Statistics 5570/6570 Spring 2013 Homework 3

Expected Completion of Content: 24 Feb
Assignment Due (by 11:59 P.M.): 19 Mar

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 help students gain familiarity with several tests for differential expression implemented in Bioconductor, and to discuss some of the statistical issues involved. 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 previous homeworkd). This homework will be graded out of 115 points.

1. Test all 54,675 genes for differential expression across obatoclax dosage levels (C = "control" [arrays 1-4], "H = high" [arrays 5-8], and "L = low" [arrays 9-12]) using RMA expression estimates and the limma/eBayes approach, in a model accounting only for dosage level.

   (a) (10 points) Report a histogram of the "raw" P-values (one for each gene, testing H_0: $\mu_C = \mu_H = \mu_L$).

   (b) (5 points) Comment briefly on what the shape of this histogram suggests.

   (c) (5 points) What percentage of all genes are called differentially expressed here when the FDR is controlled at 0.05 (using Benjamin-Hochberg adjustment)?

2. (20 points) Test for differential expression using the same H_0 as in Exercise 1, but this time use a limma/eBayes model that also accounts for cell line (R = "RS4:11" [arrays 1-2, 5-6, 9-10] and S = "SEM-K2" [arrays 3-4, 7-8, 11-12]). Report (a), (b), and (c) as in Exercise 1.

3. Consider a filter to focus on the 13,373 genes with expression above 50 in at least 20% of cases and CV above 0.15 (referring to un-logged RMA expression levels).

   (a) (8 points) Give visual evidence to show that this would be a reasonable filter for these data.

   (b) (5 points) Comment briefly on why filtering can be useful in gene expression analysis.

4. (20 points) Repeat Exercise 2 above, but test only the 13,373 genes that pass the filter of Exercise 3 above. In addition:

   (d) (5 points) Construct and comment briefly on a heatmap of the significant genes (at FDR .05, there should be 20). (While this will involve a different set of genes, it is similar to Exercise 4.1a of Homework 2.)

5. (5 points) Compare what you reported for Exercises 2 and 4, and comment briefly on the effect filtering had here.

6. (7 points) In your own words, what is the purpose of the FDR adjustment used here?

7. (15 points) Using either the 13,373 genes that passed the filter or all 54,675 of them (you choose, maybe depending on your computational resources), apply a Random Forest approach to identify the most important genes for differentiating the three dosage groups ("control", "high", and "low"). Do one of the following:

   (a) Construct a useful plot of the genes' importances, and also construct a heatmap of the most important genes (you choose how many), with useful column-side colors (and column names), and comment briefly on what these show.

   (b) Using the `varSelRF` package, perform a Random Forest variable selection, and construct a scatterplot showing the expression levels of the selected genes, with points colored to identify the three groups, and also a parallel coordinates plot. Comment briefly on what these show.

8. (5 points) Discuss briefly the relative performance of the limma/eBayes and Random Forest approaches on these data. (Comparing Exercises 4d and 7 may be helpful.)