Class 13: RNA Mini Sequence

Jenny

The authors report on differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOXA1

##RNASeq input data

Again I need two things colData countData

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

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

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

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

```
                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
```

There is an unwanted first column "length" in the countData. I will need to remove this first before going on to further analysis

> Q. Complete the code below to filter countData to exclude genes (i.e. rows) where we have 0 read count across all samples (i.e. columns).

```
countData <- as.matrix(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
```

```r
all(colnames(counts)== rownames(colData))
```

```
Warning in colnames(counts) == rownames(colData): longer object length is not a
multiple of shorter object length
```

```
[1] FALSE
```

Time to use DESeq

```r
library(DESeq2)
```

```
Loading required package: S4Vectors
```

```
Loading required package: stats4
```

```
Loading required package: BiocGenerics
```

```


Attaching package: 'BiocGenerics'
```

```
The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs
```

```
The following objects are masked from 'package:base':

    anyDuplicated, append, as.data.frame, basename, cbind, colnames,
    dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
    grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
    order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
    rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
    union, unique, unsplit, which.max, which.min
```

```


Attaching package: 'S4Vectors'
```

```
The following objects are masked from 'package:base':

    expand.grid, I, unname

Loading required package: IRanges


Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':

    windows

Loading required package: GenomicRanges

Loading required package: GenomeInfoDb

Loading required package: SummarizedExperiment

Loading required package: MatrixGenerics

Loading required package: matrixStats

Warning: package 'matrixStats' was built under R version 4.2.2


Attaching package: 'MatrixGenerics'

The following objects are masked from 'package:matrixStats':

    colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
    colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
    colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
    colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
    colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
    colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
    colWeightedMeans, colWeightedMedians, colWeightedSds,
    colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
    rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
    rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
    rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
    rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
    rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
    rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
    rowWeightedSds, rowWeightedVars
```

```
Loading required package: Biobase
```

```
Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
Attaching package: 'Biobase'
```

```
The following object is masked from 'package:MatrixGenerics':

    rowMedians
```

```
The following objects are masked from 'package:matrixStats':

    anyMissing, rowMedians
```

First step would be to setup the object required by DESeq

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

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

Run the analysis

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

```
estimating size factors
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

```
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>
ENSG00000186092    0.0000             NA        NA        NA         NA
ENSG00000279928    0.0000             NA        NA        NA         NA
ENSG00000279457   29.9136       0.179257  0.324822  0.551863 0.58104205
ENSG00000278566    0.0000             NA        NA        NA         NA
ENSG00000273547    0.0000             NA        NA        NA         NA
ENSG00000187634  183.2296       0.426457  0.140266  3.040350 0.00236304
                       padj
                  <numeric>
ENSG00000186092          NA
ENSG00000279928          NA
ENSG00000279457 0.68707978
ENSG00000278566          NA
ENSG00000273547          NA
ENSG00000187634 0.00516278
```

```
res = results(dds, contrast=c("condition", "hoxa1_kd", "control_sirna"))
```

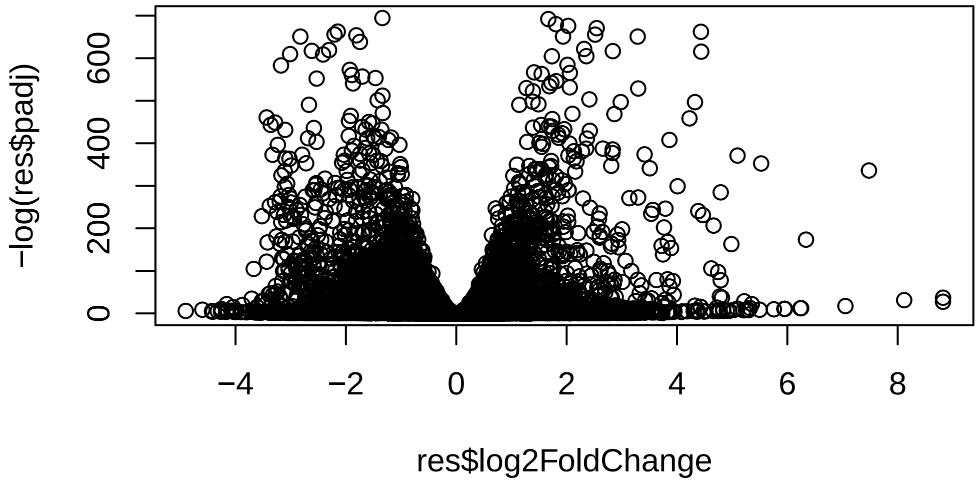
> Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

```r
summary(res)
```

```
out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4349, 27%
LFC < 0 (down)     : 4393, 27%
outliers [1]       : 0, 0%
low counts [2]     : 1221, 7.6%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Volcano Plot

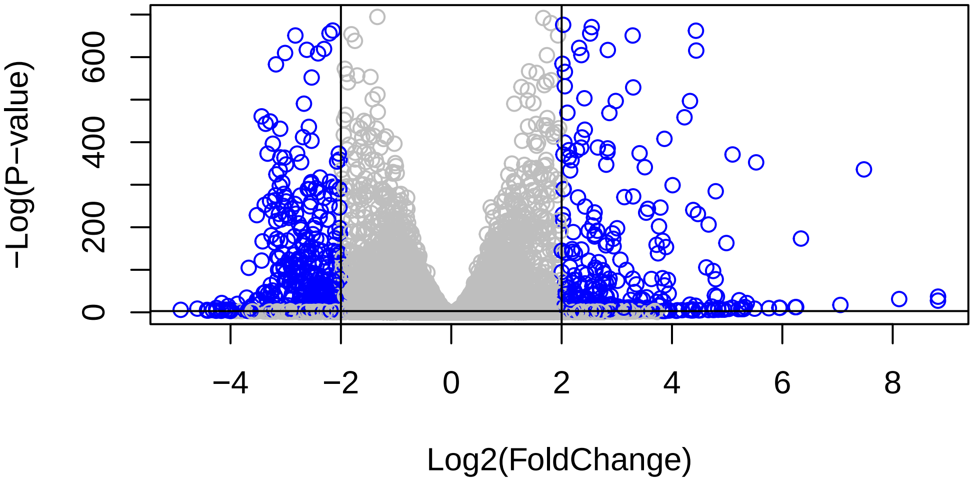
```r
plot( res$log2FoldChange, -log(res$padj) )
```



I want to add some color. Take a fold-change threshold of -2/+2 and an alpha p-adj (P-value) threshold of 0.05

Q. Improve this plot by completing the below code, which adds color and axis labels

```
mycols <- rep("gray", nrow(res) )
mycols[ abs(res$log2FoldChange) > 2 ] <- "blue"
mycols[res$padj > 0.05] <- "gray"
plot( res$log2FoldChange, -log(res$padj), col= mycols , xlab="Log2(FoldChange)", ylab="-Lo
abline(v=c(-2,+2))
abline(h=-log(0.05))
```



Adding gene annotation

I am going to add the database identifiers I need for pathway analysis here

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

```
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=rownames(res),
                    keytype="ENSEMBL",
                    column="SYMBOL",
                    multiVals="first")
```

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

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

```
'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 8 columns
                 baseMean log2FoldChange     lfcSE      stat     pvalue
                <numeric>      <numeric> <numeric> <numeric>  <numeric>
ENSG00000186092    0.0000             NA        NA        NA         NA
ENSG00000279928    0.0000             NA        NA        NA         NA
ENSG00000279457   29.9136       0.179257  0.324822  0.551863 0.58104205
ENSG00000278566    0.0000             NA        NA        NA         NA
ENSG00000273547    0.0000             NA        NA        NA         NA
```

```
ENSG00000187634  183.2296       0.426457  0.140266  3.040350 0.00236304
                      padj      symbol      entrez
                 <numeric> <character> <character>
ENSG00000186092         NA       OR4F5       79501
ENSG00000279928         NA          NA          NA
ENSG00000279457 0.68707978          NA          NA
ENSG00000278566         NA          NA          NA
ENSG00000273547         NA          NA          NA
ENSG00000187634 0.00516278      SAMD11      148398
```

> Q. Finally for this section let's reorder these results by adjusted p-value and save them to a CSV file in your current project directory.

Save my results so far to a CSV file

```r
res = res[order(res$pvalue),]
write.csv(res, file="deseq_results.csv")
```

Pathway Analysis

Again we will use the 'gage()' package & function with a focus first on KEGG &

```r
library(gage)
```

```r
library(gageData)

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

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

Recall that 'gage()' wants only a vector of importance as input that has names in ENTREZ ID format

```r
foldchanges = res$log2FoldChange
names(foldchanges) = res$entrez
head(foldchanges)
```

```
     1266     54855      1465     51232      2034      2317 
-2.422719  3.201955 -2.313738 -2.059631 -1.888019 -1.649792 
```

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

```
head(keggres$less , 5)
```

```
                                     p.geomean stat.mean        p.val
hsa04110 Cell cycle               7.077982e-06 -4.432593 7.077982e-06
hsa03030 DNA replication          9.424076e-05 -3.951803 9.424076e-05
hsa03013 RNA transport            1.121279e-03 -3.090949 1.121279e-03
hsa04114 Oocyte meiosis           2.563806e-03 -2.827297 2.563806e-03
hsa03440 Homologous recombination 3.066756e-03 -2.852899 3.066756e-03
                                        q.val set.size         exp1
hsa04110 Cell cycle               0.001160789      124 7.077982e-06
hsa03030 DNA replication          0.007727742       36 9.424076e-05
hsa03013 RNA transport            0.061296597      150 1.121279e-03
hsa04114 Oocyte meiosis           0.100589607      112 2.563806e-03
hsa03440 Homologous recombination 0.100589607       28 3.066756e-03
```

Generate a colored pathway firgure for hsa04110 Cell cycle

```
library(pathview)
```

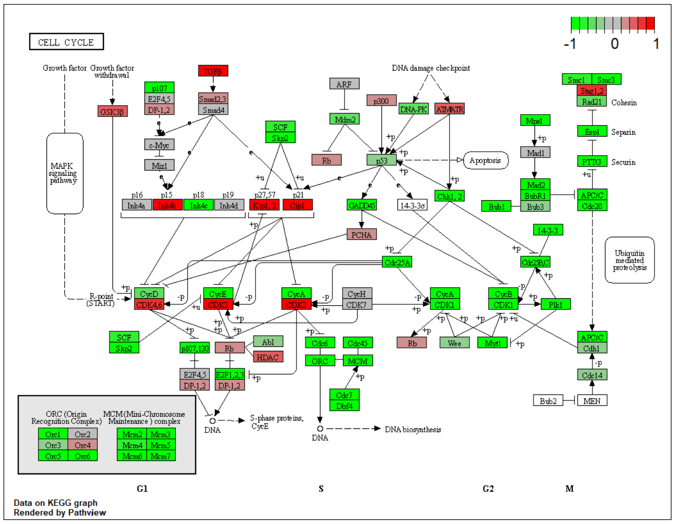
```
##############################################################################
Pathview is an open source software package distributed under GNU General
Public License version 3 (GPLv3). Details of GPLv3 is available at
http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
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).
##############################################################################
```

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

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

```
Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13
```

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



```r
keggrespathways <- rownames(keggres$greater)[1:5]

# Extract the 8 character long IDs part of each string
keggresids = substr(keggrespathways, start=1, stop=8)
keggresids
```

```
[1] "hsa04740" "hsa04640" "hsa00140" "hsa04630" "hsa04976"
```

```r
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

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

```
Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13
```

```
Info: Writing image file hsa04740.pathview.png
```

```
Info: some node width is different from others, and hence adjusted!
```

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

```
Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04976.pathview.png
```

> Q. Can you do the same procedure as above to plot the pathview figures for the top 5 down-reguled pathways?

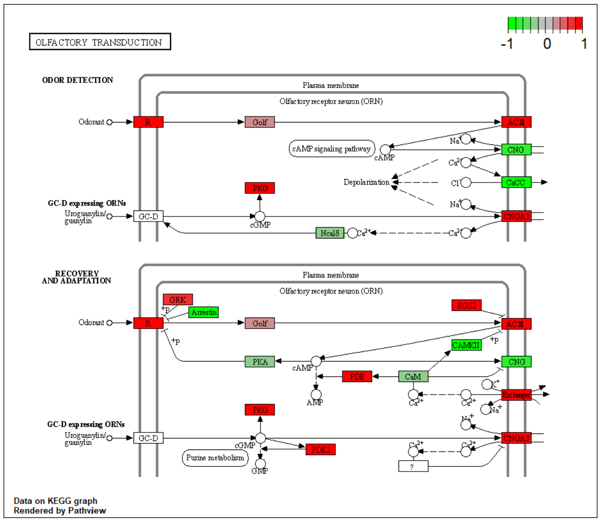
```r
keggrespathway <- rownames(keggres$less)[1:5]

# Extract the 8 character long IDs part of each string
keggresids = substr(keggrespathways, start=1, stop=8)
keggresids
```

```
[1] "hsa04740" "hsa04640" "hsa00140" "hsa04630" "hsa04976"
```

```r
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

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

```
Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04740.pathview.png

Info: some node width is different from others, and hence adjusted!

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/Users/Jennifer/Downloads/BIMM143/CLASS13

Info: Writing image file hsa04976.pathview.png
```



Gene Ontology (GO)

```
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(foldchanges, gsets=gobpsets, same.dir=TRUE)

lapply(gobpres, head)
```

```
$greater
                                               p.geomean stat.mean        p.val
GO:0007156 homophilic cell adhesion         1.624062e-05  4.226117 1.624062e-05
GO:0048729 tissue morphogenesis             5.407952e-05  3.888470 5.407952e-05
GO:0002009 morphogenesis of an epithelium   5.727599e-05  3.878706 5.727599e-05
GO:0030855 epithelial cell differentiation  2.053700e-04  3.554776 2.053700e-04
GO:0060562 epithelial tube morphogenesis    2.927804e-04  3.458463 2.927804e-04
GO:0048598 embryonic morphogenesis          2.959270e-04  3.446527 2.959270e-04
                                                 q.val set.size         exp1
GO:0007156 homophilic cell adhesion         0.07103646      138 1.624062e-05
GO:0048729 tissue morphogenesis             0.08350839      483 5.407952e-05
GO:0002009 morphogenesis of an epithelium   0.08350839      382 5.727599e-05
GO:0030855 epithelial cell differentiation  0.15370245      299 2.053700e-04
```

```
GO:0060562 epithelial tube morphogenesis   0.15370245      289 2.927804e-04
GO:0048598 embryonic morphogenesis         0.15370245      498 2.959270e-04

$less
                                            p.geomean stat.mean        p.val
GO:0048285 organelle fission             6.626774e-16 -8.170439 6.626774e-16
GO:0000280 nuclear division              1.797050e-15 -8.051200 1.797050e-15
GO:0007067 mitosis                       1.797050e-15 -8.051200 1.797050e-15
GO:0000087 M phase of mitotic cell cycle 4.757263e-15 -7.915080 4.757263e-15
GO:0007059 chromosome segregation        1.081862e-11 -6.974546 1.081862e-11
GO:0051301 cell division                 8.718528e-11 -6.455491 8.718528e-11
                                                q.val set.size         exp1
GO:0048285 organelle fission             2.620099e-12      386 6.626774e-16
GO:0000280 nuclear division              2.620099e-12      362 1.797050e-15
GO:0007067 mitosis                       2.620099e-12      362 1.797050e-15
GO:0000087 M phase of mitotic cell cycle 5.202068e-12      373 4.757263e-15
GO:0007059 chromosome segregation        9.464127e-09      146 1.081862e-11
GO:0051301 cell division                 6.355807e-08      479 8.718528e-11

$stats
                                           stat.mean     exp1
GO:0007156 homophilic cell adhesion         4.226117 4.226117
GO:0048729 tissue morphogenesis             3.888470 3.888470
GO:0002009 morphogenesis of an epithelium   3.878706 3.878706
GO:0030855 epithelial cell differentiation  3.554776 3.554776
GO:0060562 epithelial tube morphogenesis    3.458463 3.458463
GO:0048598 embryonic morphogenesis          3.446527 3.446527
```

Reactome Analysis

```r
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: 8146"
```

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

Q. What pathway has the most significant "Entities p-value"? Do the most significant pathways listed match your previous KEGG results? What factors could cause differences between the two methods? The pathway that has the most significant “Entities p-value” is the Endosomal / Vascuolar pathway. Does not match my previous KEGG results because for that one, the most signifigant one was the cell cycle. The factors that could have caused the differences between the two methods could have been the data used and the IDs we have obtained / used