Statistics 5570/6570 Spring 2014 Homework 2

Expected Completion of Content: 7 Feb
Assignment Due (by 11:59 P.M.): 19 Feb

Directions: The following exercises are to be completed in R, following the instructions given for homework on the course syllabus. Neatness and format (including R code in an appendix) will contribute 5 points to the score. The main purpose of this assignment is to help students gain familiarity with several quality check, clustering, and visualization tools implemented in Bioconductor. See the course notes to get started with the R code. For these exercises, refer to the GSE36149 data (including the README file) available on the course website (same data as for Homework 1). This Homework will be graded out of 85 points.

1. (a) (5 points) Give overlayed histograms (or smoothed histograms) comparing the distributions of intensities on the twelve arrays, including a legend to identify the arrays (numbered 1-12 is fine).

   (b) (5 points) Why might this plot be useful?

2. (a) (8 points) Generate images of the the high-dose group SEM-K2 cell line replicate 2 array (HS2, should be array 8 in your AffyBatch object), using two different color representations, such as the image with color representing residuals after fitting the RMA probe-level model to all twelve arrays. (Do not include the default image like you did for another array in Exercise 2b of Homework 1.)

   (b) (5 points) Briefly discuss and compare what these images show.

3. Obtain the RMA-preprocessed gene expression levels using all 12 arrays (recall the `eset.rma` object from Exercise 5 of Homework 1), and perform a principal components analysis, using only the expression levels for the following 25 genes (probeset IDs):

```
gn.list <-
  c( "213005_s_at", "203695_s_at", "1552664_at", "219211_at"  , "238476_at"   ,
     "209392_at"  ,  "223441_at" ,  "217678_at", "202644_s_at", "221840_at"   ,
     "221234_s_at", "202643_s_at", "230671_at" , "201212_at"  , "1558662_s_at",
     "223376_s_at", "202988_s_at", "228372_at" , "225922_at"  , "227354_at"   ,
     "228592_at"  , "226545_at"  , "244163_at" , "205127_at"  , "210279_at"   )
```

   (a) (12 points) Create a scree plot and a biplot [with sample points labeled according to treatment/line combination – 'CR', 'CS', 'HR', 'HS', 'LR', 'LS' for arrays 1-2, 3-4, 5-6, 7-8, 9-10, and 11-12, respectively].

   (b) (5 points) Comment briefly on what these plots suggest.

4. (40 points; 20 each option) Do <u>two</u> of the following three options, starting with the `eset.rma` object:

   <u>Option 4.1</u>

   (a) (8 points) Using only the expression values for the 25 genes specified in Exercise 3, give a heatmap to simultaneously represent between-gene and between-array Euclidean distances, with row and column dendrograms from "agnes" clustering

with complete linkage. Be sure to use a "nice" color palette, and add column side colors (and column labels or names) to indicate the treatment/line combination of each of the 12 arrays ('CR', 'CS', 'HR', 'HS', 'LR', 'LS' as in Exercise 3a above). Make the heatmap colors be scaled by row.

(b) (2 points) Same as part (a), but make the heatmap colors be scaled by column.

(c) (5 points) Which scaling strategy (row/column) is more reasonable here, and why?

(d) (5 points) According to the appropriate heatmap, why might this subset of genes be of interest in this study?

Option 4.2

(a) (8 points) Using only the expression levels for the 25 genes specified in Exercise 3, perform a HOPACH clustering of the arrays (not the genes), using correlation distance. Using a "nice" color palette, give a heatmap with dotted lines to represent HOPACH clusters. Also provide a brief interpretation of this plot.

(b) (2 points) How many "main" clusters does this approach give here, and how do they relate to treatment/line combinations?

(c) (5 points) Use the `bootplot` function to give a visual representation of bootstrap-estimated cluster membership probabilities. With relatively few clusters, using the `ord="none"` option may produce more meaningful output. (If you can, modify the function to use a "nicer" color palette than the default rainbow palette.) Be sure to use column names as appropriate to emphasize meaningful interpretation in this plot.

(d) (5 points) Give a brief interpretation of the plot in part (c).

Option 4.3

(a) (5 points) Create an M-A plot to compare RMA expression levels on the control group RS4:11 cell line replicate 1 (CR1) array with the expression levels on the high-dose group RS4:11 cell line replicate 1 (HR1), for all 54,675 genes. Do this "by hand," explicitly defining M and A, and creating a simple scatterplot. (Note that RMA expression levels are already on the log scale, and these two arrays should be the first and fifth arrays, respectively, in your AffyBatch object.)

(b) (10 points) Using methods discussed in class, create two alternative M-A plots to more effectively visualize the differences between the two arrays. (For these plots, use a different color palette than the one in class - greens instead of blues, for example.)

(c) (5 points) Comment briefly on how these plots in (b) are more effective than the plot in (a).

5. (10 points) Discuss briefly some concerns that could arise in clustering and visualization of gene expression data (i.e., of what should you beware?).