lab13

Table of contents

setup

```
library(DESeq2)
```

```
Warning: package 'DESeq2' was built under R version 4.1.1
```

```
Warning: package 'S4Vectors' was built under R version 4.1.2
```

```
Warning: package 'BiocGenerics' was built under R version 4.1.1
```

```
Warning: package 'IRanges' was built under R version 4.1.1
```

```
Warning: package 'GenomicRanges' was built under R version 4.1.2

Warning: package 'GenomeInfoDb' was built under R version 4.1.1

Warning: package 'SummarizedExperiment' was built under R version 4.1.1

Warning: package 'MatrixGenerics' was built under R version 4.1.1

Warning: package 'matrixStats' was built under R version 4.1.2

Warning: package 'Biobase' was built under R version 4.1.1
```

```
library(ggplot2)
library(gage)
```

```
Warning: package 'gage' was built under R version 4.1.1
```

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

```
Warning: package 'pathview' was built under R version 4.1.1
```

```
theme_set(theme_bw())
```

Read the countData and colData

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

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

Q. Do they match?

No they dont.

```
row.names(colData)
```

```
[1] "SRR493366" "SRR493367" "SRR493368" "SRR493369" "SRR493370" "SRR493371"
```

```
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
```

Need to get rid of the length column, will mess up Deseq analysis

```
countData <- countData[,- 1]
# countData[, colData$id] this is a better way
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
```

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

```
[1] TRUE
```

Q. Remove zero count genes

```
# rowSums(countData[]) sums each row and prints the rowname with the sum
# rowSums(countData[]) > 0 creates a logical vector with all the rownames, with TRUE if th
```

```
# wrapping in countData[logical vector,] gives the dataframe with rows that fufill the log
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

Q. How many genes left?

```
nrow(counts)
```

```
[1] 15975
```

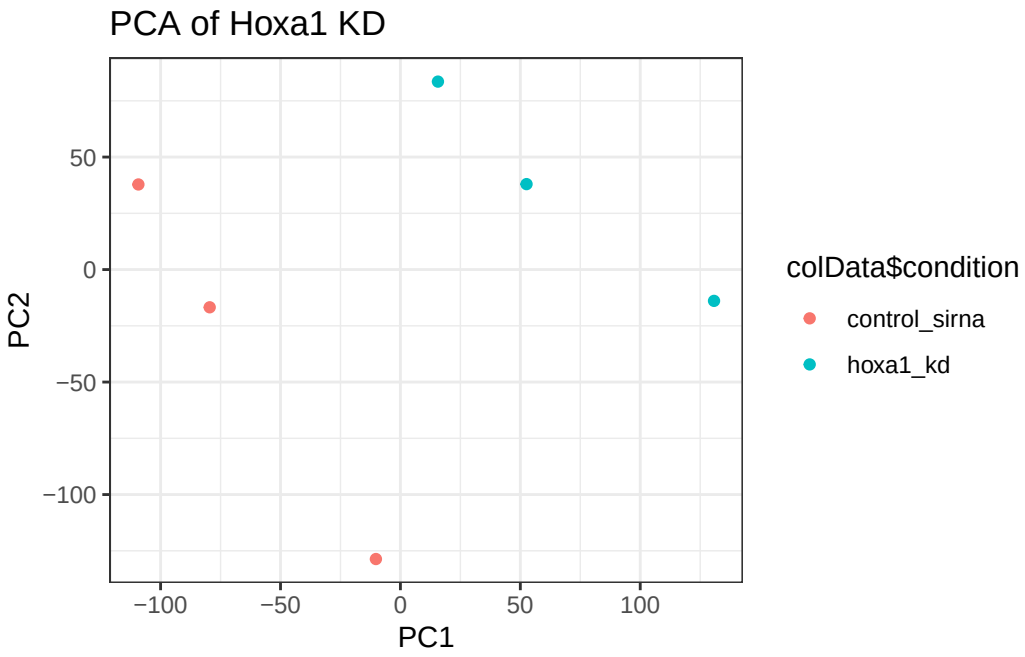
PCA as Quality control

BaseR is `prcomp()` function which often needs the data to be transposed and scaled.

```
pca <- prcomp(t(counts), scale = TRUE)
```

Plot

```
x <- as.data.frame(pca$x) # get in data frame format
ggplot(data = x) +
  aes(x = PC1, y = PC2, group = colData$condition, color = colData$condition) +
  geom_point() +
  labs(title = "PCA of Hoxa1 KD")
```



Q. How much variance captured in 2 PCs?

Summary of PCA

```
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
```

81.82% of variance captured by two PCs.

DESEQ analysis

```
library(DESeq2)
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
```

```
results <- results(dds)
head(results)
```

```
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
```

Summary plot

Plot the genes

```
library(EnhancedVolcano)
```

```
Warning: package 'EnhancedVolcano' was built under R version 4.1.1
```

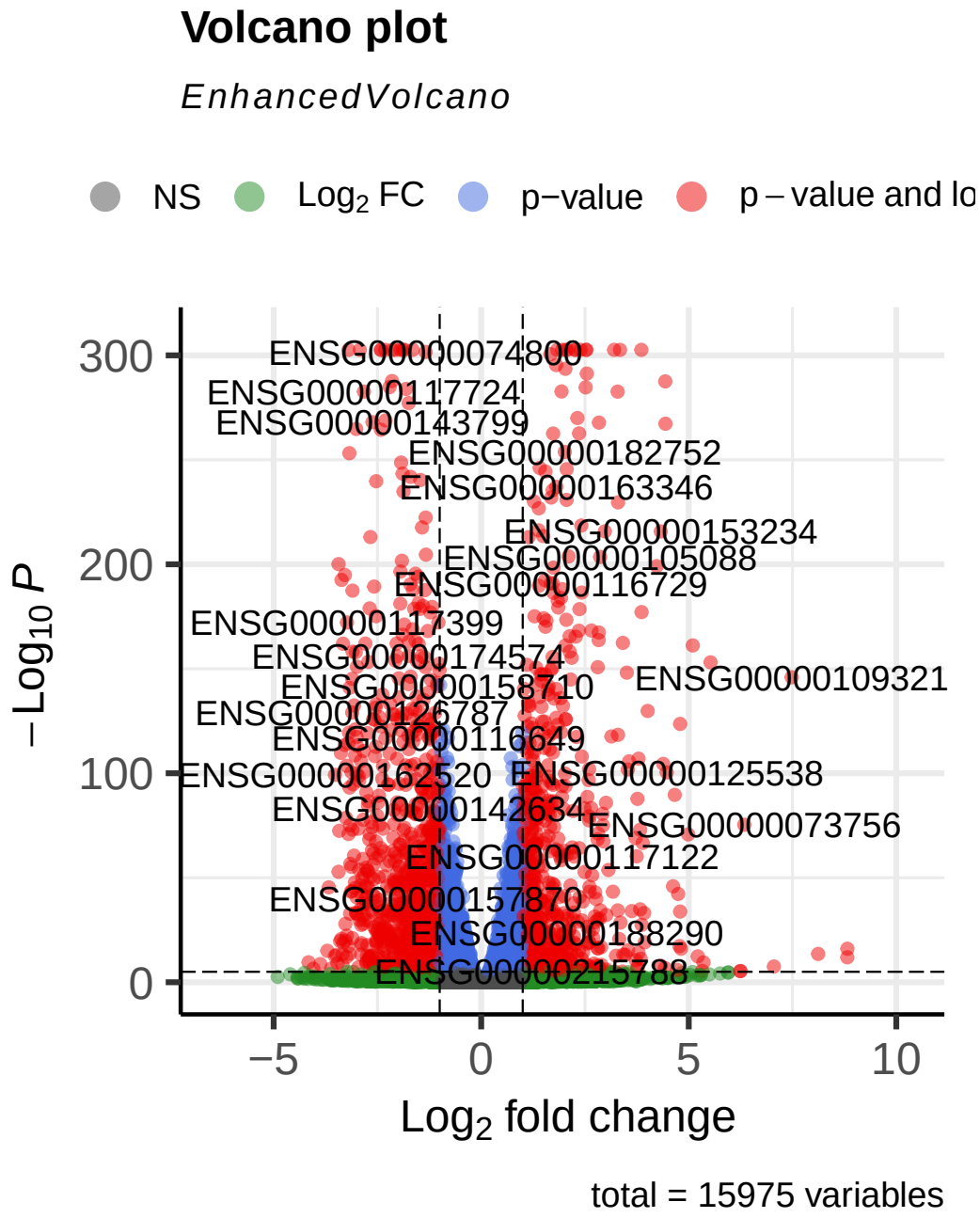
```
Loading required package: ggrepel
```

```
Warning: package 'ggrepel' was built under R version 4.1.1
```

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

```
results <- as.data.frame(results)
EnhancedVolcano(results, lab = rownames(results), x = "log2FoldChange", y = "padj")
```

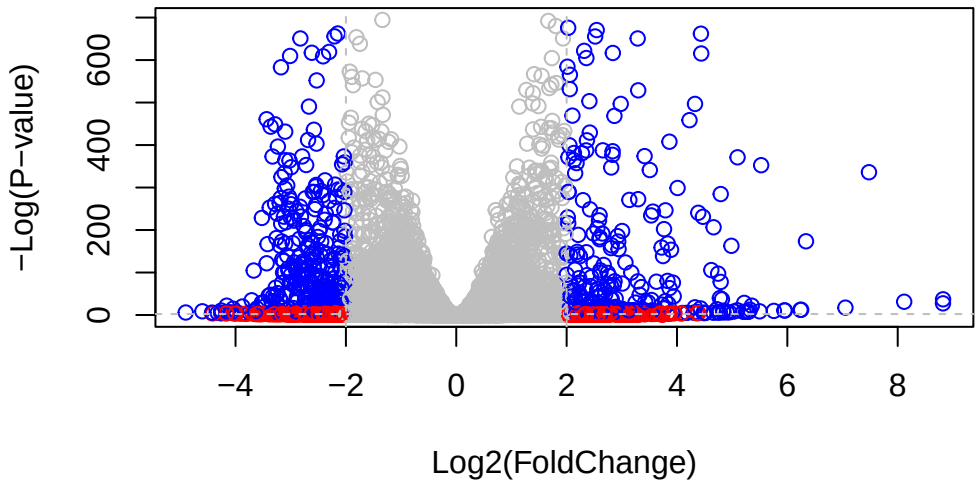
```
Warning: One or more p-values is 0. Converting to 10^-1 * current lowest non-
zero p-value...
```



in base R with colors

```
mycols <- rep("gray", nrow(results))
mycols[abs(results$log2FoldChange) > 2]  <- "red"

inds <- (results$padj < 0.01) & (abs(results$log2FoldChange) > 2 )
mycols[inds] <- "blue"
plot( results$log2FoldChange, -log(results$padj), col=mycols, ylab="-Log(P-value)", xlab="
```




```
integer(0)
```

Add annotations

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

```
Loading required package: AnnotationDbi
```

```
Warning: package 'AnnotationDbi' was built under R version 4.1.1
```

```
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"
```

use mapIds to add gene symbols, entrez ID

```
# add gene symbol
results$symbol <- mapIds(org.Hs.eg.db, keys = row.names(results), column = "SYMBOL", keyty
```

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

```
head(results$symbol)
```

```
[1] "WASH9P"  "SAMD11"  "NOC2L"   "KLHL17"  "PLEKHN1" "PERM1"
```

```
# add entrezid
results$entrez <- mapIds(org.Hs.eg.db, keys = row.names(results), column = "ENTREZID", key
```

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

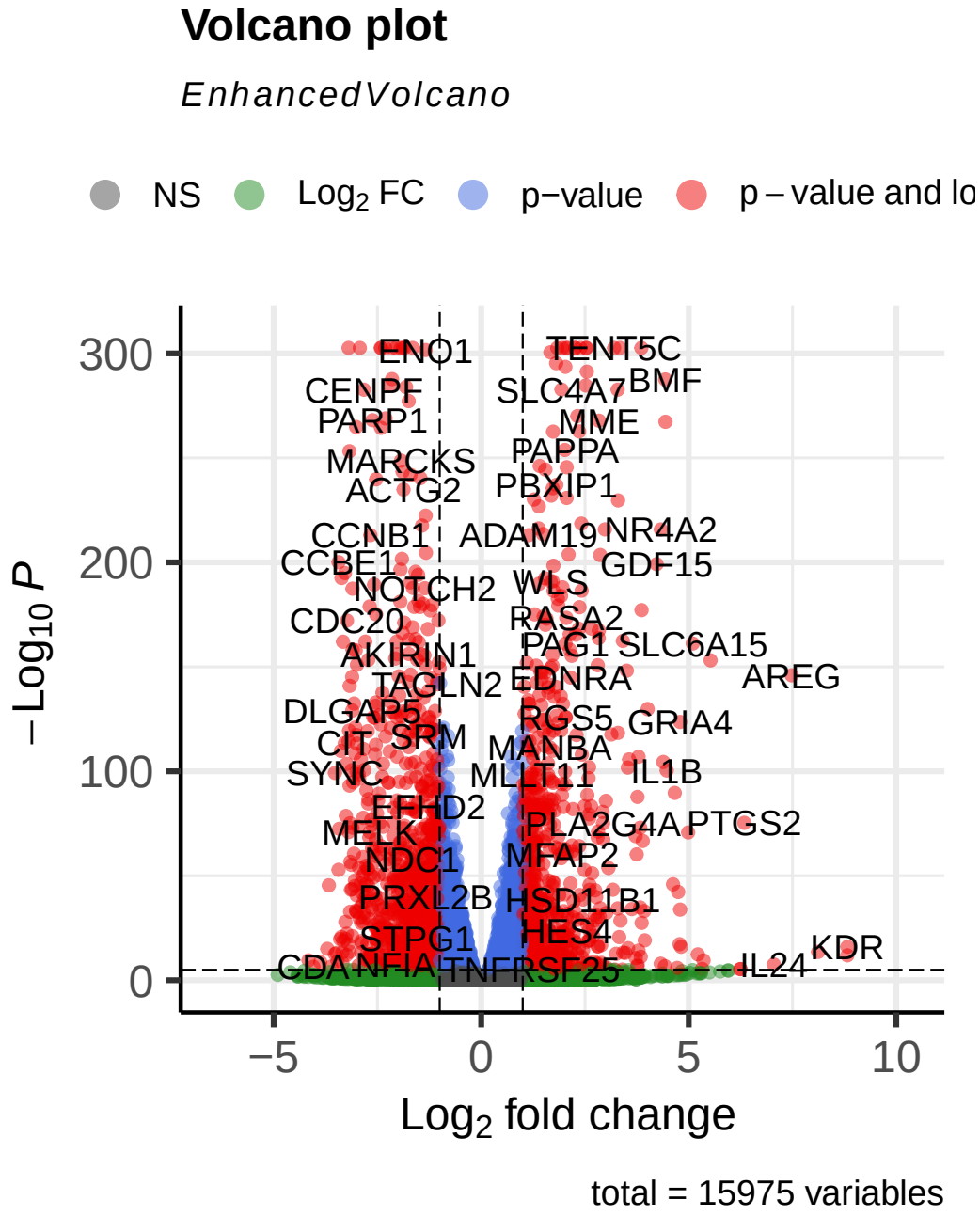
```
head(results)
```

```
                  baseMean log2FoldChange      lfcSE        stat       pvalue
ENSG00000279457   29.91358     0.17925708 0.32482157   0.5518632 5.810421e-01
ENSG00000187634  183.22965     0.42645712 0.14026582   3.0403495 2.363037e-03
ENSG00000188976 1651.18808    -0.69272046 0.05484654 -12.6301576 1.439895e-36
ENSG00000187961  209.63794     0.72975561 0.13185990   5.5343255 3.124282e-08
ENSG00000187583   47.25512     0.04057653 0.27189281   0.1492372 8.813664e-01
ENSG00000187642   11.97975     0.54281049 0.52155985   1.0407444 2.979942e-01
                       padj  symbol    entrez
ENSG00000279457 6.865548e-01  WASH9P 102723897
ENSG00000187634 5.157181e-03  SAMD11    148398
ENSG00000188976 1.765489e-35   NOC2L     26155
ENSG00000187961 1.134130e-07  KLHL17    339451
ENSG00000187583 9.190306e-01 PLEKHN1     84069
ENSG00000187642 4.033793e-01   PERM1     84808
```

New volcano:

```
library(EnhancedVolcano)
results <- as.data.frame(results)
EnhancedVolcano(results, lab = results$symbol, x = "log2FoldChange", y = "padj")
```

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



Pathway analysis

Create the input for gage() - a vector of fold changes with entrez IDs as the names

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

```
data(kegg.sets.hs)
keggres <- gage(foldchange, gsets = kegg.sets.hs)
head(keggres$less)
```

```
                                           p.geomean stat.mean
hsa04110 Cell cycle                        8.995727e-06 -4.378644
hsa03030 DNA replication                   9.424076e-05 -3.951803
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 -3.765330
hsa03013 RNA transport                     1.246882e-03 -3.059466
hsa03440 Homologous recombination          3.066756e-03 -2.852899
hsa04114 Oocyte meiosis                    3.784520e-03 -2.698128
                                                  p.val       q.val
hsa04110 Cell cycle                        8.995727e-06 0.001889103
hsa03030 DNA replication                   9.424076e-05 0.009841047
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 0.009841047
hsa03013 RNA transport                     1.246882e-03 0.065461279
hsa03440 Homologous recombination          3.066756e-03 0.128803765
hsa04114 Oocyte meiosis                    3.784520e-03 0.132458191
                                           set.size         exp1
hsa04110 Cell cycle                             121 8.995727e-06
hsa03030 DNA replication                         36 9.424076e-05
hsa05130 Pathogenic Escherichia coli infection   53 1.405864e-04
hsa03013 RNA transport                          144 1.246882e-03
hsa03440 Homologous recombination                28 3.066756e-03
hsa04114 Oocyte meiosis                         102 3.784520e-03
```

Look at the pathways:

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

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

```
Info: Working in directory C:/Users/lhodg/Documents/Research/BGGN213/lab13
```

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

Insert image:

?

GO pathways

```
data(go.sets.hs)
data(go.subs.hs)
```

```
 # the biological subprocess of go is selected with go.subs.hs$BP
gobpsets = go.sets.hs[go.subs.hs$BP]
 # gene sets
gobpres = gage(foldchange, gsets=gobpsets, same.dir=TRUE)
 # results from gage changing genes overlapping with the gene sets
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
```

Check GO codes for how they got the annotation, is it verified by experiment or computationally inferred??

Reactome

```
sig_genes <- results[results$padj <= 0.05 & !is.na(results$padj), "symbol"]
# reactome uses gene symbol, not entrez id

print(paste("Total number of significant genes:", length(sig_genes)))
```

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

```
# write out text file for reactome to take
write.table(sig_genes, file="significant_genes.txt", row.names=FALSE, col.names=FALSE, quo
```

From Reactome browser:

> 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?

Endosomal/Vacuolar pathway

Not all the pathways match. Both methods give cell cycle-related processes as a significant result. Each method uses different methods of annotating genes and verifying which gene belongs to which pathway.