BIOST 551/STAT 551: Autumn Quarter 2014
Final Exam
Due by 12:00 p.m., Friday, December 12, 2014

---

1. Collaboration is not allowed for the final exam. The exam is due Friday, December 12, 2014 by noon (12:00 p.m.). Exams should be typed and emailed to the instructor (tathornt@uw.edu).

Question	Points	Score
Question 1	40	
Question 2	30	
Question 3	30	

1. **[40 points] Genome-Wide Association Study (GWAS) of Transferrin Serum**

   Iron is essential for a number of biochemical functions including oxygen transport and oxidative phosphorylation. Excessive iron can cause various disorders, such as iron-overload-related liver diseases, whereas iron deficiency can lead to anemia. Iron status can be assessed by measuring the levels of serum transferrin in the blood. Transferrin is a beta globulin in blood plasma capable of combining with ferric ions and is essential for transporting iron in the body. The PLINK files "Transferrin.bed", "Transferrin.fam", and "Transferrin.bim" contain genome-screen data for a sample of individuals with European ancestry, and the phenotype file "Tr.pheno" contains transferrin measurements for sample individuals that have been adjusted for relevant covariates.

   (a) How many individuals are included in the study? How many SNPs are available? How many individuals have transferrin measurements? Give a description of the data.

   (b) Provide a description of the transferrin phenotype and include relevant descriptive statistics as well as a figure illustrating the distribution of the transferrin phenotype.

   (c) Perform a GWAS of transferrin serum with PLINK. For the association analysis, use the following quality control threshold filters: minor allele frequency > 0.05, at least a 99% genotyping call rate (less than 1% missing), and HWE p-values greater than 0.001. Give the statistical model that was used for the PLINK association analysis and the hypotheses of interest. What are the assumptions of this model?

   (d) Provide a Manhattan plot and a Q-Q plot of the PLINK association results.

   (e) Report SNPs that have highly significant associations with transferrin serum. You should also determine if these SNPs are in genes and discuss if the genes are plausible candidates for transferrin serum.

   (f) A 2009 *American Journal of Human Genetics* article entitled "Variants in TF and HFE Explain ~40% of Genetic Variation in Serum-Transferrin Levels Genome" by Benyamin et al. has been posted on the course website. Give a brief description of the transferrin association study conducted for this article, and compare your association results to transferrin GWAS results from this article.

2. **[30 points] Power of a Replication Study for Transferrin**

   A follow up GWAS study of transferrin is being proposed for an an independent sample of 1,100 individuals of European ancestry. Before deciding on whether or not to approve the study, a funding agency would like to know the power of replicating the top findings that you identified in question 1.

   (a) Calculate the power of replicating the association results for the top 10 SNPs from your PLINK association analysis at a genome-wide significant level $\alpha = 5 \times 10^{-8}$ (type-I error) for a sample of 1,100 individuals. For the power calculations, you can use the proportion of transferrin serum variance that is explained for each of the top 10 SNPs from your transferrin GWAS analysis with PLINK.

(b) For the top 10 SNPs, make a scatterplot of the power for a sample of size 1,100 individuals calculated in 2(a) versus the absolute value of the estimated effect sizes of the SNPs from the transferrin serum GWAS study.

(c) Now calculate the minimum sample size required to have a power of at least 0.90 at a type-1 error of 1×10^{-7} for each of the top 10 SNPs.

3. **[30 points] Mixed Model Association Mapping for Transferrin**
The file "MLM-Transferrin.txt" is available on the course website. This file contains association results from a mixed linear model (MLM) GWAS of the same genotype and transferrin phenotype data used for question 1 above. There are 6 data columns in this file: 1. CHR (chromosome number), 2. SNP (rs number), 3. BP (base pair position), 4. N (number of genotyped individuals), 5. BETA (effect size of SNP from the MLM), 6. SE (standard error of the effect size estimate), 7. chi2.1df (the χ^2 test statistic for association); 8. P1df (association p-value). The MLM for transferrin serum incorporated an empirical genetic relatedness matrix calculated from the SNP genotype data and two variance components to account for additive polygenic and unique (or non-shared) environmental effects. The heritability estimate of transferrin serum from the MLM was 0.496.

(a) Give the MLM that was used for the association analysis and the hypotheses of interest. What are the model assumptions of the MLM?

(b) Provide Manhattan and Q-Q plots for the GWAS of transferrin serum with the MLM.

(c) Compare the association results for the 10 SNPs identified with the MLM to the top 10 SNPs from the PLINK association analysis. Discuss similarities and/or differences.

(d) Which of the association models for transferrin serum, the MLM or the model used in question 1 with PLINK, appears to be more appropriate for this sample and why? Discuss and provide evidence to support your answer.