

lab 4

Simon Jorstedt

2024-12-12

Question 1

Run all the R code and reproduce the graphics. Go carefully through the R code and explain in your words what each step does. HINT Recall what a design/model matrix is from linear regression

This chunk gives an initial look at the dataset `airway`, which provides expression values in genes on four human airway muscle cell lines treated with a compound. We see that the data is organised as a matrix or assay with 63677 genes along the rows, and 8 different experiments along the columns. What follows below is a section of the assay,

```
## ----airway-SummarizedExperiment-----  
library(airway)      # An 'ExperimentData' package...  
data(airway)         # ...with a sample data set...  
airway               # ...that is a SummarizedExperiment  
  
## class: RangedSummarizedExperiment  
## dim: 63677 8  
## metadata(1): ''  
## assays(1): counts  
## rownames(63677): ENSG000000000003 ENSG000000000005 ... ENSG00000273492 ENSG00000273493  
## rowData names(10): gene_id gene_name ... seq_coord_system symbol  
## colnames(8): SRR1039508 SRR1039509 ... SRR1039520 SRR1039521  
## colData names(9): SampleName cell ... Sample BioSample  
  
head(assay(airway))  # contains a matrix of counts
```

```
##           SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520  
## ENSG000000000003      679       448       873       408       1138       1047       770  
## ENSG000000000005        0         0         0         0         0         0         0  
## ENSG000000000419      467       515       621       365       587       799       417  
## ENSG000000000457      260       211       263       164       245       331       233  
## ENSG000000000460        60        55        40        35        78        63        76  
## ENSG000000000938        0         0         2         0         1         0         0  
##           SRR1039521  
## ENSG000000000003      572  
## ENSG000000000005        0  
## ENSG000000000419      508  
## ENSG000000000457      229  
## ENSG000000000460        60  
## ENSG000000000938        0
```

```
head(rowRanges(airway)) # information about the genes...
```

```
## GRangesList object of length 6:
## $ENSG000000000003
## GRanges object with 17 ranges and 2 metadata columns:
##      seqnames      ranges strand | exon_id exon_name
##      <Rle>        <IRanges> <Rle> | <integer> <character>
## [1]      X 99883667-99884983   - |   667145 ENSE00001459322
## [2]      X 99885756-99885863   - |   667146 ENSE00000868868
## [3]      X 99887482-99887565   - |   667147 ENSE00000401072
## [4]      X 99887538-99887565   - |   667148 ENSE00001849132
## [5]      X 99888402-99888536   - |   667149 ENSE00003554016
## ...      ...      ...      ... |   ...      ...
## [13]     X 99890555-99890743   - |   667156 ENSE00003512331
## [14]     X 99891188-99891686   - |   667158 ENSE00001886883
## [15]     X 99891605-99891803   - |   667159 ENSE00001855382
## [16]     X 99891790-99892101   - |   667160 ENSE00001863395
## [17]     X 99894942-99894988   - |   667161 ENSE00001828996
## -----
## seqinfo: 722 sequences (1 circular) from an unspecified genome
## ...
## <5 more elements>
```

```
colData(airway)[, 1:3] # ...and samples
```

```
## DataFrame with 8 rows and 3 columns
##      SampleName      cell      dex
##      <factor> <factor> <factor>
## SRR1039508 GSM1275862 N61311 untrt
## SRR1039509 GSM1275863 N61311 trt
## SRR1039512 GSM1275866 N052611 untrt
## SRR1039513 GSM1275867 N052611 trt
## SRR1039516 GSM1275870 N080611 untrt
## SRR1039517 GSM1275871 N080611 trt
## SRR1039520 GSM1275874 N061011 untrt
## SRR1039521 GSM1275875 N061011 trt
```

```
## coordinated subsetting
untrt <- airway[, airway$dex == 'untrt']
head(assay(untrt))
```

```
##      SRR1039508 SRR1039512 SRR1039516 SRR1039520
## ENSG000000000003      679      873      1138      770
## ENSG000000000005       0       0       0       0
## ENSG000000000419      467      621      587      417
## ENSG000000000457      260      263      245      233
## ENSG000000000460       60       40       78       76
## ENSG000000000938       0       2       1       0
```

```
colData(untrt)[, 1:3]
```

```
## DataFrame with 4 rows and 3 columns
##           SampleName      cell      dex
##           <factor> <factor> <factor>
## SRR1039508 GSM1275862 N61311      untrt
## SRR1039512 GSM1275866 N052611      untrt
## SRR1039516 GSM1275870 N080611      untrt
## SRR1039520 GSM1275874 N061011      untrt
```

```
## -----airway-colData-----
library(airway)          # An 'ExperimentData' package...
data(airway)             # ...with a sample data set...
colData(airway)[, 1:3]   # ...represented as a SummarizedExperiment
```

```
## DataFrame with 8 rows and 3 columns
##           SampleName      cell      dex
##           <factor> <factor> <factor>
## SRR1039508 GSM1275862 N61311      untrt
## SRR1039509 GSM1275863 N61311      trt
## SRR1039512 GSM1275866 N052611      untrt
## SRR1039513 GSM1275867 N052611      trt
## SRR1039516 GSM1275870 N080611      untrt
## SRR1039517 GSM1275871 N080611      trt
## SRR1039520 GSM1275874 N061011      untrt
## SRR1039521 GSM1275875 N061011      trt
```

```
## -----airway-assay-----
head(assay(airway))
```

	SRR1039508	SRR1039509	SRR1039512	SRR1039513	SRR1039516	SRR1039517	SRR1039520
## ENSG000000000003	679	448	873	408	1138	1047	770
## ENSG000000000005	0	0	0	0	0	0	0
## ENSG000000000419	467	515	621	365	587	799	417
## ENSG000000000457	260	211	263	164	245	331	233
## ENSG000000000460	60	55	40	35	78	63	76
## ENSG000000000938	0	0	2	0	1	0	0
## SRR1039521							
## ENSG000000000003	572						
## ENSG000000000005	0						
## ENSG000000000419	508						
## ENSG000000000457	229						
## ENSG000000000460	60						
## ENSG000000000938	0						

Below, we perform a differential expression analysis, based on the Negative Binomial distribution, using the DESeq2 package. This involves estimating size factors and dispersions, as well as fitting and testing the model. We then extract the results and order them from largest to smallest absolute log fold change. Below we present a table illustrating a subset of the results of the model. Among other things we report the p-value and log fold change for each gene.

```
## ----airway-toptable-----
library(DESeq2)      # package implementing statistical methods
dds <-              # data and experimental design
  DESeqDataSet(airway, design = ~ cell + dex)
dds <- DESeq(dds)    # initial analysis

## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

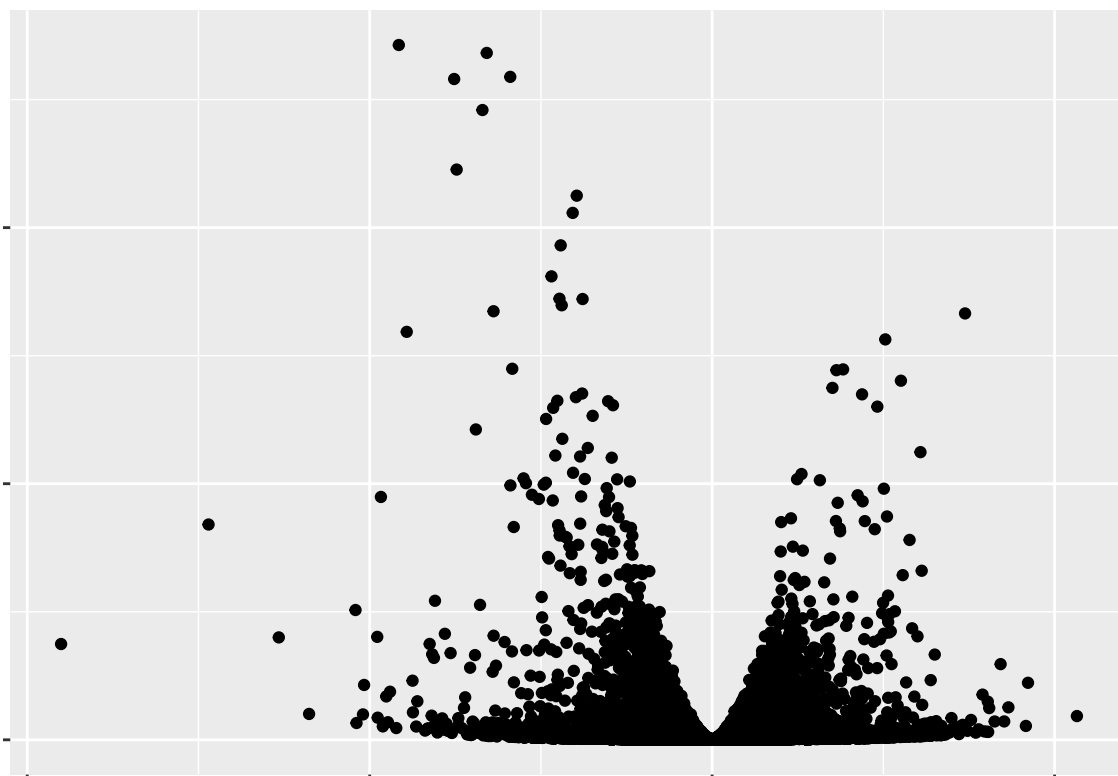
res <- results(dds) # summary results
ridx <-            # order from largest to smallest absolute log fold change
  order(abs(res$log2FoldChange), decreasing=TRUE)
res <- res[ridx,]
head(res)          # top-table

## log2 fold change (MLE): dex untrt vs trt
## Wald test p-value: dex untrt vs trt
## DataFrame with 6 rows and 6 columns
##           baseMean log2FoldChange  lfcSE      stat      pvalue      padj
##           <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
## ENSG00000179593  67.24305      -9.50597  1.054503 -9.01465 1.97505e-19 1.25304e-17
## ENSG00000109906 385.07103      -7.35263  0.536389 -13.70764 9.13762e-43 2.25537e-40
## ENSG00000250978  56.31819      -6.32738  0.677797 -9.33521 1.00788e-20 7.20631e-19
## ENSG00000132518   5.65465      -5.88511  1.324044 -4.44480 8.79726e-06 1.00006e-04
## ENSG00000128285   6.62474       5.32590  1.257815  4.23425 2.29314e-05 2.37869e-04
## ENSG00000127954 286.38412      -5.20716  0.493082 -10.56044 4.54548e-26 5.05559e-24
```

Below we finally plot the negative logarithmized p-value against the log fold change.

```
## ----airway-viz-----
library(ggplot2)
ggplot(as.data.frame(res),
  aes(x=log2FoldChange, y=-10 * log10(pvalue))) +
  geom_point()

## Warning: Removed 30208 rows containing missing values or values outside the scale range
## (`geom_point()`).
```



```
## ----airway-mapids-----
library(org.Hs.eg.db)
ensid <- head(rownames(res))
select(org.Hs.eg.db, ensid, c("SYMBOL", "GENENAME"), "ENSEMBL")
```

```
## 'select()' returned 1:1 mapping between keys and columns
```

	ENSEMBL	SYMBOL	GENENAME
## 1	ENSG00000179593	ALOX15B	arachidonate 15-lipoxygenase type B
## 2	ENSG00000109906	ZBTB16	zinc finger and BTB domain containing 16
## 3	ENSG00000250978	<NA>	<NA>
## 4	ENSG00000132518	GUCY2D	guanylate cyclase 2D, retinal
## 5	ENSG00000128285	MCHR1	melanin concentrating hormone receptor 1
## 6	ENSG00000127954	STEAP4	STEAP4 metalloredutase

```
## ----shiny-BAMSpector, eval=FALSE-----
# app <- system.file(package="BiocUruguay2015", "BAMSpector")
# shiny::runApp(app)

## ----shiny-MAPlotExplorer, eval=FALSE-----
# app <- system.file(package="BiocUruguay2015", "MAPlotExplorer")
# shiny::runApp(app)

## ----sessionInfo-----
#sessionInfo()
```

Question 2

In the presented analysis, there are no plots of raw paired data. In the section where the contrasts are defined and the three contrasts. Present the variables versus each other original, log scaled and MAplot for each considered pair both before and after normalization. A cluster analysis is performed on the page but not reported. Present plots and also draw heatmaps.