PubH 7445: Homework Assignment 7

October 28, 2019

Due Wednesday, November 13, 2019

Please hand in a print-out of your answer and R code, and also email your R code to Souvik (sealx017@umn.edu).

1. Use the CLLbatch dataset for this homework. Set up this affyBatch object as we did in class (i.e. match disease status and create an annotated data frame to hold the phenotypic data and metadata).

   (a) Delete the microarrays labeled `CLL1` and `CLL11`

   (b) Use the RMA algorithm to get probe level summaries of gene expression.

   (c) Filter the results from RMA to exclude all probe sets that are Affymetrix controls, probe sets that don't have a gene ontology assignment for biological process and probe sets that are the 20%, 40%, 60% and 80% least variable.

   (d) For each of the datasets from the previous part, use *t*-tests and the moderated *t*-tests of the limma approach to test for differences between groups defined by disease status.

   (e) Make volcano plots for all 8 analyses from the previous part and highlight (e.g. use a different color) those probe sets for which there is at least a 1.5 fold difference and the false discovery rate for declaring a probe set is different is less than 20% (which is high).

   (f) Obtain the symbols and Entrez Gene ID for all genes that differ with some level of filtering for either analysis. For both of the 2 approaches from the previous part, are there any genes that differ in terms of their level of expression for every level of filtering? Are there any genes that differ for more than one level of filtering?

   (g) For each filtered dataset test for over-representation of GO pathways using at least 2 different cutoffs for the FDR for the gene list.

2. Use biomaRt to obtain the ENSEMBL gene ID, the gene symbol, the GO ID, and GO term name and GO category for the following genes: ENSBTAG00000000005, ENSBTAG00000000008, ENSBTAG00000000009, ENSBTAG00000000010, and ENSBTAG00000000011. How many of the GO descriptors include the word "lung"?

3. The Affymetrix human 133 series of microarrays had multiple arrays which were referenced by a letter. For this exercise obtain the database versions of the A and B arrays `hgu133a.db` and `hgu133b.db`.

   (a) How many probes are on each of these arrays?

   (b) How many distinct ENSEMBL gene IDs map to a probe set on each array?

   (c) How many distinct gene symbols map to a probe set on each array?

(d) Are there any genes (using gene symbols to define a gene) that have a probe set on array B but not on array A?

(e) For each array, which gene (using gene symbols to define a gene) that is represented by a probe set on that array has the most distinct ENSEMBL IDs?

(f) For each array, make a table of the frequency distribution of the number of times each gene is represented on the array (using gene symbols to define a gene).

(g) For each array determine which gene (as represented by a symbol) has the most probe sets and how many probe sets does this gene have?