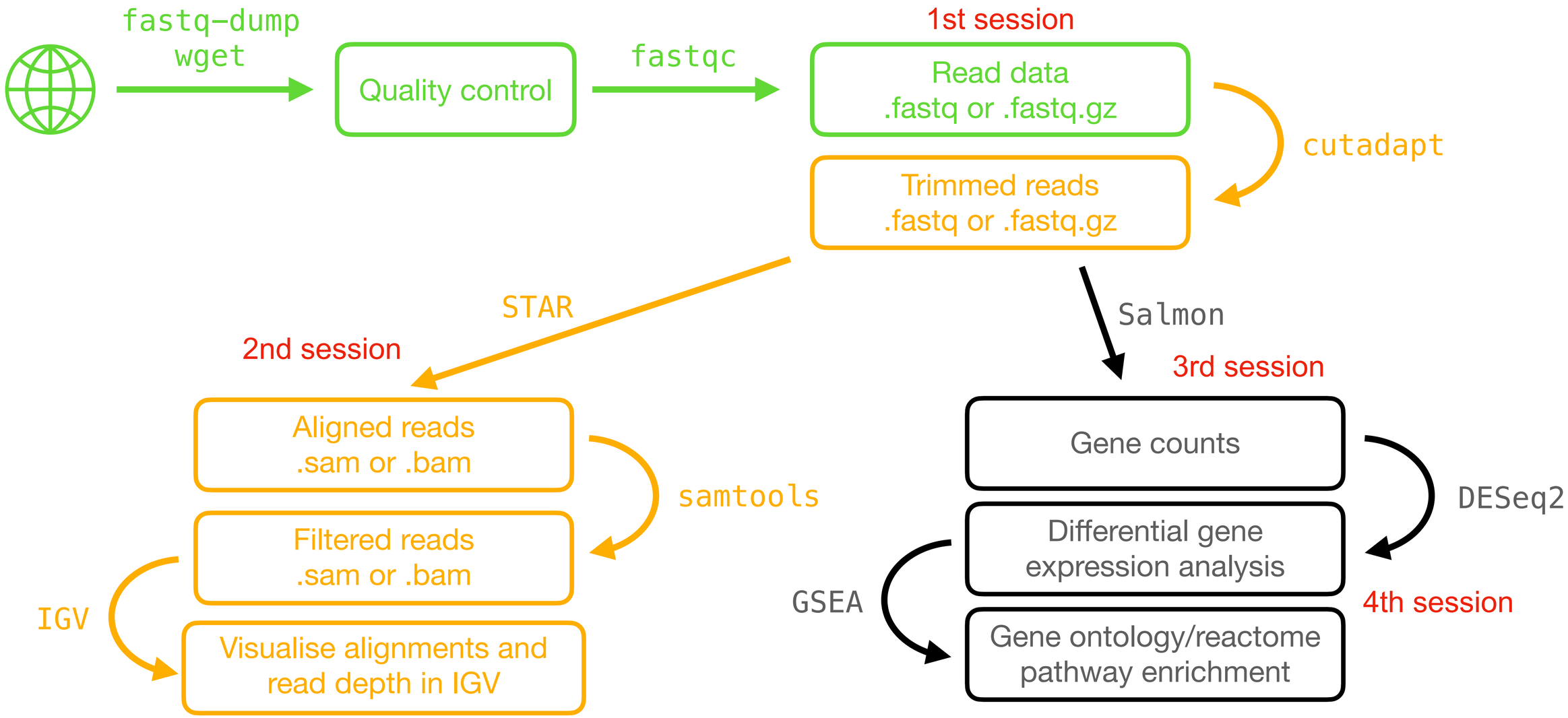
Overview of the course



✦ Introduction

- Differential gene expression using Deseq2
- Pathway analysis
- Introduction to R

✦ Practical

- The output of salmon quants
- tximport.r
- DGE.r
- PathwayAnalysis.r

Differential gene Expression (DGE)

Aim:

Determine whether the mean expression levels of different sample groups are significantly different.

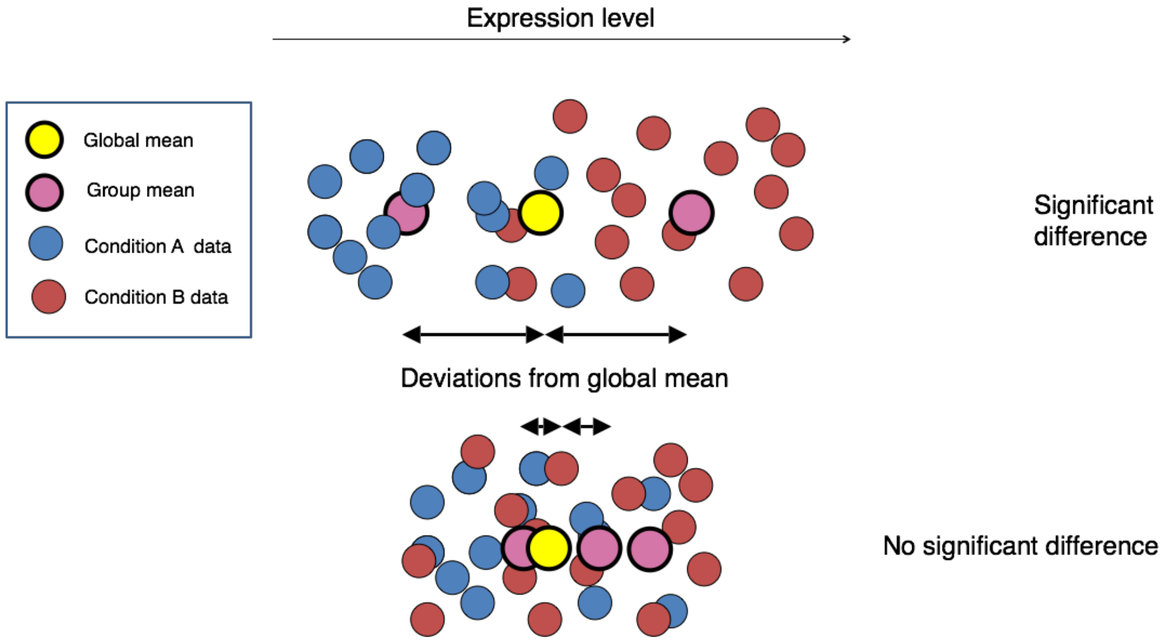
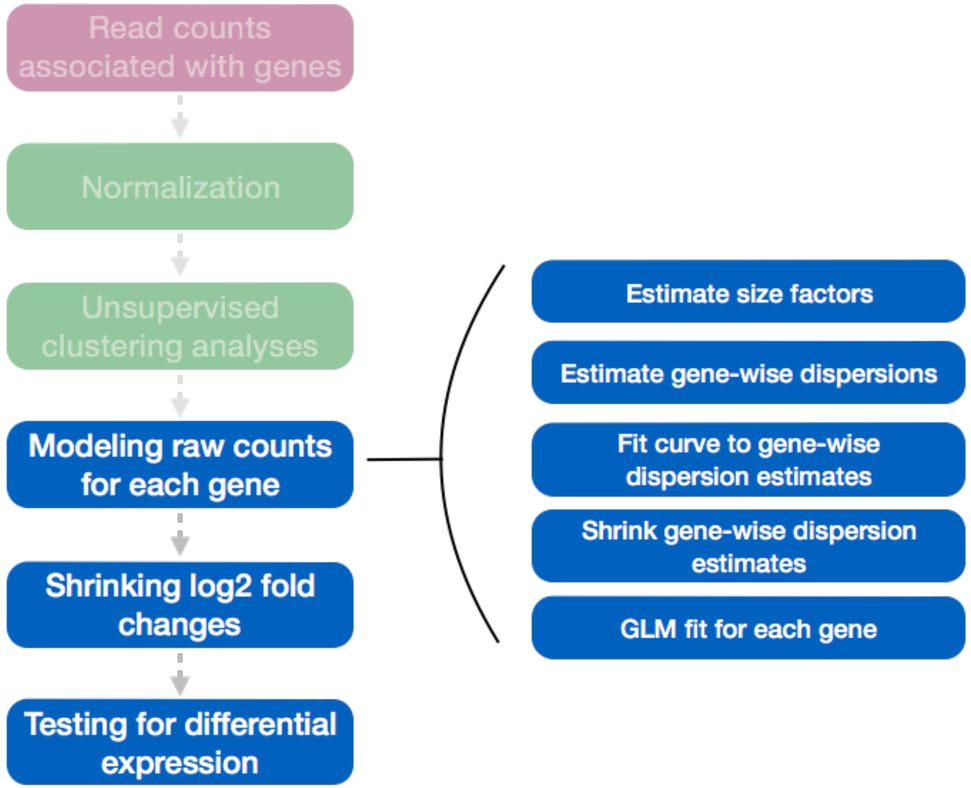


Image credit: Paul Pavlidis, UBC

https://angus.readthedocs.io/en/2019

DEseq2

- The counts generated by RNAseq data shows overdispersion (variance > mean).
- DESeq2 uses the negative binomial distribution to make estimates and perform statistical inference on differences.
- DESeq2 performs hypothesis testing using the Wald test or Likelihood Ratio Test.



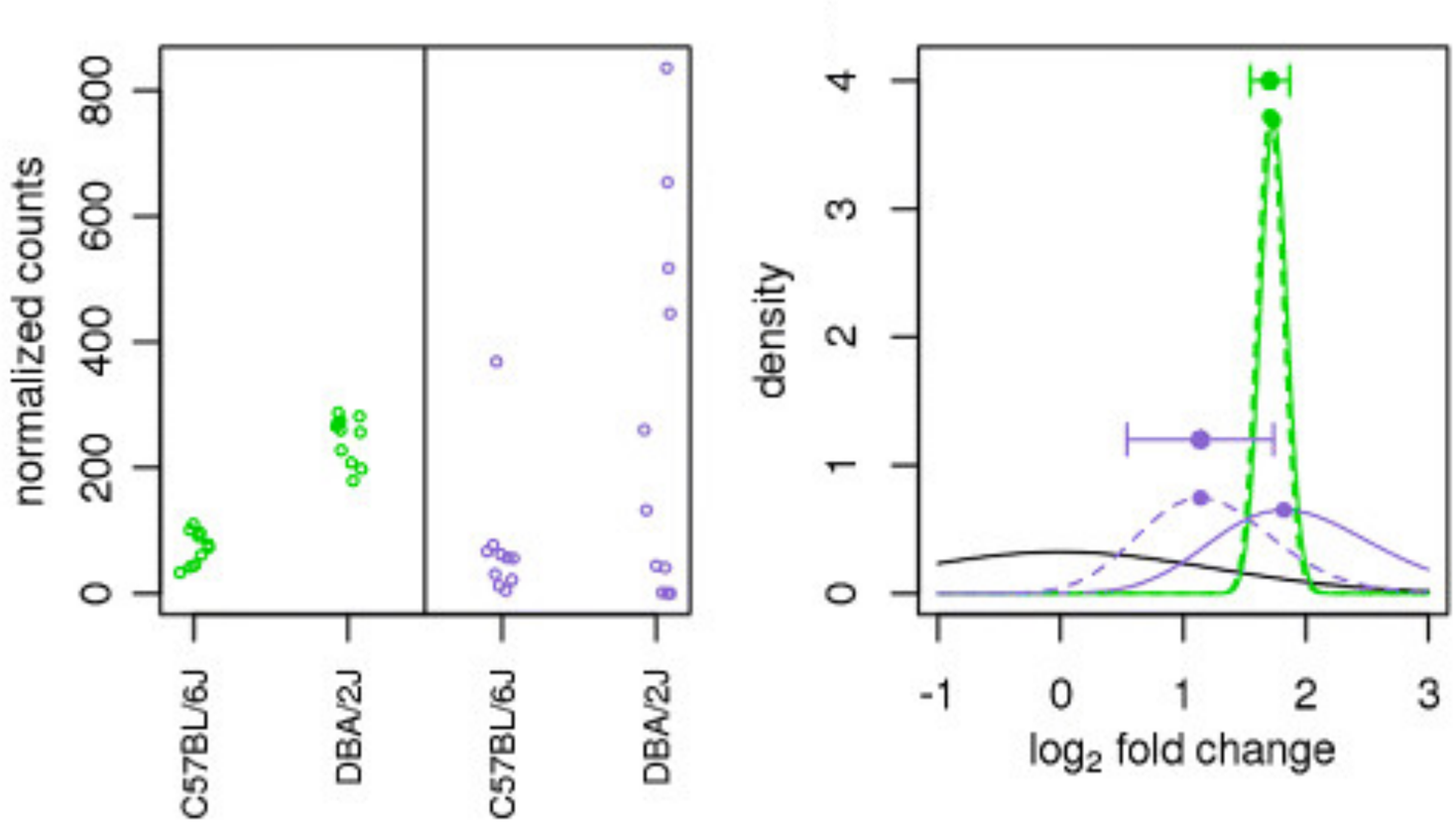
https://angus.readthedocs.io/en/2019

Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**:550. 10.1186/s13059-014-0550-

Shrunken log2 foldchanges (LFC)

To generate more accurate log2 foldchange estimates, DESeq2 allows for the shrinkage of the LFC estimates when the information for a gene is low, which could include:

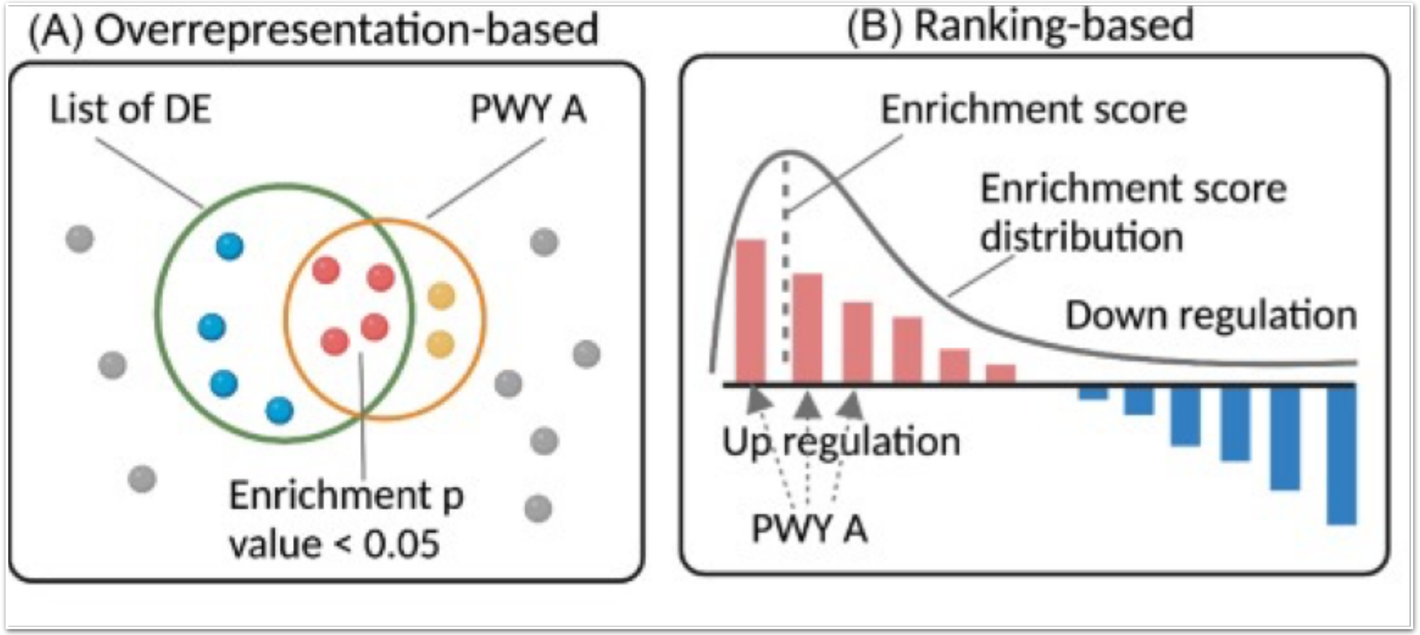
- Low counts
- High dispersion values



Love, M.I., Huber, W., Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, **15**:550. 10.1186/s13059-014-0550-

Pathway Analysis

- The output of DGE analysis is a list of genes.
- Pathway analysis allows a better understanding of the biological insight of the significant genes.



Zhao K and Rhee S Y . Trends in Genetics, 2023, https://doi.org/10.1016/j.tig.2023.01.003

Online tools for DGE

Provide a list of Gene or Gene ID and they generate graphs and reports
No bioinformatics skills are required

- https://metascape.org/gp/index.html#/main/step1 
- https://maayanlab.cloud/Enrichr/ 
- https://genemania.org 
- http://www.webgestalt.org/ 

Extract the raw counts from the quants file

We will be using R and R studio

- **R**: programming language and environment for data manipulation, statistical computing, and graphical display
- **R:** Free and open source https://www.r-project.org/

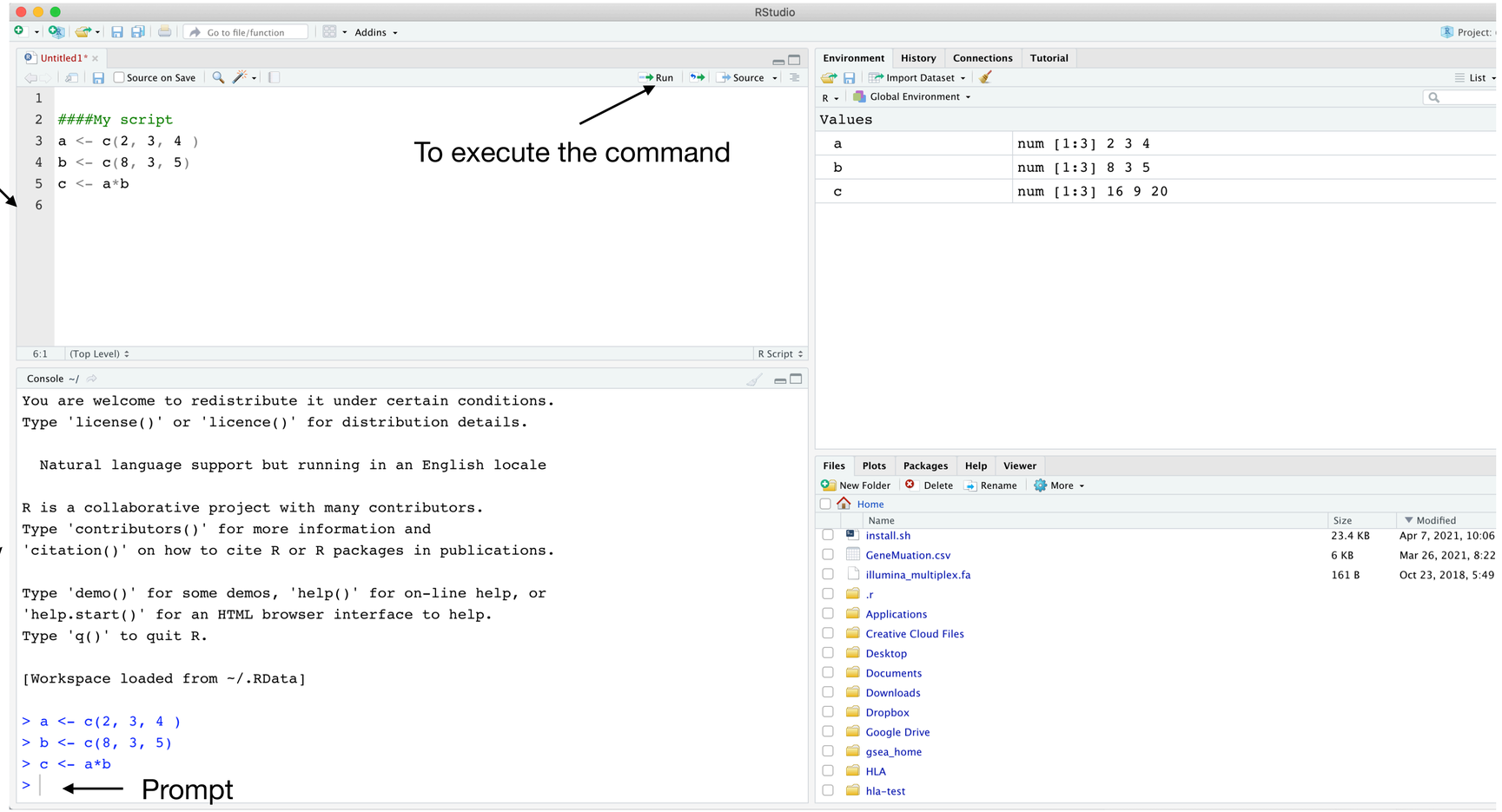
- **R studio:** Free and open source IDE (Integrated Development Environment) for R. Available for Mac, Windows and Linux

Rstudio interface

Script editor

To execute the command

Object, history and Enviroment



R Console

Tree of folders, graph window, Packages, help and Viewrt

Installing packages in R..... demo only

To install libraries in R, don’t run all done for this course

The first package that we will be using is the "tximport". This is a Bioconductor package

https://bioconductor.org/packages/release/bioc/html/tximport.html

```
if (!require("BiocManager", quietly = TRUE))
install.packages("BiocManager")
BiocManager::install("tximport")

install.packages("dplyr")
#The installation will be one time only, then we call them using the library()
```

Best practices when using R

- Document everything you do so your code is reproducible.
- Use the # sign to comment. In R anything after the # is ignored.
- Write the code in the script editor rather than R console.
- Keep track of all the libraries used in the analysis together with the R version.

Practical… 1st part

✦View the Salmon output

✦Convert the TPM counts to raw gene count using tximport.

Output of salmon quant

In the last session, we generated the counts with salmon quant

- navigate to the output directory

  cd /mnt/beegfs/workshop/**<SSO>**/results/salmon_quants

- View the content of the folder (what command do we use to list the files?)

- For those who did not get the chance to generate the count please copy the results to your own workspace

cp -r /mnt/beegfs/workshop/DGE_results_codes/salmon_quants   /mnt/beegfs/workshop/**<SSO>/**

quants file

```
$ ls -lh
drwxrwxr-x 5 jesu2166 hassan    6 May  5 12:10 5aza_rep1_sample_quant
drwxrwxr-x 5 jesu2166 hassan    6 May  5 12:15 5aza_rep2_sample_quant
drwxrwxr-x 5 jesu2166 hassan    6 May  5 12:19 DMSO_rep1_sample_quant
drwxrwxr-x 5 jesu2166 hassan    6 May  5 12:24 DMSO_rep2_sample_quant
```

Navigate to one of the folders

```
$ cd 5aza_rep1_sample_quant
$ ls -lh
total 12M
drwxrwxr-x 2 jesu2166 hassan  12 May  5 12:10 aux_info
-rw-rw-r-- 1 jesu2166 hassan 393 May  5 12:06 cmd_info.json
-rw-rw-r-- 1 jesu2166 hassan 575 May  5 12:10 lib_format_counts.json
drwxrwxr-x 2 jesu2166 hassan   1 May  5 12:10 libParams
drwxrwxr-x 2 jesu2166 hassan   1 May  5 10:59 logs
-rw-rw-r-- 1 jesu2166 hassan 12M May  5 12:10 quant.sf
```

View the content of the quant.sf

```
$less quant.sf
```

Name	Length	EffectiveLength	TPM	NumReads	
ENST00000631435.1		12	12.000	0.000000	0.000
ENST00000415118.1		8	7.000	0.000000	0.000
ENST00000448914.1		13	13.000	0.000000	0.000
ENST00000434970.2		9	9.000	0.000000	0.000
ENST00000632524.1		11	11.000	0.000000	0.000
ENST00000633009.1		20	19.000	0.000000	0.000
ENST00000634070.1		18	17.000	0.000000	0.000
ENST00000632963.1		20	19.000	0.000000	0.000
ENST00000633030.1		19	18.000	0.000000	0.000
.					
.					
.					
.					
ENST00000620516.4		1179	1145.496	12.961913	5.427

TPM: Transcripts Per Million

Press q to exit the less command

R script to extract the raw count and save it

- Before running R we need copy the scripts for today's session to your work directory and create an output directory to save the results
- cd /mnt/beegfs/workshop/**<SSO>**/
- copy the scripts: note the use of the willd card and the dot at the end of the command line
- cp /mnt/beegfs/workshop/DGE_results_codes/*.r .
- mkdir /mnt/beegfs/workshop/**<SSO>**/results/counts
- Use the NoMachine
- $ srun -c 4 --mem 8000 --X11 --pty /bin/bash
- Are you sure you want to continue connecting (yes/no/[fingerprint])? yes
- SSO@linux020's password: type your SSO password
- $ module load RSTUDIO TXIMPORT R_NGS_ANALYSIS/1.0 APEGLM/3.17 DESEQ2/1.38.0
- $ rstudio

- Open the tximport.r to the Rstudio
- File > Open File …> tximport.r
- setwd ("/mnt/beegfs/workshop/**<SSO>**/results")
- change **<SSO>** to your own credentials

R script…. tximport.r

```
library("tximport")

library("dplyr")

## assign your working directory

setwd (“/mnt/beegfs/workshop/<SSO>/results/")

## List all directories containing data

samples <- list.files(path = "./salmon_quants", full.names = T, pattern="_sample_quant$")

## Obtain a vector of all filenames including the path

files <- file.path(samples, "quant.sf")

## list all the files to the console

files

##assign a shorter name for each element

> names(files) <- list.files(“salmon_quants”)
```

R script…. tximport.r

```
## Read the annotation file

tx2gene <- read.csv("/mnt/beegfs/workshop/DGE_results_codes/tx2gene_ens109.csv")

## if you want to view this file type head (tx2gene)

# Run tximport

txi <- tximport(files, type="salmon", tx2gene=tx2gene[,c("TXNAME", "GENEID")], countsFromAbundance="lengthScaledTPM",
ignoreTxVersion = TRUE)

## Check the output

head(txi[["counts"]])

## Extract the counts, round the values and change the list to a data frame

counts <- txi$counts %>% round() %>% data.frame()

## Change the column names

colnames (counts) <- c("5aza_rep1", "5aza_rep2", "DMSO_rep1", "DMSO_rep2" )

View(counts)

### Save the results for the next session using the write.table command

write.table(counts, "counts/Count_fromtximport_Salmon.txt", sep ="\t", quote = F)

####keep a note of the packages and their versions

sessionInfo()
```

How to run the script

- Highlight the line and press the **run** button. Don’t hit Enter (it deletes the command)
- The variables will be stored in the upper right-hand panel of studio

  and the output will be printed on the console.

Check the results:

They will be saved under /mnt/beegfs/workshop/**<SSO>**/results/counts

Differential gene Expression

Practical 2

- Open the DGE.r into the Rstudio
- File > Open File …> DGE.r
- setwd ("/mnt/beegfs/workshop/**<SSO>**/results/")
- change **<SSO>** to your own credentials

Output: when running the script, some results will be seen in the console, and some results will be saved.

The csv files will be saved at: /mnt/beegfs/workshop/**<SSO>**/results/counts

The pdf files will be saved at: /mnt/beegfs/workshop/**<SSO>**/results

DGE.r…… the main points

```
Read the count data generated in the previous code
#Load the Sample information (metadata).

counts <- read.table(file = "counts/Count_fromtximport_Salmon.txt", header= TRUE, check.names = F)
metadata <- read.csv(file= "/mnt/beegfs/workshop/DGE_results_codes/SampleInfo_ngs.csv", row.names = 1)

metadata$condition <- factor(metadata$condition,
                             levels = c("Control", “Treatment"))

#check if the sample name is matching the metadata
all(rownames(metadata)%in% colnames(counts))

all(rownames(metadata)== colnames(counts))

#If the order of rows and columns is  not the same, try do the following
counts<- counts[, row.names(metadata)]
```

DGE

##design the matrix

```
design <- as.formula(~condition)
model<- model.matrix(design, data= metadata)
keep <- rowSums(counts)>5
countdata<- counts[keep,]
countdata<- as.matrix(countdata)

#### Create DESeq2Dataset object
dds.raw <- DESeqDataSetFromMatrix(countData = countdata,
                                  colData = metadata,
                                 design = design)
#Perform the differential gene expression
dds <- DESeq(dds.raw)

res_05<- results(dds,alpha= 0.05)

resLFC_05 <- lfcShrink(dds, coef="condition_Treatment_vs_Control", type="apeglm", res= res_05)
```

Map gene symbol and ENTREZID to ENSEMBL

```
DEG <- as.data.frame(resLFC_05) # assign the results to data frame degenes

# assign symbol(common names) and ENTREZ id according to ENSEMBL id
DEG$symbol<- mapIds(org.Hs.eg.db,
                    keys= rownames(DEG),
                    column = "SYMBOL",
                    keytype = "ENSEMBL",
                    multiVals = "first")

DEG$entrez<- mapIds(org.Hs.eg.db,
                    keys= rownames(DEG),
                    column = "ENTREZID",
                    keytype = "ENSEMBL",
                    mutiVals= "first")

# remove genes that don't have a common name and those with duplicated gene name
DEG_symbol<- DEG[is.na(DEG$symbol)== FALSE,]
dim(DEG_symbol)

DEG_symbol<- DEG_symbol[!duplicated(DEG_symbol$symbol),]
```

Pathway analysis

Practical 3

The script is PathwayAnalysis.r, do exactly what you did before for the other R scripts

```
DEG05_symbol <- read.csv("counts/DEGs_5uMaza_treatment_significant.csv"  )

geneset <- as.character(DEG05_symbol$entrez)

#GO over-representation analysis
ora_go <- enrichGO(gene          = geneset,
                   universe      = NULL,              # all available genes in db
                   OrgDb         = org.Hs.eg.db,      # Hs: homo sapiens
                   ont           = "BP",              # One of MF, BP, CC*
                   pAdjustMethod = "BH",
                   pvalueCutoff  = 0.01,
                   qvalueCutoff  = 0.05,
                   readable      = TRUE)
```

Before the end of this session

- Save your code
- Quite Rstudio
- cancel your job for the Rstudio session otherwise, your group will be charged as long as the job is running
- scancel job ID
- or use: scancel -u <SSO> #this will cancel all your jobs