

## Lab 2: Simple Microarray Data Analysis

### Due Thursday, 3/7 in class

The aim of this lab is to introduce you to some simple yet powerful ideas involving permutation testing in the context of microarray data analysis. It will be based on some of the data analyzed in the paper S. Dudoit *et al* **Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments** August 2000, which can be found at:

<http://www.stat.Berkeley.EDU/users/terry/zarray/Html/papersindex.html>

The paper is there in .pdf format as well as a zipped PostScript version. Hopefully, you can read one of these two without necessarily printing it. If you can't on your own computer system, this is possible at the Stat dept terminals.

In the file `/class/data/s131TPS/ko8.lab` you will find a 994 (genes) by 16 (samples) matrix of  $\log_2$  (Red/Green) values. The first 8 red values in each row come from liver mRNA samples from 8 control mice of the same strain, abbreviated WT, and the second 8 red values in each row are from liver mRNA samples from 8 treatment mice whose Apo AI gene has been knocked out, abbreviated KO. In all 16 cases, the green values correspond to a pool of liver mRNA from the 8 control mice.

Each set of 16 values corresponds to a single spot on the glass slide, and we have data from just 994 of the 6,348 cDNA spots, hereafter called genes. The preprocessing (including image analysis and normalization) that was carried out is described in the paper.

1. Go to the web page:

[www.stat.Berkeley.EDU/users/terry/zarray/Data/ApoA1/Images/jpegindex.html](http://www.stat.Berkeley.EDU/users/terry/zarray/Data/ApoA1/Images/jpegindex.html)

Images of all 8 knock-out experiments are shown. They are the overlay image of R over G. You can click on each image to see the original size version. In addition, you can save an image and view it using "imview" on the Stat dept system.

Single gene analysis, I.

2. For a random gene of your choice and for gene #993, calculate the two sample  $t$ -statistics which might be used to test the null hypothesis that the mean  $\log(\text{ratio})$  for KO mice is the same as that for WT mice. Assume that these  $t$ -statistics do indeed have a  $t$ -distribution, and obtain the  $p$ -values for testing this hypothesis from  $t$ -tables.
3. Can you give a reason why we might not want to trust  $p$ -values from the  $t$ -tables in this context?

All gene analysis, I.

4. Use Matlab to calculate all 994  $t$ -statistics and plot them in a histogram. Are there any apparent extreme values?
5. Obtain a normal qq-plot of your 994  $t$ -statistics.

Single gene analysis, II.

6. Obtain a single random permutation of the assignment of 16 mice to 8 “treatment” and 8 “control”, ignoring their actual status as KO or WT. For this assignment, calculate the two new values of the  $t$ -statistics for your gene in 2 above and for gene #993. Repeat, with a new random permutation.
7. Explain why there are 12,870 such assignments of 16 mice to 8 “treatment” and 8 “control”.
8. Use the Matlab program supplied to estimate the full permutation distribution of the  $t$ -statistic for the two genes you are using from 500 random permutations. Estimate the permutation  $p$ -value for the observed  $t$  for those genes, and compare them to the values you got from the  $t$ -tables.

All gene analysis, II. Harder.

9. Calculate all 994  $t$ -statistics for 500 random assignments, and estimate the full permutation distribution of the smallest and largest of the 994  $t$ -statistics.
10. Are there any genes whose  $t$ -statistics look unusual in relation to the permutation distributions for the smallest and largest of 994  $t$ -statistics?
11. Summarize your conclusions.