Statistical Analysis of Genetic Data
Math 155, Spring 2008
Jo Hardin
HW #7

Due Friday April 11

1. Using the code below, generate 10,000 random genes from two groups (of 10 samples each). Consider the first 10 samples (columns) as cancer data and the remaining 10 samples (columns) as healthy data. Perform a two-sample t-test for each gene (row) of your data. Sort the p-values from smallest to largest. I say to explain because I want you to tell me what the number means (rather than simply writing down a single numerical answer to each question).

   (a) How many of the genes are "significant" according to some α level you choose? Explain.

   (b) What is the family-wise error rate? Explain.

   (c) What is the false discovery rate? Explain.

   (d) How many of the genes are "significant" according to some α level using the Bonferroni adjustment? Šidák adjustment? Holm step-down method? Explain.

```
rand.gns <- matrix(rnorm(200000),ncol=20)   # note the number is 200,000
dim(rand.gns)
p.vals <- c()
for (i in 1:10000){
   p.vals <- c(p.vals,t.test(rand.gns[i,1:10],rand.gns[i,11:20])$p.value)}
p.vals.sort <- sort(p.vals)
sum(p.vals < .05)
```

2. Repeat the above questions with your data. Don't worry about models or weights or repeated measures, but do try to make the comparison meaningful. For example, if you want to compare humans to primates, and the first 6 columns are humans and the next 13 columns are primates, your code would be:

```
prim.gns <- prim.normalized$M    # the M matrix from your normalization
dim(prim.gns)
p.vals <- c()
for (i in 1:nrow(prim.gns)){
   p.vals <- c(p.vals,t.test(prim.gns[i,1:6],prim.gns[i,7:19])$p.value)}
p.vals.sort <- sort(p.vals)
sum(p.vals < .05)
```

   (a) Answer (a) - (d) from above.

(b) How is your data different from the randomly generated data in #1? There are two very substantial differences that affect error rates. Were the error rates affected? That is, are the answers to (a) - (d) substantially different in #1 and #2? How?

3. Over email I've sent you a tab delimited datafile, `hw7_testdata.txt`. The data contain 3 replicates of a simulated experiment with two conditions (control and treatment). The first three columns are the control samples, the second three columns are the treatment samples (each entry is the ratio of sample / reference). The genes names **test 2x**, **test 4x**, and **test 8x** are known to be induced by 2, 4, and 8 fold respectively, but have measurement error introduced into the ratio values based on actual experimental error. There are 100 of these truly induced genes. All the other genes are known to have no differences between the two experimental conditions. Fill out the following table:

	% of top 25 genes that are true	% of top 100 genes that are true
Sort by average fold induction		
Sort by median fold induction		
Sort by t-test p-value		
Genes with $p < 0.05$ then by average fold induction		
Genes that have an average fold induction $> 2x$, then sorted by p-value		

	Actual FDR	% true genes identified
SAM with 75% FDR		
SAM with 25% FDR		