RNA-seq analysis - Final Exercise

Juan R Gonzalez

BRGE - Bioinformatics Research Group in Epidemiology
Barcelona Institute for Global Health, ISGlobal
http://brgeisglobal.org

Recount is a repository of RNAseq data available at `https://jhubiostatistics.shinyapps.io/recount/`. Go to the TCGA tab and download one of the Ranged Summarized Experiment (RSE) objects for any of the available tumors in the gene column (select any from version 2). The file is named `rse_gen_XXX.Rdata` where XXX stand for the selected cancer. Let us assume I have donwloaded the file `rse_gen_pancreas.Rdata`. Load the data into R by `load(''rse_gen_pancreas.Rdata'')`. An object called `rse_gene` will be available into your R session.

The aim of this exercise is to perform a complete RNAseq analysis in which we aim to compare individuals having early vs advance tumoral stages (variable GROUP).

The variable `gdc_cases.diagnoses.tumor_stage` contains such information in different categories. The variable GROUP can be created by using this code

```
stage <- rse_gene$gdc_cases.diagnoses.tumor_stage

ids.early <- grep(paste("stage i$", "stage ia$", "stage ib$" ,
                        "stage ic$", "stage ii$", "stage iia$",
                        "stage iib$" , "stage iic$",
                         sep="|"), stage)
ids.late <- grep(paste("stage iii$", "stage iiia$", "stage iiib$" ,
                       "stage iiic$", "stage iv$", "stage iva$",
                       "stage ivb$" , "stage ivc$",
                        sep="|"), stage)
colData(rse_gene)$GROUP <- rep(NA, ncol(rse_gene))
colData(rse_gene)$GROUP[ids.early] <- "early"
colData(rse_gene)$GROUP[ids.late] <- "late"
```

TO DELIVER:

Write a report indicating the tumor you have selected and making a short description of the data (e.g., number of samples, features, ...). Then describe the analyses you have carried out, interpret the results and include some discussion making reference to the tables and figures you think are necessaries to be showed in the report. Remember that data visualization is important.

The analyses should include data normalization, differential expression and enrichment analyses. Select the methods you think are appropiated to your data. Standard enrichment methods (i.e GO, KEGG) will allow to get 9 over 10. The use of more sophisticated approaches will have the possibility of getting an extra point.

Create the report using R Mardown and set `echo=TRUE` in the entire document. Upload ONLY the pdf to the Campus Virtual (anything else will be evaluated).