

class12R

Katelyn Brown PID 15891811

2/24/2022

Upload data.

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

1. Input the counts and metadata files.

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

```
countData <- read.csv(countFile, row.names = 1)
head(countData)
```

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

Q. Remove first column of countData.

```
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
```

Q. Check if colData matches the row names.

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

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

Q. Remove genes that have 0 read count across all samples.

```
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
```

2. Run DESeq.

```
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
```

```
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: 15975 6
## metadata(1): version
## assays(4): counts mu H cooks
## rownames(15975): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
##      ENSG00000271254
## rowData names(22): baseMean baseVar ... deviance maxCooks
## colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
## colData names(2): condition sizeFactor
```

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.

```
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)    : 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
```

3. 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"       "UCSCCKG"
## [26] "UNIPROT"
```

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

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

```
head(res$symbol)
```

```
## ENSG00000279457 ENSG00000187634 ENSG00000188976 ENSG00000187961 ENSG00000187583
##      "WASH9P"      "SAMD11"      "NOC2L"      "KLHL17"      "PLEKHN1"
## ENSG00000187642
##      "PERM1"
```

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

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

```
head(res$entrez)
```

```
## ENSG00000279457 ENSG00000187634 ENSG00000188976 ENSG00000187961 ENSG00000187583
##      "102723897"      "148398"      "26155"      "339451"      "84069"
## ENSG00000187642
##      "84808"
```

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

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

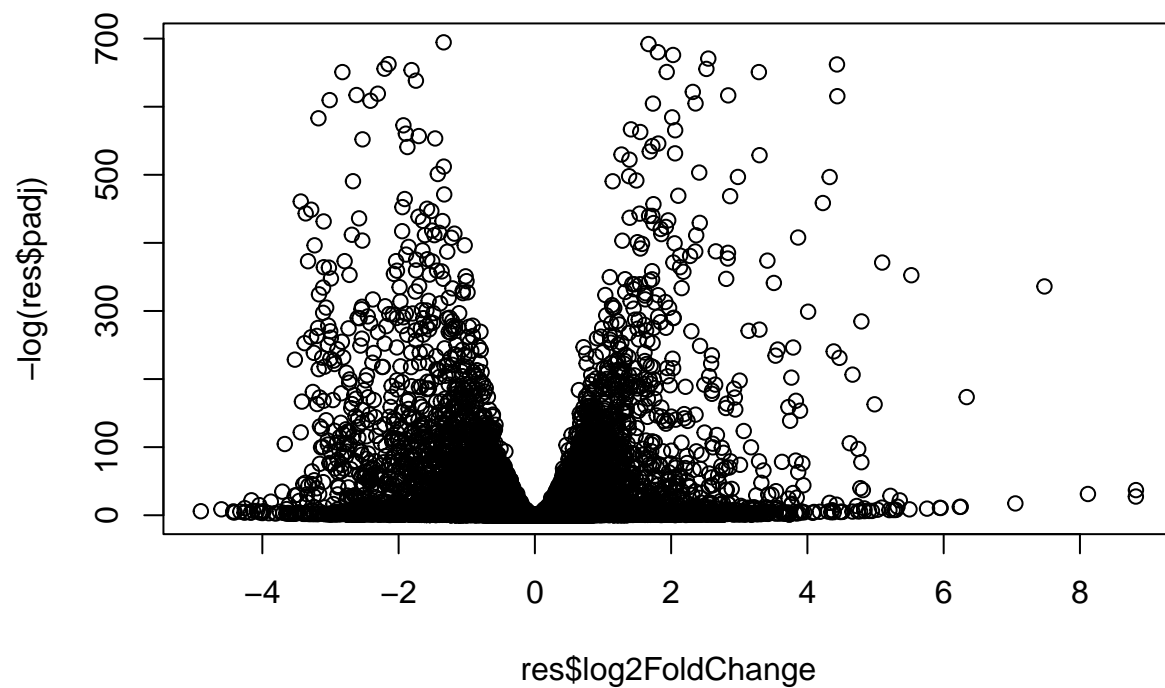
```
head(res$genenames)
```

```
## NULL
```

4. Make volcano plot.

Common summary figure that gives an overview of the results.

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



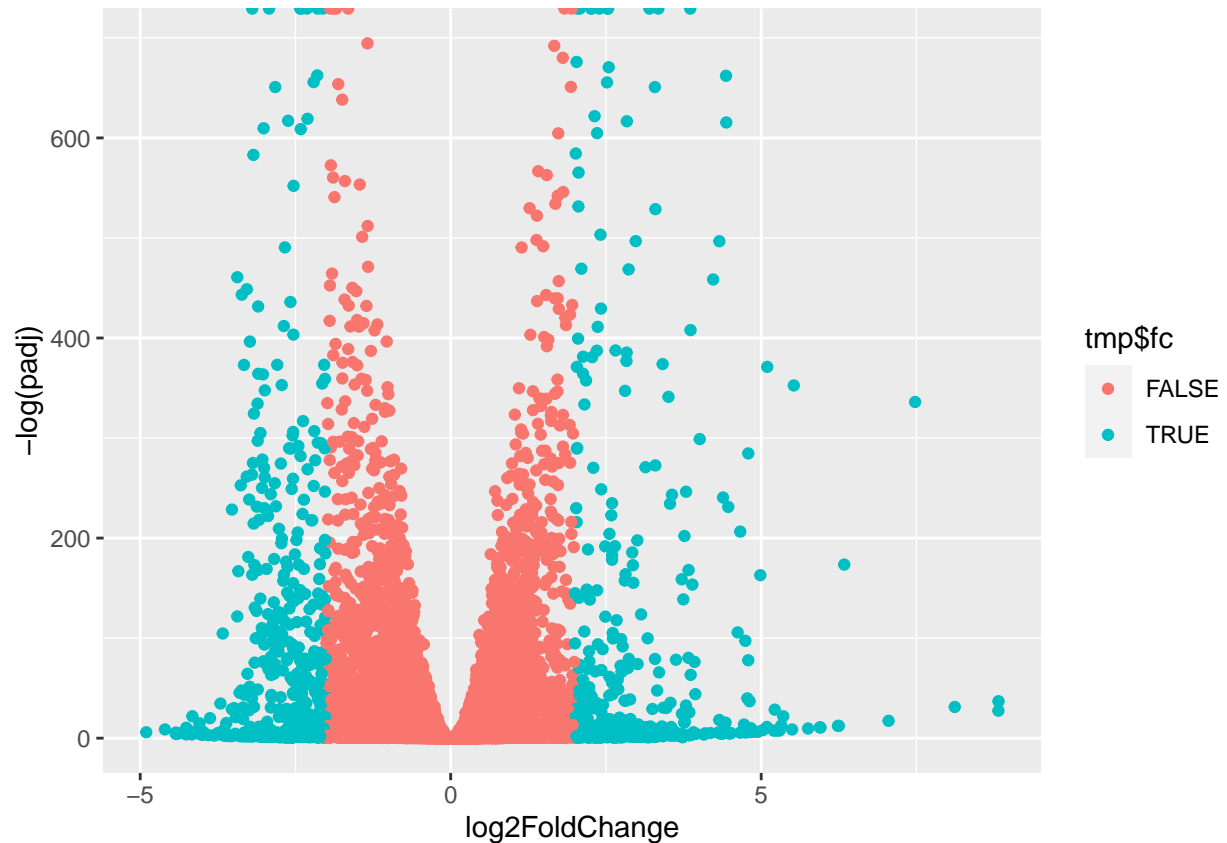
Try ggplot.

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

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

```
## Warning: Use of 'tmp$fc' is discouraged. Use 'fc' instead.
```

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



Use Enhanced Volcano.

```
library(EnhancedVolcano)
```

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

```
## Registered S3 methods overwritten by 'ggalt':
```

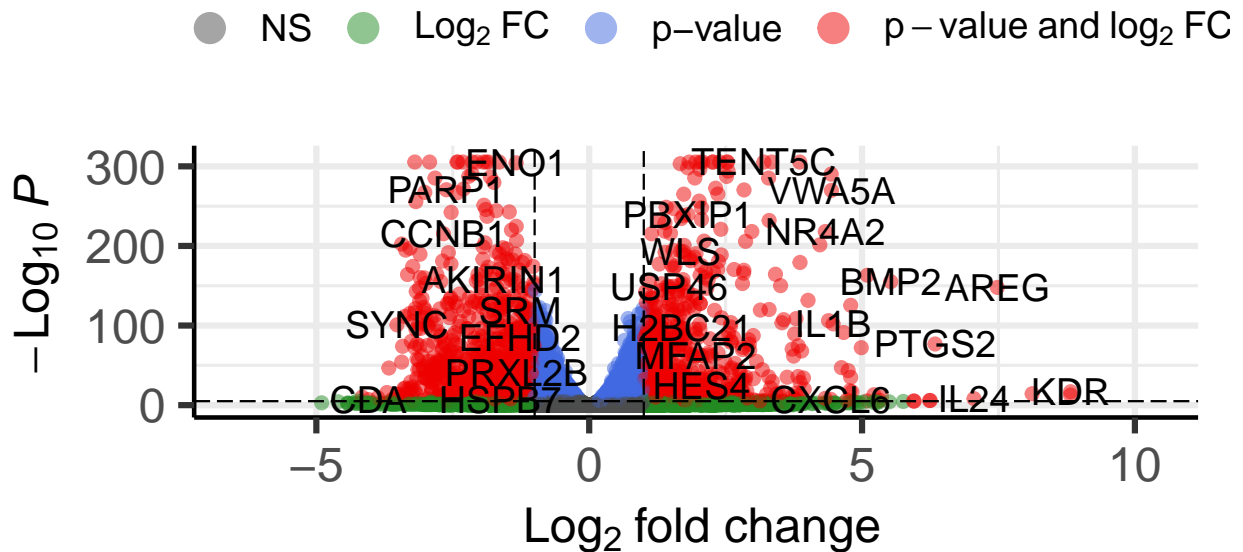
```
##   method                      from
##   grid.draw.absoluteGrob      ggplot2
##   grobHeight.absoluteGrob     ggplot2
##   grobWidth.absoluteGrob      ggplot2
##   grobX.absoluteGrob          ggplot2
##   grobY.absoluteGrob          ggplot2
```

```
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...
```

Volcano plot

EnhancedVolcano



Save results.

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

5. Pathway analysis.

Try to bring back the biology and help with the interpretation of the results. Try to answer the question: which pathways and functions feature heavily in our differentially expressed genes?

```
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)
```

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

```
kegg.sets.hs <- kegg.sets.hs[sigmet.idx.hs]  
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"
```

Make a vector and use gage.

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

```
## 102723897 148398 26155 339451 84069 84808  
## 0.17925708 0.42645712 -0.69272046 0.72975561 0.04057653 0.54281049
```



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

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

Look at down and up pathways.

```
head(keggres$less)
```

```
##
## 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
## hsa04114 Oocyte meiosis 3.784520e-03 -2.698128 3.784520e-03
## hsa00010 Glycolysis / Gluconeogenesis 8.961413e-03 -2.405398 8.961413e-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
## hsa04114 Oocyte meiosis 0.121861535 102 3.784520e-03
## hsa00010 Glycolysis / Gluconeogenesis 0.212222694 53 8.961413e-03
```

```
head(keggres$greater)
```

```
##
## p.geomean stat.mean p.val
## hsa04640 Hematopoietic cell lineage 0.002822776 2.833362 0.002822776
## hsa04630 Jak-STAT signaling pathway 0.005202070 2.585673 0.005202070
## hsa00140 Steroid hormone biosynthesis 0.007255099 2.526744 0.007255099
## hsa04142 Lysosome 0.010107392 2.338364 0.010107392
## hsa04330 Notch signaling pathway 0.018747253 2.111725 0.018747253
## hsa04916 Melanogenesis 0.019399766 2.081927 0.019399766
##
## q.val set.size exp1
## hsa04640 Hematopoietic cell lineage 0.3893570 55 0.002822776
## hsa04630 Jak-STAT signaling pathway 0.3893570 109 0.005202070
## hsa00140 Steroid hormone biosynthesis 0.3893570 31 0.007255099
## hsa04142 Lysosome 0.4068225 118 0.010107392
## hsa04330 Notch signaling pathway 0.4391731 46 0.018747253
## hsa04916 Melanogenesis 0.4391731 90 0.019399766
```

Use pathview to visualize results from cell cycle.

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

Look at top 5 pathways.

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

keggresids <- substr(keggrespathways, start=1, stop=8)
keggresids
```

```
## [1] "hsa04640" "hsa04630" "hsa00140" "hsa04142" "hsa04330"
```

Use pathview to visualize top 5 pathways.

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

```
## Info: Writing image file hsa04142.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 /Users/katybrown/Downloads/BIMM 143 R/class12
```

```
## Info: Writing image file hsa04330.pathview.png
```

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

```
keggresdown <- rownames(keggres$less)[1:5]
keggresdownids <- substr(keggresdown, start = 1, stop = 8)
keggresdownids
```

```
## [1] "hsa04110" "hsa03030" "hsa03013" "hsa03440" "hsa04114"
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

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

```
## Info: Working in directory /Users/katybrown/Downloads/BIMM 143 R/class12
```

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

Gene Ontology

```
data(go.sets.hs)
data(go.subs.hs)
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
## G0:0007156 homophilic cell adhesion      8.519724e-05  3.824205 8.519724e-05
## G0:0002009 morphogenesis of an epithelium 1.396681e-04  3.653886 1.396681e-04
## G0:0048729 tissue morphogenesis          1.432451e-04  3.643242 1.432451e-04
## G0:0007610 behavior                      2.195494e-04  3.530241 2.195494e-04
## G0:0060562 epithelial tube morphogenesis 5.932837e-04  3.261376 5.932837e-04
## G0:0035295 tube development              5.953254e-04  3.253665 5.953254e-04
##
##          q.val set.size      exp1
## G0:0007156 homophilic cell adhesion      0.1951953    113 8.519724e-05
## G0:0002009 morphogenesis of an epithelium 0.1951953    339 1.396681e-04
## G0:0048729 tissue morphogenesis          0.1951953    424 1.432451e-04
## G0:0007610 behavior                      0.2243795    427 2.195494e-04
## G0:0060562 epithelial tube morphogenesis 0.3711390    257 5.932837e-04
## G0:0035295 tube development              0.3711390    391 5.953254e-04
##
## $less
##
##          p.geomean stat.mean      p.val
## G0:0048285 organelle fission              1.536227e-15 -8.063910 1.536227e-15
## G0:0000280 nuclear division              4.286961e-15 -7.939217 4.286961e-15
## G0:0007067 mitosis                      4.286961e-15 -7.939217 4.286961e-15
## G0:0000087 M phase of mitotic cell cycle 1.169934e-14 -7.797496 1.169934e-14
## G0:0007059 chromosome segregation        2.028624e-11 -6.878340 2.028624e-11
## G0:0000236 mitotic prometaphase         1.729553e-10 -6.695966 1.729553e-10
##
##          q.val set.size      exp1
## G0:0048285 organelle fission              5.841698e-12    376 1.536227e-15
## G0:0000280 nuclear division              5.841698e-12    352 4.286961e-15
## G0:0007067 mitosis                      5.841698e-12    352 4.286961e-15
## G0:0000087 M phase of mitotic cell cycle 1.195672e-11    362 1.169934e-14
## G0:0007059 chromosome segregation        1.658603e-08    142 2.028624e-11
## G0:0000236 mitotic prometaphase         1.178402e-07     84 1.729553e-10
##
## $stats
##
##          stat.mean      exp1
## G0:0007156 homophilic cell adhesion      3.824205 3.824205
## G0:0002009 morphogenesis of an epithelium 3.653886 3.653886
## G0:0048729 tissue morphogenesis          3.643242 3.643242
## G0:0007610 behavior                      3.530241 3.530241
## G0:0060562 epithelial tube morphogenesis 3.261376 3.261376
## G0:0035295 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)
```

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 endosomal/vacuolar pathway has the most significant entities p-value, and the cell cycle/mitotic pathway has a significant entities p-value. The cell cycle pathway matches the KEGG results. Differing databases can cause the differences between the methods.