

### **Exam III Outline – Gene Expression Analysis**

1. Different ways of coding explanatory variables in linear models
2. Relationship between two-sample t-test and a linear model with coded variables.
3. Reading in raw data (.CEL files) from GEO and processing the data using robust multi-array average (RMA). **(omitted)**
4. Reading in processed data from the Gene Expression Omnibus (GEO) using the *getGEO* function from the *GEOquery* library. **(omitted)**
5. Understanding the format of the gene expression matrix and clinical variables.
6. Evaluating whether a single gene is differentially expressed, using the two-sample t-test, constructing a boxplot and calculating its fold-change and *p*-value.
7. Identifying differentially expressed genes using the *limma* package and understanding the false discovery rate (FDR)
8. Generating heatmaps
9. Converting a probe name to the corresponding gene symbol and vice versa using the platform (GPL) data provided by GEO
10. Classification using *k*-nearest neighbors (*knn*), including leave-one-out classification, optimization, and making predictions in a test dataset.
11. Perform a gene set enrichment analysis using DAVID, based on a list of genes
12. Additional data analysis and R programming concepts (based on your Piazza questions)