Class 12: RNA-Seq analysis mini-project

Taylor F. (A59010460)

2/25/2022

1. Input our counts and metadata files

- Check the format and fix if necessary

```
library(DESeq2)
library(ggplot2)
library(AnnotationDbi)
```

Input counts and metadata

```
#Read in the data and set the first column to be the row names
countData0 <- read.csv("GSE37704_featurecounts.csv", row.names = 1)
metaData <- read.csv("GSE37704_metadata.csv", row.names = 1)
#head(countData0)
head(metaData)
```

```
##                 condition
## SRR493366 control_sirna
## SRR493367 control_sirna
## SRR493368 control_sirna
## SRR493369      hoxa1_kd
## SRR493370      hoxa1_kd
## SRR493371      hoxa1_kd
```

```
#We need to get rid of the first column of countData
countData0 <- as.matrix(countData0[,-1])
head(countData0)
```

```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000186092         0         0         0         0         0         0
## ENSG00000279928         0         0         0         0         0         0
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000278566         0         0         0         0         0         0
## ENSG00000273547         0         0         0         0         0         0
## ENSG00000187634       124       123       205       207       212       258
```

```
#Now let's remove the rows that sum to 0; we can achieve that by looking through countData0, and only k
countData <- countData0[rowSums(countData0) > 0,]
head(countData)
```

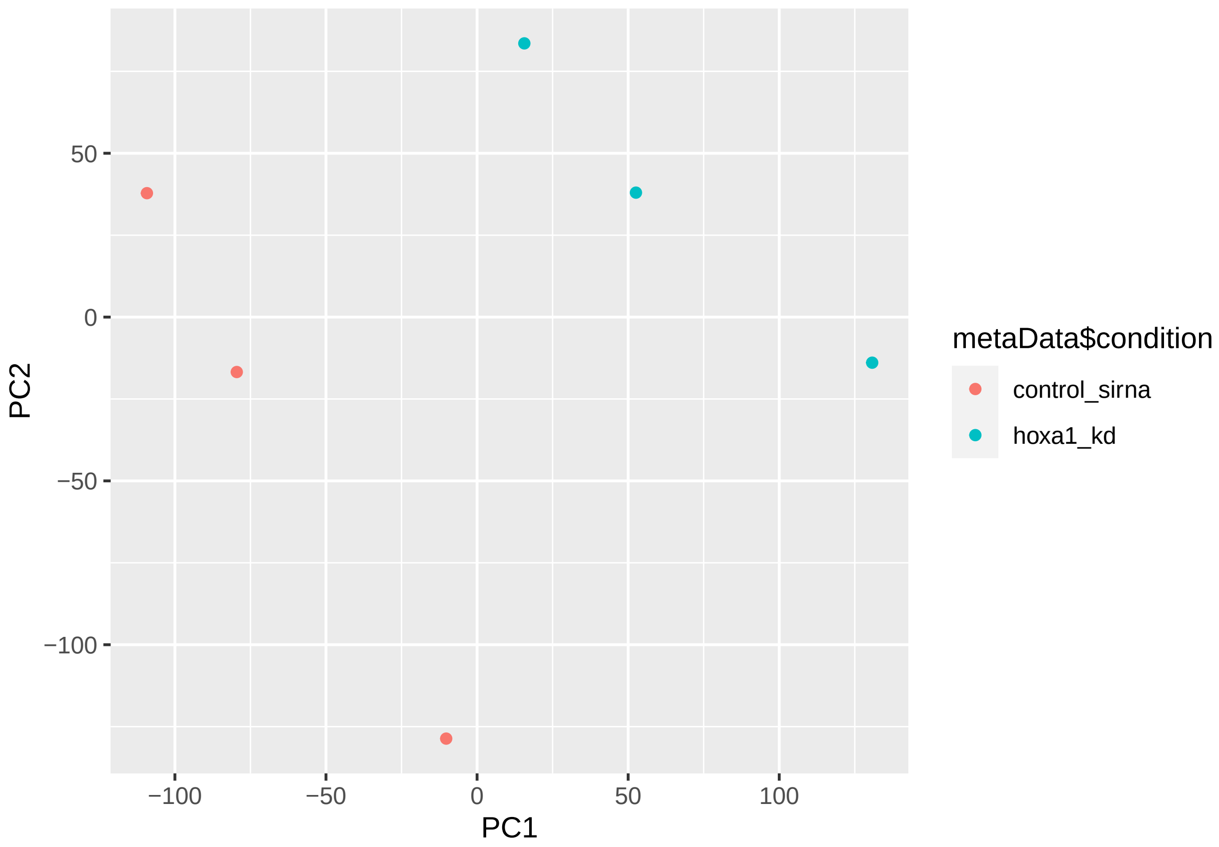
```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000187634       124       123       205       207       212       258
## ENSG00000188976      1637      1831      2383      1226      1326      1504
## ENSG00000187961       120       153       180       236       255       357
## ENSG00000187583        24        48        65        44        48        64
## ENSG00000187642         4         9        16        14        16        16
```

Running a PCA

```
#Do a PCA on countData and transpose it using t()
pca <- prcomp(t(countData), scale = TRUE)

summary(pca)
```

```
## Importance of components:
##                            PC1     PC2      PC3      PC4      PC5       PC6
## Standard deviation     87.7211 73.3196 32.89604 31.15094 29.18417 6.648e-13
## Proportion of Variance  0.4817  0.3365  0.06774  0.06074  0.05332 0.000e+00
## Cumulative Proportion   0.4817  0.8182  0.88594  0.94668  1.00000 1.000e+00
```

```
ggplot(as.data.frame(pca$x), aes(PC1, PC2, col = metaData$condition)) + geom_point()
```



2. Run differential expression analysis

- Setup that object required by DESeq()
- Run DESeq()

DESeq Analysis

Like lots of bioconductor functions, it want our data in an organized way.

```
dds <- DESeqDataSetFromMatrix(countData = countData,
                              colData = metaData,
                              design = ~condition)
```

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

```
#Run DESeq on dds
dds <- DESeq(dds)
```

```
## estimating size factors
```

```
## estimating dispersions
## gene-wise dispersion estimates
```

```
## mean-dispersion relationship
```

```
## final dispersion estimates
```

```
## fitting model and testing
```

```
#Calculate results of the dds
res <- results(dds)

head(res)
```

```
## log2 fold change (MLE): condition hoxa1 kd vs control sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 6 rows and 6 columns
##                  baseMean log2FoldChange     lfcSE       stat      pvalue
##                 <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
## ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43990e-36
## ENSG00000187961  209.6379      0.7297556 0.1318599   5.534326 3.12428e-08
## ENSG00000187583   47.2551      0.0405765 0.2718928   0.149237 8.81366e-01
## ENSG00000187642   11.9798      0.5428105 0.5215598   1.040744 2.97994e-01
##                        padj
##                   <numeric>
## ENSG00000279457 6.86555e-01
## ENSG00000187634 5.15718e-03
## ENSG00000188976 1.76549e-35
## ENSG00000187961 1.13413e-07
## ENSG00000187583 9.19031e-01
## ENSG00000187642 4.03379e-01
```

3. Add some annotation

- Gene names and Entrez IDs

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

```
##
```

```
res$symbol <- mapIds(org.Hs.eg.db, keys = row.names(countData), keytype = "ENSEMBL", column = "SYMBOL",
```

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

```
res$entrez <- mapIds(org.Hs.eg.db, keys = row.names(countData), keytype = "ENSEMBL", column = "ENTREZID
```

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

```
res$name <- mapIds(org.Hs.eg.db, keys = row.names(countData), keytype = "ENSEMBL", column = "GENENAME",
```

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

```
head(res)
```

```
## log2 fold change (MLE): condition hoxa1 kd vs control sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 6 rows and 9 columns
##                  baseMean log2FoldChange     lfcSE       stat      pvalue
##                 <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
## ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43990e-36
## ENSG00000187961  209.6379      0.7297556 0.1318599   5.534326 3.12428e-08
## ENSG00000187583   47.2551      0.0405765 0.2718928   0.149237 8.81366e-01
## ENSG00000187642   11.9798      0.5428105 0.5215598   1.040744 2.97994e-01
##                        padj      symbol      entrez                   name
##                   <numeric> <character> <character>            <character>
## ENSG00000279457 6.86555e-01      WASH9P   102723897 WAS protein family h..
## ENSG00000187634 5.15718e-03      SAMD11      148398 sterile alpha motif ..
## ENSG00000188976 1.76549e-35       NOC2L       26155 NOC2 like nucleolar ..
## ENSG00000187961 1.13413e-07      KLHL17      339451 kelch like family me..
## ENSG00000187583 9.19031e-01     PLEKHN1       84069 pleckstrin homology ..
## ENSG00000187642 4.03379e-01       PERM1       84808 PPARGC1 and ESRR ind..
```

4. Create a volcano plot

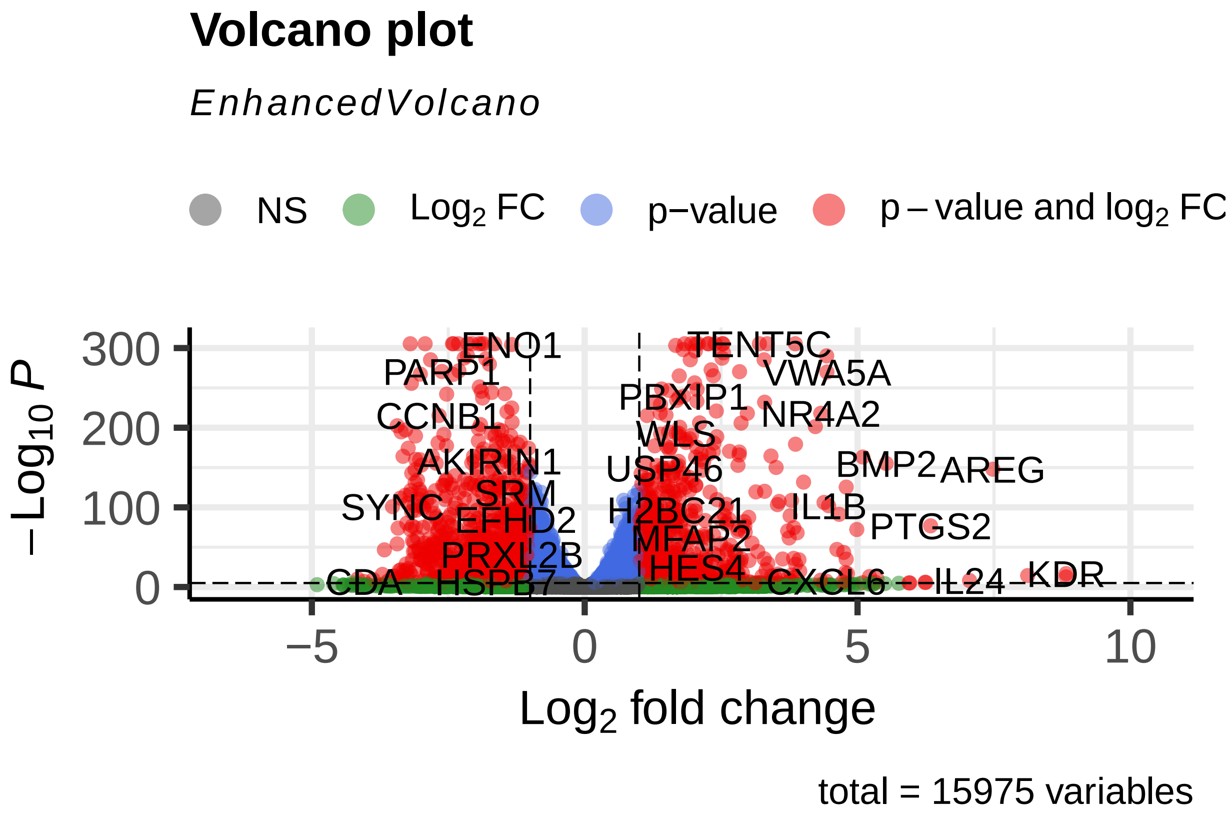
```
library(EnhancedVolcano)
```

```
## Loading required package: ggrepel
```

```
## Registered S3 methods overwritten by 'ggalt':
##   method                  from
##   grid.draw.absoluteGrob  ggplot2
##   grobHeight.absoluteGrob ggplot2
##   grobWidth.absoluteGrob  ggplot2
##   grobX.absoluteGrob      ggplot2
##   grobY.absoluteGrob      ggplot2
```

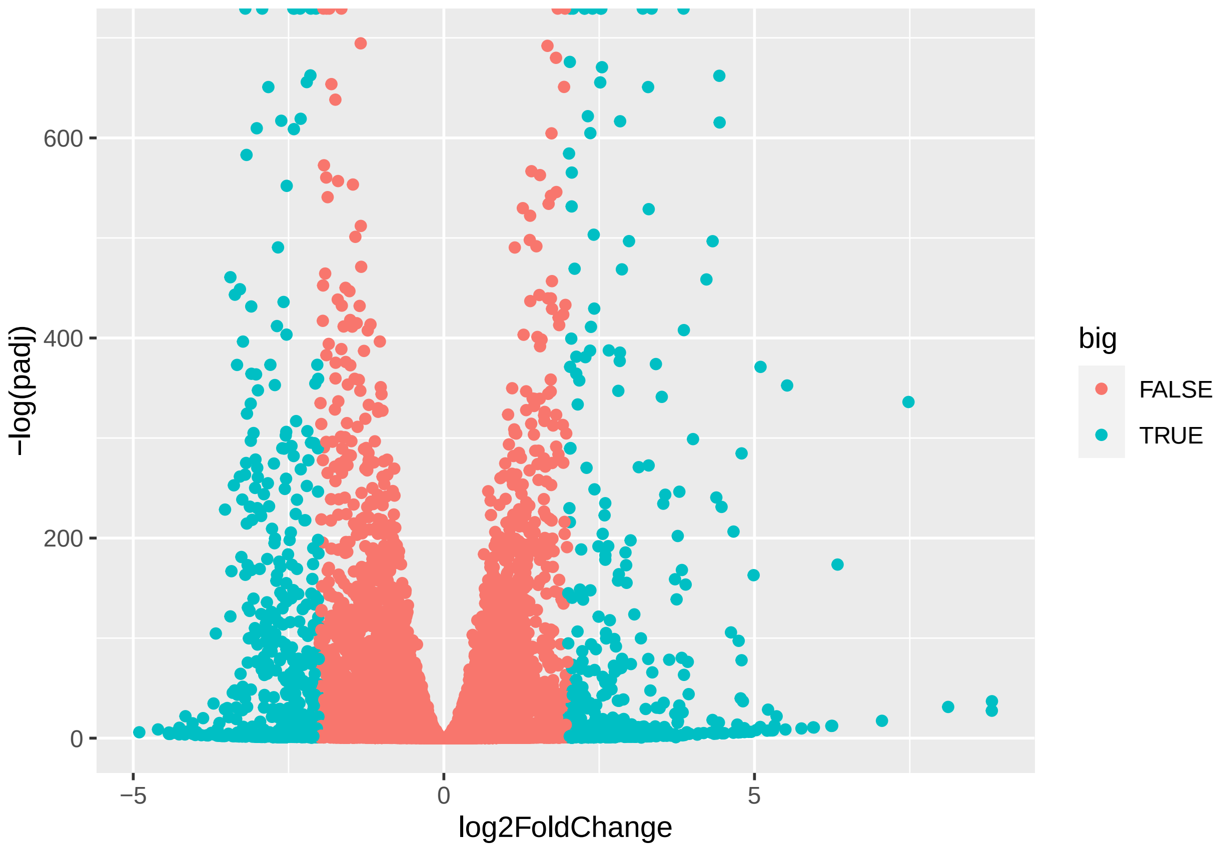
```
x <- as.data.frame(res)
x$big <- abs(res$log2FoldChange) > 2

EnhancedVolcano(x, lab = x$symbol, x = 'log2FoldChange', y = 'pvalue')
```

```
## Warning: One or more p-values is 0. Converting to 10^-1 * current lowest non-
## zero p-value...
```



```
ggplot(x, aes(log2FoldChange, -log(padj), col = big)) + geom_point()
```

```
## Warning: Removed 1237 rows containing missing values (geom_point).
```



5. Pathway analysis

```
#Load relevant packages
#Load the packages
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).
## ##############################################################################
```

```
library(gage)
```

```
##
```

```
library(gageData)

foldchange <- res$log2FoldChange
names(foldchange) <- res$entrez
```

Now we bring in the kegg dataset

```
data(kegg.sets.hs)
data(sigmet.idx.hs)

# Focus on signaling and metabolic pathways only
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]

keggres = gage(foldchange, gsets=kegg.sets.hs)
```

```
head(keggres$less, 4)
```

```
##                                      p.geomean stat.mean        p.val
## hsa04110 Cell cycle               8.995727e-06 -4.378644 8.995727e-06
## hsa03030 DNA replication          9.424076e-05 -3.951803 9.424076e-05
## hsa03013 RNA transport            1.246882e-03 -3.059466 1.246882e-03
## hsa03440 Homologous recombination 3.066756e-03 -2.852899 3.066756e-03
##                                         q.val set.size         exp1
## hsa04110 Cell cycle               0.001448312      121 8.995727e-06
## hsa03030 DNA replication          0.007586381       36 9.424076e-05
## hsa03013 RNA transport            0.066915974      144 1.246882e-03
## hsa03440 Homologous recombination 0.121861535       28 3.066756e-03
```

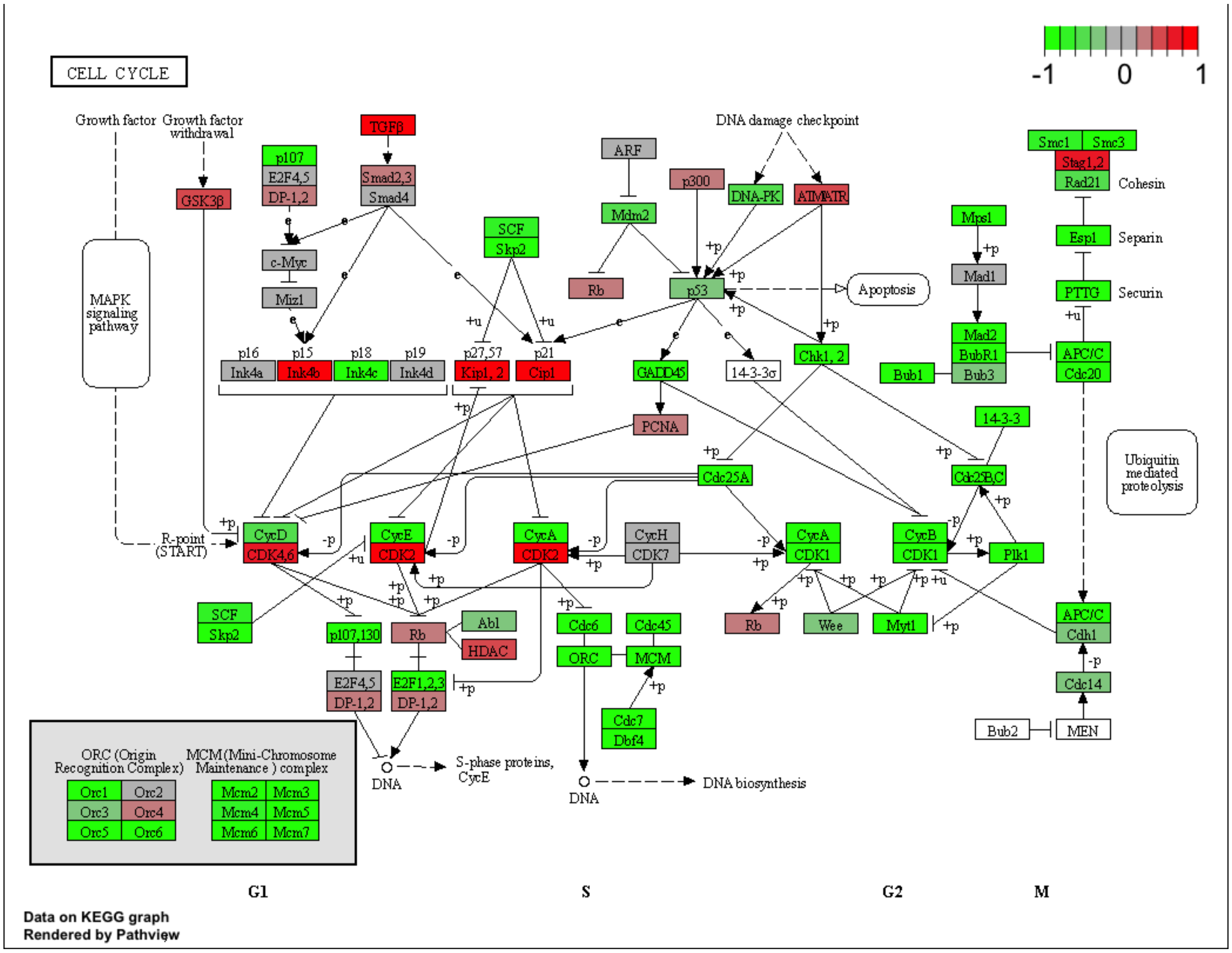
Let's pull up one of these kegg pathways with our DEGs shown.

```
pathview(gene.data = foldchange, pathway.id = "hsa04110")
```

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

```
## Info: Working in directory /Users/tforman/BGGN213/Class12
```

```
## Info: Writing image file hsa04110.pathview.png
```



Gene Ontology, Reactome

```
#Gene Ontology
data(go.sets.hs)
data(go.subs.hs)

# Focus on Biological Process subset of GO
gobpsets = go.sets.hs[go.subs.hs$BP]

gobpres = gage(foldchange, gsets=gobpsets, same.dir=TRUE)

lapply(gobpres, head)
```

```
## $greater
##                                              p.geomean stat.mean        p.val
## GO:0007156 homophilic cell adhesion       8.519724e-05  3.824205 8.519724e-05
## GO:0002009 morphogenesis of an epithelium 1.396681e-04  3.653886 1.396681e-04
## GO:0048729 tissue morphogenesis           1.432451e-04  3.643242 1.432451e-04
## GO:0007610 behavior                       2.195494e-04  3.530241 2.195494e-04
## GO:0060562 epithelial tube morphogenesis  5.932837e-04  3.261376 5.932837e-04
## GO:0035295 tube development               5.953254e-04  3.253665 5.953254e-04
##                                               q.val set.size         exp1
## GO:0007156 homophilic cell adhesion       0.1951953      113 8.519724e-05
## GO:0002009 morphogenesis of an epithelium 0.1951953      339 1.396681e-04
## GO:0048729 tissue morphogenesis           0.1951953      424 1.432451e-04
## GO:0007610 behavior                       0.2243795      427 2.195494e-04
## GO:0060562 epithelial tube morphogenesis  0.3711390      257 5.932837e-04
## GO:0035295 tube development               0.3711390      391 5.953254e-04
##
## $less
##                                             p.geomean stat.mean        p.val
## GO:0048285 organelle fission             1.536227e-15 -8.063910 1.536227e-15
## GO:0000280 nuclear division              4.286961e-15 -7.939217 4.286961e-15
## GO:0007067 mitosis                       4.286961e-15 -7.939217 4.286961e-15
## GO:0000087 M phase of mitotic cell cycle 1.169934e-14 -7.797496 1.169934e-14
## GO:0007059 chromosome segregation        2.028624e-11 -6.878340 2.028624e-11
## GO:0000236 mitotic prometaphase          1.729553e-10 -6.695966 1.729553e-10
##                                                 q.val set.size         exp1
## GO:0048285 organelle fission             5.841698e-12      376 1.536227e-15
## GO:0000280 nuclear division              5.841698e-12      352 4.286961e-15
## GO:0007067 mitosis                       5.841698e-12      352 4.286961e-15
## GO:0000087 M phase of mitotic cell cycle 1.195672e-11      362 1.169934e-14
## GO:0007059 chromosome segregation        1.658603e-08      142 2.028624e-11
## GO:0000236 mitotic prometaphase          1.178402e-07       84 1.729553e-10
##
## $stats
##                                           stat.mean     exp1
## GO:0007156 homophilic cell adhesion        3.824205 3.824205
## GO:0002009 morphogenesis of an epithelium  3.653886 3.653886
## GO:0048729 tissue morphogenesis            3.643242 3.643242
## GO:0007610 behavior                        3.530241 3.530241
## GO:0060562 epithelial tube morphogenesis   3.261376 3.261376
## GO:0035295 tube development                3.253665 3.253665
```

```
#Reactome
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj), "symbol"]
print(paste("Total number of significant genes:", length(sig_genes)))
```

```
## [1] "Total number of significant genes: 8147"
```

```
write.table(sig_genes, file="significant_genes.txt", row.names=FALSE, col.names=FALSE, quote=FALSE)
```

6. Save our results

```
write.csv(res, "DESeq_results.csv")
```

7. Go to Joshua Tree