RNA seq

Key steps

Raw counts	Normalization	Model or comparisons	DGEA	Downstream analysis
Prefiltering lowly expressed genes > Visualize the raw data (PCA) > Keep in mind which genome version was used	Observe how the data is modified after normalization > Visualize the data (PCA) > Check the dispersion	Select the relevant comparisons and choose the model to be used (GLM, etc)	Perform the differential gene expressions analysis and use FDR correction (automatic in DESeq2)	GO analysis: check biological terms overrepresentation. Typically BP, CC or MF. GSEA: similar to the GO analysis but this one takes into consideration the FC. KEGG: Check if the set of genes are involved in particular pathways.

1 - Raw counts

- ★ DESeq2 requires an object containing the counts and the metadata that describes the properties of each sample
- ★ There are several ways to load the data for DESeq2


```
library("DESeq2")
ddsSE <- DESeqDataSet(se, design = ~ cell + dex)
ddsSE
```

```
ddsHTSeq <- DESeqDataSetFromHTSeqCount(sampleTable = sampleTable,
```

```
smallestGroupSize <- 3
keep <- rowSums(counts(dds) >= 10) >= smallestGroupSize
dds <- dds[keep,]
```

Taken from the Vignette of DESeq2

1 - Raw counts

- ★ Remove lowly expressed genes
- ★ Take as reference the smallest groups of your data set, i.e., if you have 2 treated and 4 controls, it should be 2

```
smallestGroupSize <- 3
keep <- rowSums(counts(dds) >= 10) >= smallestGroupSize
dds <- dds[keep,]
```

Example of DESeq object

```
## class: DESeqDataSet
## dim: 14599 7
## metadata(1): version
## assays(1): counts
## rownames(14599): FBgn0000003 FBgn0000008 ... FBgn0261574 FBgn0261575
## rowData names(0):
## colnames(7): treated1 treated2 ... untreated3 untreated4
## colData names(2): condition type
```

Taken from the Vignette of DESeq2

2 - Normalization, 3 - models and 4 - DGEA

```
dds <- DESeq(dds)
res <- results(dds)
res
```

- ★ DESeq performs the normalization, dispersion estimates and the DGEA in one step!
- ★ It's important to set what group should be taken as "Reference" `dds$condition <- relevel(dds$condition, ref = "untreated")`
- ★ If more than two groups, you can select which comparison to be made in two ways:

```
res <- results(dds, name="condition_treated_vs_untreated")
res <- results(dds, contrast=c("condition","treated","untreated"))
```

- ★ Calling "resultsNames" can show all the comparisons

Taken from the Vignette of DESeq2

2 - Normalization, 3 - models and 4 - DGEA

★ Here is an example of the output from calling the "results" command

```
## log2 fold change (MLE): condition treated vs untreated
## Wald test p-value: condition treated vs untreated
## DataFrame with 8148 rows and 6 columns
##               baseMean log2FoldChange     lfcSE       stat    pvalue      padj
##              <numeric>      <numeric> <numeric>  <numeric> <numeric> <numeric>
## FBgn0000008   95.28865     0.00399148  0.225010  0.0177391 0.9858470  0.996699
## FBgn0000017 4359.09632    -0.23842494  0.127094 -1.8759764 0.0606585  0.289604
## FBgn0000018  419.06811    -0.10185506  0.146568 -0.6949338 0.4870968  0.822681
## FBgn0000024    6.41105     0.21429657  0.691557  0.3098756 0.7566555  0.939146
## FBgn0000032  990.79225    -0.08896298  0.146253 -0.6082822 0.5430003  0.848881
## ...               ...            ...       ...        ...       ...       ...
## FBgn0261564   1160.028     -0.0857255  0.108354 -0.7911643 0.4288481  0.789246
## FBgn0261565    620.388     -0.2943294  0.140496 -2.0949303 0.0361772  0.206423
## FBgn0261570   3212.969      0.2971841  0.126742  2.3447877 0.0190379  0.133380
## FBgn0261573   2243.936      0.0146611  0.111365  0.1316493 0.8952617  0.977565
## FBgn0261574   4863.807      0.0179729  0.194137  0.0925784 0.9262385  0.986726
```

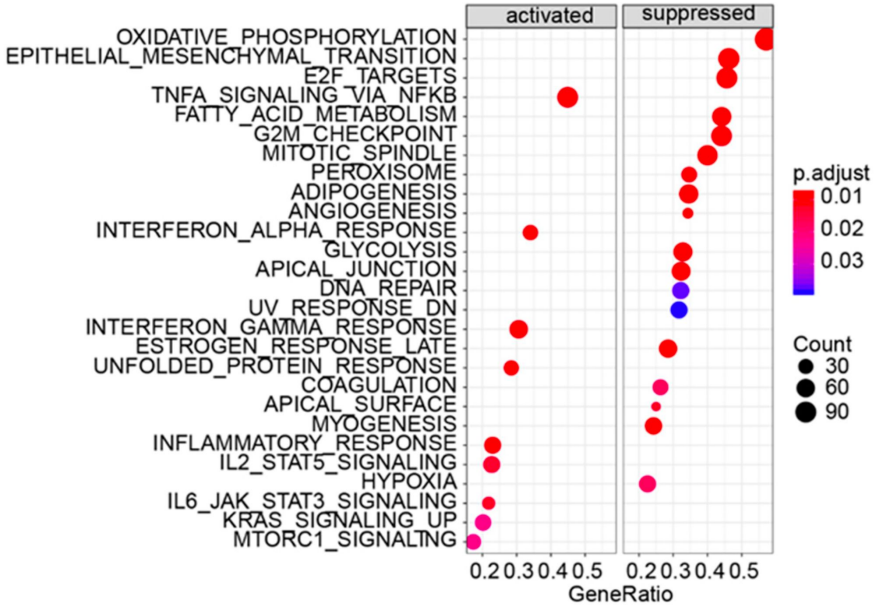
★ log2FoldChange is calculated as treated/untreated

★ The padj is the p-value after correction for FDR

Taken from the Vignette of DESeq2

5 - Downstream analysis

- ★ The idea is to identify special properties of the regulated genes that scape human eyes
- ★ Popular analysis may include GO, GSEA or KEGG
  - They require the fold changes in expression alongside with the a gene identifier
  - They also require annotation of the organism
  - Popular software packages are clusterProfiler



Taken from
https://doi.org/10.3390/biomedicines11020263