Northeastern University
College of Science

Module 10 - Homework

Problem 1 (20 points) Preprocessing a data set

Install the "ArrayExpress" package from Bioconductor. Load the yeast microarray data using R commands:

```
library(ArrayExpress)
yeast.raw = ArrayExpress('E-MEXP-1551')
```

(a) Preprocess the raw data set into an expression data set using: the “mas” background correction method, the “quantiles” normalization method, “pmonly” pm correction method and “medianpolish” summary method. Give the R command here for doing this task.

(b) Print out the mean expression values for the first five genes across all samples.

(c) How many genes and how many samples are in the preprocessed expression data set?

Problem 2 (30 points) Searching Annotations

(a) What is the annotation package for the yeast data set in question 1?

Install the annotation package from Bioconductor.

(b) Search the 1769308_at gene GO numbers related to Molecular Function (MF). How many GO numbers do you get?

(c) Find the GO parents of the GO IDs in part (b). How many GO parents are there?

(d) Find the GO children of the GO IDs in part (b). How many GO children are there?

Problem 3 (30 points) Gene filtering on B-cell ALL patients

We work with the patients in stages "B2","B3".

(a) We look for genes expressed differently in stages B2 and B3. Use genefilter to program the Wilcoxon test and the Welch t-test separately for each gene. For each test, we select the genes with p-value<0.001. To save computational time, we set exact=F in the Wilcoxon test function.

(b) Compute a Venn diagram for the Wilcoxon test and the t-test, and plot it.

(c) How many pass the Wilcoxon filter? How many passes both filters?

(d) What is the annotation package for the ALL data set? Find the GO numbers for "oncogene".

(e) How many genes passing the filters in (a) are oncogenes?

Problem 4 (20 points)

Stages of B-cell ALL in the ALL data. Use the limma package to answer the questions below.

(a) Select the persons with B-cell leukemia which are in stage B1, B2, and B3.

(b) Use the linear model to test the hypothesis of all zero group means. Use "topTable()" to report the **top five** genes with nonzero means in **B3 group**.

(c) Use two contrasts to perform analysis of variance to test the null hypothesis of equal group means. Do this with a false discovery rate of 0.01. **How many** differentially expressed genes are found? Use "topTable()" to report the top five genes that express differently among the three groups.