ITX Integrative Transcriptomics
Dr. K. Nieselt, Faculty of Science

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

EyeTN workshop **March 30 - April 1 2015**

Differential Expression

During the following practical sessions you will learn how you can find differentially expressed genes within your RNA-seq data.

1. **Finding differentially expressed genes with `DESeq`**

   After counting the reads for the different transcripts (e.g. using `htseq-count`) we are going to use `R` and the package `DESeq` for the differential expression analysis.

R and DESeq

`R` is a free software environment for statistical computing and graphics. It can be downloaded at `http://www.r-project.org`. If you want to work with `R` we also recommend `RStudio` (`http://www.rstudio.com/`), which is a graphical user interface for `R`. You can then enter `R` either through the terminal (by typing R) or simply by opening `RStudio`.
Within `R` you will need the package `DESeq` that can be easily installed with the following commands:

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

All the following steps are performed in `R`.

a) Loading the data

- Make sure that the library for `DESeq` is loaded: `library("DESeq")`
- Load the table that contains the count data
  `countTable = read.table(file = "combinedTable.tsv", header = T, row.names = 1)`
- specify the conditions for the experiments (ordering of the conditions has to match with the experiment columns in your data)
  `condition = factor(c("0","0","2","2",...))`

b) Building the CountDataSet

After we loaded the data table and specified the conditions we can construct the central data structure:

```
cds = newCountDataSet(countTable, condition)
```

c) Normalisation

In order to be able to compare the different experiments, we need to estimate the effective library size. If the counts of non-differentially expressed genes in one sample are, on average,

twice as high as in an other (because the library was sequenced twice as deeply), the size factor for the first sample should be twice that of the other sample. This procedure is considered as the normalisation and can be done by the following command:

```
cds = estimateSizeFactors(cds)
```

d) Standard comparison between two experimental conditions

Next we test whether there is differential expression between conditions "0" and "2" (i.e. time point 0h and 2h) for example. For this we simply call the function `nbionmTest`.

```
res = nbinomTest( cds, "0", "2" )
```

Take a short look at your results with

```
head(res)
```

Now we want to filter for significant genes, according to some chosen threshold for the false discovery rate (FDR)

```
resSig = res[ res$padj < 0.05, ]
```

and list the resulting genes:

```
head( resSig[ order(resSig$pval), ] )
```

2. Finding differentially expressed genes with `Mayday`

For the differential expression analysis with `Mayday` we will give a live demo. Here, our statistics are based on the expression matrix from `Mayday SeaSight`.

Further information and tutorials about `Mayday` can be obtained from our website:
`http://it.inf.uni-tuebingen.de/?page_id=248`.

After the live demo, we will compute differentially expressed genes using the t-test and filter with FDR.

3. Comparison of differentially expressed genes found using `Mayday` with those found using DESeq

After we have computed the differentially expressed genes between time point 0h and 2h using the t-test and FDR ($p < 0.05$), we will compare the results from 1d. and 2.