Class12: RNA-Seq Mini Project

Kiley Hooker (PID: A15441609)

2/24/2022

Here we will work on a complete differential expression analysis project. We will use DESeq2 for this.

```
library(DESeq2)
library(ggplot2)
library(AnnotationDbi)
library(org.Hs.eg.db)
```

1. Input the vounts and metadata files

```
countData <- read.csv("GSE37704_featurecounts.csv", row.names = 1)
colData <- read.csv("GSE37704_metadata.csv")
```

Inspect these objects.

```
colData
```

```
##          id     condition
## 1 SRR493366 control_sirna
## 2 SRR493367 control_sirna
## 3 SRR493368 control_sirna
## 4 SRR493369      hoxa1_kd
## 5 SRR493370      hoxa1_kd
## 6 SRR493371      hoxa1_kd
```

```
head(countData)
```

```
##                 length SRR493366 SRR493367 SRR493368 SRR493369 SRR493370
## ENSG00000186092    918         0         0         0         0         0
## ENSG00000279928    718         0         0         0         0         0
## ENSG00000279457   1982        23        28        29        29        28
## ENSG00000278566    939         0         0         0         0         0
## ENSG00000273547    939         0         0         0         0         0
## ENSG00000187634   3214       124       123       205       207       212
##                 SRR493371
## ENSG00000186092         0
## ENSG00000279928         0
## ENSG00000279457        46
## ENSG00000278566         0
## ENSG00000273547         0
## ENSG00000187634       258
```

> Q. Complete the code below to remove the troublesome first column from countData

```
countData <- countData[,-1]
head(countData)
```

```
##                 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
```

> Q. Check on corespondence of colData and countData

```
all(colData$id == colnames(countData))
```

```
## [1] TRUE
```

> Q. Filter countData to exclude genes (i.e. rows) where we have 0 read count across all samples (i.e. columns).

```
counts <- countData[rowSums(countData) != 0, ]
head(counts)
```

```
##                 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
```

Run DESeq analysis

The steps here are to first setup the object required by DESeq using the `DESeqDataSetFromMatrix()` function. This will store the counts and metadata (i.e. colData) along with the design of the experiment (i.e. where in the metadata we have the description of what the columns of counts correspond to).

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

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

Now I can run my differential expression with `DESeq()`

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

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

Now get my results out of this dds object

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

```
## log2 fold change (MLE): condition hoxa1 kd vs control sirna 
## Wald test p-value: condition hoxa1 kd vs control sirna 
## DataFrame with 15975 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
## ...                   ...            ...       ...        ...         ...
## ENSG00000273748  35.30265       0.674387  0.303666   2.220817 2.63633e-02
## ENSG00000278817   2.42302      -0.388988  1.130394  -0.344117 7.30758e-01
## ENSG00000278384   1.10180       0.332991  1.660261   0.200565 8.41039e-01
## ENSG00000276345  73.64496      -0.356181  0.207716  -1.714752 8.63908e-02
## ENSG00000271254 181.59590      -0.609667  0.141320  -4.314071 1.60276e-05
##                        padj
##                   <numeric>
## ENSG00000279457 6.86555e-01
## ENSG00000187634 5.15718e-03
## ENSG00000188976 1.76549e-35
## ENSG00000187961 1.13413e-07
## ENSG00000187583 9.19031e-01
## ...                     ...
## ENSG00000273748 4.79091e-02
## ENSG00000278817 8.09772e-01
## ENSG00000278384 8.92654e-01
## ENSG00000276345 1.39762e-01
## ENSG00000271254 4.53648e-05
```

> Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

```
summary(res)
```

```
##
## out of 15975 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 4349, 27%
## LFC < 0 (down)     : 4396, 28%
## outliers [1]       : 0, 0%
## low counts [2]     : 1237, 7.7%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Add annotation

> Q. Use the mapIDs() function multiple times to add SYMBOL, ENTREZID and GENENAME annotation to our results by completing the code below.

```
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"
```

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

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

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

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

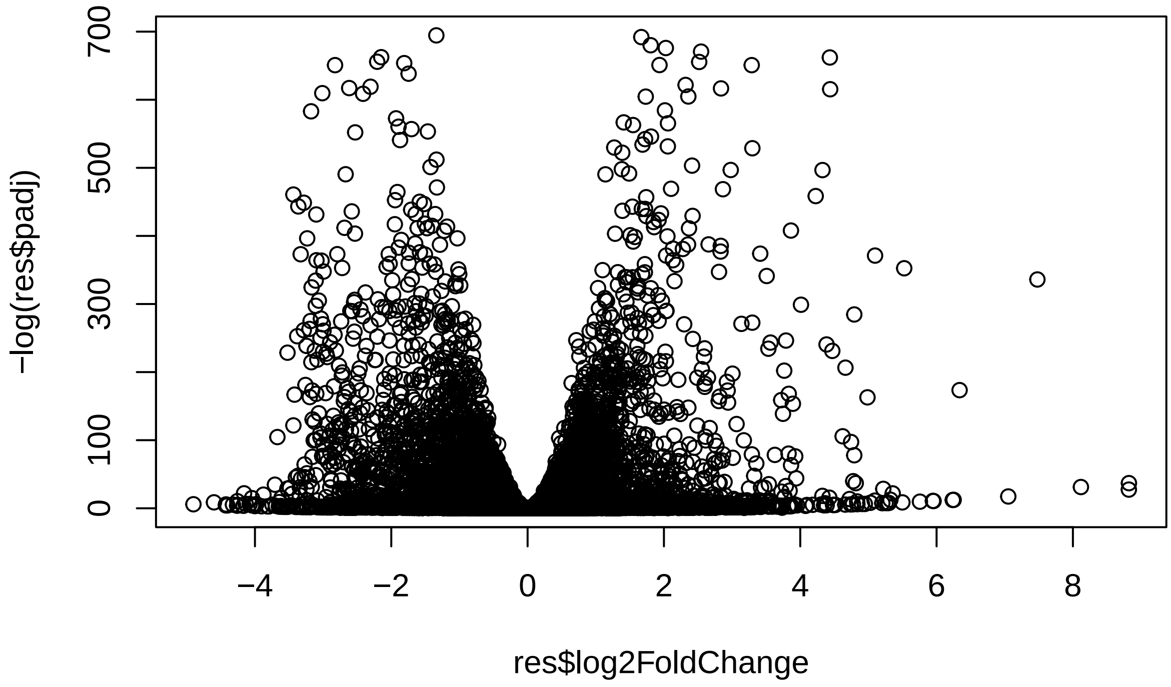
```
res$name <- mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="GENENAME",
                    multiVals="first")
```

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

Volcano Plot

Common summary figure that gives a nice overview of our resutls.

```
plot(res$log2FoldChange, -log(res$padj))
```

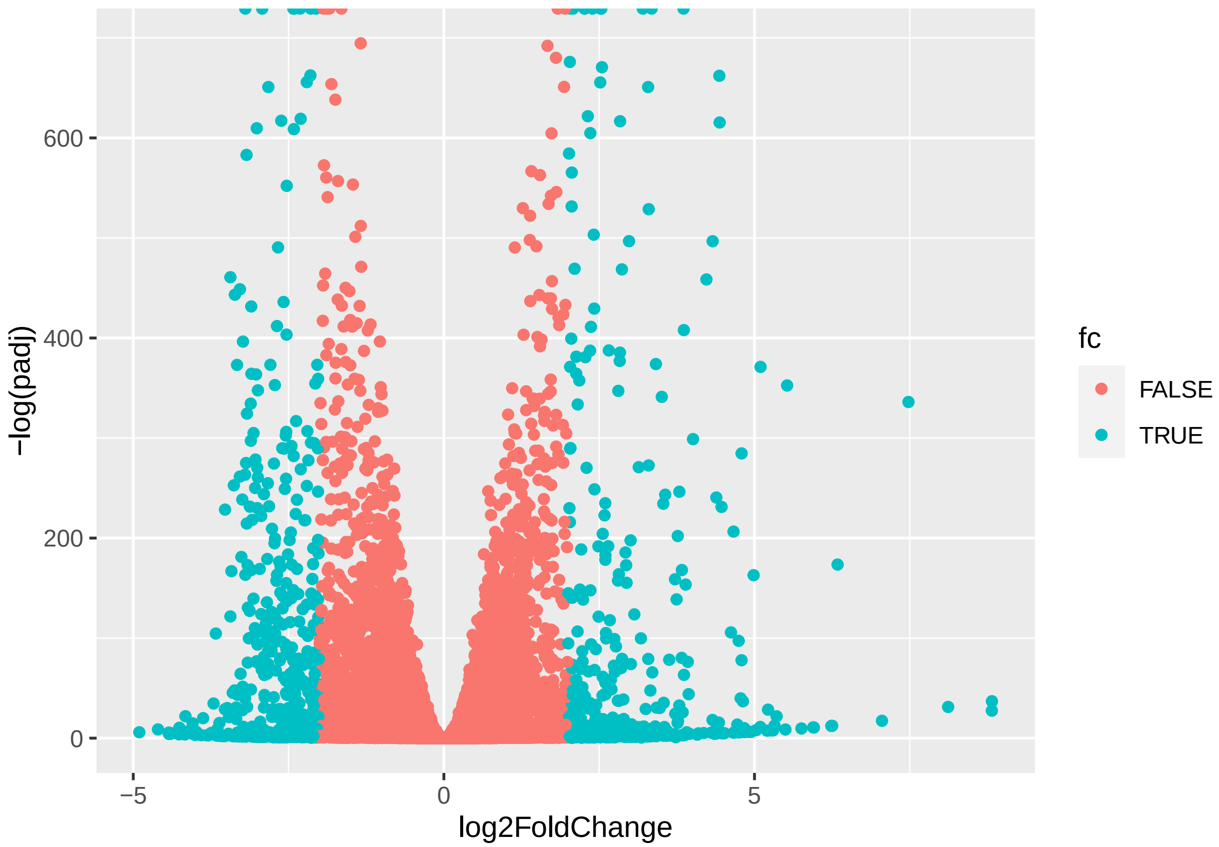


Try ggplot for this.

```
tmp <- as.data.frame(res)
tmp$fc <- abs(res$log2FoldChange) > 2

ggplot(tmp) +
  aes(log2FoldChange, -log(padj), col=fc) +
  geom_point()
```

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



Try the EnhancedVolcano package from bioconductor.

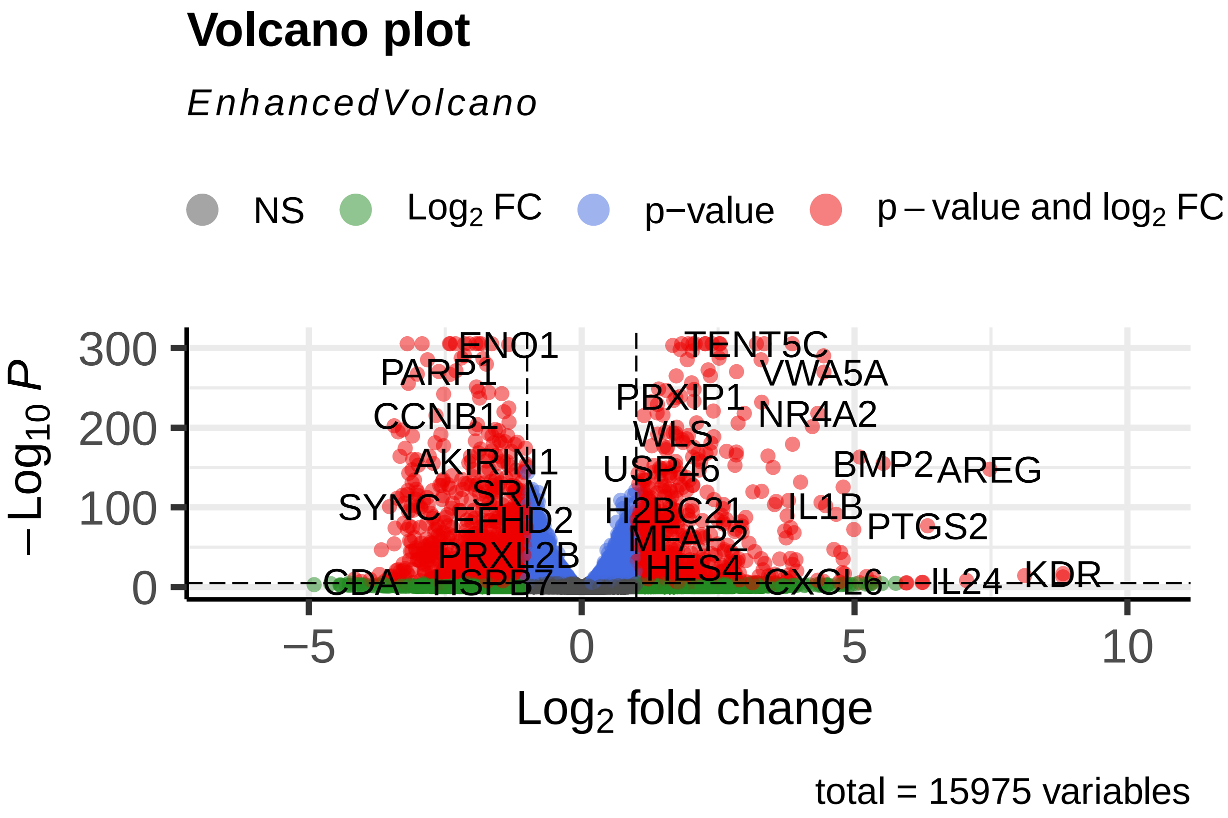
```
library(EnhancedVolcano)
```

```
## 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
```

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

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



Pathway analysis and gene set enrichment

Here we try to bring back the biology and help with the interpretation results. We try to answer the question: which pathways and functions feature heavily in our differentially expressed genes.

Recall that we need a "vector of importance" as input for GAGE that has ENTREZ ids set as the names attribute.

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

```
library(pathview)
library(gage)
library(gageData)

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

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

Look at the first 3=2 down-regulated pathways

```r
# Look at the first few down (less) pathways
head(keggres$less, 2)
```

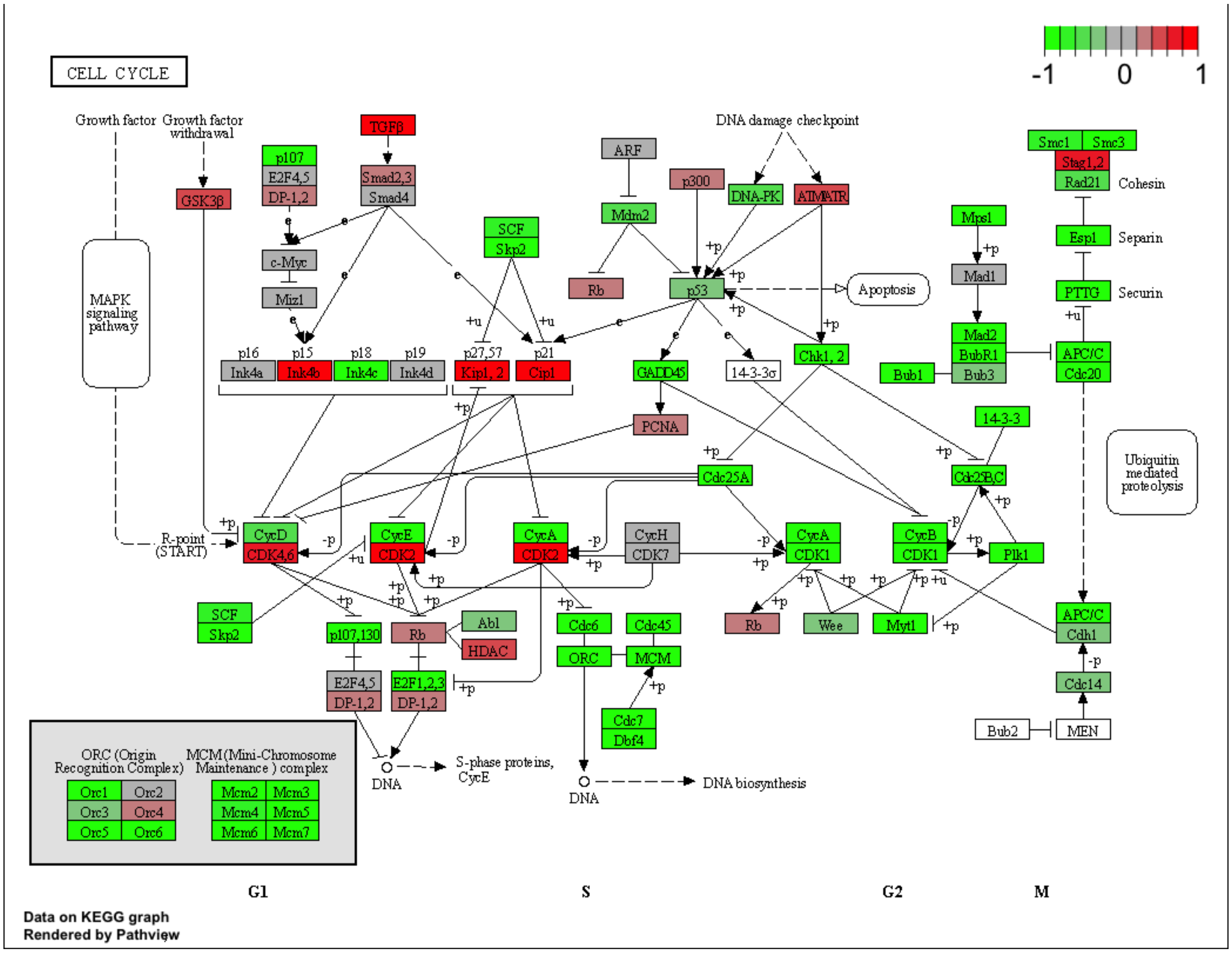
```
##                              p.geomean stat.mean        p.val       q.val
## hsa04110 Cell cycle      8.995727e-06 -4.378644 8.995727e-06 0.001889103
## hsa03030 DNA replication 9.424076e-05 -3.951803 9.424076e-05 0.009841047
##                          set.size         exp1
## hsa04110 Cell cycle           121 8.995727e-06
## hsa03030 DNA replication       36 9.424076e-05
```

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

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

## Info: Working in directory /Users/kileyhooker/Desktop/BIMM143/week8

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



Gene Ontology (GO)

```r
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

We can use Reactome either as an R package (just like above) or we can use the website. The website needs a file of “gene importance” just like gage above.

Reactome is database consisting of biological molecules and their relation to pathways and processes. Reactome, such as many other tools, has an online software available (https://reactome.org/) and R package available (https://bioconductor.org/packages/release/bioc/html/ReactomePA.html).

```
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)
```

Save my results

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

```
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats4    stats     graphics  grDevices utils     datasets  methods
## [8] base
##
## other attached packages:
##  [1] gageData_2.32.0             gage_2.44.0
##  [3] pathview_1.34.0             EnhancedVolcano_1.12.0
##  [5] ggrepel_0.9.1               org.Hs.eg.db_3.14.0
##  [7] AnnotationDbi_1.56.2        ggplot2_3.3.5
##  [9] DESeq2_1.34.0               SummarizedExperiment_1.24.0
## [11] Biobase_2.54.0              MatrixGenerics_1.6.0
## [13] matrixStats_0.61.0          GenomicRanges_1.46.1
## [15] GenomeInfoDb_1.30.1         IRanges_2.28.0
## [17] S4Vectors_0.32.3            BiocGenerics_0.40.0
##
## loaded via a namespace (and not attached):
##  [1] bitops_1.0-7           bit64_4.0.5            ash_1.0-15
##  [4] RColorBrewer_1.1-2     httr_1.4.2             Rgraphviz_2.38.0
##  [7] tools_4.1.2            utf8_1.2.2             R6_2.5.1
## [10] vipor_0.4.5            KernSmooth_2.23-20     DBI_1.1.2
## [13] colorspace_2.0-2       withr_2.4.3            tidyselect_1.1.1
## [16] ggrastr_1.0.1          ggalt_0.4.0            bit_4.0.4
## [19] compiler_4.1.2         extrafontdb_1.0        graph_1.72.0
## [22] cli_3.1.1              DelayedArray_0.20.0    labeling_0.4.2
## [25] KEGGgraph_1.54.0       scales_1.1.1           proj4_1.0-11
## [28] genefilter_1.76.0      stringr_1.4.0          digest_0.6.29
## [31] rmarkdown_2.11         XVector_0.34.0         pkgconfig_2.0.3
## [34] htmltools_0.5.2        extrafont_0.17         fastmap_1.1.0
## [37] highr_0.9              maps_3.4.0             rlang_1.0.0
## [40] rstudioapi_0.13        RSQLite_2.2.10         generics_0.1.2
## [43] farver_2.1.0           BiocParallel_1.28.3    dplyr_1.0.8
## [46] RCurl_1.98-1.6         magrittr_2.0.2         GO.db_3.14.0
## [49] GenomeInfoDbData_1.2.7 Matrix_1.3-4           Rcpp_1.0.8
## [52] ggbeeswarm_0.6.0       munsell_0.5.0          fansi_1.0.2
## [55] lifecycle_1.0.1        stringi_1.7.6          yaml_2.2.2
## [58] MASS_7.3-54            zlibbioc_1.40.0        grid_4.1.2
## [61] blob_1.2.2             parallel_4.1.2         crayon_1.4.2
## [64] lattice_0.20-45        Biostrings_2.62.0      splines_4.1.2
## [67] annotate_1.72.0        KEGGREST_1.34.0        locfit_1.5-9.4
## [70] knitr_1.37             pillar_1.7.0           geneplotter_1.72.0
## [73] XML_3.99-0.8           glue_1.6.1             evaluate_0.14
## [76] png_0.1-7              vctrs_0.3.8            Rttf2pt1_1.3.10
## [79] gtable_0.3.0           purrr_0.3.4            cachem_1.0.6
## [82] xfun_0.29              xtable_1.8-4           survival_3.2-13
## [85] tibble_3.1.6           beeswarm_0.4.0         memoise_2.0.1
## [88] ellipsis_0.3.2
```