05$. How many association hits are indicated at this cutoff (explain your answer)? Also add a horizontal line above which markers are significant at a Bonferroni corrected threshold for a single test $\alpha = 0.1$. How many association hits are indicated at this cutoff (again, explain your answer)? Explain to your collaborator why the associations at either cutoff may both be true positives, why both may be false positives, or one may be a true positive and one may be a false positive? Is there a reason why the collaborator should be "more confident" in the hit(s) at the more conservative cutoff (explain your answer and be careful with your reasoning!)?

9. List the number of the two SNP markers corresponding to the two MOST significant hits indicating DIFFERENT locations in the genome (again considering a model with a covariate = question 4). For these two SNPs, provide the nine dummy variables for the first FIVE individuals listed in the genotype file, where the dummy variables correspond to the the additive, dominance, and epistatic parameters (note that you do not need to estimate the β's or do any other analyses - just provide the dummy variables and provide R code for how you determined the dummy variables!). Your collaborator tells you that they think that these two hits indicate two proteins that are epistatic because they form a dimer protein complex. Provide the traditional definition of epistasis for your collaborator (that we learned in the class) and explain why epistasis cannot say anything about this protein hypothesis.

10. If instead of being case-control, the phenotype had been normal (i.e. the data fit a normal error model), describe how you would have performed a GWAS using a Bayesian approach (or more accurately a pseudo-Bayesian approach). In your answer, describe each of the critical components of such a Bayesian analysis, any assumptions that you are making, and note that you are welcome to use equations in your answer (note that you are NOT required to implement a Bayesian analysis to answer this question - just explain how you would perform the analysis!).