Lung Fibroblast in response to HOXA1

James Lee

12/6/2020

1. Differential Expression Analysis

```
metaFile <- "data/GSE37704_metadata.csv"
countFile <- "data/GSE37704_featurecounts.csv"

# Import metadata and take a peak
colData = read.csv(metaFile, row.names=1)
head(colData)
```

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

```
# Import countdata
countData = read.csv(countFile, 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
```

```
# Note we need to remove the odd first $length col
countData <- as.matrix(countData[,!(colnames(countData) %in% c("length"))])
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
```

```
#Eliminate rows with 0 value
countData2 = countData[(rowSums(countData)!= 0), ]
head(countData2)
```

```
##                 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 DESeq2

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

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```
dds
```

```
## class: DESeqDataSet
## dim: 19808 6
## metadata(1): version
## assays(4): counts mu H cooks
## rownames(19808): ENSG00000186092 ENSG00000279928 ... ENSG00000277475
##   ENSG00000268674
## rowData names(22): baseMean baseVar ... deviance maxCooks
## colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
## colData names(2): condition sizeFactor
```

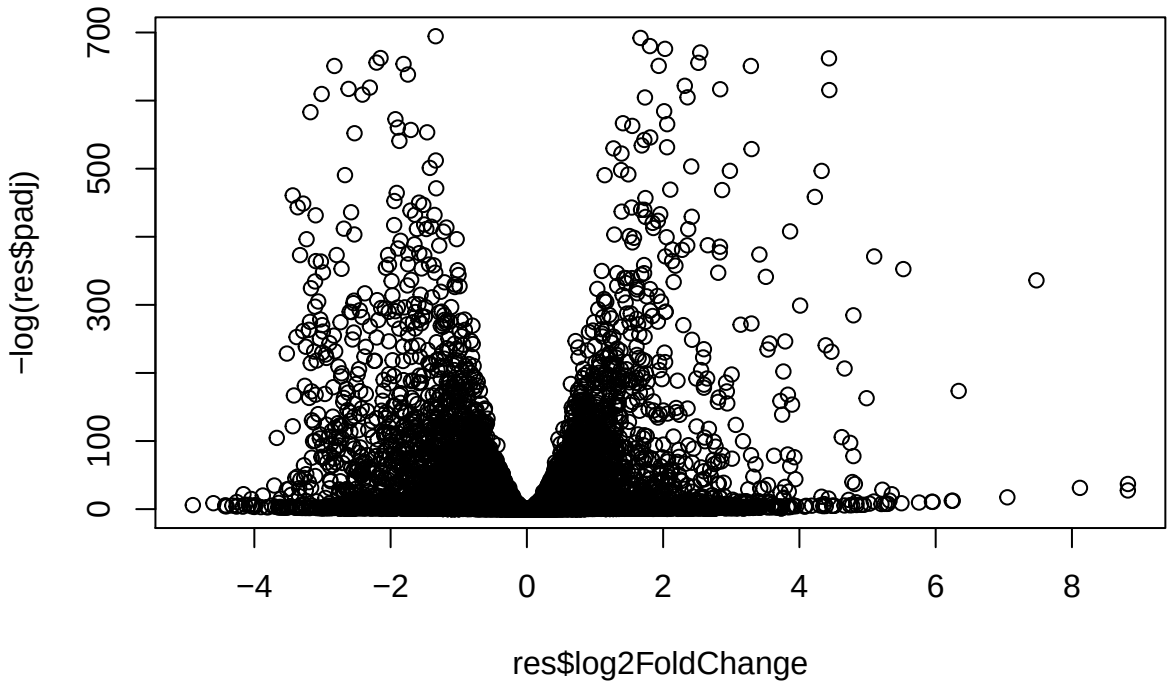
```
res = results(dds, contrast=c("condition", "hoxa1_kd", "control_sirna"))
```

```
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

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



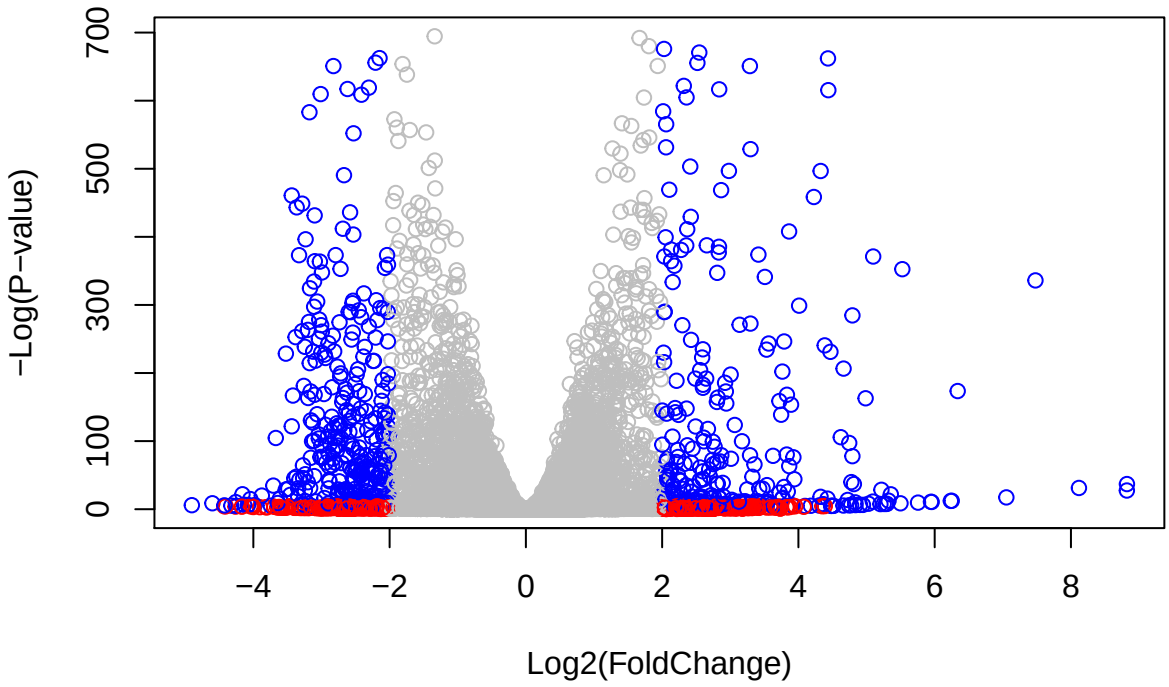
Improved volcano plot

```
# Make a color vector for all genes
mycols <- rep("gray", nrow(res) )
```

```
# Color red the genes with absolute fold change above 2
mycols[ abs(res$log2FoldChange) > 2 ] <- "red"

# Color blue those with adjusted p-value less than 0.01
#  and absolute fold change more than 2
inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"

plot( res$log2FoldChange, -log(res$padj), col=mycols, xlab="Log2(FoldChange)", ylab="-Log(P-value)" )
```



Adding gene annotation

```
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] "GO"           "GOALL"        "IPI"          "MAP"          "OMIM"
## [16] "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"         "PMID"
## [21] "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"       "UNIGENE"
## [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
```

```
head(res, 10)
```

```
## log2 fold change (MLE): condition hoxa1_kd vs control_sirna
## Wald test p-value: condition hoxa1 kd vs control sirna
## DataFrame with 10 rows and 9 columns
##                         baseMean     log2FoldChange             lfcSE
##                        <numeric>          <numeric>         <numeric>
## ENSG00000186092                0                 NA                NA
## ENSG00000279928                0                 NA                NA
## ENSG00000279457 29.9135794276176   0.17925708367269  0.324821565250145
## ENSG00000278566                0                 NA                NA
## ENSG00000273547                0                 NA                NA
## ENSG00000187634 183.229649921658  0.426457118403306  0.140265820376892
## ENSG00000188976 1651.18807619944 -0.692720464846366 0.0548465415913946
## ENSG00000187961 209.637938486147  0.729755610585225  0.131859899969345
## ENSG00000187583 47.2551232589398 0.0405765278756312  0.271892808601774
## ENSG00000187642 11.9797501642461  0.542810491577363  0.521559849534146
##                            stat               pvalue                 padj
##                       <numeric>            <numeric>            <numeric>
## ENSG00000186092              NA                   NA                   NA
## ENSG00000279928              NA                   NA                   NA
## ENSG00000279457 0.551863246932648    0.581042050747032    0.687079780133182
## ENSG00000278566              NA                   NA                   NA
## ENSG00000273547              NA                   NA                   NA
## ENSG00000187634  3.04034951107421  0.00236303749730996   0.0051627802806621
## ENSG00000188976 -12.6301576133481 1.43989540156582e-36 1.76740572002514e-35
## ENSG00000187961  5.53432552849563  3.1242824807768e-08 1.13536117540347e-07
```

```
## ENSG00000187583 0.149237223611387    0.881366448669148    0.918988027114106
## ENSG00000187642  1.04074439790984    0.297994191720983    0.403817230025208
##                      symbol      entrez
##                 <character> <character>
## ENSG00000186092       OR4F5       79501
## ENSG00000279928          NA          NA
## ENSG00000279457          NA          NA
## ENSG00000278566          NA          NA
## ENSG00000273547          NA          NA
## ENSG00000187634      SAMD11      148398
## ENSG00000188976       NOC2L       26155
## ENSG00000187961      KLHL17      339451
## ENSG00000187583     PLEKHN1       84069
## ENSG00000187642       PERM1       84808
##                                                                    name
##                                                             <character>
## ENSG00000186092         olfactory receptor family 4 subfamily F member 5
## ENSG00000279928                                                      NA
## ENSG00000279457                                                      NA
## ENSG00000278566                                                      NA
## ENSG00000273547                                                      NA
## ENSG00000187634                 sterile alpha motif domain containing 11
## ENSG00000188976 NOC2 like nucleolar associated transcriptional repressor
## ENSG00000187961                              kelch like family member 17
## ENSG00000187583                 pleckstrin homology domain containing N1
## ENSG00000187642             PPARGC1 and ESRR induced regulator, muscle 1
```

```
#Export the datawframe to csv file
res = res[order(res$pvalue),]
write.csv(res, "deseq_results.csv")
```

2. Pathway Analysis

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

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

# Examine the first 3 pathways
head(kegg.sets.hs, 3)
```

```
## $`hsa00232 Caffeine metabolism`
## [1] "10"   "1544" "1548" "1549" "1553" "7498" "9"
##
## $`hsa00983 Drug metabolism - other enzymes`
##  [1] "10"     "1066"   "10720"  "10941"  "151531" "1548"   "1549"   "1551"
##  [9] "1553"   "1576"   "1577"   "1806"   "1807"   "1890"   "221223" "2990"
## [17] "3251"   "3614"   "3615"   "3704"   "51733"  "54490"  "54575"  "54576"
## [25] "54577"  "54578"  "54579"  "54600"  "54657"  "54658"  "54659"  "54963"
## [33] "574537" "64816"  "7083"   "7084"   "7172"   "7363"   "7364"   "7365"
## [41] "7366"   "7367"   "7371"   "7372"   "7378"   "7498"   "79799"  "83549"
```

```
## [49] "8824"   "8833"   "9"      "978"
##
## $`hsa00230 Purine metabolism`
##   [1] "100"    "10201"  "10606"  "10621"  "10622"  "10623"  "107"    "10714"
##   [9] "108"    "10846"  "109"    "111"    "11128"  "11164"  "112"    "113"
##  [17] "114"    "115"    "122481" "122622" "124583" "132"    "158"    "159"
##  [25] "1633"   "171568" "1716"   "196883" "203"    "204"    "205"    "221823"
##  [33] "2272"   "22978"  "23649"  "246721" "25885"  "2618"   "26289"  "270"
##  [41] "271"    "27115"  "272"    "2766"   "2977"   "2982"   "2983"   "2984"
##  [49] "2986"   "2987"   "29922"  "3000"   "30833"  "30834"  "318"    "3251"
##  [57] "353"    "3614"   "3615"   "3704"   "377841" "471"    "4830"   "4831"
##  [65] "4832"   "4833"   "4860"   "4881"   "4882"   "4907"   "50484"  "50940"
##  [73] "51082"  "51251"  "51292"  "5136"   "5137"   "5138"   "5139"   "5140"
##  [81] "5141"   "5142"   "5143"   "5144"   "5145"   "5146"   "5147"   "5148"
##  [89] "5149"   "5150"   "5151"   "5152"   "5153"   "5158"   "5167"   "5169"
##  [97] "51728"  "5198"   "5236"   "5313"   "5315"   "53343"  "54107"  "5422"
## [105] "5424"   "5425"   "5426"   "5427"   "5430"   "5431"   "5432"   "5433"
## [113] "5434"   "5435"   "5436"   "5437"   "5438"   "5439"   "5440"   "5441"
## [121] "5471"   "548644" "55276"  "5557"   "5558"   "55703"  "55811"  "55821"
## [129] "5631"   "5634"   "56655"  "56953"  "56985"  "57804"  "58497"  "6240"
## [137] "6241"   "64425"  "646625" "654364" "661"    "7498"   "8382"   "84172"
## [145] "84265"  "84284"  "84618"  "8622"   "8654"   "87178"  "8833"   "9060"
## [153] "9061"   "93034"  "953"    "9533"   "954"    "955"    "956"    "957"
## [161] "9583"   "9615"
```

```
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
```

```
# Get the results
keggres = gage(foldchanges, gsets=kegg.sets.hs)
```

```
attributes(keggres)
```

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

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

```
##                                       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
## hsa00010 Glycolysis / Gluconeogenesis 4.360092e-03 -2.663825 4.360092e-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
## hsa00010 Glycolysis / Gluconeogenesis 0.119175854       65 4.360092e-03
```

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

```
## Info: Downloading xml files for hsa04110, 1/1 pathways..

## Info: Downloading png files for hsa04110, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

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

```
# A different PDF based output of the same data
pathview(gene.data=foldchanges, pathway.id="hsa04110", kegg.native=FALSE)
```

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

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

```
## Focus on top 5 upregulated pathways here for demo purposes only
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"
```

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

```
## Info: Downloading xml files for hsa04740, 1/1 pathways..

## Info: Downloading png files for hsa04740, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

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

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

## Info: Downloading xml files for hsa04640, 1/1 pathways..

## Info: Downloading png files for hsa04640, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa04640.pathview.png

## Info: Downloading xml files for hsa00140, 1/1 pathways..

## Info: Downloading png files for hsa00140, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa00140.pathview.png

## Info: Downloading xml files for hsa04630, 1/1 pathways..

## Info: Downloading png files for hsa04630, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa04630.pathview.png

## Info: Downloading xml files for hsa04976, 1/1 pathways..

## Info: Downloading png files for hsa04976, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

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

```
## Focus on top 5 downregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$less)[1:5]

# Extract the 8 character long IDs part of each string
keggresids = substr(keggrespathways, start=1, stop=8)
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

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

## Info: Downloading xml files for hsa03030, 1/1 pathways..

## Info: Downloading png files for hsa03030, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa03030.pathview.png

## Info: Downloading xml files for hsa03013, 1/1 pathways..

## Info: Downloading png files for hsa03013, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa03013.pathview.png

## Info: Downloading xml files for hsa04114, 1/1 pathways..

## Info: Downloading png files for hsa04114, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa04114.pathview.png

## Info: Downloading xml files for hsa03440, 1/1 pathways..

## Info: Downloading png files for hsa03440, 1/1 pathways..

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

## Info: Working in directory C:/Users/james/Desktop/UCSD/BIMM143/week09/project-DESeq

## Info: Writing image file hsa03440.pathview.png
```

3. 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 0.15191286      299 2.053700e-04
## GO:0060562 epithelial tube morphogenesis   0.15191286      289 2.927804e-04
## GO:0048598 embryonic morphogenesis         0.15191286      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
```

4. Reactome Analysis

```
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"
```

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

Then, to perform pathway analysis online go to the Reactome website (https://reactome.org/PathwayBrowser/#TOOL=AT). Select "choose file" to upload your significant gene list. Then, select the parameters "Project to Humans", then click "Analyze".

> 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?

They do not match because of the different datasets and sizes