Statistics 5570/6570 Spring 2014 Homework 1

Expected completion of material in class: 17 Jan
Assignment due (by 11:59 P.M.): 5 Feb (~~procrastinate~~ START SOON!)

Directions: The following exercises are to be completed in R, following the instructions given for homework on the course syllabus. Neatness and format (including appendix) will contribute 5 points to the score. The main purpose of this assignment is to reinforce the interpretation of gene expression technology in general and to help students get started with R and Bioconductor. For these exercises, refer to the GSE36149 data (including the README file) available on the course website. This assignment will be graded out of 75 points.

1. (20 points) Write 1-2 paragraphs (about half a page) summarizing gene expression technology and its purpose. (You may use resources other than the class notes.)

2. Create an AffyBatch object in R for the GSE36149 data. (Consider `abatch.raw <- ReadAffy()`.)

   (a) (3 points) How many "genes" (probesets) are represented on each array?

   (b) (3 points) Create the image of the array for the control group RS4:11 cell line replicate 1 (CR1, should be array 1 in your AffyBatch object).

   (c) (3 points) Plot the PM (perfect match) intensities for this same CR1 array vs. the PM intensities for the high-dose group RS4:11 cell line replicate 1 (HR1, should be array 5 in your AffyBatch object). Make both axes be on the log scale, and with a 45-degree reference line (intercept 0, slope 1).

   (d) (3 points) What percentage of probes on the CR1 array have PM $\leq$ the PM on the HR1 array? (HINT: Create a vector of TRUE/FALSE evaluating this for each probe, and take the mean of this vector.)

3. (6 points) In your own words, explain what preprocessing is for gene expression studies. Describe and comment on the purpose of each of the three steps.

4. Using **all twelve** GSE36149 arrays, obtain the following AffyBatch objects in R (the `bg.correct` and `normalize` functions will be useful here, and should be used in sequence):

Object Name	Contains
abatch.raw:	"raw" intensities from the .cel files - returned by `ReadAffy()`
abatch.bg:	RMA-background-corrected intensities - returned by `bg.correct.rma(abatch.raw)`
abatch.bg.qn:	quantile-normalized RMA-background-corrected intensities - returned by `normalize(abatch.bg, method="quantiles")`

Then, using the CR1 array data from these different AffyBatch objects (should be array 1 in each), create the following plots with both axes on the log scale, and with a 45-degree reference line (intercept 0, slope 1):

(a) (6 points) "raw" PM vs. RMA-background-corrected PM

(b) (6 points) RMA-background-corrected PM vs. quantile-normalized RMA-background-corrected PM

5. Using the AffyBatch objects from Exercise 4 and the ExpressionSet object resulting from using the full `rma` procedure (consider `eset.rma <- rma(affybatch.raw)`), create histograms (using the `hist` function) of the following quantities for the CR1 array:

   (a) (2 points) "raw" PM, on the log2 scale (from abatch.raw)

   (b) (2 points) background-corrected PM, on the log2 scale (from abatch.bg)

   (c) (2 points) quantile-normalized background-corrected PM, on the log2 scale (from abatch.bg.qn)

   (d) (4 points) gene expression values from the `rma` procedure (from eset.rma)

6. (10 points) Based on the results you obtained in Exercises 4 and 5, comment on the apparent effect of each of the three steps of RMA preprocessing for the data on this one array (CR1).