BIOS 497/897-001

HW5: differential expression analysis with edgeR

Note: The edgeR package requires limma package. If did not install limma package before, you need to install limma package at the same time.

> source("http://bioconductor.org/biocLite.R")

> biocLite("edgeR")

> biocLite("limma")

The R code for using edgeR

> library(edgeR)

> library(stats)

> setwd("/Users/chizhang/homework/HW5")

> set.seed(133)

> y <- as.matrix(read.table("RNA\_seq\_data.txt"))

> tags <- as.matrix(read.table("tags.txt"))

> rownames(y)=tags

> lib.sizes <- c(sum(y[,1]),sum(y[,2]),sum(y[,3]),sum(y[,4]),sum(y[,5]), sum(y[,6]))

> d<-DGEList(counts=y, group=factor(c("C","C","C","T","T","T")), remove.zeros = TRUE)

> d<-estimateCommonDisp(d)

> ms<-exactTest(d)

> result=topTags(ms, n=39656, adjust.method= "fdr", sort.by="logFC")

Questions:

1. What do the following two R statements work for?

> tags <- as.matrix(read.table("tags.txt"))

> rownames(y)=tags

A: To associate the gene IDs with their corresponding numbers of reads.

1. How do you calculate the total number of reads? (Hints: you may use R command to calculate.)

A: The following command can calculate the total number of reads:

> lib.sizes <- c(sum(y[,1]),sum(y[,2]),sum(y[,3]),sum(y[,4]),sum(y[,5]), sum(y[,6]))

However, please note that this step is useless for edgeR, because the function DGEList() can calculate the total numbers of reads by itself.

1. What are the top five genes that have the largest Fold Change (absolute values)? What are their Log(Fold Change) and P-values?

A: the following command can save top five genes into the “result” variable.

> result=topTags(ms, n=5, adjust.method= "fdr", sort.by="logFC")

4. To plot the volcano plot

A: any one of the following commands can plot the volcano plot:

>plot(result$table[,1], -log2(result$table[,3]), xlab="log(FC)", ylab="-log(P-value)")

or

>plot(result$table$logFC, -log2(result$table$Pvalue), xlab="log(FC)", ylab="-log(P-value)")