GS01 0163 Assignment 4 Fall 2005
Due Date: 13 October 2005

1. This assignment continues Assignment 3, using R and BioConductor to explore a subset of the Singh prostate cancer data. For this problem, we will work with data that has already been background corrected using "rma". Normalize the background-corrected data using four different methods: "constant", "invariantset", "quantiles" and "loess". For each method, prepare boxplots and histograms (density plots) of the normalized data. Based on these plots, which method would you choose?

   (Implementation Note: Since each of these objects is large, you may need to remove (`rm`) some objects or play with the `memory.limit` command to complete this problem. Keep the next problem in mind when doing this. Also, remember that "loess" normalization takes much longer than the other methods.)

2. This is a continuation of the previous problem. For each of the four normalization methods, quantify the data using the "medianpolish" summarization method. Use `simpleCluster` to produce dendrograms. Does the choice of normalization method affect clustering based on the highest expressing genes?

3. This is a continuation of Problem 1. Select two arrays (one from each dChip cluster), and construct a data frame with eight columns. Each column should contain the PM intensity measurements from a different normalization method for one of the arrays. Prepare M-versus-A plots (using `mva.pairs`) of this data frame. (You can speed things up by selecting 10% of the PM features, and get adequate plots.) Explain any interesting qualitative features of these plots. Do these plots change your conclusion to the previous problem? Why or why not?

4. In this problem, we will work with the data that has already been background-corrected using "rma" and normalized using "quantiles". Use the `expresso` function to select the "pmonly" features and summarize them using 4 different methods: the "avgdiff" method of MAS4.0, the "liwong" method of dChip, the "mas" method of MAS5.0, and the "medianpolish" method of RMA. Prepare boxplots of the processed data for each method. (Note: because normalization produces an `exprSet`, you will have to manually extract the `exprs`, log transform them (in most cases: read about RMA!), and convert them to a data frame before creating the boxplot.) Do these plots lead you to prefer one method over the others? Why or why not?

5. This is a continuation of the previous problem. For each of the four summarization methods, use `simpleCluster` to produce dendrograms. Does the choice of summarization method affect clustering based on the highest expressing genes?

6. We have now systematically tried different background-correction, normalization, and summarization methods to look at 20 samples from the prostate cancer study. We were largely motivated by the observation that hierarchical clustering of the samples was driven by some feature of the data that was not related to the biological contrast between cancer and normal samples. In your opinion, can this unusual clustering be

fixed by changing the way we process the data? If yes, explain which processing method you would use. If no, then describe how we should proceed with this data set.