VI. Bioinformatics in R (presentation)

Center for Health Data Science, University of Copenhagen

21 October, 2021

Bioconductor

Bioconductor provides tools for computational biology and bioinformatics analysis in R - it is open source and open development and it has an active user community.
Mostly when we install R-packages we use `install.packages('name_of_package')`. When we use this command we refer to the CRAN repository of packages, however sometimes we want a package from `Bioconductor` instead. For this we use the command `BiocManager::install('name_of_package')`. In order to use this installer, you need to download the R-package `BiocManager` e.g. `install.packages('BiocManager')`.

Gene Expression Analysis in R with `DEseq2`

DEseq2 is one of the many packages/frameworks which exists for analysis of bulk gene expression data in R. For more information on DEseq2, please have a look at the original publication here.

Other highly used packages for differential expression analysis *DEA* are:

- limma
- edgeR
- NOIseq

`DEseq2` has many advantages over classical models and post hoc tests, as it is specifically developed for handling common issues and biases in expression data, including differences in sequencing depth and highly variable dispersion of counts between genes.
In brief, `DEseq2` fits a generalized linear model (GLM) for each gene in the dataset. In the case where we compare two groups i.e. treatment vs control, the GLM fit returns coefficients indicating the overall expression strength of a gene, along with the log2 fold change between groups. `DEseq2` adjusts variable gene dispersion estimates using an empirical Bayes approach which borrows information across genes and shrinks gene-wise dispersions towards a common dispersion trend to increase accuracy of differential expression testing.

About the Dataset

The dataset used for this presentation was acquired from the following github tutorial on RNAseq analysis: https://combine-australia.github.io/RNAseq-R/06-rnaseq-day1.html.

RNA sequencing data generated from luminal and basal cell sub-populations in the mammary gland of three groups of mice:
- Control
- Pregnant
- Lactating

The objective of the original study (found here) was to identify genes specifically expressed in lactating mammary glands, the gene expression profiles of luminal and basal cells from different developmental stages were compared.

Load R-packages:

```
# Data Wrangling
# install.packages("tidyverse")
# install.packages("readxl")
library(tidyverse)
library(readxl)

# For Plotting
# install.packages("ggplot2")
library(ggplot2)


# For DEA
# install.packages("BiocManager")
# BiocManager::install("DESeq2")
library(DESeq2)
library(dplyr)
```

Importing Data

Reading in data:

```
exprDat <- read_excel("MouseRNAseq.xlsx")
exprInfo <- read_excel("MouseSampleInfo.xlsx")

# Look at the data:
head(exprDat, n=5)
```

```
## # A tibble: 5 x 13
##   GeneName MCL1.DG MCL1.DH MCL1.DI MCL1.DJ MCL1.DK MCL1.DL MCL1.LA MCL1.LB
##   <chr>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 Xkr4         438     300      65     237     354     287       0       0
## 2 Rp1            1       1       0       0       0       0      10       3
## 3 Sox17        106     182      82     105      43      82      16      25
## 4 Mrpl15       309     234     337     300     290     270     560     464
## 5 Lypla1       652     515     948     935     928     791     826     862
## # ... with 4 more variables: MCL1.LC <dbl>, MCL1.LD <dbl>, MCL1.LE <dbl>,
## #   MCL1.LF <dbl>
```

```
dim(exprDat)
```

```
## [1] 23151    13
```

```
head(exprInfo)
```

```
## # A tibble: 6 x 4
##   SampleName CellType Status   CellType.colors
##   <chr>      <chr>    <chr>    <chr>
## 1 MCL1.DG    basal    control  #79ADDC
## 2 MCL1.DH    basal    control  #79ADDC
## 3 MCL1.DI    basal    pregnant #79ADDC
## 4 MCL1.DJ    basal    pregnant #79ADDC
## 5 MCL1.DK    basal    lactate  #79ADDC
## 6 MCL1.DL    basal    lactate  #79ADDC
```

Convert character columns to factor types:

```
exprInfo <- exprInfo %>%
  mutate(CellType = as.factor(CellType),
         Status = as.factor(Status))

head(exprInfo)
```

```
## # A tibble: 6 x 4
##   SampleName CellType Status   CellType.colors
##   <chr>      <fct>    <fct>    <chr>
## 1 MCL1.DG    basal    control  #79ADDC
## 2 MCL1.DH    basal    control  #79ADDC
## 3 MCL1.DI    basal    pregnant #79ADDC
## 4 MCL1.DJ    basal    pregnant #79ADDC
## 5 MCL1.DK    basal    lactate  #79ADDC
## 6 MCL1.DL    basal    lactate  #79ADDC
```

Initial Data Check & Filtering:

Let's try to sample 16 (n) random genes and plot their count distribution.

```
# Sample 16 random rows
expr16 <- exprDat %>%
  sample_n(.,16)
expr16
```

```
## # A tibble: 16 x 13
##    GeneName      MCL1.DG MCL1.DH MCL1.DI MCL1.DJ MCL1.DK MCL1.DL MCL1.LA MCL1.LB
##    <chr>           <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 Retsat            479     479     813     911     475     405     547     536
##  2 Arhgap45           42      35      37      43      12      21       4       4
##  3 Mir144              0       0       0       0       0       0       0       0
##  4 Gm1673             38      50      43      58      16      24      54      70
##  5 Flrt1              33      30      57      31      38      40       3       6
##  6 H2bc23              4       6      14      13      18       9      26      46
##  7 Mir153              0       0       0       0       0       0       0       0
##  8 Tmsb15a             0       2       0       0       0       0       0       0
##  9 D630003M21Rik      40      33      19      20      13       7       5      11
## 10 Rpl14            3832    4057    3282    3135    1853    1703    6631    5357
## 11 Rnf19a           6025    4448    4460    3975    4650    4327    2173    2704
## 12 Olfr1135            0       0       0       0       0       0       0       0
## 13 Ptges3l            94      93     143     107     112     118      74      88
```

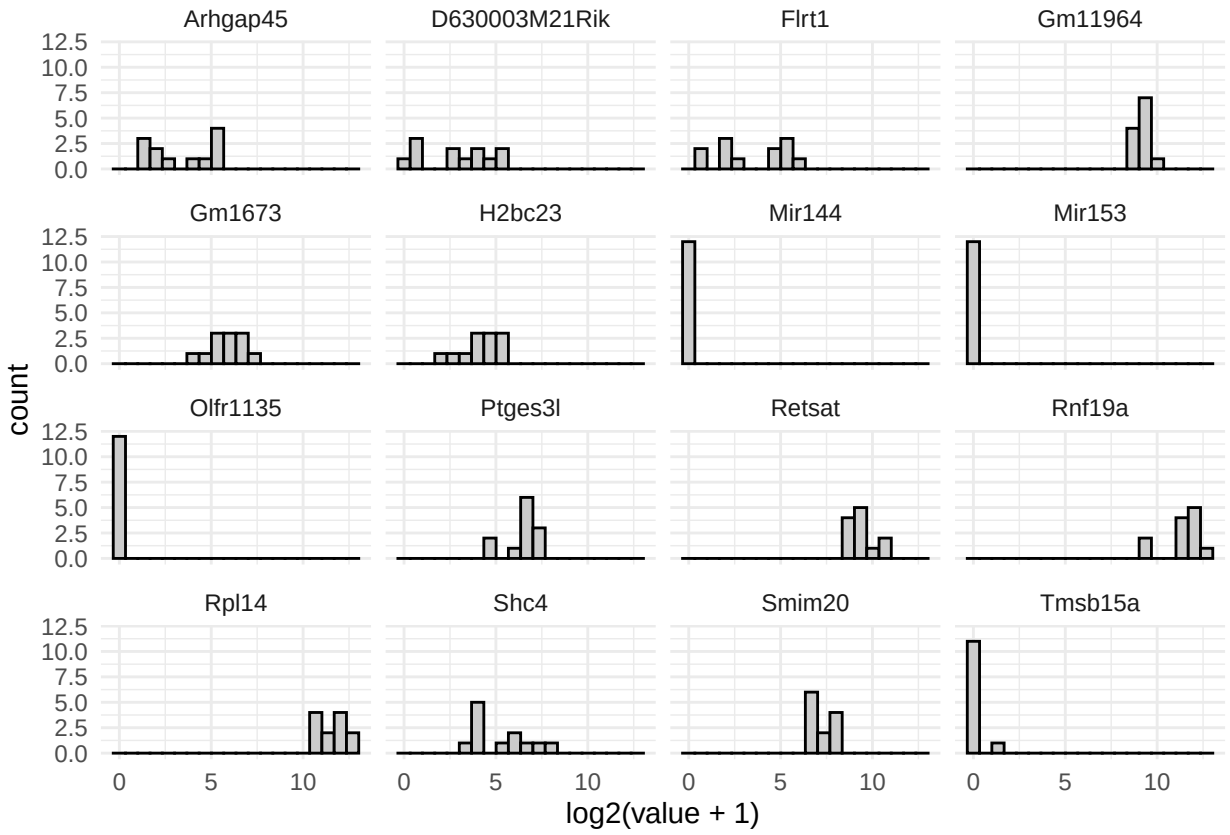
```
## 14 Shc4                45      52     227     187      79      90      16      13
## 15 Smim20             120      83     119     121     111     126     257     250
## 16 Gm11964            640     519     774     744     514     465     706     732
## # ... with 4 more variables: MCL1.LC <dbl>, MCL1.LD <dbl>, MCL1.LE <dbl>,
## #   MCL1.LF <dbl>
```

```
# Extract genenames
GeneName <- expr16$GeneName

# Gather counts
expr16 <- expr16 %>%
  dplyr::select(-GeneName) %>%
  t() %>%
  as_tibble() %>%
  rename_at(vars(names(.)), ~GeneName) %>%
  gather()

# Give it a look:
expr16
```

```
## # A tibble: 192 x 2
##    key    value
##    <chr>  <dbl>
##  1 Retsat   479
##  2 Retsat   479
##  3 Retsat   813
##  4 Retsat   911
##  5 Retsat   475
##  6 Retsat   405
##  7 Retsat   547
##  8 Retsat   536
##  9 Retsat  1478
## 10 Retsat  1825
## # ... with 182 more rows
```

Plot:

```
ggplot(expr16, aes(log2(value+1))) +
  geom_histogram(color="black", fill="grey80", bins=20) +
  theme_minimal() +
  facet_wrap(~key)
```



We will filter out low expressed genes. There are many strategies for doing so, but here we will filter out genes that have less than 3 counts in at least n samples. We select n as the smallest number of biologically meaningful groups. In this case, it is 3.

```
table(exprInfo$CellType, exprInfo$Status)
```

```
##
##           control lactate pregnant
##   basal         2       2        2
##   luminal       2       2        2
```

```
# 2 samples in each group
```

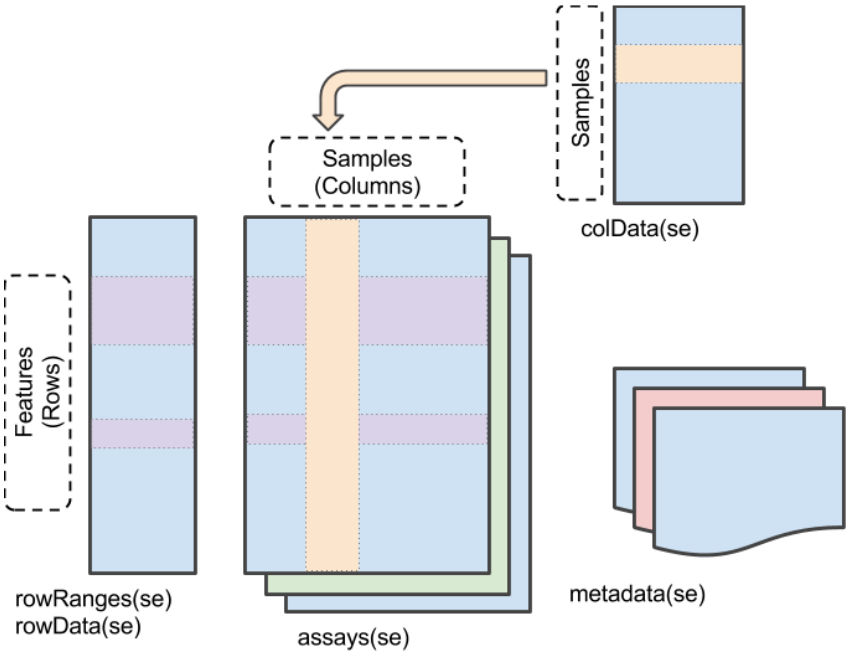
```
# First, we count the number of times a count value in a sample is greater or equal to 3. Then Filter r

exprDat <- exprDat %>%
  mutate(nzeros = rowSums(dplyr::select(.,-GeneName)<=3)) %>% # count number of
  filter(nzeros <= 4) %>%
  dplyr::select(-nzeros)

# How many genes do we have left:
dim(exprDat)
```

```
## [1] 13956    13
```

Differential Expression Analysis- DESeq2

We will now make a DESeq2 object. For this we use the function `DESeqDataSetFromMatrix` from the DEseq2 package.

DESeq object is a type of SummarizedExperiment container used to store the input values, intermediate calculations and results of an analysis of differential expression. The rows typically represent Genes (genomic ranges) of interest and the columns represent samples.



First, Convert exprDat to a dataframe and make GeneNames column into rownames:

```
# Pull out GeneNames and EntrezGeneID for later use
GeneNames <- exprDat %>%
  dplyr::select(GeneName)

exprDat <- exprDat %>%
  column_to_rownames(., var = "GeneName")
```

Make a DESeq2 object: As input we give our count matrix, our gene IDs and our meta data (exprInfo). Additionally we include a design for DE contrasts. In this case we add CellType (luminal or basal) and Status (control, pregnant or lactating).

```
exprObj <- DESeqDataSetFromMatrix(countData = exprDat,
                                  colData = exprInfo,
                                  design= ~CellType+Status)
exprObj
```

```
## class: DESeqDataSet
## dim: 13956 12
## metadata(1): version
## assays(1): counts
## rownames(13956): Sox17 Mrpl15 ... G530011O06Rik Gm47283,
## rowData names(0):
## colnames(12): MCL1.DG MCL1.DH ... MCL1.LE MCL1.LF
```

```
## colData names(4): SampleName CellType Status CellType.colors
```

Preliminary analysis:

There are multiple biases in RNAseq experiment: library size, genes length, genes GC composition, etc. Library size is the most well-known bias. For the purpose of DEA - genes length and GC composition are not so important because it is supposed to be about the same for the gene across different samples.

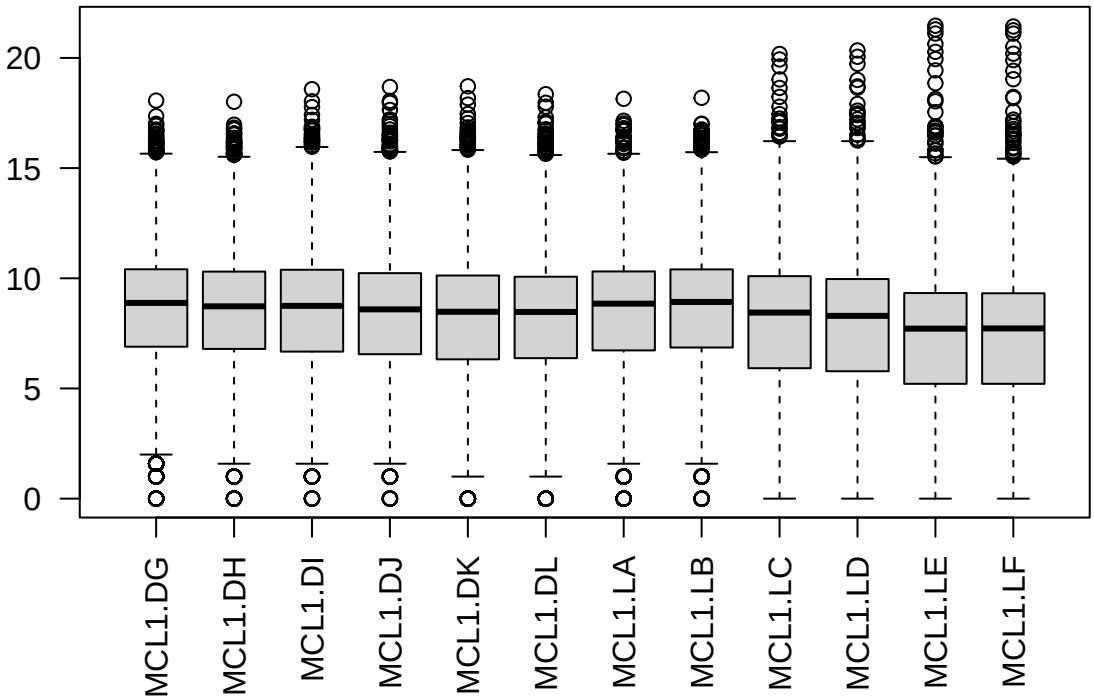
Let's have a look at the library sizes:

```
colSums(assay(exprObj))
```

```
##  MCL1.DG  MCL1.DH  MCL1.DI  MCL1.DJ  MCL1.DK  MCL1.DL  MCL1.LA  MCL1.LB
## 22594352 21113247 23444520 22054554 21008636 19535887 19674759 20911407
##  MCL1.LC  MCL1.LD  MCL1.LE  MCL1.LF
## 21666004 21449077 24408388 24355204
```

The count distributions may be dominated by a few genes with very large counts. These genes will drive plotting e.g. heatmaps, PCA analysis etc. Let's see if we have any "outlier" genes in our dataset and at the same time inspect the sample library sizes. For convenience I am using the base R boxplot function:

```
#boxplot(assay(exprObj), las=2)
boxplot(log2(assay(exprObj)+1), las=2)
```



As you can see we do not have any extreme outliers, but we do see some differences between libraries.

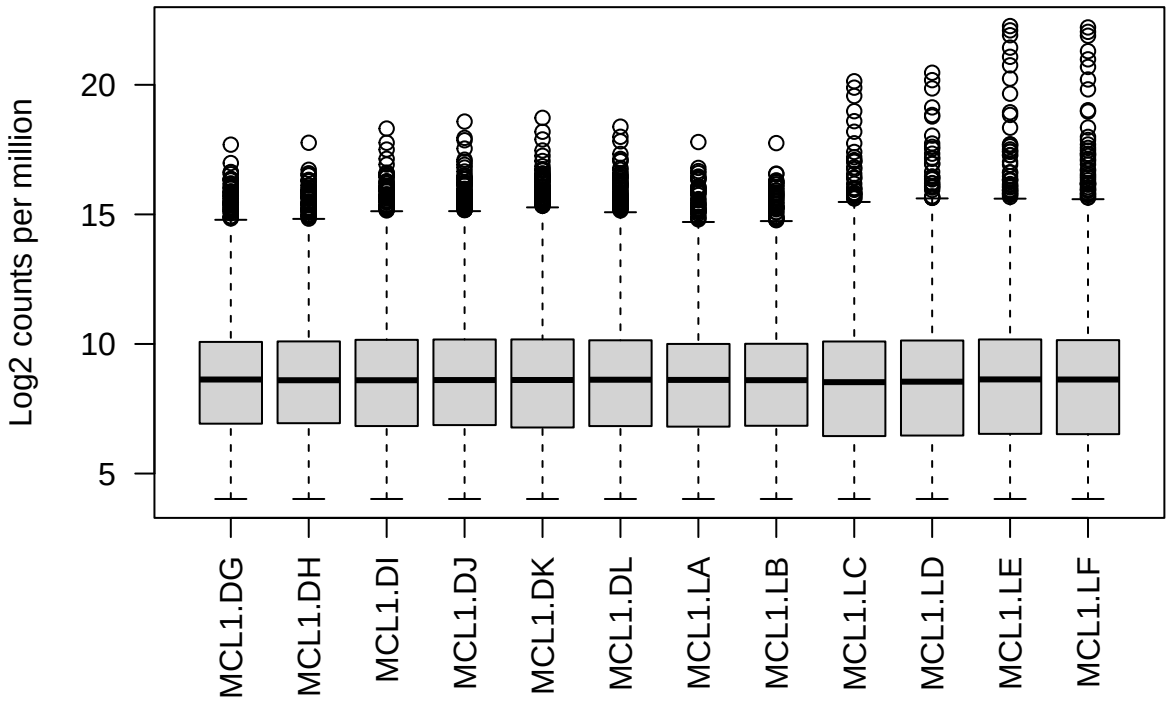
Next, we will apply the "vst" function to do a couple of things

- normalize library size to obtain counts per million mapped reads
- log2 transform the data to get more normally distributed data
- apply variance stabilizing transformation which we will discuss below.

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

Let's plot normalized data.

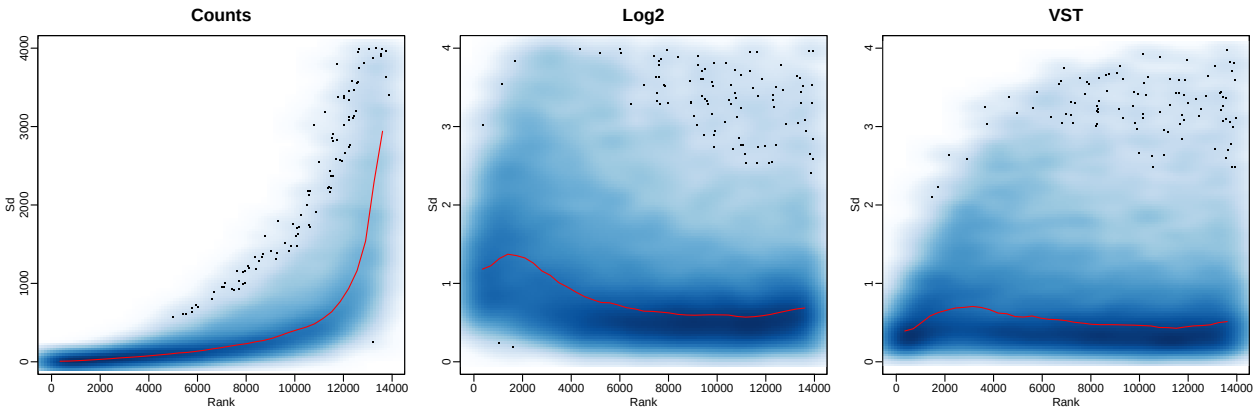
```
par(mfrow=c(1,1))
boxplot(assay(exprObjvst), xlab="", ylab="Log2 counts per million",las=2)
```



Variance stabilizing transformation:

In RNA-Seq data, genes with larger average expression have on average larger observed variances (sd) across samples. This is know as data heteroscedasticity. Expression varies from sample to sample more than other genes with lower average expression.

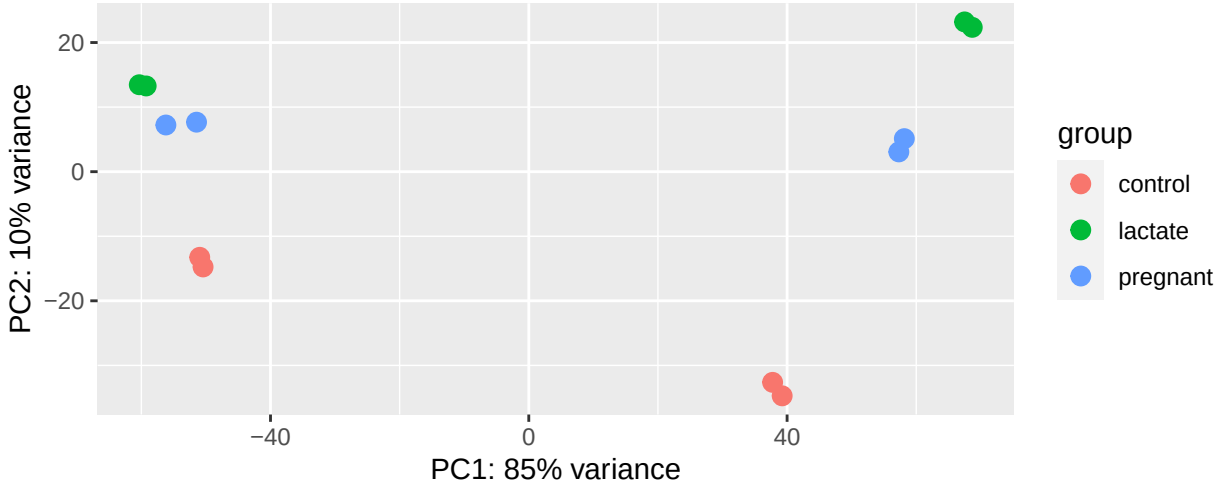
```
# BiocManager::install("vsn")
library(vsn)
rawSd <- vsn::meanSdPlot(as.matrix(assay(exprObj)), plot=FALSE)
logSd <- vsn::meanSdPlot(log2(as.matrix(assay(exprObj))+1), plot=FALSE)
vstSd <- vsn::meanSdPlot(as.matrix(assay(exprObjvst)), plot=FALSE)

par(mfrow = c(1,3),  mar = c(3, 3, 3, 1), mgp = c(1.5, 0.5, 0))
smoothScatter(x = rawSd$px, y = rawSd$py, ylab = "Sd",xlab = "Rank", main = "Counts", cex.main=1.5, yli
lines(x = rawSd$rank, y = rawSd$sd, add = TRUE, col = "red")
smoothScatter(x = logSd$px, y = logSd$py, ylab = "Sd",xlab = "Rank", main = "Log2", cex.main=1.5, ylim =
lines(x = logSd$rank, y = logSd$sd, add = TRUE, col = "red")
smoothScatter(x = vstSd$px, y = vstSd$py, ylab = "Sd",xlab = "Rank", main = "VST", cex.main=1.5, ylim =
lines(x = vstSd$rank, y = vstSd$sd, add = TRUE, col = "red")
```



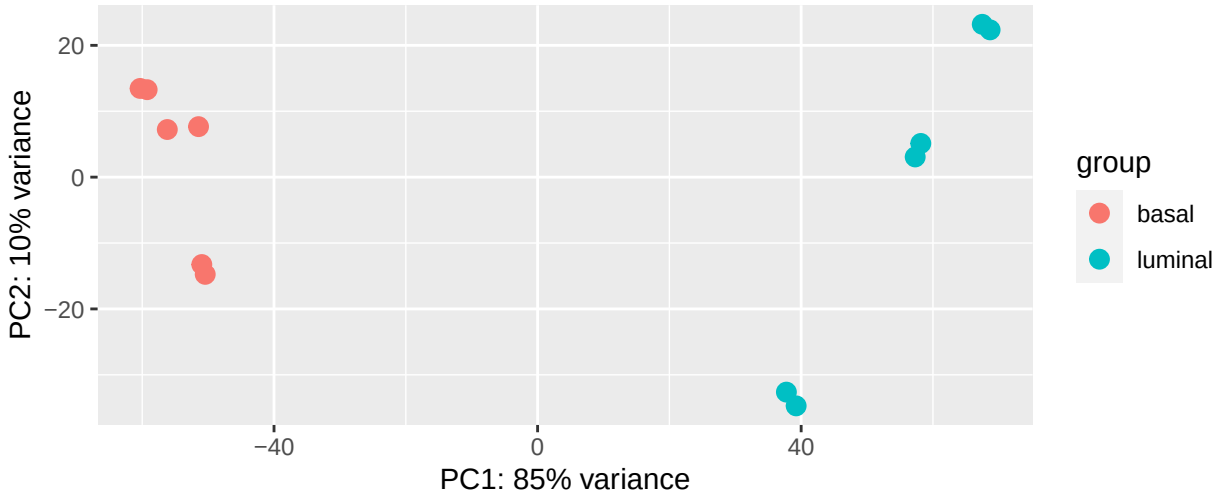
Principal Component Analysis

Before performing DEA it is a good idea to explore how samples cluster together based on there gene expression profile. The expectation here is that samples from the same group (treatment vs control, condition A vs condition B, etc.) will cluster together. A principal component analysis (PCA) plot can also help us to identify outlier samples which might need to be removed from the analysis. We use our vst counts for principal component analysis:

```
plotPCA(exprObjvst,intgroup="Status")
```




```
plotPCA(exprObjvst,intgroup="CellType")
```



DESeq function for DEA

Next, we use `DEseq()` to estimate dispersion, gene-wise and mean-dispersion, fitting model(s):

```
exprObj <- DESeq(exprObj)
```

Testing

Have a look at the group comparisons:

```
resultsNames(exprObj)
```

```
## [1] "Intercept"                  "CellType_luminal_vs_basal"
## [3] "Status_lactate_vs_control"  "Status_pregnant_vs_control"
```

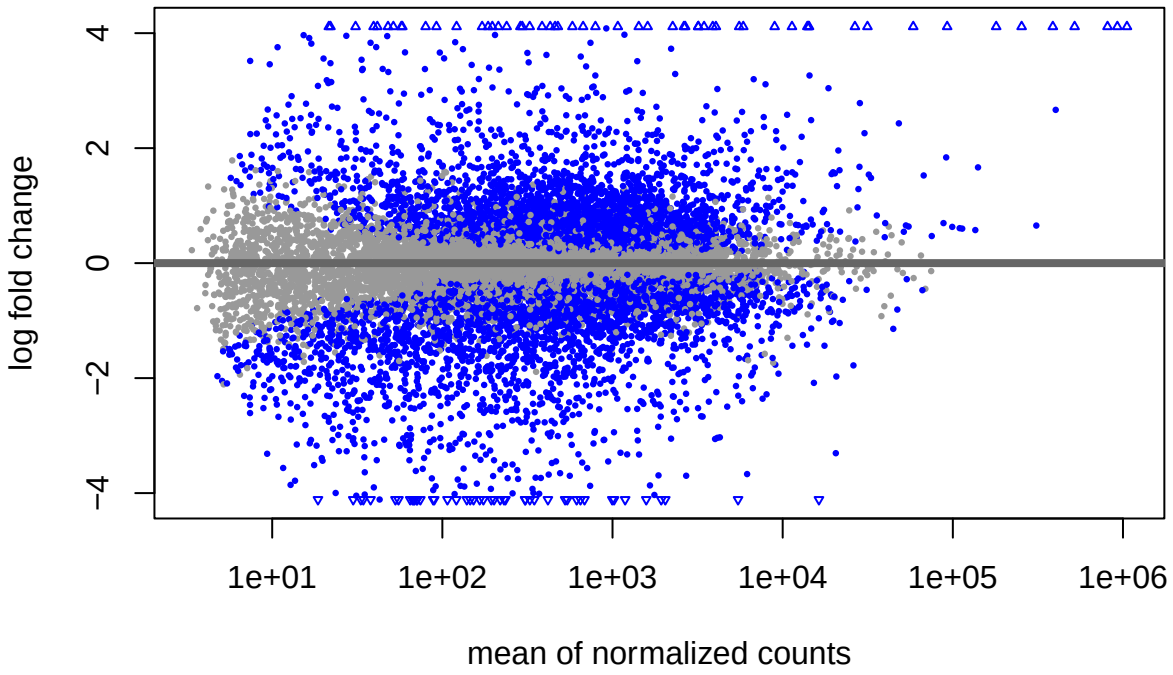
Test for DE genes between the three groups of mice, adjusted for cell type:

(I) lactating and control mice:

```
resLC <- results(exprObj, contrast = c("Status", "lactate", "control"), independentFiltering = FALSE)
```

Summary and plot of DE analysis results:

```
DESeq2::plotMA(resLC)
```




```
summary(resLC)
```

```
##
## out of 13956 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 3352, 24%
## LFC < 0 (down)     : 3319, 24%
## outliers [1]       : 0, 0%
## low counts [2]     : 0, 0%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Below we perform the same steps as above to get the DE genes between (II) pregnant and control mice and (III) lactating and pregnant mice:

(II) pregnant and control mice:

```
resPC <- results(exprObj, contrast = c("Status", "pregnant", "control"),  independentFiltering = FALSE)

#DESeq2::plotMA(resPC)
#summary(resPC)
```

(III) lactating and pregnant mice:

```
resLP <- results(exprObj, contrast = c("Status", "lactate", "pregnant"),  independentFiltering = FALSE)

#DESeq2::plotMA(resLP)
#summary(resLP)
```

We filter the results of the DEA to only include those genes which are differentially expressed based on logFC (>= 1.0 or <= -1.0) and adjusted p-value (< 0.01).

Firstly, bind the three DE genesets together and convert to a tibble. Then, add a column with GeneNames. Lastly, filter and arrange rows (genes) based on logFC and adjusted p-values.

```
resDE <- rbind(resLC, resPC, resLP) %>%
  as_tibble() %>%
  mutate(GeneName = rep(GeneNames$GeneName, 3)) %>%
  filter((log2FoldChange >= 1.0 | log2FoldChange <= -1.0) & padj <= 0.01) %>%
  arrange(padj, desc(abs(log2FoldChange)))

# DE genes
dim(resDE)
```

```
## [1] 3719    7
```

```
length(unique(resDE$GeneName))
```

```
## [1] 2514
```

Heatmap Visualization

To visually inspect if DE genes identified in our DESeq2 analysis successfully separate the three groups of mice (control, pregnant and lactating), we will make a heatmap. For this we use the `heatmap` function.

It will not make sense to include all DE genes in this heatmap (almost 3000 unique genes). Instead pick the top 100 most significant DE genes, based on adj. p-value and logFC.

Make a vector of unique GeneNames (top100):

```
# Make a vector of unique EntrezGeneIDs (top100):

top100 <- resDE[1:100,] %>%
  pull(GeneName) %>%
  unique()

length(top100)
```

```
## [1] 83
```

The expression counts themselves (not logFC) are needed for the heatmap. We use the topDE vector to extract these from the vst normalized DESeq2 object.

```
head(assay(exprObjvst), n=5)
```

```
##          MCL1.DG   MCL1.DH   MCL1.DI   MCL1.DJ   MCL1.DK   MCL1.DL   MCL1.LA
## Sox17   6.807137  7.526950  6.612493  7.004070  6.172622  6.831531  5.249017
## Mrpl15  8.078000  7.836346  8.283146  8.283294  8.339938  8.264094  8.876529
## Lypla1  9.065419  8.864289  9.678733  9.821289  9.918659  9.714720  9.405766
## Tcea1  10.312516 10.336055 10.513555 10.300662 10.229824 10.130987 10.071321
## Rgs20   4.643890  4.482110  5.206922  4.704652  4.525065  4.528381  7.356000
##          MCL1.LB   MCL1.LC   MCL1.LD  MCL1.LE   MCL1.LF
## Sox17   5.491345  5.456965  5.055592 4.828645  5.449353
```

```
## Mrpl15 8.544276  8.985745  8.607371 9.151063  9.283624
## Lypla1 9.380779  9.411505  9.526833 9.940947 10.017603
## Tcea1  9.904348 10.063042 10.027368 9.845480  9.808164
## Rgs20  7.249534  7.365392  6.572828 6.250945  5.739676
```

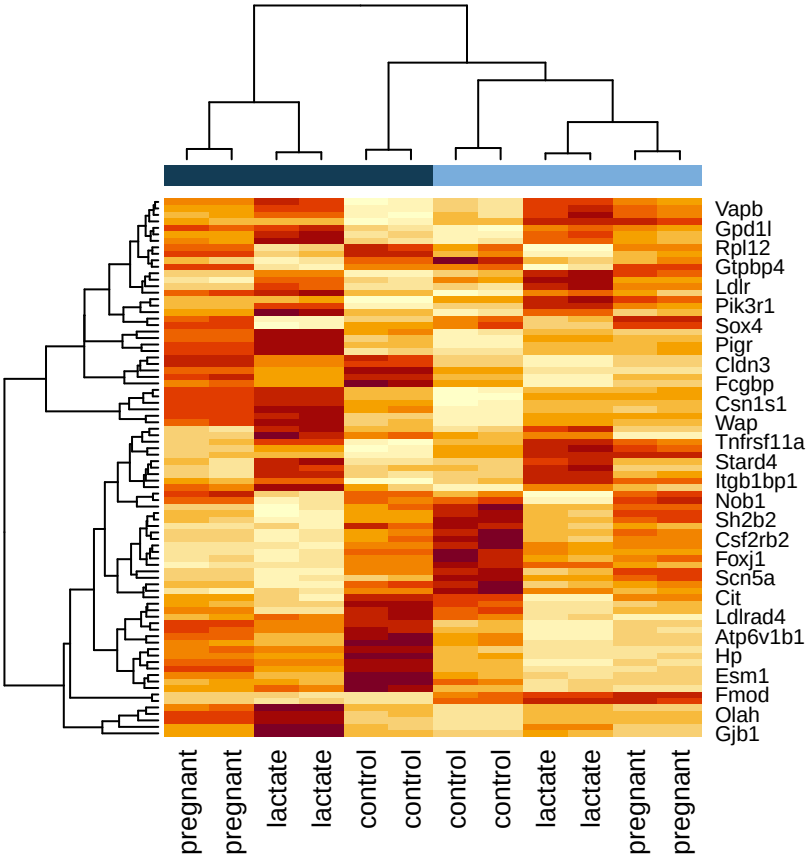
```
resVST <- assay(exprObjvst) %>%
  as.data.frame() %>%
  rownames_to_column(var = "GeneName")  %>%
  as_tibble() %>%
  filter(GeneName %in% top100)
```

The heatmap function in base R wants gene expression data as a matrix (a dataframe with numeric values only). We extract the GeneNames column and convert the tibble into a matrix:

```
HPnames <- resVST %>%
  pull(GeneName)


HPdat <- resVST %>%
  dplyr::select(-GeneName) %>%
  as.matrix()
```

We use the `heatmap` function to generate a heatmap. We can modify the look of the heatmap as desired, e.g. add column colors, row labels, change color scheme etc.

```
heatmap(HPdat,
        ColSideColors=exprInfo$CellType.colors,
        labCol=exprInfo$Status,
        labRow = HPnames,
        cexCol=1.2,
        cexRow = 1.0)
```


Vapb
Gpd1l
Rpl12
Gtpbp4
Ldlr
Pik3r1
Sox4
Pigr
Cldn3
Fcgbp
Csn1s1
Wap
Tnfrsf11a
Stard4
Itgb1bp1
Nob1
Sh2b2
Csf2rb2
Foxj1
Scn5a
Cit
Ldlrad4
Atp6v1b1
Hp
Esm1
Fmod
Olah
Gjb1
pregnant
pregnant
lactate
lactate
control
control
control
control
lactate
lactate
pregnant
pregnant