GS01 0163 Assignment 4 Fall 2007
Due Date: Thursday, 18 October 2007

1. This assignment uses data from the study of MLL that was described in our first lectures on dChip, and which comes from a paper by Armstrong et al, *Nature Genetics*, 2002;**30**:41–47. The data can be found at: `http://www.broad.mit.edu/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=63`
   (a) Begin by downloading the CEL files for the ALL and the MLL samples. In addition, download a copy of the file `scaling_factors_and_fig_key.txt`. How many CEL files are there? How many arrays are U95A and how many are U95Av2?
   (b) Get a copy of the original publication. Briefly summarize the main results of the paper.
2. Prepare a "data list file" and a "sample info file" that can be used to load the U95A (but not the U95Av2) arrays into dChip. Your sample info file should include a column that describes whether the sample came from a patient with ALL or a patient with MLL. Analyze the data in dChip, producing a list of genes that are differentially expressed between ALL samples and MLL samples. How mnay genes are on the list? Use a permutation test to estimate the false discovery rate (FDR) for your list of genes.
3. We are now going to analyze the same data in R. Process the data using `just.rma` (or `justRMA` if you prefer the widget-based GUI for loading data) to quantify the files. What kind of object is produced, and what does it contain?
4. Perform two-sample t-tests for each gene.
   (a) Which group of samples has a higher expression for a gene if the t-statistic is positive?
   (b) Plot a histogram of the t-statistics.
   (c) How many genes have a t-statistic greater than 4 in absolute value? How many have a t-statistic greater than 3.5?
   (d) How many genes are significant using a Bonferroni correction and a significance level of 5%?
5. This problem is a continuation of the previous problem.
   (a) Plot a histogram of the p-values.
   (b) Fit a beta-uniform mixture (BUM) model to the set of p-values. Does the model appear to fit the data well?
   (c) How many genes are called significant if you use the BUM model to bound the false discovery rate (FDR) at 5%?
   (d) What p-value cutoff corresponds to FDR= 5%?
6. Perform a Wilcoxon test for each gene.
   (a) If the rank-sum statistic for a gene is very large, which group of samples has a higher level of expression for that gene?
   (b) Plot a histogram of the Wilcoxon statistics, and overlay a plot of the theoretical Wilcoxon distribution. Based on the plot, are there likely to be any differentially expressed genes? Are more genes overexpressed in ALL or in MLL samples?

(c) Use the empirical Bayes method to compute posterior probabilities for differential expression as a function of the Wilcoxon rank-sum statistics. By trial and error, find the largest value of the prior probability of not being differentially expressed that ensures that the computed posterior probabilities are all positive.

(d) Using the prior probability from the previous part, determine how many genes have a posterior probability of being differentially expressed that is at least 90%.

7. Analyze the data using the tail-rank test. Since all the MLL samples have a similar genetic abnormality, we will assume they are more homogeneous and use them as the baseline group. Using a target specificity of 95% and a confidence level of 90%, how many genes are found to be significant by the tail rank test?

8. At this point, we have analyzed the ALL and MLL samples multiple ways, and produced 4 different lists of differentially expressed genes (dChip, t-test, Wilcoxon test, tail-rank test). Construct a table that counts the overlap between each pair of analyses. (Note that you will have to export the results from dChip and read them into R in order to complete this problem.)