# Lab 14 Pathway Analysis from RNA-Seq Results

Jazz Zhang (A16149005)

**Differential expression analysis**

```r
library(DESeq2)
```

```r
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
```

```r
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
```

```r
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
countData = countData[rowSums(countData==0)<ncol(countData), ]
head(countData)
```

```
                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
```

```r
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
```

2

```
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
```

```
res = results(dds)
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
```
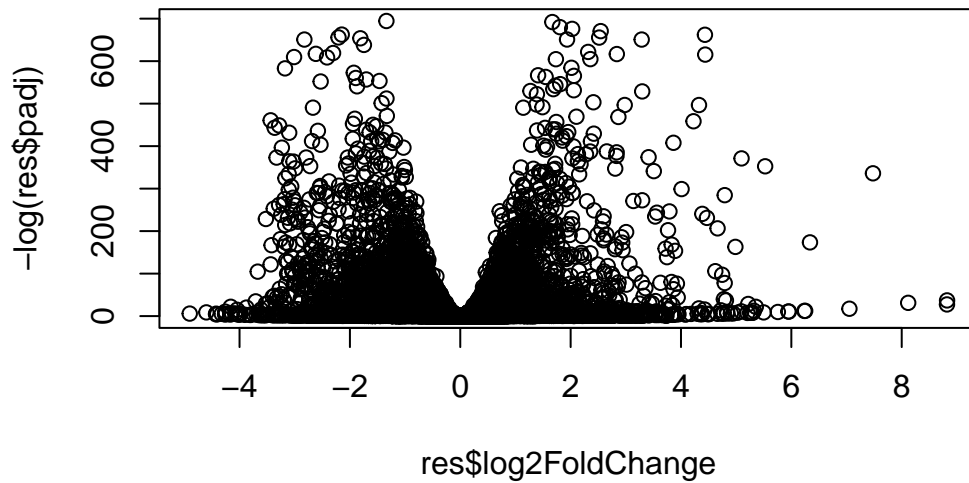
```
plot( res$log2FoldChange, -log(res$padj) )
```



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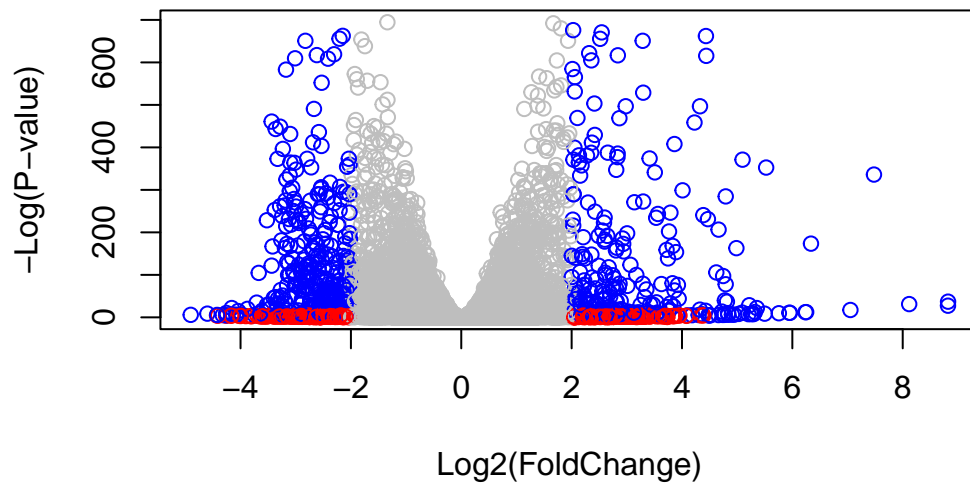
```r
library("AnnotationDbi")
library("org.Hs.eg.db")

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"
```

```r
res$symbol = mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="SYMBOL",
                    multiVals="first")

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

```
                  column="ENTREZID",
                  multiVals="first")

res$name =    mapIds(org.Hs.eg.db,
                  keys=row.names(res),
                  keytype="ENSEMBL",
                  column="GENENAME",
                  multiVals="first")

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   29.913579      0.1792571 0.3248216    0.551863 5.81042e-01
ENSG00000187634  183.229650      0.4264571 0.1402658    3.040350 2.36304e-03
ENSG00000188976 1651.188076     -0.6927205 0.0548465  -12.630158 1.43990e-36
ENSG00000187961  209.637938      0.7297556 0.1318599    5.534326 3.12428e-08
ENSG00000187583   47.255123      0.0405765 0.2718928    0.149237 8.81366e-01
ENSG00000187642   11.979750      0.5428105 0.5215598    1.040744 2.97994e-01
ENSG00000188290  108.922128      2.0570638 0.1969053   10.446970 1.51282e-25
ENSG00000187608  350.716868      0.2573837 0.1027266    2.505522 1.22271e-02
ENSG00000188157 9128.439422      0.3899088 0.0467163    8.346304 7.04321e-17
ENSG00000237330    0.158192      0.7859552 4.0804729    0.192614 8.47261e-01
                       padj      symbol      entrez                      name
                  <numeric> <character> <character>               <character>
ENSG00000279457 6.86555e-01          NA          NA                        NA
ENSG00000187634 5.15718e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76549e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13413e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.19031e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03379e-01       PERM1       84808 PPARGC1 and ESRR ind..
ENSG00000188290 1.30538e-24        HES4       57801 hes family bHLH tran..
ENSG00000187608 2.37452e-02       ISG15        9636 ISG15 ubiquitin like..
ENSG00000188157 4.21963e-16        AGRN      375790                     agrin
ENSG00000237330          NA      RNF223      401934 ring finger protein ..
```

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

**Pathway analysis**

```
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       2317
-2.422719   3.201955 -2.313738 -2.059631 -1.888019 -1.649792
```

```
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             8.995727e-06 -4.378644 8.995727e-06
hsa03030 DNA replication        9.424076e-05 -3.951803 9.424076e-05
hsa03013 RNA transport          1.375901e-03 -3.028500 1.375901e-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          0.073840037      144 1.375901e-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
```

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

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

```
Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14
```

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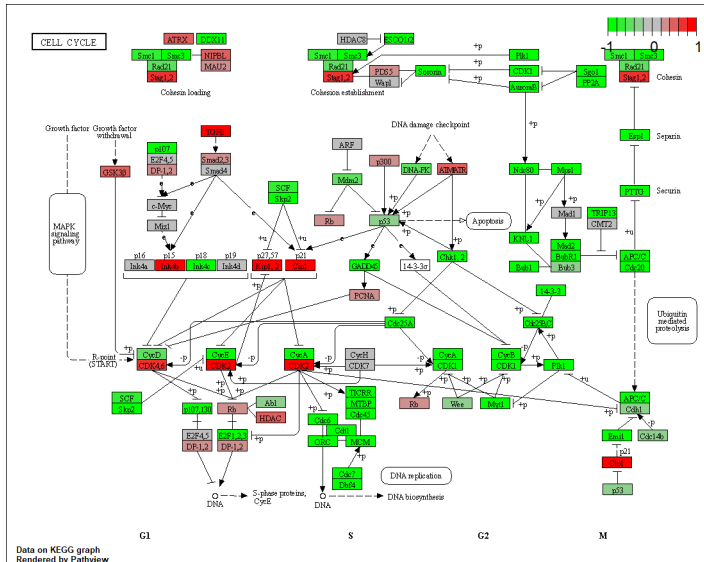


```r
# Top 5 upregulated pathways
keggrespathways_up <- rownames(keggres$greater)[1:5]

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

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

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

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

```
Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14
```

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

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

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 C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa04330.pathview.png

HEMATOPOIETIC CELL LINEAGE

Data on KEGG graph
Rendered by Pathview

JAK-STAT SIGNALING PATHWAY

Data on KEGG graph
Rendered by Pathview

STEROID HORMONE BIOSYNTHESIS

Data on KEGG graph
Rendered by Pathview

13

```
# Top 5 downregulated pathways
keggrespathways_down <- rownames(keggres$less)[1:5]

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

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

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

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

```
Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa04110.pathview.png

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa03440.pathview.png

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

Info: Working in directory C:/Users/jz644/OneDrive/ /UCSD/BGGN_213/lab14

Info: Writing image file hsa04114.pathview.png
```
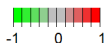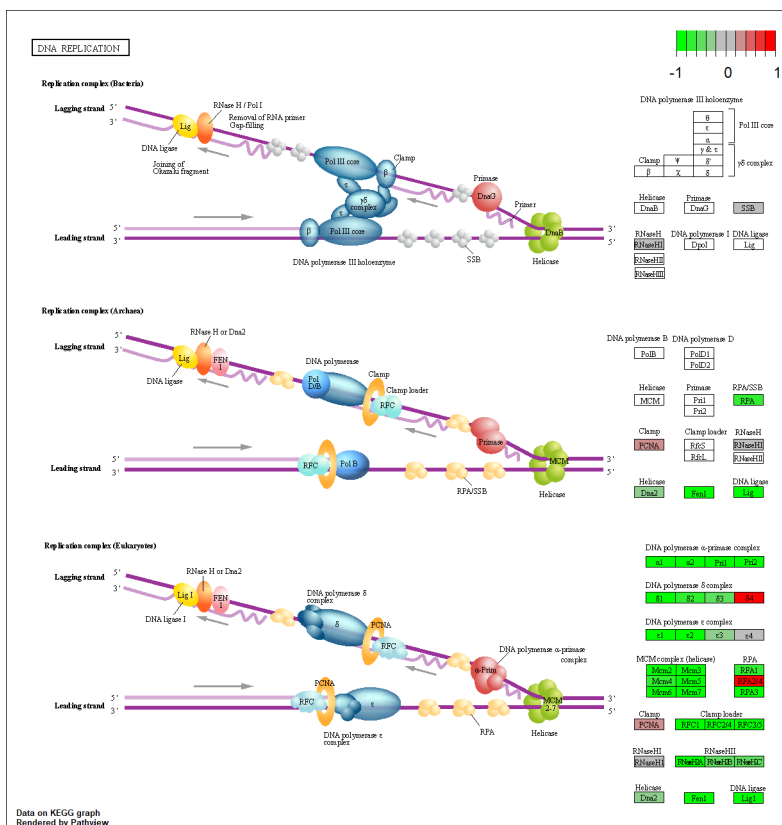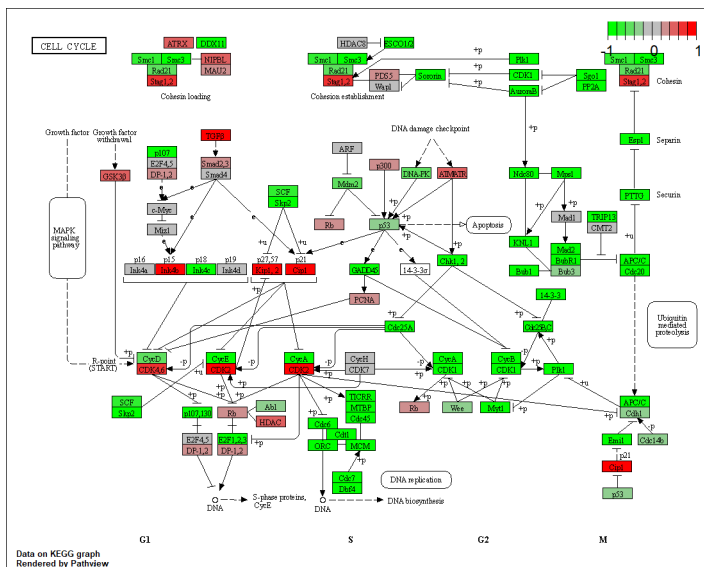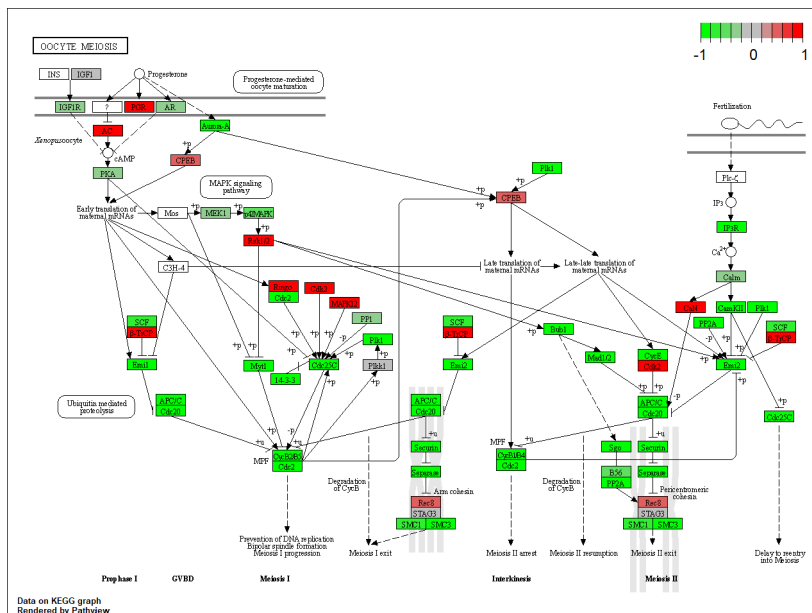
Data on KEGG graph
Rendered by Pathview



Data on KEGG graph
Rendered by Pathview

## NUCLEOCYTOPLASMIC TRANSPORT

Import
Importin
NLS
NPC
Cytoplasmic fibrils
Cytoplasm
Cytoplasmic ring
Central channel
Spoke complex
Lumenal ring
Nucleoplasmic ring
Nuclear basket
Lumen
Nucleus
NLS
Ran GTP
Importin

Export
Exportin
NES
Pi
Ran GTP
DDX19 Nup98
Nup358 complex
Nup214
Rae1
Nup62 complex
Nup107-160 complex
ELYS-153
Tpr
Exportin
Ran GTP
NES

mRNA Export
Upf1
Upf2
Tap
Ref/Aly
PARP
AAAAA
Cytoplasm
Lumen
Nucleus
NPC
mRNA surveilance pathway
SRm160
Pinin
EJC
Upf3
Tap
p15
TREX Ref/Aly
CBC m7G
AAAAA

PYM EJC AGG

**Nuclear Pore complex (NPC)**

Cytoplasmic fibrils

| ALADIN | hCG1 | Gle1 | DDX19 | Rae1 | Nup98 | Nup214 | Nup88 |

Nup358 complex

| RanBP2 | RanGAP | UBC9 | SUMO |

Cytoplasmic ring / Nucleoplasmic ring (Symmetrical maps)

| Nup160 | Nup85 | Sec13 | Nup107 | Nup133 | Nup96 | | Seh1 | Nup43 | Nup37 | ELYS |
| | | | | | Nup145 | | | | | |

Central channel

| Nup62 | Nup58/45 | Nup54 |

Spoke complex

| Nup205 | Nup188 | Nup155 | Nup93 | Nup53 |
| | | | | Nup59 |

Lumenal ring

| NDC1 | gp210 | pom121 | pom152 | pom34 | pom33 |

Nuclear basket

| Tpr | Nup50 | Nup153 | Senp2 |
| Nup2 | Nup1 | Nup60 |

**Nuclear transport complex**

Importin

| IPOA | IPOB |

Adaptor proteins

| SPN1 |

Exportin

| XPO | Ran |

| eEF1A |
| PHAX | CBC |
| NMD3 |

**Exon-junction complex (EJC)**

EJC inner core

| Y14 | Mago/H | MLN51 | EIF4A3 |

EJC outer shell

| ACIN1 | SAP18 | RNPS1 | Pinin | Ref/Aly |

Transiently interacting factors

| Upf1 | Upf2 | Upf3 |
| Tap | p15 | UAF36 | SRm160 | PYM |

**Transcription-export (TREX) complex**

THO subcomplex

| THOC1 | THOC2 | THOC5 | THOC6 | THOC7 | TEX1 |

Data on KEGG graph
Rendered by Pathview

## HOMOLOGOUS RECOMBINATION

Prokaryotic type
Eukaryotic type

Double strand break

5' to 3' resection (Escherichia coli)
(Saccharomyces cerevisiae) (Mammals)

RecJ SSB
RecB RecC RecD

Recognition
ATM
+p

MRX complex
Rad50 Mre11 XRS2

MRN complex
Rad50 Mre11 Nbs1

CtIP
BARD1 BRCA1 BRIP1
TOPBP1

BRCA2-DSS1

RecJ SSB RecBCD RPA Rad51
NBA1 BRE BRCC36

Filament formation

RecF RecA
RecO RecR
RecA

Filament formation
(S. cerevisiae)
Rad55 Rad57

(Common)
RPA
Rad51
Rad52

Abraxas
RAP80

PALB2
DSS1 BRCA2 SVCP3

RecFOR
RecA filament

Rad51B Rad51C
Rad51D XRCC3 Rad51B
XRCC2 Rad51C Rad51C Rad51D XRCC2
Rad51 paralogs

Rad51 paralogs

Rad51 filament

Strand invasion
DNA synthesis
Pol

Strand invasion
DpoIII

Strand invasion
(S. cerevisiae) (Common)
Rad59 Rad54

Rad52

Holliday junction intermediate

Rad51

Rad54 D-loop

Second end capture
Holliday junction intermediate

DNA synthesis
(Common)
polδ

Branch migration and resolution of Holliday junction
RuvA or RecG
RuvB RuvC

Branch migration and resolution of Holliday junction
RuvA or RecG
RuvB RuvC

Branch migration and resolution of Holliday junction
(Common)
BLM Mus81
TOP3 Eme1

Strand displacement

Strand displacement

RecFOR pathway

Replication restart
PriC
PriA PriB DnaT

Crossover or

Flap removal and annealing
(Common)
Ligation
BLM

PriA

RecBC pathway

Non-crossover
DSBR
Double-strand break repair

SDSA
Synthesis-dependent strand annealing

BIR
Break-induced replication

Data on KEGG graph
Rendered by Pathview

## Gene ontology

```
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 | 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 | 1.925222e-04 | 3.565432 | 1.925222e-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                      0.1967577        426 1.925222e-04
GO:0060562 epithelial tube morphogenesis 0.3565320        257 5.932837e-04
GO:0035295 tube development              0.3565320        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                      3.565432 3.565432
GO:0060562 epithelial tube morphogenesis 3.261376 3.261376
GO:0035295 tube development              3.253665 3.253665
```

**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: 8147"
```

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

```
            col.names=FALSE,
            quote=FALSE)
```

Cell cycle, mitotic: it matches the previous KEGG result. Reactome analysis generates a more comprehensive analysis of enriched pathways than KEGG, and therefore returns a more detailed pathway.