



# Statistical Analysis of Expression Data

Practical for FOS course Molecular Data Science: from disease mechanisms to personalized medicine, Tuesday October 24, 2019

*Rodrigo Coutinho de Almeida (Molecular Epidemiology)*

In this practical you will analyze RNA-seq data to identify differentially expressed genes (DEGs) in arthritic cartilage. For this purpose, gene expression profiles of articular cartilage samples were generated with Illumina HiSeq2000 and HiSeq4000. Cartilage samples were collected from patients undergoing total joint replacement surgery due to primary hip or knee osteoarthritis (OA), both, from macroscopically unaffected cartilage regions (**Preserved**) and from affected cartilage regions (**Lesioned**).

Analyses will be run in **R**.

At the end of the practical send your work to `r.coutinho_de_almeida@lumc.nl` as Word document with your name(s) on top of the first page. The file name should contain your surname(s) and date.

First we need to load the appropriate analysis package. For differential gene expression analysis and normalization we will use the **DESeq2** package. Start **R** as administrator and copy/paste the following commands to make sure both packages are installed and loaded:

```
# Load Libraries
> library(DESeq2)
> library(biomaRt)
> library(pheatmap)
> library(ggplot2)
> library(enrichR)
```

Next, we need to load the actual data:

```
> load(url("https://raw.githubusercontent.com/molepi/Molecular-Data-Science/master/ExpressionData/RNASeq_PairedOA_FOS.Rdata")) # Load the expression data
```

This will load the expression matrix of the OA paired samples (PairsOA) and the phenotype data (pheno). Inspect the loaded data by running the following commands:

```
> head(pheno) # show phenotype data
```

## Question 1.

- What information is provided?
- How many samples are in the data set? How many unique individuals?
- How many lesioned and how many preserved OA?
- How many lesioned and how many preserved OA are in batches?
- How many Atrophic and how many Normotrophic patients are in the dataset?

## Question 2.

The next steps consist of data and normalization.

- Plot the non-normalized with the commands bellow and explain why is important to normalize your RNA-seq data?

```
> boxplot(PairsOA_all)
```

```
> hist(as.matrix(PairsOA_all))
```

- b) Now normalize the data and plot. Compare the results with the previous plot and explain what happened to your dataset.

```
> ddsGenes <- DESeqDataSetFromMatrix(countData=PairsOA_all,
                                     colData=pheno, design = ~1)

rldGenes <- rlogTransformation(ddsGenes, fitType='local', blind = F)
MatrldGenes <- assay(rldGenes)

hist(MatrldGenes, main = "Distribution of genes in OA cartilage", col = "salmon",
     xlab = "Log2 Normalized DESEQ2", prob = F)

hist(MatrldGenes, main = "Distribution of genes in OA cartilage", col = "salmon",
     xlab = "Log2 Normalized DESEQ2", prob = T, ylim = c(0,0.45))
lines(density(rldGenes),lwd = 2, col = "blue")

boxplot(MatrldGenes) #plot normalize data
```

- c) Inspect your data using Principle Component Analysis (PCA), run the following command and report what you see.

```
> plotPCA(rldGenes, intgroup=c("Batch"))
```

- d) In case you identified a batch effect, run the following command and plot. Explain what happened to your data. Did you remove all the batch effects?

```
> NoBatch <- limma::removeBatchEffect(assay(rldGenes), rldGenes$Batch)

> pc <- prcomp(t(NoBatch))

percenVar <- pc$sdev^2/sum(pc$sdev^2)

plotData <- data.frame(pheno, pc$x[,1:3])

p <- ggplot(plotData, aes(x = PC1, y = PC2, colour = Batch))

p + geom_point(size = 3) +
  scale_color_manual(values = c("salmon", "blue"))+
  xlab(paste0("PC1: ", round(percenVar[1] * 100), "% variance")) +
  ylab(paste0("PC2: ", round(percenVar[2] * 100), "% variance"))
```

Now we are ready to perform the actual differential expression analysis. We're particularly interested in generic gene expression differences between pairs of preserved and OA lesioned cartilage. To analyse this we will perform a pairwise test analysis.

Which is done by the following commands:

```

> ddsGenes <- DESeqDataSetFromMatrix(countData=PairsOA_all,
                                     colData=pheno,
                                     design = ~Patient + Status) #Here you build the
model take the pairs of patients into account
ddsGenes <- DESeq(ddsGenes)

resGenes <- results(ddsGenes)
resGenes <- results(ddsGenes, contrast=c("Status", "Lesion", "Preserved")) #Fold
change will be set for Lesion
resGenes <- as.data.frame(resGenes)
resGenes <- resGenes[complete.cases(resGenes),] #take out NA
resGenes <- resGenes[order(resGenes$padj),] #Order according with the P-values
(FDR)
resGenes$FC <- 2^resGenes$log2FoldChange #Include an extra column with the fold
changes

#Subset the signifcant genes
sigGen <- resGenes[resGenes$padj < 0.05,]

#Change Transcripts ID to Genes Symbols
ensembl=useMart("ensembl")
ensembl = useDataset("hsapiens_gene_ensembl",mart=ensembl)
attributes <- listAttributes(ensembl)
transcriptsids= as.character(rownames(sigGen))

geneIDs <- getBM(attributes = c('ensembl_gene_id', 'external_gene_name'),
                 filters = 'ensembl_gene_id',
                 values = transcriptsids,
                 mart = ensembl)

#Merge and preprocess the data
SignGeneNames <- merge(sigGen,geneIDs,by.x = "row.names",by.y = ,"ensembl_gene_id")
colnames(SignGeneNames)
SignGeneNames <- SignGeneNames[,c(1,9,8,3,6,7)]
colnames(SignGeneNames)[1:2] <- c("Emsembl_ID", "GeneNames")
SignGeneNames <- SignGeneNames[order(SignGeneNames$padj),]

```

**Question 3.** Inspect the output:

- Do you think the batch effect was removed in the differential expression analysis?  
*tip:* Remember that you are doing a pairwise analysis
- Run `head(resGenes, n)` to inspect the top  $n$  genes in your differential expression analysis. How many genes after correction and how many genes before the adjustment you observed?
- What is the range of differential expression?  
*R tips:* `> range(resGenes$FC)`
- Which gene is the most downregulated? Which one is the most upregulated gene? Are they significant after correction for multiple test?  
*R tips:* `> result[which(resGenes$FC %in% range(resGenes$FC)), ]`
- Do you think these changes are biologically relevant?
- How many genes remain if you only consider those with a Fold-change of at least 2?
- Are the top differentially expressed genes known to be involved in OA? If so, does the direction of differential expression support the reported role in OA?
- How about genes not known to be involved in OA, does their in- or decrease in expression make sense in relation to OA?

#### Question 4.

To add biological context to your results you can perform a pathway enrichment analysis. Using the following command you will check how your differentially expressed genes are enriched in one branch of the Biological Process from the Gene Ontology database (GO).

```
dbs <- c("GO_Biological_Process_2017") #Subset a database for the exercise

enriched <- enrichr(SignGeneNames$GeneNames, dbs) #Perform the enrichment analysis

head(enriched[["GO_Biological_Process_2017"]]) #Check the results

#Here you will preprocess the results to visualize
GOBioProc <- as.data.frame(enriched$GO_Biological_Process_2017)
GOBioProc$log.Adjpvalue <- -log10(as.numeric(GOBioProc$Adjusted.P.value))
GOBioProc$log.pvalue <- -log10(as.numeric(GOBioProc$P.value))
GOBioProc$Term <- factor(GOBioProc$Term, levels =
GOBioProc$Term[order(GOBioProc$P.value, decreasing = T)])

#Subset only pathways significant after correction for multiple testing
GOBioProc_FDR <- GOBioProc[GOBioProc$Adjusted.P.value < 0.05,]

#Plot the top 10 pathways
ggplot(data=GOBioProc_FDR[1:10,], aes(x=Term, y=log.pvalue)) +
  geom_bar(stat="identity", position="identity", width = 0.5) + coord_flip() +
  theme(axis.text.x = element_text(size = 10, angle = 45,hjust = 1, vjust = 1)) +
  labs( x="GO Biological process", y="-log10(Adj P-value)")
```

- a) How many pathways are significant after correction for multiple testing?
- b) Which pathway is the most significant? Is this a known pathway in osteoarthritis?
- c) How many genes is overlapping in the most significant pathway?
- d) How many genes are not significant after correction for multiple test?

*R tips:* you can use this code to subset:

```
GOBioProc[GOBioProc$P.value < 0.05 & GOBioProc$Adjusted.P.value > 0.05,]
```