Bulk RNASeq Analysis Laboratory

Objective

1. To get familiarised with R scripting
2. Understanding the public data available through an article
3. Find differentially expressed genes in PrimaryColon Vs Normal and Metastasis Vs PrimaryColon

Background Tasks

Please refer the github page complete analysis pipeline steps and the corresponding scripts uploaded in the link https://github.com/IBEXCluster/B322.
The above link has a folder "counts" where the files needed for differential expression task is available.
make sure you create a directory for this hands on ibex either on your home directory or under your scratch folder.
copy the meta data and counts files for GSE50760 in your working directory.
Go through edgeR user manual for differential expression analysis.
Refer IntroductionToR slides for exploratory data analysis.

For this exercise, you need to login to ibex and load the following modules on command line

```
###Loading modules
module load R/3.6.0/gnu-6.4.0
module load RStudio_Desktop/1.1.383

###To invoke RStudio to work with R
# type the below word, on the command line
rstudio &
```

Libraries needed

```
### create a new file and start writing your script
##Load the following R libraries in your script
library(limma)
library(edgeR)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(tidyverse)
library(ggplot2)
library(gridExtra)
library(ggrepel)
```

```
library(reshape2)
library(GGally)
library(EnhancedVolcano)
```

ENSEMBL Id to Gene Symbol Code

```
####Code to replace ensembl ids in the count table, with gene symbol
## Map from Ensembl gene id to gene symbol
ensg <- sub("\\..*", "", rownames(cts))  # remove version number

sym <- mapIds(org.Hs.eg.db, keys=ensg,
          column="SYMBOL", keytype="ENSEMBL")

gene <- data.frame(ENSGID=ensg, SYMBOL=sym, stringsAsFactors=F)
#use the above vector for you raw count rownames
```

Hands on Questions

Question 1 :
Explore the data
1a. remove genes with no expression for all samples
1b. boxplot of library size for Tissue group and Subjects
1c. filter genes by expression before normalization
1d. Look at the difference in cpm of raw expression and filtered expression values by density plots

Question 2 :
Data normalisation by TMM method

Question 3 :
Run multidimensionality reduction like PCA or MDS
Find out the outlier or not properly grouped samples and remove them for downstream analysis

Question 4 :
Differential expression for two contrasts a. PrimaryColon Vs Normal, b. MetastasisColon Vs PrimaryColon

Question 5 :
Find differentially expressed genes at p value < 0.01 NS lfc=log2(4)

Question 6 :
Find the functional analysis of top 10 differentially expressed genes in case a and case b

(Use either DAVID or gprofiler for this)

Question 7 :

Handson Results (for submission)

Write a 3 page report on Population transcriptomic analysis explaining the workflow starting from input data to differential expression analysis.
Explain in detail what does each part of the pipe line does,understand why you have to do that, tools used, output files, etc
For differential expression analysis, make sure you could observe the DE genes, same as the authors of our reference paper.

Submission Rule

We are looking for a report from a team of two students.
If you can't make a group with your fellow student, we expect a report per student.
Reports not following these rules will not be considered for evaluation.
Any queries related this lab should be addressed via B322 slack channel
https://kaust-ibex.slack.com/

Deadline for Submission

Students are requested to send the report to manjula.thimma@kaust.edu.sa, mentioning student(s) name and id
Deadline for submission: March 28th 5.00pm