Class 13: RNAseq with DESeq2

Josie (A11433761)

Data import

```
counts<-read.csv("airway_scaledcounts.csv", row.names=1)
metadata<-read.csv("airway_metadata.csv")
head(counts)
```

```
                SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
ENSG00000000003        723        486        904        445       1170
ENSG00000000005          0          0          0          0          0
ENSG00000000419        467        523        616        371        582
ENSG00000000457        347        258        364        237        318
ENSG00000000460         96         81         73         66        118
ENSG00000000938          0          0          1          0          2
                SRR1039517 SRR1039520 SRR1039521
ENSG00000000003       1097        806        604
ENSG00000000005          0          0          0
ENSG00000000419        781        417        509
ENSG00000000457        447        330        324
ENSG00000000460         94        102         74
ENSG00000000938          0          0          0
```

```
head(metadata)
```

```
          id     dex celltype     geo_id
1 SRR1039508 control   N61311 GSM1275862
2 SRR1039509 treated   N61311 GSM1275863
3 SRR1039512 control  N052611 GSM1275866
4 SRR1039513 treated  N052611 GSM1275867
5 SRR1039516 control  N080611 GSM1275870
6 SRR1039517 treated  N080611 GSM1275871
```

Q1: How many transcripts/genes are in the `counts` object? There are 38694 in this dataset

```r
nrow(counts)
```

```
[1] 38694
```

Q2: How many control samples are there?

```r
sum(metadata$dex=="control")
```

```
[1] 4
```

OR…

```r
table(metadata$dex)
```

```

control treated 
      4       4 
```

Compare control vs treated 1. Split the "counts" into `control.counts` and `treated.counts`

```r
control.inds<-metadata$dex=="control"
```

Syntax with df[ROWs,COLs]

```r
control.counts<-counts[,control.inds]
```

```r
treated.counts<-counts[,metadata$dex=="treated"]
```

2. Calculate mean counts per gene for `control` and `treated`. Then compare.

Let's call `control.mean` and `treated.mean`

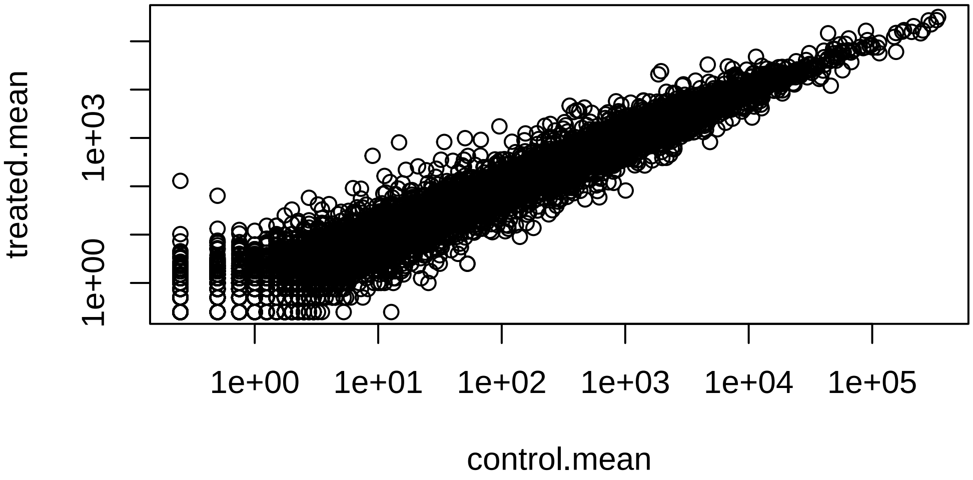
```r
#I can use `apply` function to apply `mean()` over the rows and columns of any data.frame

control.mean<-apply(control.counts, 1,mean)
treated.mean<-apply(treated.counts, 1, mean)
```

```
meancounts <- data.frame(control.mean, treated.mean)
plot(meancounts, log="xy")
```

```
Warning in xy.coords(x, y, xlabel, ylabel, log): 15032 x values <= 0 omitted
from logarithmic plot
```

```
Warning in xy.coords(x, y, xlabel, ylabel, log): 15281 y values <= 0 omitted
from logarithmic plot
```



We use log2 transforms here because it make the math easier. Log2(1)=0, so if treated/control =1, the log2 says there is no change.

```
log2(10/10)
```

```
[1] 0
```

```
log2(20/10)
```

```
[1] 1
```

```
log2(5/10)
```

```
[1] -1
```

```
log2(40/10)
```

```
[1] 2
```

```
log2(2.5/10)
```

```
[1] -2
```

Let's calculate log2 fold change and add it to our table

```
meancounts$log2fc<-log2(meancounts$treated.mean/meancounts$control.mean)
head(meancounts)
```

```
                control.mean treated.mean      log2fc
ENSG00000000003       900.75       658.00 -0.45303916
ENSG00000000005         0.00         0.00         NaN
ENSG00000000419       520.50       546.00  0.06900279
ENSG00000000457       339.75       316.50 -0.10226805
ENSG00000000460        97.25        78.75 -0.30441833
ENSG00000000938         0.75         0.00        -Inf
```

Filter out all genes with zero counts in either control or treated

```
to.rm<-rowSums(meancounts[,1:2]==0)>0
mycounts<-meancounts[!to.rm,]#"!" inverts
```

```
nrow(mycounts)
```

```
[1] 21817
```

Q: How many "down" regulated genes do we have at the log2 fold change value of -2

```
sum(mycounts$log2fc < -2)
```

```
[1] 367
```

Q: How many "up" regulated at log2FC> +2

```
sum(mycounts$log2fc > 2)
```

```
[1] 250
```

Do we trust these results? We are missing the stats

DESeq analysis

```
library(DESeq2)
```

DESeq, like many BiocManager packages, wants our input data in a very specific format

```
dds<-DESeqDataSetFromMatrix(countData=counts, colData=metadata, design= ~dex)
```

```
converting counts to integer mode
```

```
Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
design formula are characters, converting to factors
```

The main function in DESeq2 is called `DESeq()`

```
dds<- DESeq(dds)
```

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

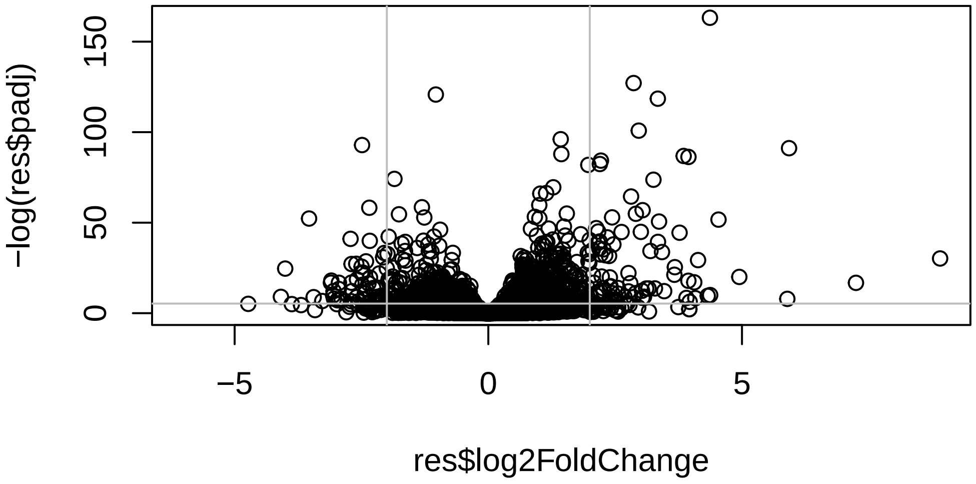
```r
res<-results(dds)
```

```r
head(res)
```

```
log2 fold change (MLE): dex treated vs control
Wald test p-value: dex treated vs control
DataFrame with 6 rows and 6 columns
                  baseMean log2FoldChange     lfcSE      stat    pvalue
                 <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000000003 747.194195     -0.3507030  0.168246 -2.084470 0.0371175
ENSG00000000005   0.000000             NA        NA        NA        NA
ENSG00000000419 520.134160      0.2061078  0.101059  2.039475 0.0414026
ENSG00000000457 322.664844      0.0245269  0.145145  0.168982 0.8658106
ENSG00000000460  87.682625     -0.1471420  0.257007 -0.572521 0.5669691
ENSG00000000938   0.319167     -1.7322890  3.493601 -0.495846 0.6200029
                     padj
                <numeric>
ENSG00000000003  0.163035
ENSG00000000005        NA
ENSG00000000419  0.176032
ENSG00000000457  0.961694
ENSG00000000460  0.815849
ENSG00000000938        NA
```

A common figure that plots logFC vs P-value

```r
plot(res$log2FoldChange, -log(res$padj))
abline(v=c(-2,2), col="grey")
abline(h=-log(0.005),col="grey")
```

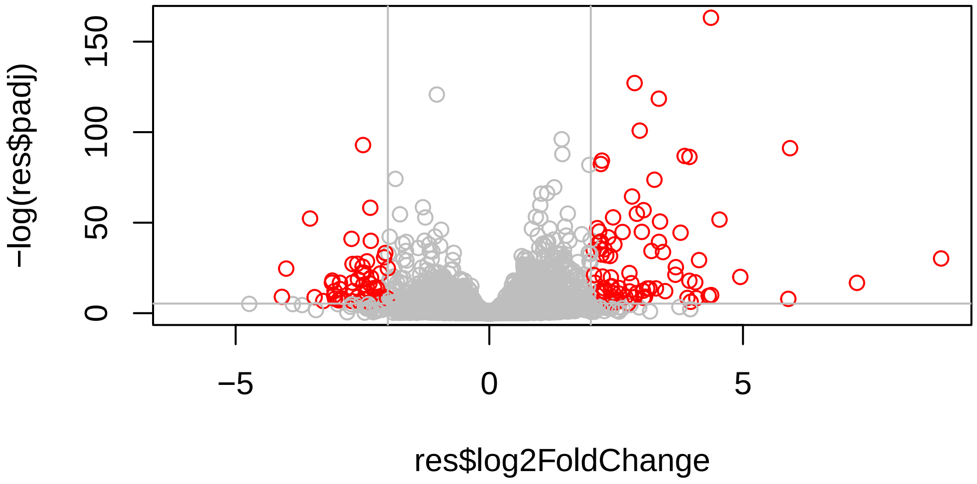


```
mycols<-rep("grey", nrow(res))
mycols[res$log2FoldChange> 2]<-"red"

mycols[res$log2FoldChange< -2]<-"red"

mycols[res$padj>0.005]<-"grey"

plot(res$log2FoldChange, -log(res$padj), col=mycols)
abline(v=c(-2,2), col="grey")
abline(h=-log(0.005),col="grey")
```



```r
write.csv(res, file="myresults.csv")
```

Gene annotation

```r
library("AnnotationDbi")
library("org.Hs.eg.db")
```

```r
columns(org.Hs.eg.db)
```

```
 [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
 [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
[11] "GENETYPE"     "GO"           "GOALL"        "IPI"          "MAP"
[16] "OMIM"         "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"
[21] "PMID"         "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
[26] "UNIPROT"
```

```r
res$symbol <- mapIds(org.Hs.eg.db,
                     keys=row.names(res), # Our gene names
                     keytype="ENSEMBL",        # The format of our gene names
                     column="SYMBOL",          # The new format we want to add
                     multiVals="first")
```

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

```r
head(res)
```

```
log2 fold change (MLE): dex treated vs control
Wald test p-value: dex treated vs control
DataFrame with 6 rows and 7 columns
                  baseMean log2FoldChange     lfcSE      stat    pvalue
                 <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000000003 747.194195     -0.3507030  0.168246 -2.084470 0.0371175
ENSG00000000005   0.000000             NA        NA        NA        NA
ENSG00000000419 520.134160      0.2061078  0.101059  2.039475 0.0414026
ENSG00000000457 322.664844      0.0245269  0.145145  0.168982 0.8658106
ENSG00000000460  87.682625     -0.1471420  0.257007 -0.572521 0.5669691
ENSG00000000938   0.319167     -1.7322890  3.493601 -0.495846 0.6200029
                     padj      symbol
                <numeric> <character>
ENSG00000000003  0.163035      TSPAN6
ENSG00000000005        NA        TNMD
ENSG00000000419  0.176032        DPM1
ENSG00000000457  0.961694       SCYL3
ENSG00000000460  0.815849       FIRRM
ENSG00000000938        NA         FGR
```

Pathway analysis

```r
library(pathview)
```

```
##############################################################################
Pathview is an open source software package distributed under GNU General
Public License version 3 (GPLv3). Details of GPLv3 is available at
http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
formally cite the original Pathview paper (not just mention it) in publications
```

```
or products. For details, do citation("pathview") within R.

The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
license agreement (details at http://www.kegg.jp/kegg/legal.html).
##############################################################################
```

```r
library(gage)
```

```r
library(gageData)

data(kegg.sets.hs)
```

```r
res$entrez <- mapIds(org.Hs.eg.db,
                     keys=row.names(res), # Our gene names
                     keytype="ENSEMBL",        # The format of our gene names
                     column="ENTREZID",          # The new format we want to add
                     multiVals="first")
```

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

```r
foldchanges<-res$log2FoldChange
names(foldchanges)<-res$entrez
head(foldchanges)
```

```
       7105       64102        8813       57147       55732        2268
-0.35070302          NA  0.20610777  0.02452695 -0.14714205 -1.73228897
```

```r
keggres = gage(foldchanges, gsets=kegg.sets.hs)
attributes(keggres)
```

```
$names
[1] "greater" "less"    "stats"
```

```r
head(keggres$less, 3)
```

```
                                       p.geomean stat.mean        p.val
hsa05332 Graft-versus-host disease 0.0004250461 -3.473346 0.0004250461
hsa04940 Type I diabetes mellitus  0.0017820293 -3.002352 0.0017820293
hsa05310 Asthma                    0.0020045888 -3.009050 0.0020045888
                                         q.val set.size         exp1
hsa05332 Graft-versus-host disease 0.09053483       40 0.0004250461
hsa04940 Type I diabetes mellitus  0.14232581       42 0.0017820293
hsa05310 Asthma                    0.14232581       29 0.0020045888
```

```
pathview(gene.data=foldchanges, pathway.id="hsa05310")
```

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

```
Info: Working in directory /Users/josierivera/Library/CloudStorage/OneDrive-Personal/Documen
```

```
Info: Writing image file hsa05310.pathview.png
```

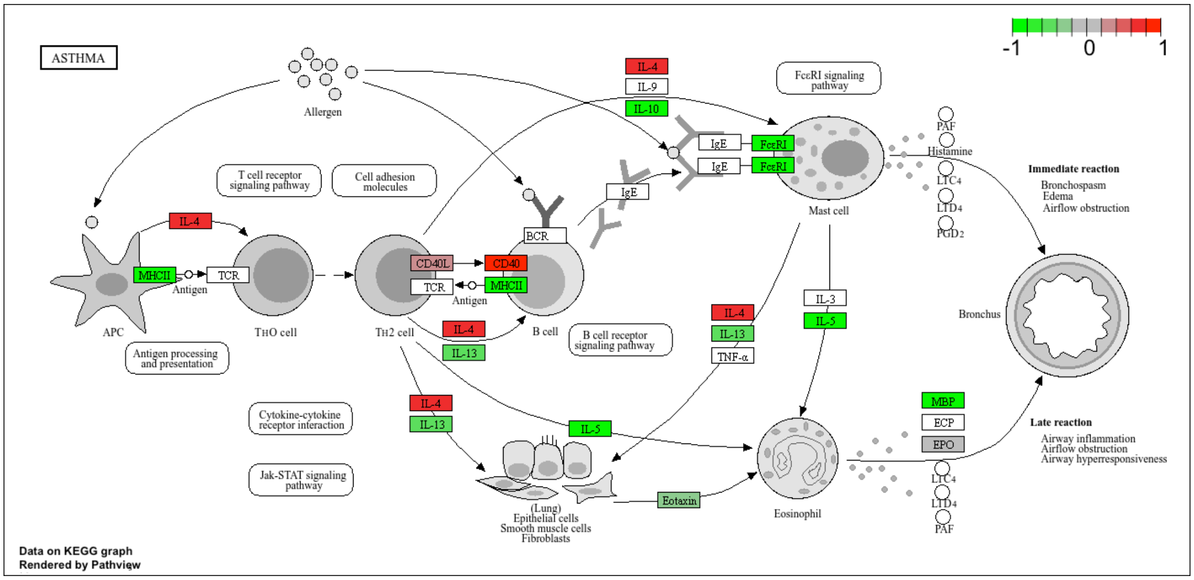


Figure 1: A pathway figure

```
write.csv(res, file="myresults.csv")
```