VI. Bioinformatics in R (solution)

Center for Health Data Science, University of Copenhagen

01 September, 2021

In this exercise you will perform differential expression analysis using the `DESeq2` package for R on an RNAseq expression dataset from airway smooth muscle cells. You will effectively be performing the same work flow of analysis as the one provided in **presentation 6**, with some variation along the way.

About the Dataset
Common medicine for treatment of astma are beta2-agonists and glucocorticosteroids, which mainly target the airway smooth muscle.
The dataset `airway` contains mRNA profiles from bulk RNAseq of smooth muscle cells from four male donors. The cell lines were treated with dexamethasone and albuterol, or were left untreated (controls). See original data and summary of intent here.

PART 1 - Preliminary Analysis

Install and load packages

For this exercise you will need five R-packages: `tidyverse`, `ggplot2`, `readxl`, `BiocManager`, `DESeq2` and `EDASeq`. You should already have these installed!
Copy the chunk below to your .R or .Rmd file to load packages.
N.B If you do get an error when trying to load one of the packages below, you can try to install the missing package using `install.packages('my.package')`.

```
#install.packages("tidyverse")
#install.packages("ggplot2")
#install.packages("readxl")
#install.packages("BiocManager")
#BiocManager::install("DESeq2")
#BiocManager::install("EDASeq")

library(tidyverse)
library(ggplot2)
library(readxl)
library(BiocManager)
library(DESeq2)
library(EDASeq)
```

---

1. In the exercises folder you will find two files: `airway_counts.xlsx` (RNAseq count data) and `airway_metadata.xlsx` (sample information).
   Read in these two files and name them ***airDat*** and ***airMet***, respectively.
   N.B. If you are not working from within the exercise folder, you need to remember to either (I) set the full/true path of the files or (II) copy these files to your current working directory.

```
airDat <- read_xlsx("airway_counts.xlsx")
#airDat

airMet <- read_xlsx("airway_metadata.xlsx")
#airMet
```

---

2. What kind of information do you have in the first four columns of the airDat? How many samples and genes are there in your count data?

```
# Gene information
airDat %>%
  dplyr::select(1:4) %>%
  head(., n=3)
```

```
## # A tibble: 3 x 4
##   Ensgene         GeneSymbol    GC Length
##   <chr>           <chr>      <dbl>  <dbl>
## 1 ENSG00000000003 TSPAN6      40.4  1564.
## 2 ENSG00000000005 TNMD        40.8  1353
## 3 ENSG00000000419 DPM1        40.2  1080.
```

```
# Number of genes
nrow(airDat)
```

```
## [1] 29264
```

```
# Number of samples
ncol(airDat)-4
```

```
## [1] 8
```

3. In your metadata you have four variables, all characters, convert *condition* and *celltype* into factors for further analysis.

```
airMet <- airMet %>%
  mutate(condition=as.factor(condition), celltype=as.factor(celltype))
```

Copy and run the code below. This line of code will give you a new column in your **airDat** tibble named *nzeros* (sum of 0 across samples for each gene).
First we temporarily de-select the columns with gene information, then we count 0s across samples.

```
airDat <- airDat %>%
  mutate(nzeros = rowSums(dplyr::select(., -Ensgene, -GeneSymbol, -GC, -Length)==0))
```

4. We do not want any genes where less than half of the samples have a count above 0. Filter out the genes for which this is the case, and remove the column *nzeros* from your tibble after filtering. Check how many genes you are left with.

```
airDat <- airDat %>%
  filter(nzeros <= 4) %>%
  dplyr::select(-nzeros)

dim(airDat)
```

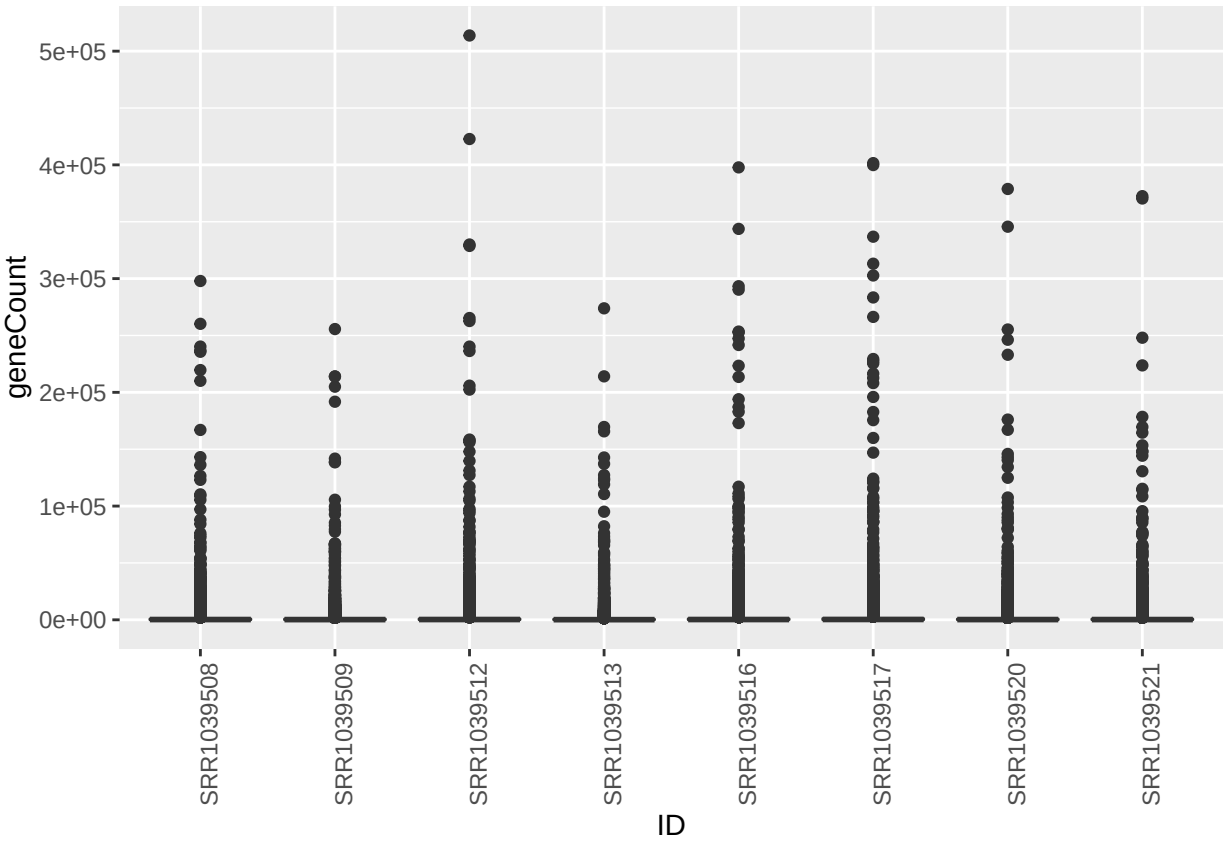
```
## [1] 17468    12
```

5. To get en idea of sample library sizes make a boxplot of gene counts, one for each sample. You will need to `gather()` the gene counts across samples into one column and the sample IDs into another column. Assign this output to a new variable named `airPlot`.
   HINT 1: First remove the columns with information on genes (there are four) and next use this: `gather(key = ID, value=geneCount)` to gather the counts into one column.
   HINT 2: Use `geom_boxplot()` to make a boxplot with ggplot2. Extra: To tilt your x-axis labels 90 degrees, add `theme(axis.text.x = element_text(angle = 90))` to your ggplot code.

```
airPlot <- airDat %>%
  dplyr::select(-Ensgene, -GeneSymbol, -GC, -Length) %>%
  gather(key = ID, value=geneCount)


airPlot
```

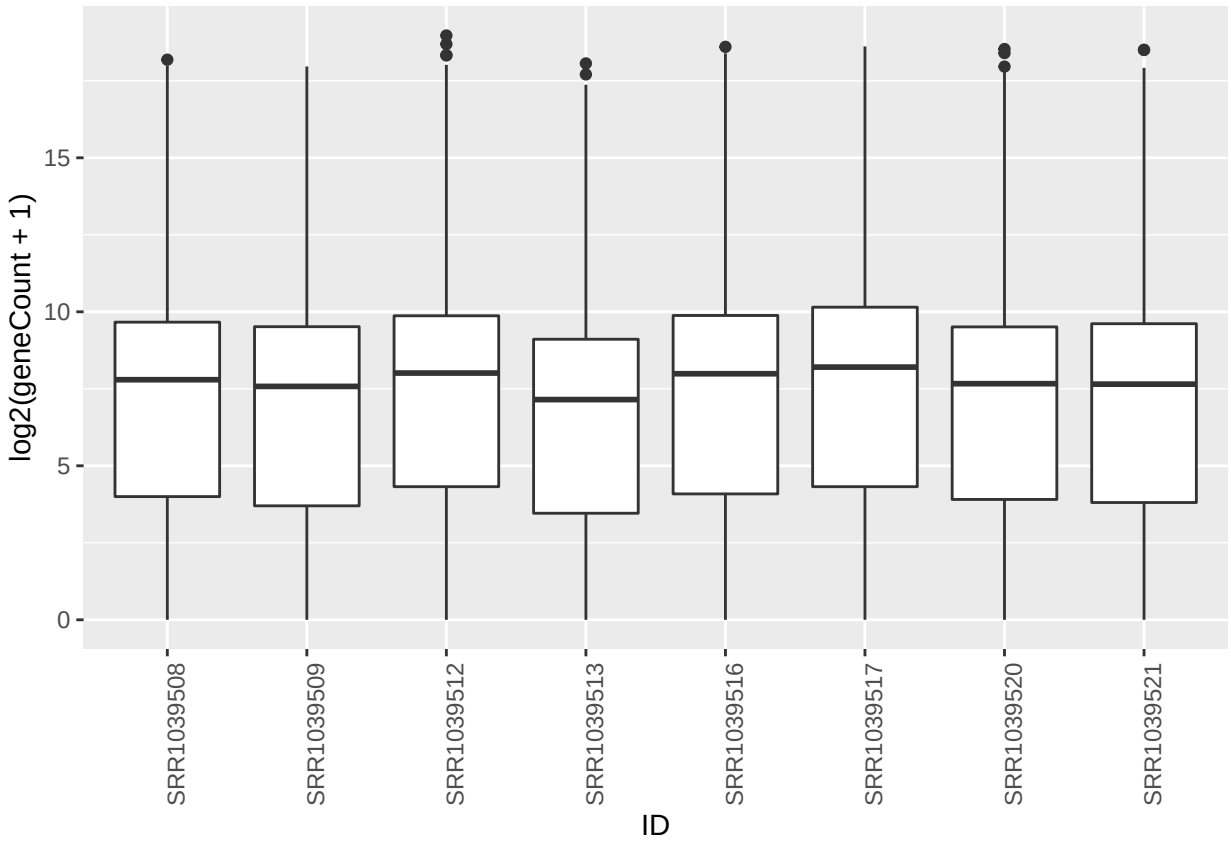
```
## # A tibble: 139,744 x 2
##    ID         geneCount
##    <chr>          <dbl>
##  1 SRR1039508       679
##  2 SRR1039508       467
##  3 SRR1039508       260
##  4 SRR1039508        60
##  5 SRR1039508      3251
##  6 SRR1039508      1433
##  7 SRR1039508       519
##  8 SRR1039508       394
##  9 SRR1039508       172
## 10 SRR1039508      2112
## # ... with 139,734 more rows
```

```
p1 <- ggplot(airPlot, aes(ID, geneCount)) + geom_boxplot() + theme(axis.text.x = element_text(angle = 90

p1
```



6. You will see that boxplots are squeezed due to difference in count range between genes. Which type of data transformation could you use to overcome this problem?
   Remake the boxplot with transformed counts.
   HINT remember to add 1.0 pseudo count to all counts before transformation to handle counts which are 0.

```
p2 <- ggplot(airPlot, aes(ID, log2(geneCount+1))) + geom_boxplot() + theme(axis.text.x = element_text(a

p2
```



Adjusting for GC content and gene length:

It is well-established that gene length and GC-content will affect the number of reads obtained. It is therefore standard practice to check for such biases and correct accordingly. You already have information on length and GC content in the **airDat** tibble.

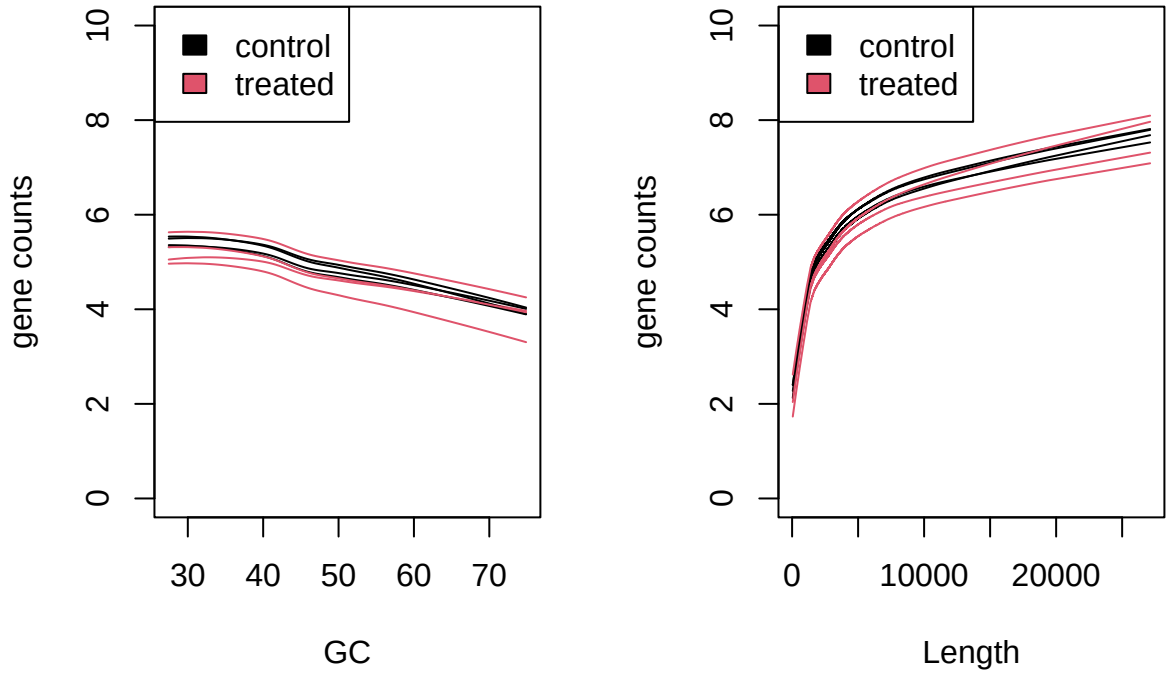
The `EDAseq` package can help you correct for gene length and GC content. `EDAseq` provides a `biasPlot` showing you if GC content and gene length are a problem. To use this function your data must be in a format that `biasPlot` accepts, e.g. and `EDAseq` object.

7. Copy the code from both chunks below and run it. Have a look at the object you created, named *EDAobj*, and importantly, look at the plots created, does there appear to be a GC and/or gene length bias in your data?
   HINT 1: An unbias dataset should not show any trend in GC content or gene length across gene counts.

```
# Converts tibble to matrix of expression counts and removes gene information:
airDatM <- airDat %>%
  dplyr::select(-Ensgene, -GeneSymbol, -GC, -Length) %>%
  as.matrix() %>%
  unname()

# Make EDASeq object
EDAobj <- newSeqExpressionSet(airDatM,
                              phenoData=AnnotatedDataFrame(data.frame(condition=airMet$condition,
                                                                      celltype=airMet$celltype)),
                              featureData=AnnotatedDataFrame(data.frame(GC=airDat$GC,
```

```
                                                    Length=airDat$Length)))
```

```
# Bias plots
par(mfrow=c(1,2))
biasPlot(EDAobj, "GC", log=TRUE, ylim=c(0,10))
biasPlot(EDAobj, "Length", log=TRUE, ylim=c(0,10))
```

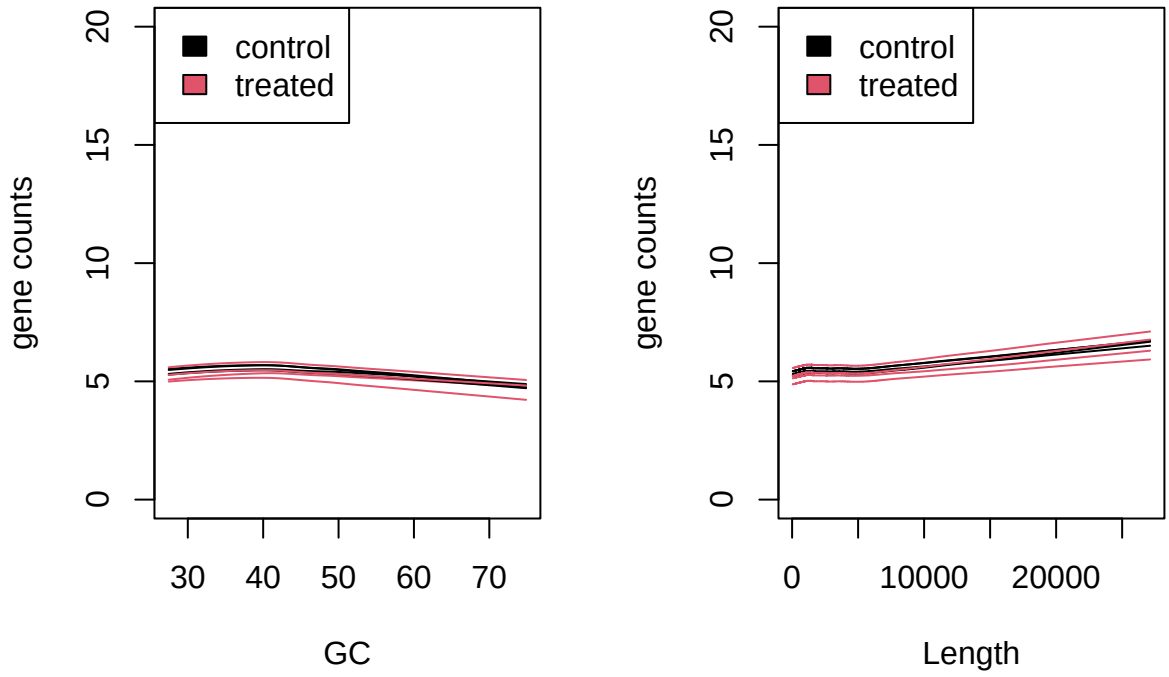


If you observe a bias in your data, you can correct for this with the function `withinLaneNormalization` from `EDASeq`. You can have a look at the EDASeq documentation under point 5 for more info on this function.

8. Copy and run the code in below to normalize your *EDAobj* object for GC content and gene length. Remake the `biasPlot` from above with the *EDAobjNorm* object. Are the biases removed?

```
EDAobjNorm <- withinLaneNormalization(EDAobj, "GC", which="full", offset=TRUE) %>%
  withinLaneNormalization(., "Length", which="full", offset=TRUE)
```

```
par(mfrow=c(1,2))
biasPlot(EDAobjNorm, "GC", log=TRUE, ylim=c(0,20))
biasPlot(EDAobjNorm, "Length", log=TRUE, ylim=c(0,20))
```



PART 2 - Differential Expression Analysis with DESeq2

Now we we begin our differential expression analysis in `DESeq2`. First, we use the function `DESeqDataSetFromMatrix` to make a `DESeq2 object` (just as shown in the presentation).

9. Fill in the DESeq2 object below. We want our design matrix to include *celltype* and *condition* from the metadata, (**airMet**). Fill in the *countData*, *colData* and *design* (rowData is optional).
   HINT Your countData must only contain counts, no gene IDs etc., so you should filter/slice these these columns out before addition the counts it to the object.

```
exprObj <- DESeqDataSetFromMatrix(countData = ,
                                  rowData = ,
                                  colData = ,
                                  design=~)
```

```
exprObj <- DESeqDataSetFromMatrix(countData = airDat[,-c(1:4)],
                                  rowData = airDat[,1:4],
                                  colData = airMet,
                                  design=~celltype+condition)
exprObj
```

```
## class: DESeqDataSet
## dim: 17468 8
## metadata(1): version
## assays(1): counts
## rownames: NULL
## rowData names(4): Ensgene GeneSymbol GC Length
## colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521
## colData names(4): id condition celltype geo_id
```

910. If we are to be correct, you should "feed" the function `DESeqDataSetFromMatrix` the normalized counts

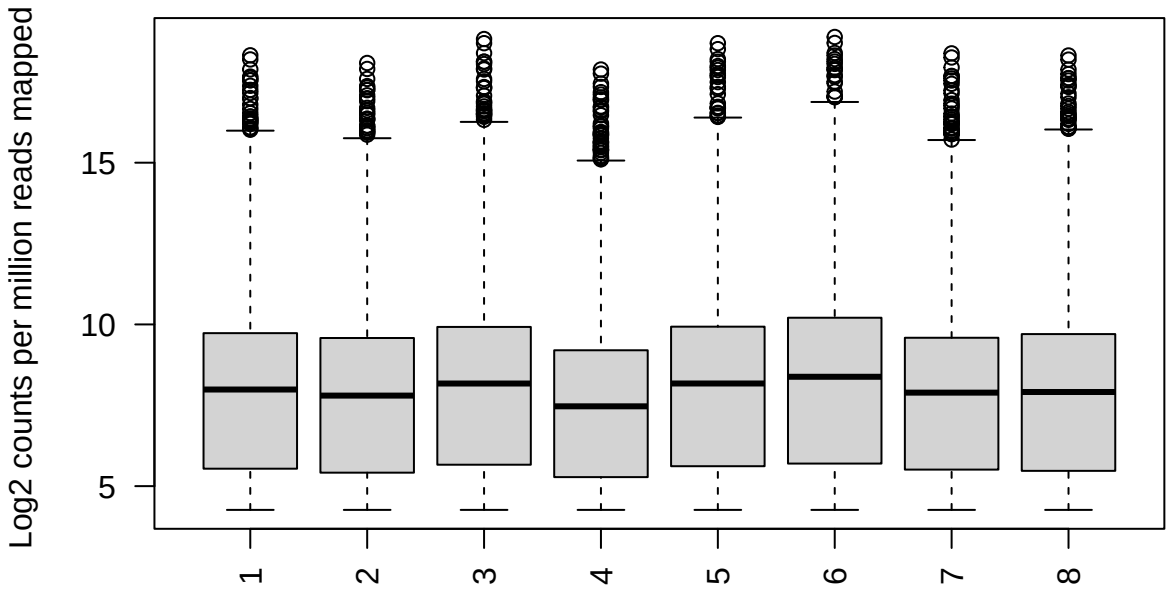
from the EDASeq analysis, where you corrected for GC content and gene Length. This is slightly more tricky, so we have provided you with the code needed for this. Copy and run the code below.

```
exprObj <- DESeqDataSetFromMatrix(countData = counts(EDAobjNorm),
                                  colData = pData(EDAobjNorm),
                                  design = ~celltype+condition)

normFactors <- exp(-1 * offst(EDAobjNorm))
normFactors <- normFactors / exp(rowMeans(log(normFactors)))
normalizationFactors(exprObj) <- normFactors
```

In question 5., we made a ggplot2 boxplots to see the difference in library sizes and variances of samples. Based on this plot it seems like variance stabilizing transformation might improve our dataset for DEA analysis. Copy and run the code below to perform vst transformation.
Extra (optional): Make a boxplot with the vst counts.

```
exprObjvst <- vst(exprObj,blind=FALSE)
```

```
boxplot(assay(exprObjvst), xlab="", ylab="Log2 counts per million reads mapped ",las=2)
```



We would like to use a PCA plot to inspect if the vst transformation has improved the clustering of our samples by group. For this we will first perform principal component analysis using both the 'raw' counts and vst-transformed counts. We extract counts from the DESeq2 objects using the function `assay()`.

11. Copy and run the two chunks of code below.
    In the second chunk we are using three different functions: `t()`, `dist()` and `cmdscale()`. Use `?` to figure out what each of them do! What inputs do they take? and what are the default values?

```
# un-transformed counts
unTrf <- assay(exprObj)
head(unTrf, n=3)
```

```
##     1   2   3   4    5    6   7   8
## 1 679 448 873 408 1138 1047 770 572
## 2 467 515 621 365  587  799 417 508
## 3 260 211 263 164  245  331 233 229
```

```
# vst transformed counts
vstTrf <- assay(exprObjvst)
head(vstTrf, n=3)
```

```
##          1        2        3        4         5         6        7        8
## 1 9.499667 8.954623 9.810949 8.806491 10.177670 10.074366 9.624205 9.284815
## 2 9.001044 9.053105 9.352068 8.633747  9.263922  9.685102 8.849182 9.185863
## 3 8.206549 7.958539 8.232267 7.616493  8.143245  8.531044 8.096834 8.063889
```

```
# un-transformed counts
unTrf <- unTrf %>% t() %>%
    dist() %>% cmdscale(., eig=TRUE, k=2)

unTrf <- tibble(PC1=unTrf$points[,1],PC2=unTrf$points[,2])

# vst transformed counts

vstTrf <- vstTrf %>% t() %>%
    dist() %>% cmdscale(., eig=TRUE, k=2)

vstTrf <- tibble(PC1=vstTrf$points[,1],PC2=vstTrf$points[,2])
```

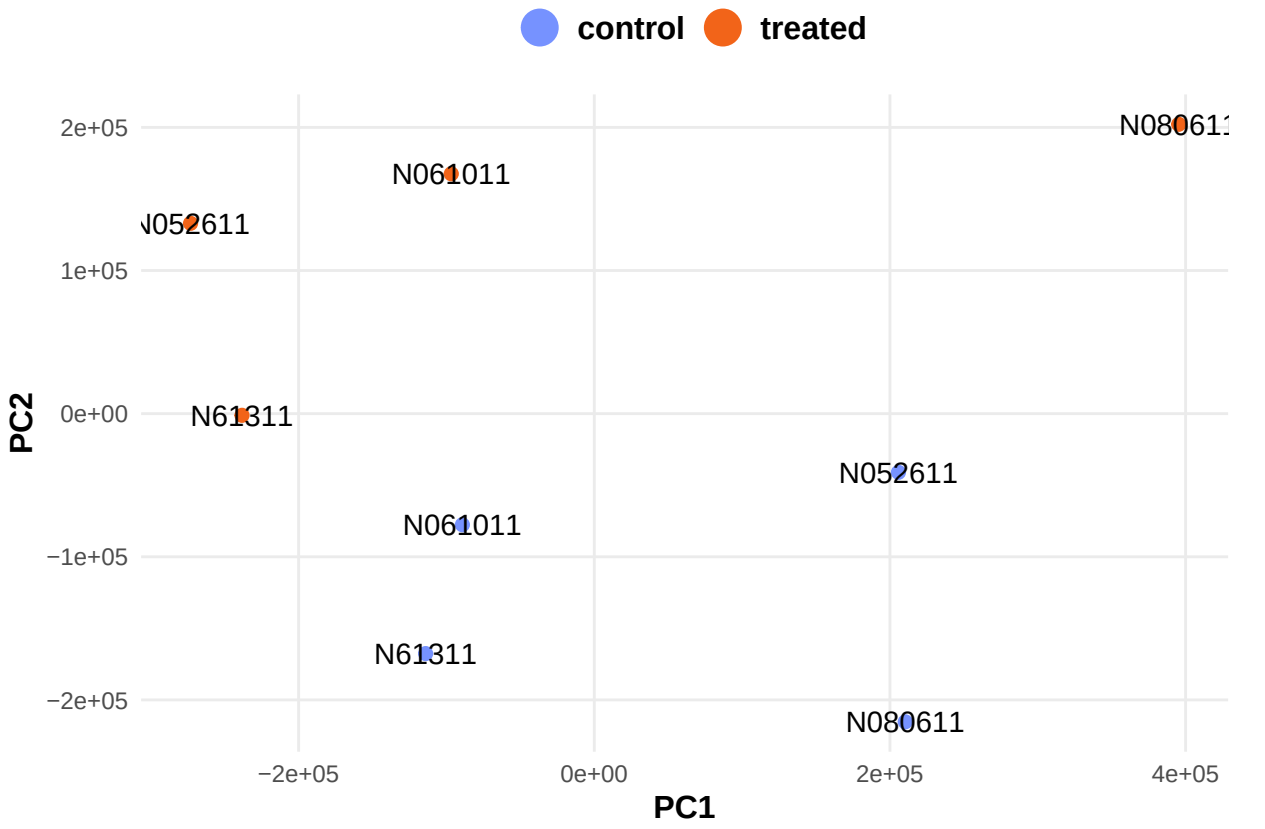
You now have two datasets containing the first two principal components (PC1 & PC2) for both 'raw' (un-transformed) counts, *unTrf*, and vst-transformed counts, *vstTrf*.

12. Make a PCA plot for each set using ggplot2. Color the samples by condition and label them by *celltype* (**airMet** has this information).
    HINT A PCA plot is just a `geom_point()` plot with x=PC1 and y=PC2. Would you say that the transformation improved the partitioning of control and treated samples?

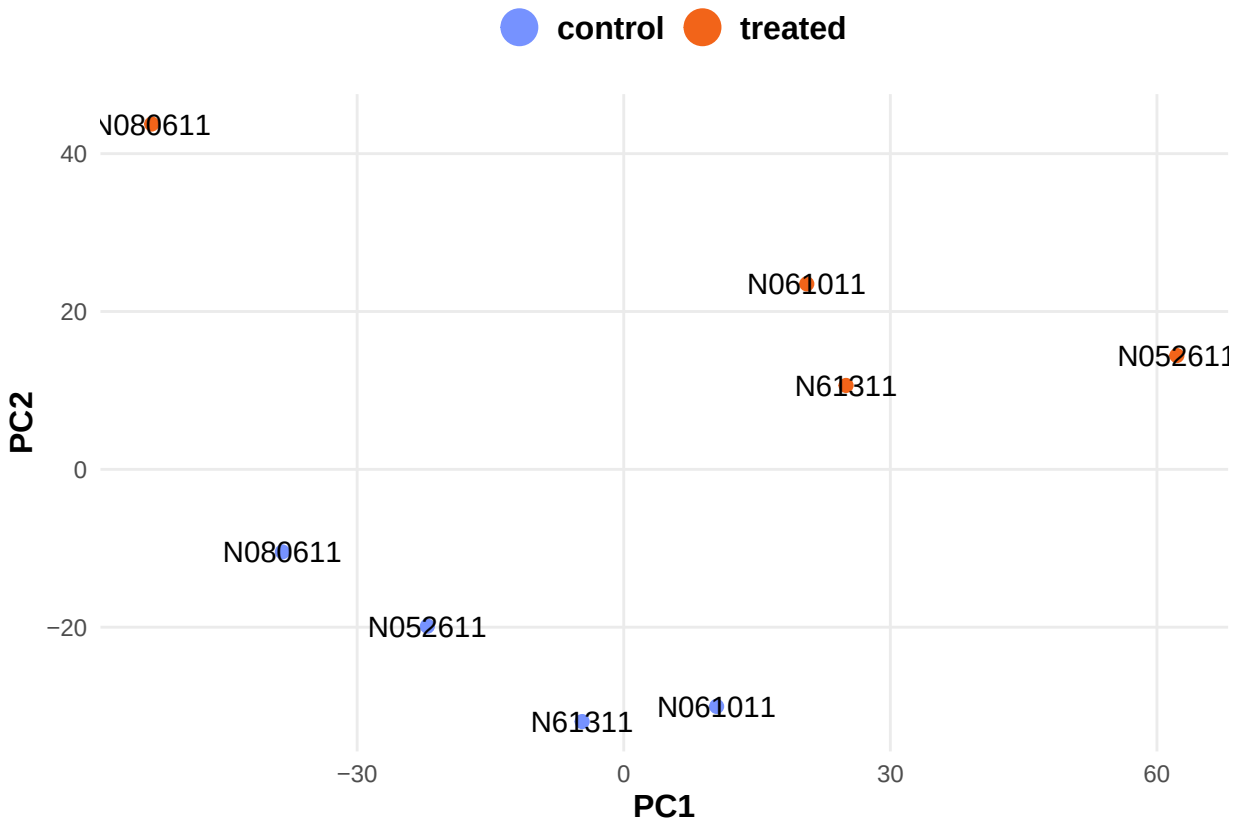
```
# PCA Plot
# Takes as arguments:
# countDat = a dataframe of expression counts
# groupVar = a factor vector of groups (conditions) to color points by (may be the same or different as
# labelVar = a vector of IDs for labeling (may be the same or different as groupVar)
# colVar = a vector of colors (one color for each group)
# ------------------------------------------------------------------------------------------------------


pcaPlot <- function(countDat, groupVar, labelVar, colVar) {

  # Make basic plot
  pcap <- ggplot(data=countDat) + geom_point(aes(x=PC1,y=PC2,color= groupVar), size=2) +
    geom_text(aes(x=PC1,y=PC2, label=labelVar)) +

    # Improve look of plot:
    theme_minimal() + theme(panel.grid.minor = element_blank()) +
    scale_color_manual(values = colVar) +
    theme(legend.title=element_blank(), legend.text = element_text(size = 12, face="bold"), legend.posi
    guides(colour = guide_legend(override.aes = list(size=6)))
  return(pcap)
}
```

```
# Make plot
p3 <- pcaPlot(unTrf, airMet$condition, airMet$celltype, c("#7692FF", "#F26419"))
```

```
p4 <- pcaPlot(vstTrf, airMet$condition, airMet$celltype, c("#7692FF", "#F26419"))

p3
```




```
p4
```



Next, we use `DEseq()` to estimate dispersion, gene-wise and mean-dispersion, fitting model(s). Copy the code below and run it.

```
# Fitting gene-wise glm models:
exprObj <- DESeq(exprObj)
```

We now have our model(s) ready and we want to contrast our condition groups, e.g. treated vs control.

13. Use the `DESeq2` function `results()` to do the post hoc test (just like we did in the presentation). Figure out what arguments it takes. As a minimum you will have to specify a DESeq2 model object (denoted by '.' below) and a contrast of interest. When you have run the function, have a look at the output.

```
resTC <- results(. , contrast = c(), independentFiltering = FALSE)
```

```
# Test for DE genes between lactating mice and control mice adjusted for cell type:
# Lactating vs Control mice
resTC <- results(exprObj, contrast = c("condition", "treated", "control"), independentFiltering = FALSE)
```

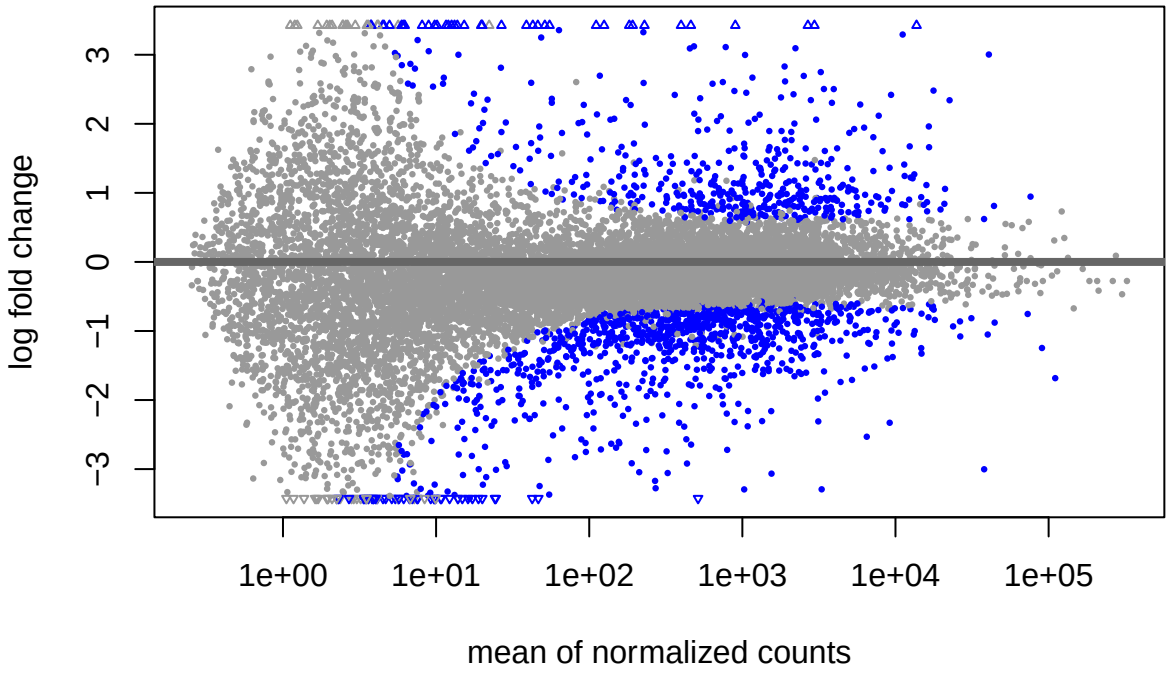
14. Use the function `summary()` to see the number of identified differentially expressed gene. Use the `plotMA()` function to visualize these.

```
# Summary and plot of DE analysis results:
summary(resTC)
```

```
##
## out of 17468 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 493, 2.8%
```

```
## LFC < 0 (down)     : 977, 5.6%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

```
DESeq2::plotMA(resTC)
```



15. Convert your results from `DESeq2` to a tibble and do the following:
    (a) Add two new columns to it (using `mutate`): *dir* indicating directionality of the logFC. The column *dir* can be made using the following syntax: `dir=factor(ifelse(log2FoldChange >= 1.0, "Up", "Down"), levels=c("Up", "Down"))` and one named *geneSymbols* with gene symbols from *airDat*.
    (b) Filter your results to only include genes with log2FoldChange of more than 1 or less than -1 and a *padj* (adjusted p-value) of less than 0.05.
    (c) Arrange by *padj* (ascending) and the **absolute** *log2FoldChange* (descending). **HINT:** use the function `abs()` to get the absolute log2FoldChange before arranging.
    (d) Extract the top 50 most significant DE genes based on *log2FoldChange* and *padj*.

```
resTC <- resTC %>%
  as_tibble() %>%
  mutate(dir=factor(ifelse(log2FoldChange >= 1.0, "Up", "Down"), levels=c("Up", "Down")),
         geneSymbol=airDat$GeneSymbol) %>%
  filter((log2FoldChange > 1.0 | log2FoldChange < -1.0) & padj < 0.05) %>%
  arrange(padj, desc(abs(log2FoldChange))) %>%
  top_n(50)
```

16. Make a bubble plot (fancy point plot) of the top 50 most significant DE genes:
    (a) The size of the point should reflect the **absolute** *log2FoldChange* and the shade of the point should reflect the significance (e.g. the padj).

(b) Remove the x-axis labels (gene symbols) and instead add these to the points themselves with `geom_text()`.

(c) Use `facet_grid(rows = vars(dir))` to wrap the top most up-regulated and down-regulated genes in each their own plot (grid).

(D) Based on your plot, which genes seem to be most effected by treatment with dexamethasone and albuterol?