class13

Steven Gan

2022-11-09

Table of contents

Section 1. Differential Expression Analysis

```
library(DESeq2)
```

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

Q1:

```
countData <- countData[,-1]
```

Q2:

```
# Delete rows that have either all 0 in WT or KD
countData = countData[((rowSums(countData[,1:3] != 0) != 0) &
                       (rowSums(countData[,4:6] != 0) != 0)),]
nrow(countData) # 14861
```

```
[1] 14861
```

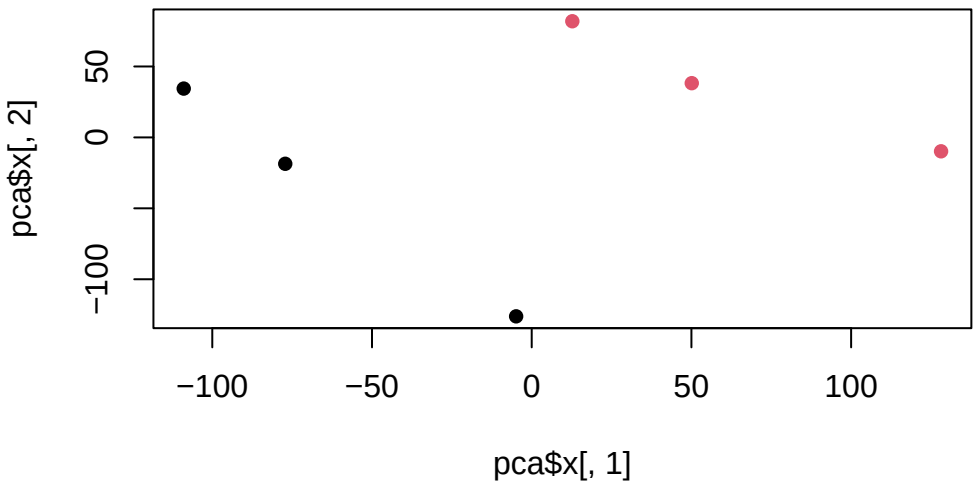
```
# Deplete rows with all 0
# countData = countData[rowSums(countData != 0) != 0,]
# nrow(countData) #15975
```

PCA analysis

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

```
Importance of components:
                           PC1     PC2      PC3      PC4      PC5        PC6
Standard deviation     85.9548 71.7134 29.43895 27.80061 26.27619 1.133e-13
Proportion of Variance  0.4972  0.3461  0.05832  0.05201  0.04646 0.000e+00
Cumulative Proportion   0.4972  0.8432  0.90153  0.95354  1.00000 1.000e+00
```

```
plot(pca$x[,1], pca$x[,2], col = as.factor(colData$condition), pch = 16)
```



Running DESeq2

```
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
```

```
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: 14861 6
metadata(1): version
assays(4): counts mu H cooks
rownames(14861): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
  ENSG00000271254
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

```
res = results(dds)
```

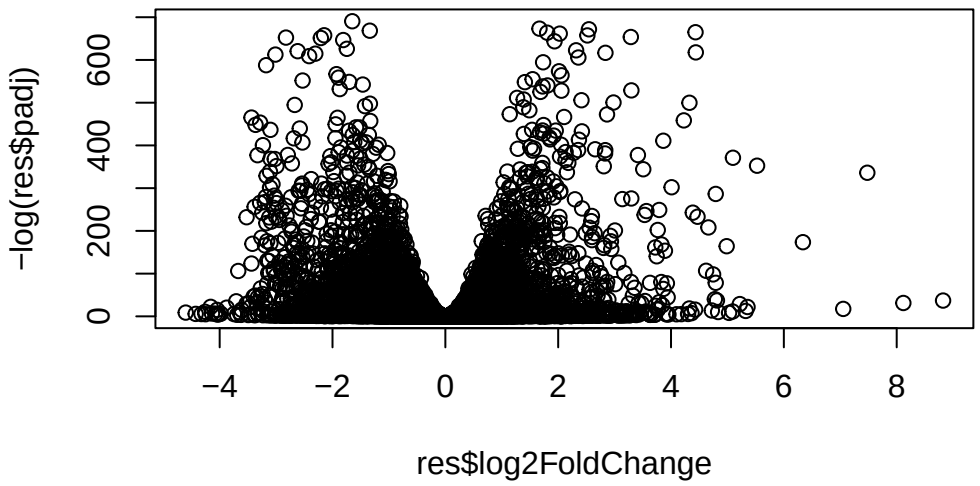
Q3:

```
summary(res)
```

```
out of 14861 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4350, 29%
LFC < 0 (down)     : 4432, 30%
outliers [1]       : 0, 0%
low counts [2]     : 577, 3.9%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Volcono plot

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



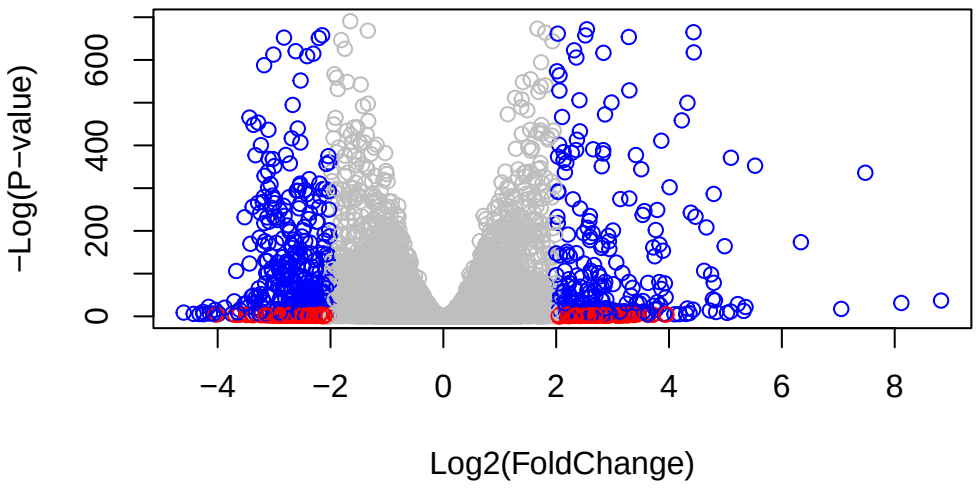
Q4:

```
mycols <- rep("gray", nrow(res))

mycols[abs(res$log2FoldChange) > 2] <- "red"

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

Q5:

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

```
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       stat      pvalue
                <numeric>      <numeric> <numeric>  <numeric>   <numeric>
ENSG00000279457   29.9136       0.179637 0.3200391   0.561296 5.74596e-01
ENSG00000187634  183.2296       0.426251 0.1384705   3.078282 2.08198e-03
ENSG00000188976 1651.1881      -0.692718 0.0548898 -12.620152 1.63506e-36
ENSG00000187961  209.6379       0.729830 0.1303027   5.601034 2.13077e-08
ENSG00000187583   47.2551       0.040098 0.2676614   0.149809 8.80916e-01
ENSG00000187642   11.9798       0.541601 0.5134384   1.054851 2.91493e-01
ENSG00000188290  108.9221       2.056774 0.1946888  10.564415 4.35675e-26
ENSG00000187608  350.7169       0.257250 0.1016757   2.530105 1.14028e-02
ENSG00000188157 9128.4394       0.389909 0.0472922   8.244686 1.65594e-16
ENSG00000131591  156.4791       0.196702 0.1437393   1.368464 1.71167e-01
                      padj      symbol      entrez                   name
                 <numeric> <character> <character>            <character>
ENSG00000279457 6.72142e-01          NA          NA                     NA
ENSG00000187634 4.42149e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.92382e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 7.54673e-08      KLHL17      339451 kelch like family me..
ENSG00000187583 9.16060e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 3.88078e-01       PERM1       84808 PPARGC1 and ESRR ind..
```

```
ENSG00000188290 3.68454e-25        HES4       57801 hes family bHLH tran..
ENSG00000187608 2.15447e-02       ISG15        9636 ISG15 ubiquitin like..
ENSG00000188157 9.41619e-16        AGRN      375790                  agrin
ENSG00000131591 2.47791e-01    C1orf159       54991 chromosome 1 open re..
```

Q6:

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

Section 2. Pathway Analysis

```
# BiocManager::install(c("pathview", "gage", "gageData"))

library(pathview)
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"
```

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

```
     1266     54855      1465     51232      2034      2150
-2.422705  3.201920 -2.313728 -2.059619 -1.888012  3.344498
```

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

```
attributes(keggres)
```

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

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

```
                                        p.geomean stat.mean        p.val
hsa04110 Cell cycle                   1.125274e-05 -4.329005 1.125274e-05
hsa03030 DNA replication              6.612499e-05 -4.049772 6.612499e-05
hsa04114 Oocyte meiosis               1.266819e-03 -3.060131 1.266819e-03
hsa03440 Homologous recombination     2.716776e-03 -2.896957 2.716776e-03
hsa03013 RNA transport                3.994003e-03 -2.677890 3.994003e-03
hsa00010 Glycolysis / Gluconeogenesis 6.292535e-03 -2.544231 6.292535e-03
                                            q.val set.size         exp1
hsa04110 Cell cycle                   0.001789186      120 1.125274e-05
hsa03030 DNA replication              0.005256937       36 6.612499e-05
hsa04114 Oocyte meiosis               0.067141415       98 1.266819e-03
hsa03440 Homologous recombination     0.107991827       28 2.716776e-03
hsa03013 RNA transport                0.127009296      141 3.994003e-03
hsa00010 Glycolysis / Gluconeogenesis 0.166752172       48 6.292535e-03
```

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

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

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

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

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

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

Q7:

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

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

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

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

Section 3. Gene Ontology (GO)

```
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
GO:0007156 homophilic cell adhesion       1.965723e-05  4.204451 1.965723e-05
GO:0016337 cell-cell adhesion             1.270107e-04  3.680842 1.270107e-04
GO:0060429 epithelium development         1.478662e-04  3.633919 1.478662e-04
GO:0048729 tissue morphogenesis           2.478884e-04  3.498804 2.478884e-04
GO:0002009 morphogenesis of an epithelium 5.681600e-04  3.270491 5.681600e-04
GO:0035295 tube development               6.300004e-04  3.238557 6.300004e-04
                                               q.val set.size         exp1
GO:0007156 homophilic cell adhesion       0.07711532      103 1.965723e-05
GO:0016337 cell-cell adhesion             0.19335971      304 1.270107e-04
GO:0060429 epithelium development         0.19335971      454 1.478662e-04
GO:0048729 tissue morphogenesis           0.24311651      385 2.478884e-04
GO:0002009 morphogenesis of an epithelium 0.31142171      311 5.681600e-04
GO:0035295 tube development               0.31142171      361 6.300004e-04

$less
                                             p.geomean stat.mean        p.val
GO:0000279 M phase                        4.089509e-17 -8.491364 4.089509e-17
GO:0048285 organelle fission              5.399903e-16 -8.210482 5.399903e-16
GO:0000280 nuclear division               1.492899e-15 -8.090360 1.492899e-15
GO:0007067 mitosis                        1.492899e-15 -8.090360 1.492899e-15
GO:0000087 M phase of mitotic cell cycle  4.536940e-15 -7.935124 4.536940e-15
GO:0007059 chromosome segregation         8.122281e-12 -7.040529 8.122281e-12
                                                 q.val set.size         exp1
GO:0000279 M phase                        1.604314e-13      489 4.089509e-17
GO:0048285 organelle fission              1.059191e-12      372 5.399903e-16
GO:0000280 nuclear division               1.464161e-12      348 1.492899e-15
GO:0007067 mitosis                        1.464161e-12      348 1.492899e-15
GO:0000087 M phase of mitotic cell cycle  3.559683e-12      358 4.536940e-15
```

```
GO:0007059 chromosome segregation        5.310618e-09      141 8.122281e-12

$stats
                                          stat.mean     exp1
GO:0007156 homophilic cell adhesion        4.204451 4.204451
GO:0016337 cell-cell adhesion              3.680842 3.680842
GO:0060429 epithelium development          3.633919 3.633919
GO:0048729 tissue morphogenesis            3.498804 3.498804
GO:0002009 morphogenesis of an epithelium  3.270491 3.270491
GO:0035295 tube development                3.238557 3.238557
```

Section 4. 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: 8180"
```

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

Q8:

"Endosomal/Vacuolar pathway" exhibited the lowest entities p value, followed by "Cell cycle, mitotic". However, in terms of entities found, the latter exhibits 410 entities, which is far more than the entities found for the former. Possible explanation for the difference results between the methods is that the reactome analysis only analysed based on the genes names with significant chances, however KEGG also considered the fold change level of genes.

Section 5. GO online (OPTIONAL)

Q9:

The rank-1 pathway is "regulation of cell migration involved in sprouting angiogenesis", followed by "negative regulation of mitotic nuclear division". Certain relations could be argued in between this result and the result from KEGG & GO. Negative mitotic nuclear division appeared in both methods, and cell migration is related to homophile cell adhesion and cell cell adhesion. Similarly to reactome analysis, GO online also ignore the fold change level and only focus on what genes were regulated.

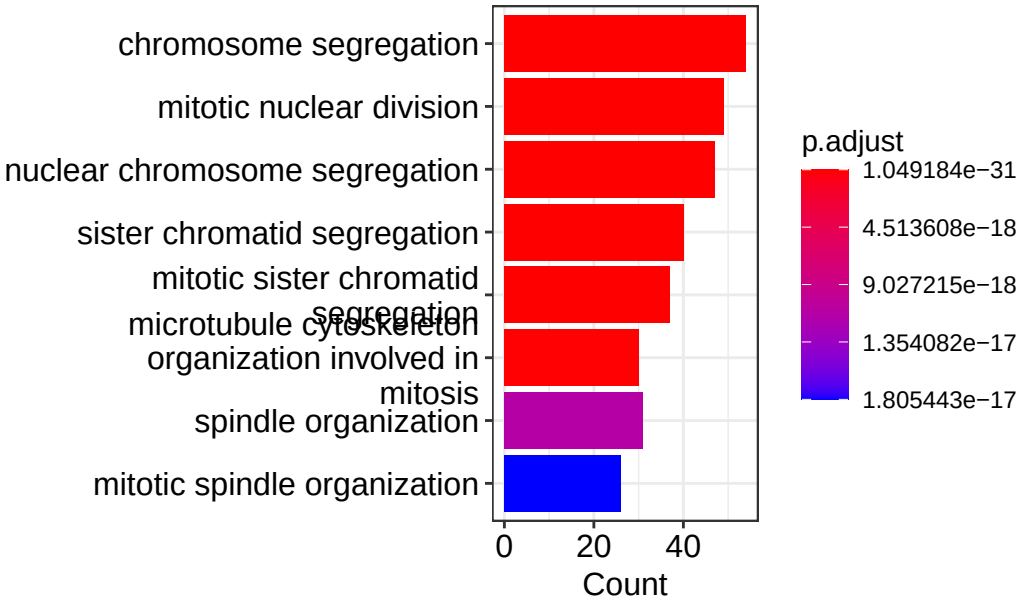
GO ploting

```
# BiocManager::install("clusterProfiler")
library(clusterProfiler)
library(tidyverse)
```

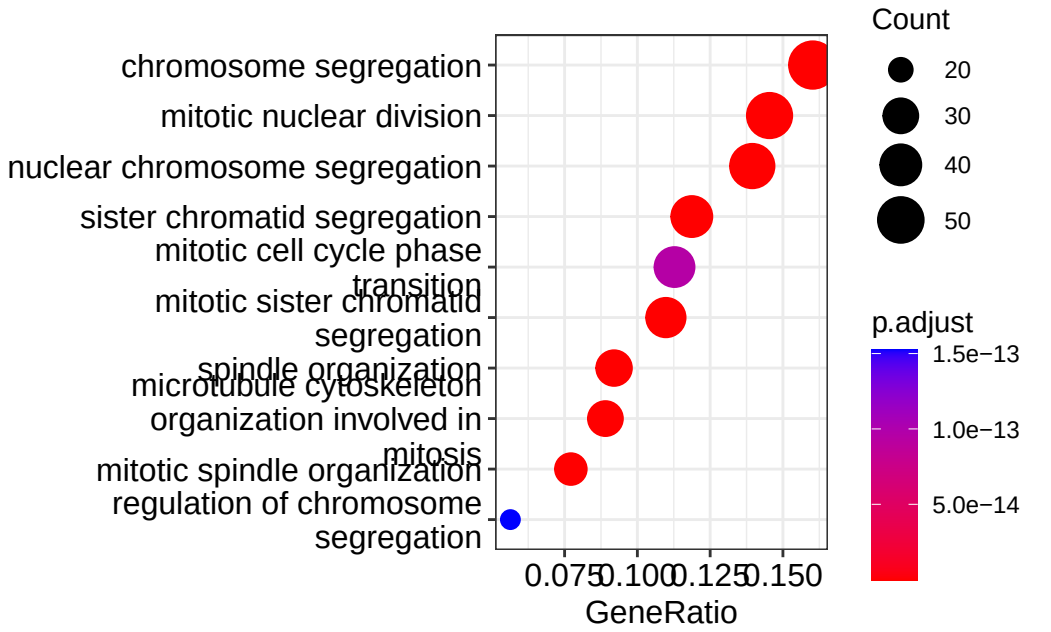
```
x <- as.data.frame(res)
y <- filter(x, log2FoldChange < -2 , padj < 0.05)
```

```
ecc <- enrichGO(rownames(y),
                OrgDb = org.Hs.eg.db,
                keyType = "ENSEMBL",
                ont = "BP")
```

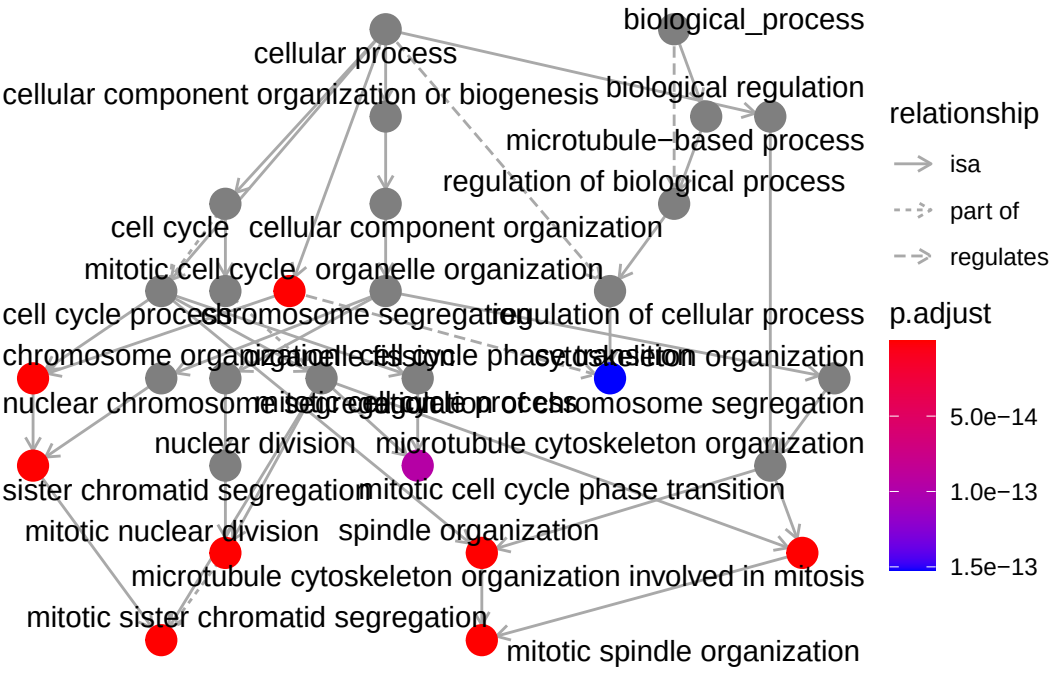
```
barplot(ecc)
```




```
dotplot(ecc)
```


chromosome segregation
mitotic nuclear division
nuclear chromosome segregation
sister chromatid segregation
mitotic cell cycle phase transition
mitotic sister chromatid segregation
spindle organization
microtubule cytoskeleton organization involved in mitosis
mitotic spindle organization
regulation of chromosome segregation
0.075
0.100
0.125
0.150
GeneRatio
Count
20
30
40
50
p.adjust
1.5e−13
1.0e−13
5.0e−14

```
goplot(ecc)
```


biological_process
cellular process
cellular component organization or biogenesis
biological regulation
microtubule−based process
regulation of biological process
cell cycle
cellular component organization
mitotic cell cycle
organelle organization
cell cycle process
chromosome segregation
regulation of cellular process
chromosome organization
cytoskeleton organization
nuclear chromosome segregation
nuclear division
microtubule cytoskeleton organization
sister chromatid segregation
mitotic cell cycle phase transition
mitotic nuclear division
spindle organization
microtubule cytoskeleton organization involved in mitosis
mitotic sister chromatid segregation
mitotic spindle organization
relationship
isa
part of
regulates
p.adjust
5.0e−14
1.0e−13
1.5e−13