**GS01 0163 Assignment 6 Fall 2006**
**Due Date: 7 December 2006**

1. Download the breast cancer cell line data set `BrCaCellLines.rda` from the class web site. This file is an R data object; you can use the `load` command to read it into R. Describe the objects that are contained in the data set. (Hint: you probably need to load one or more BioConductor libraries in order to interpret these objects. Also, remember that MIAME objects can contain other information without telling you.) How many samples are there? How many genes? Where did the data set come from? What kinds of samples are in the two channels of the arrays?

2. The `backgroundCorrect` function in the `limma` package provides several methods for performing background correction, but the default "subtract" method produces negative intensities. Select a method that ensures that all the corrected estimates remain positive and use it to perform background correction. Then use the `plotDensities` function to graph the distributions of corrected intensity vaslues for all the arrays.

3. When these data were submitted to GEO, information about the spatial position and print-tip origins of spots were lost. Use the `normalizeWithinArrays` function to perform loess (but NOT print-tip loess) normalization of the background-corrected valuies. Then plot the density functions of the normalized intensity values.

4. The `limma` Users Guide shows how to use the `boxplot` function to get another view of the distributions. Produce boxplots of the log ratios of the normalized data. Do all of the arrays appear to be on the same scale?

5. Use the function `normalizeBetweenArrays` to put all the arrays on a common scale. Produce density plots and boxplots of the results. Was this normalization step worthwhile?

6. Use the methods in `limma` to fit a linear model to the normalized log ratio data, with a factor for Treatment and a factor for CellLine. List the top ten genes.

7. The gene annotations that came with the data only include the Compugen identifiers (for the long oligos on the arrays) and the associated GenBank identifiers. We have updated these using the `AnnBuilder` package. Download the UNC.Compugen library from trhe class web site. Install it and load it. Then find the gene name, gene symbol, and UniGene cluster id of the top ten genes.

8. Use a t-test or a linear model to select the top twenty genes that distinguish the cell line HME.CC from the other three cell lines. Use linear discriminant analysis (LDA) to develop a classifier to predict which samples are from the HME.CC cell line. How well does this classifier perform?

9. Repeat Question 8 using $k$ nearest neighbors (KNN) instead of LDA.

10. Repeat Question 8 using classification and regression trees (CART) instead of LDA.

11. Repeat Question 8 using LDA, but perform leave-one-out cross-validation (including the feature selection step) to evaluate the performance..

12. Repeat Question 11 using $k$ nearest neighbors (KNN) instead of LDA.

13. Repeat Question 11 using classification and regression trees (CART) instead of LDA.