GS01 0163 Assignment 1 Fall 2005
Due Date: Tuesday, 13 September 2005

1. Use PubMed to locate the paper by Dinesh Singh and colleagues that applied microarrays to the study of prostate cancer (Cancer Cell 2002; 1:203-9). According to the paper:
   (a) What kind of microarrays did they use?
   (b) How many experimental samples were involved, and what kinds of samples were they?
   (c) Where can you find supplemental information about this study?
   (d) How were the microarray images turned into the expression values used in the paper?
   (e) Summarize the main results of the paper.

2. The supplementary data to the *Cancer Cell* paper by Singh and colleagues provides access to the raw CEL files, which are stored in batches in a set of gzipped tar files (i.e., compressed files whose extension ends in `.tar.gz`). Download all the CEL files from this study and use `WinZip` (or an equivalent program) to uncompress and extract the CEL files into one or more directories.
   (a) How many `.tar.gz` files were there? How many CEL files were in each `.tar.gz` file?
   (b) Can you tell why the CEL files were divided up this way?
   (c) Is there anything about the way the data was stored that might raise concerns about the design of the experiment? Explain your answer.
   (d) Table 1 of the paper describes the clinical and pathological characteristics of the prostate cancer patients in the study. Which of these factors would you need to know to be able to reproduce the main findings of the study? Does the supplementary information allow you to determine the values of these clinical variables for individual samples?

3. Download and install the latest version of DNA Chip Analyzer (dChip) from its official web site maintained by Cheng Li and Wing Wong. Also download and install the gene information files for the HG_U95Av2 and HG-U133A microarrays. What is the current version number for dChip? Can you tell when the gene information files were last updated?

4. Affymetrix maintains support files for its microarrays on its web site. Connect to the web site and proceed to the NetAffx Analysis Center. (This step will require you to register; registration is free but requires you to supply a valid email address and some other information.) Find and download the CDF library files for the HG_U95Av2 and HG-U133A microarrays. Describe how you located the files. How much extra stuff did you have to download in order to get copies of the CDF files?

5. At this point, you have collected everything you need from the internet to perform a full analysis of a large microarray data set. Before performing such an analysis in dChip, however, you must still prepare a “data list file” (which tells dChip where to find the CEL files you downloaded in step (2)) and a “sample info file” that allows dChip to use short names for the samples and to learn about the main properties of the samples. Prepare those files, and submit copies of them with this assignment.

6. Use dChip to analyze the Singh prostate cancer data set. Find a list of genes that are differentially expressed between prostate cancer and normal prostate. Prepare a report that describes how you processed (normalized, quantified, etc) the raw data. make sure you describe the settings used by all the modeling steps. Also describe the criteria you used to select differentially expressed genes. Include a file containing the list of genes when you submit your solutions.

7. (a) Use the `Analysis/Filter Genes` menu choice to select the most highly expressed genes on the microarray. (Be careful not to filter based on the variation or percentage of P calls.) Adjust the parameters until the number of genes passing the filter is between 3000 and 4000.

   (b) Use the `Analysis/Hierarchical Clustering` menu choice to cluster both the filtered genes and the samples. Does the sample clustering match what you think is the most important biological contrast in the data?