

---

## Exercises RNAseq Tutorial

### Download the files:

[https://www.dropbox.com/s/8soma231fskdku/RNAseq\\_Mus.txt](https://www.dropbox.com/s/8soma231fskdku/RNAseq_Mus.txt)

### Work locally on your laptop

The following instructions have been tested on Ubuntu. Mac and Windows users can use Ubuntu in the Virtual Machine (setup distributed by Stefan Wyder)

1. Copy the count table (.txt) file to your computer:  

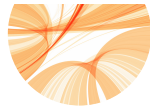
```
wget https://www.dropbox.com/s/8soma231fskdku/RNAseq_Mus.txt
```
2. All RNAseq analysis in this exercises are done within R. Therefore make sure that the newest R is installed: add deb `http://stat.ethz.ch/CRAN/bin/linux/ubuntu precise/` in your `/etc/apt/source.list` file:  

```
sudo add-apt-repository 'deb
http://stat.ethz.ch/CRAN/bin/linux/ubuntu precise/'
sudo apt-get update
sudo apt-get install r-base
```
3. Make sure that following R packages are installed: DESeq, biomaRt, goseq, GenomicFeatures, GO.db. These are all bioconductor packages.  
First the XML and curl package need to be installed:  

```
sudo apt-get install libxml2-dev
sudo apt-get install libcurl4-gnutls-dev
```

Now open R as administrator (`sudo R`) and download packages:

```
source("http://bioconductor.org/biocLite.R")
biocLite("DESeq")
biocLite("biomaRt")
biocLite("goseq")
biocLite("GenomicFeatures")
biocLite("GO.db")
biocLite("org.Mm.eg.db")
```




## Exercise 1: Experiment Setup and differential gene expression analysis

Today we will work on an RNAseq data set from a laboratory strain of mouse (*Mus musculus*). The normalized cDNA of three different life stages (newborn, juvenile and adult) and both sexes were obtained by paired-end Illumina sequencing. Three individuals and several tissues (kidney, liver, heart, muscle and brain) were pooled for each “life stage-sex- category”. The resulting reads were mapped against the reference cDNA sequences of mouse (ENSEMBL) and read counts were obtained per gene.

The provided read-count table contains the raw read counts of the 6 different “life stage-sex-categories” (N: newborn, J: juvenile, A: adult, f: female, m: male) for 31,656 genes (83 % of all ENSEMBL annotated genes).

	<i>Mus musculus</i>	
Newborn		
Juvenile	♀	♂
Adult		



In this first exercise we aim to find the genes showing differential expression between newborns and adults. For both life-stages we sequenced a male and a female, which we will use as biological replicates.

1. Read in read-count table with genes in rows and counts for each sample in columns (first row contains column names, gene names are given in the first column and used as row names):

```
countTable <- read.delim("RNAseq_Mus.txt", header=TRUE,
row.names=1)
```

Have a look at the first few rows of the table:

```
head(countTable)
```

2. Define conditions the samples come from (samples forming biological replicates are assigned to the same condition): vector with values corresponding to sample order in read-count table

```
group <- factor(c("N", "N", "J", "J", "A", "A"))
```

3. Load the DESeq library and set up the count data set:

```
library(DESeq)
cds <- newCountDataSet(countTable, group)
```



4. Normalize the count data set by estimating the size factor.

```
cds <- estimateSizeFactors(cds)
sizeFactors(cds)
```

If each column of the count table is divided by the size factor for this column, the count values are brought to a common scale.

5. Estimate variance (dispersion) of gene expression between biological replicates. If the gene expression differs between replicates by 20%, then the gene's dispersion is  $0.2^2 = 0.04$ .

```
cds <- estimateDispersions(cds)
```

plot the estimated variance (red line= fitted dispersion)

```
plotDispEsts(cds)
```

6. Get differentially expressed genes between newborns and adults:

```
res <- nbinomTest(cds, "N", "A")
head(res)
```

Columns:

id = gene ID

baseMean = mean normalized counts averaged over samples from both conditions

baseMeanA = mean normalized counts from "N"

baseMeanB = mean normalized counts from "A"

foldChange = fold change from "N" to "A"

log2FoldChange = logarithm with basis 2 of the fold change

pval = p value for the statistical significance of this change

padj = p value adjusted for multiple testing with Benjamini-Hochberg procedure (FDR)

7. Plot results with all genes differentially expressed (red: FDR < 5% in red):

```
plotMA(res, col=ifelse(res$padj>=0.05, "gray32", "red3"))
```

histogram of p-value/FDR distributions:

```
hist(res$pval, breaks=100, col="skyblue", main="")
```

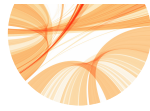
```
hist(res$padj, breaks=100, col="skyblue", main="")
```

8. Order results according strength of differential expression:

```
res.ordered <- res[order(res$padj), ]
```

9. Get list of over-expressed genes in newborns/adulsts at FDR 5%:

```
over.N <- res.ordered[res.ordered$padj<0.05 &
!is.na(res.ordered$padj) & res.ordered$log2FoldChange<0, ]
```



```
nrow(over.N)
over.A <- res.ordered[res.ordered$padj<0.05 &
!is.na(res.ordered$padj) & res.ordered$log2FoldChange>0, ]
nrow(over.A)
```

10. Write differential expression table in a text file:

```
write.table(res.ordered, "diffExp_N-A.txt", sep="\t",
row.names=FALSE)
```



How many genes are differentially expressed at a FDR of 1%?

## Exercise 2: Get additional information for differentially expressed genes

In this exercise we will use the biomaRt library to get additional information about the genes overexpressed in newborns.

1. Load biomaRt library and setup database to use

```
library(biomaRt)
mart <- useDataset("mmusculus_gene_ensembl", useMart("ensembl"))
```

2. Get vector of gene names overexpressed in newborns:

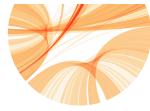
```
genes.over.N <- over.N$id
```

3. Get additional gene information from biomart: gene name, chromosomal location, gene description

```
ensembl_translation <- getBM(filters= "ensembl_gene_id",
attributes= c("ensembl_gene_id", "external_gene_id",
"chromosome_name", "description"), values=genes.over.N,
mart=mart)
```

## Exercise3: search for enriched GO terms:

In this exercise we will use goSeq to test if overexpressed genes of newborns are enriched for any GO terms.



1. Prepare goSeq database (get genes and their length). This will take some time.:

```
library(GenomicFeatures)
txdb <-
makeTranscriptDbFromBiomart(dataset="mmusculus_gene_ensembl")
txsByGene <- transcriptsBy(txdb,"gene")
lengthData <- median(width(txsByGene))
```
2. Load goSeq library and convert differential expressed genes to a goSeq vector (genes overexpressed in newborn =1, genes not overexpressed in newborns = 0)

```
library(goseq)
genes.N <- as.integer(res.ordered$padj<=0.05 &
!is.na(res.ordered$padj) & res.ordered$log2FoldChange<0)
names(genes.N)=res.ordered$id
head(genes.N)
```
3. Quantify the length bias present in the dataset with a probability weighting function (function which gives the probability that a gene will be differentially expressed based on its length alone)

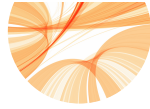
```
lengthData.N <- lengthData[names(genes.N)]
pwf.N <- nullp(genes.N, bias.data=lengthData.N)
```
4. Get enriched GO terms. Here we limit to the GO term category Biological processes (to get all categories remove the argument "test.cats"):

```
GO.N.BP <- goseq(pwf.N,"mm10","ensGene", test.cats=c("GO:BP"))
```
5. Correct for multiple testing by the Benjamini and Hochberg correction (FDR):

```
GO.N.BP$FDR_over_represented <-
p.adjust(GO.N.BP$over_represented_pvalue, method="BH")
```
6. Filter for GO terms enriched at a FDR of 5% and get a description for those GO terms using the GO.db library

```
enriched.GO.N.BP = GO.N.BP[GO.N.BP$FDR_over_represented <= 0.05,]
nrow(enriched.GO.N.BP)
head(enriched.GO.N.BP)

library(GO.db)
enriched.GO.N.BP$Description <- NA
index <- 1
for(go in enriched.GO.N.BP$category){
  enriched.GO.N.BP$Description[index] <- Term(GOTERM[[go]])
  index <- index + 1
}
head(enriched.GO.N.BP)
```



What GO terms are enriched in genes overexpressed in newborns?



And what GO terms are enriched in genes overexpressed in adults?

## Exercise 4: differential expression between females and males



What genes are overall differentially expressed between females and males (hint: samples from different life stages can be used as biological replicates)?



Are there also some GO terms enriched in genes overexpressed in females/males?