# Class 14: RNASeq mini Project

Le, Sarah (PID: A18518276)

## Table of contents



### Background

Here we work through a complete RNASeq analysis project. The input data comes from a knock-down experiment of a HOX gene.

### Data Import

Reading the counts and metadata CSV files

```
countData <- read.csv("GSE37704_featurecounts.csv",row.names = 1)
metadata <- read.csv("GSE37704_metadata.csv")
```

```
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
```

```
head(metadata)
```

```
         id     condition
1 SRR493366 control_sirna
2 SRR493367 control_sirna
3 SRR493368 control_sirna
4 SRR493369      hoxa1_kd
5 SRR493370      hoxa1_kd
6 SRR493371      hoxa1_kd
```

Q1. Complete the code below to filter countData to exclude genes (i.e. rows) where we have 0 read count across all samples (i.e. columns).

```
# Note we need to remove the odd first $length col
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
```

Q2. Complete the code below to filter countData to exclude genes (i.e. rows) where we have 0 read count across all samples (i.e. columns).

```
# Filter count data where you have 0 read count across all samples.
countData = countData[rowSums(countData)>0, ]
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
```

Some book-keeping is required as there looks to be a mis-match between metadata rows and counts columns.

```
ncol(countData)
```

```
[1] 6
```

```
nrow(metadata)
```

```
[1] 6
```

We need to remove the first column for count.

```
cleancounts <- countData[,-1]
head(cleancounts)
```

```
                SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
ENSG00000279457        28        29        29        28        46
ENSG00000187634       123       205       207       212       258
ENSG00000188976      1831      2383      1226      1326      1504
ENSG00000187961       153       180       236       255       357
ENSG00000187583        48        65        44        48        64
ENSG00000187642         9        16        14        16        16
```

```
all(colnames(cleancounts) == metadata$id)
```

```
Warning in colnames(cleancounts) == metadata$id: longer object length is not a
multiple of shorter object length
```

```
[1] FALSE
```

### Remove zero count genes

There are lost of genes with zero counts. We can remove these from further analysis.

```
head(cleancounts)
```

```
                SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
ENSG00000279457        28        29        29        28        46
ENSG00000187634       123       205       207       212       258
ENSG00000188976      1831      2383      1226      1326      1504
ENSG00000187961       153       180       236       255       357
ENSG00000187583        48        65        44        48        64
ENSG00000187642         9        16        14        16        16
```

```
to.keep.inds <- rowSums(cleancounts) >0
nonzero_counts <- cleancounts[to.keep.inds,]
```

### DESeq Analysis

Load the package.

```
library(DESeq2)
```

Setup DESeq.

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

```
Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
design formula are characters, converting to factors
```

Run DESeq.

```
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(3): id condition sizeFactor
```

Get results.

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

> Q3. 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.

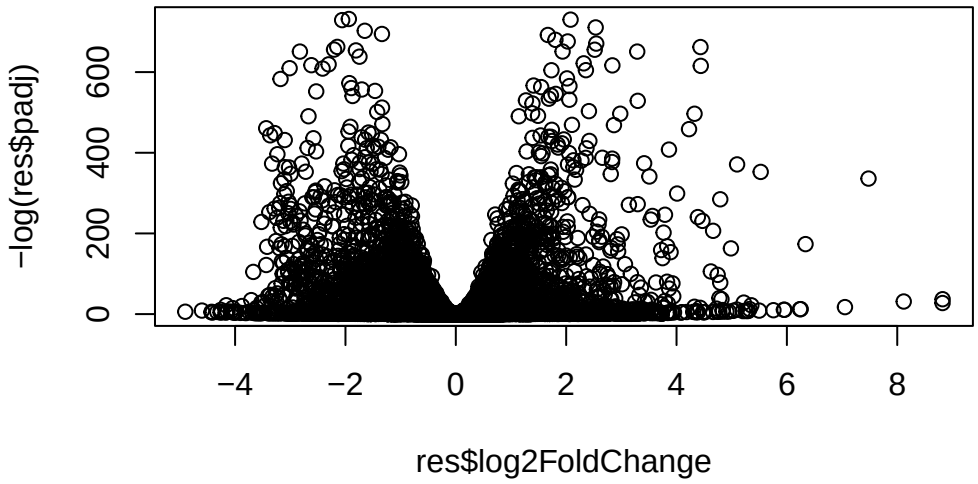
```
summary(res)
```

```
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
```

## Data Visualization

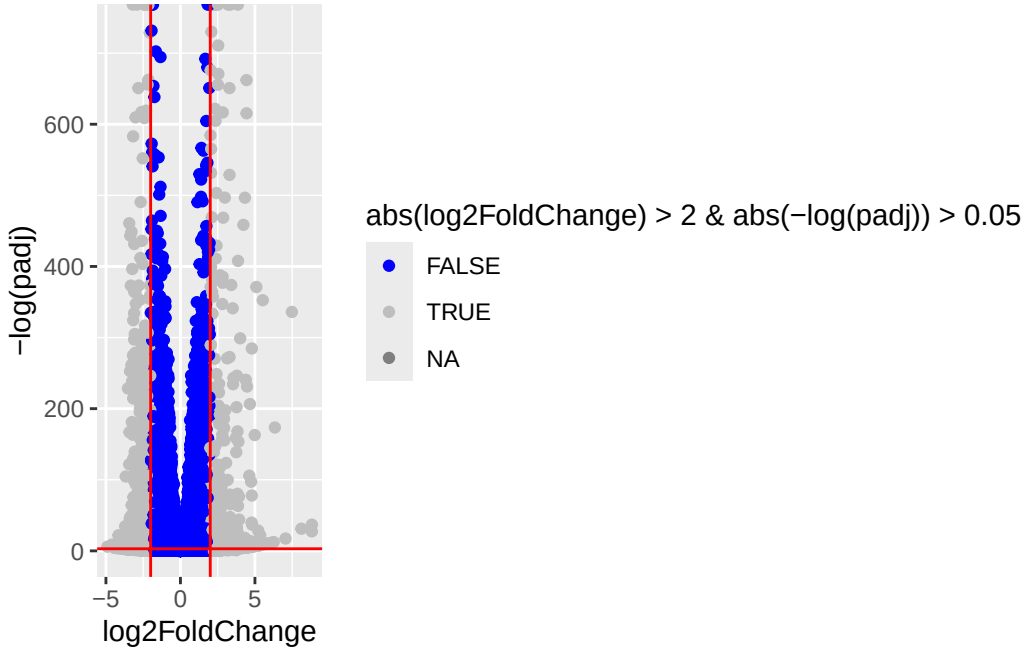
Volcano plot.

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




```
library(ggplot2)

ggplot(res) +
  aes(log2FoldChange, -log(padj),
      color = abs(log2FoldChange) > 2 & abs(-log(padj)) > 0.05) +
  geom_point() +
  scale_color_manual(values = c("blue", "grey")) +
  geom_vline(xintercept = c (-2,2), col="red") +
  geom_hline(yintercept = -log(0.05), col="red")
```

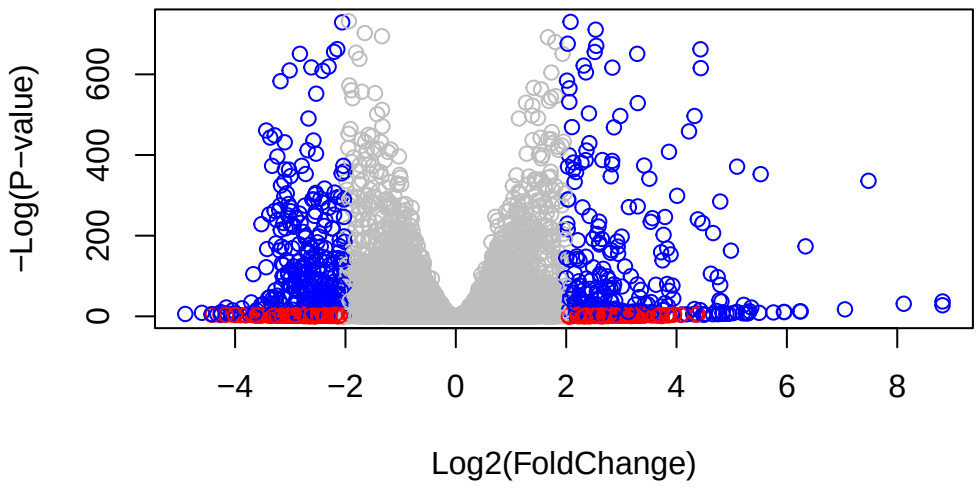
```
Warning: Removed 1237 rows containing missing values or values outside the scale range
(`geom_point()`).
```




Add threshold lines for fold-change and p-value and color our subset of genes that make these threshold cut-offs in the plot.

Q4. Improve this plot by completing the below code, which adds color and axis labels

```
# 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
```




## Add annotation

Add gene symbols and entrez ids.

> Q5. Use the mapIDs() function multiple times to add SYMBOL, ENTREZID and GENENAME annotation to our results by completing the code below.

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

```
res$symbol <- mapIds(x=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(x=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   29.913579      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 ..
```

Q6. Finally for this section let's reorder these results by adjusted p-value and save them to a CSV file in your current project directory.

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

## Pathway Analysis

## KEGG Pathways

Run gage analysis with KEGG.

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

```
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"
```

We need a named vector of fold-change values as input for gage.

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

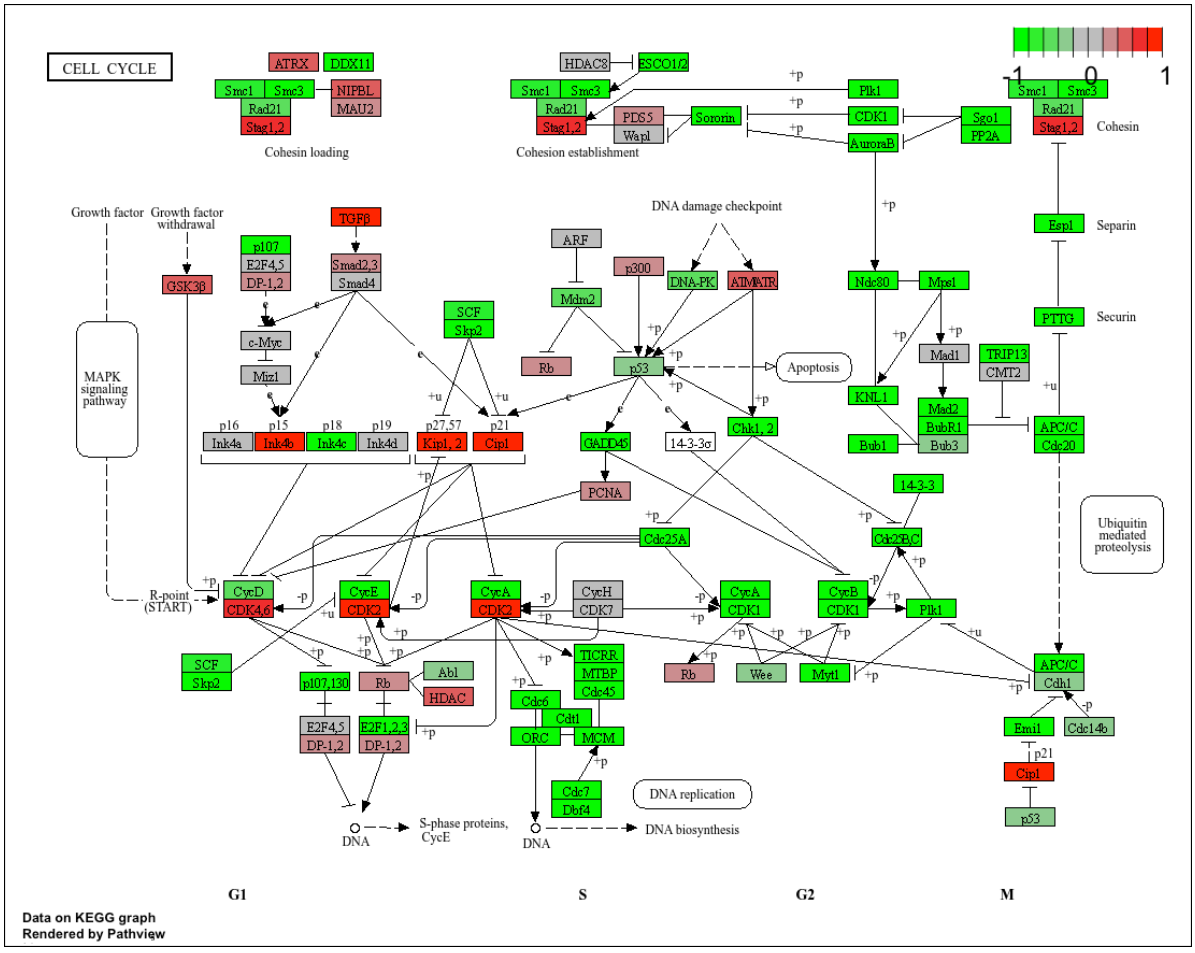
```
     1266     54855      1465      2034      2150      6659
-2.422719  3.201955 -2.313738 -1.888019  3.344508  2.392288
```

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

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

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14
```

```
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
```

```
Warning: reconcile groups sharing member nodes!
```

```
     [,1] [,2]
[1,] "9"  "300"
[2,] "9"  "306"
```

```
Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14
```

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

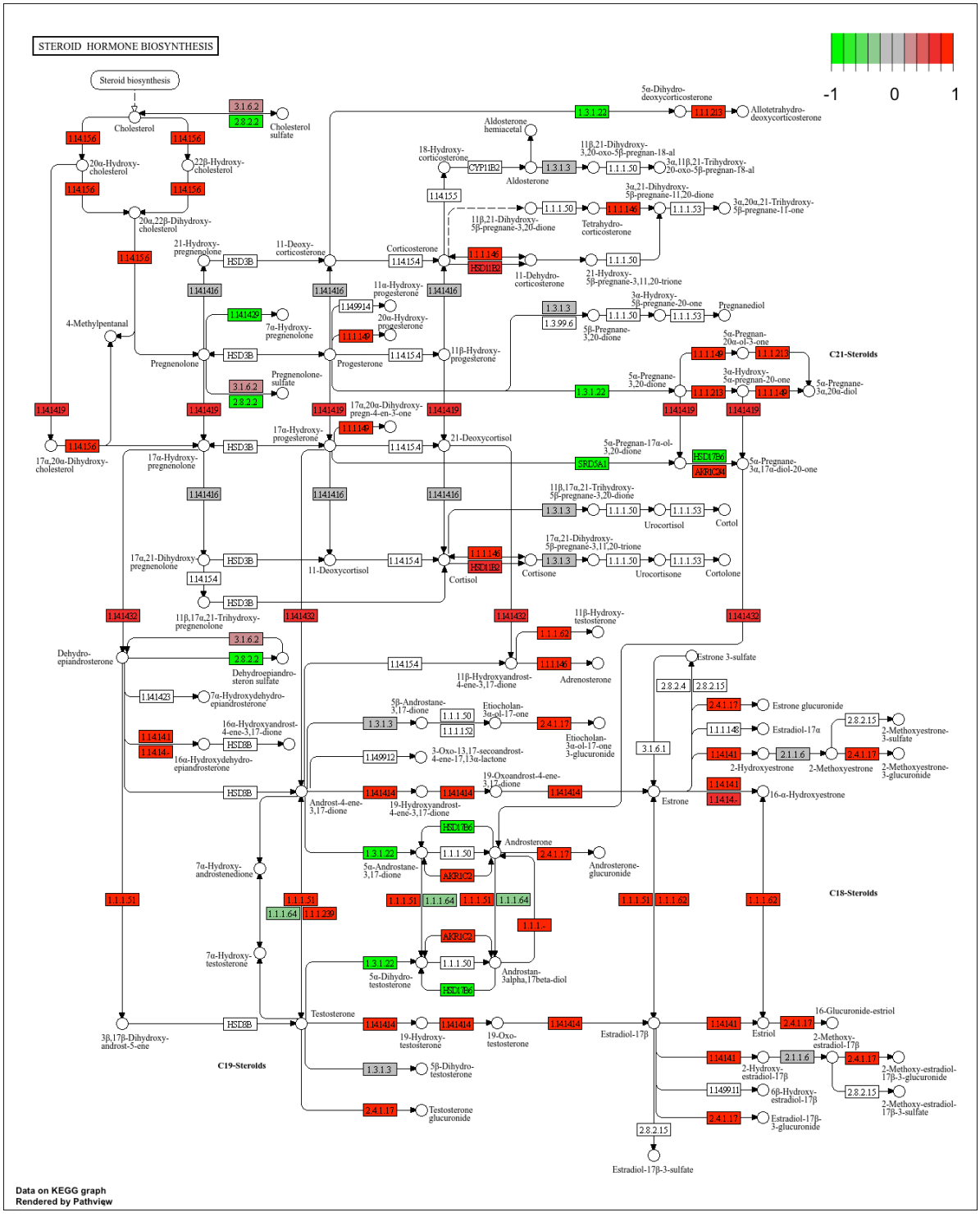
```
## Focus on top 5 upregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$greater)[1:5]

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

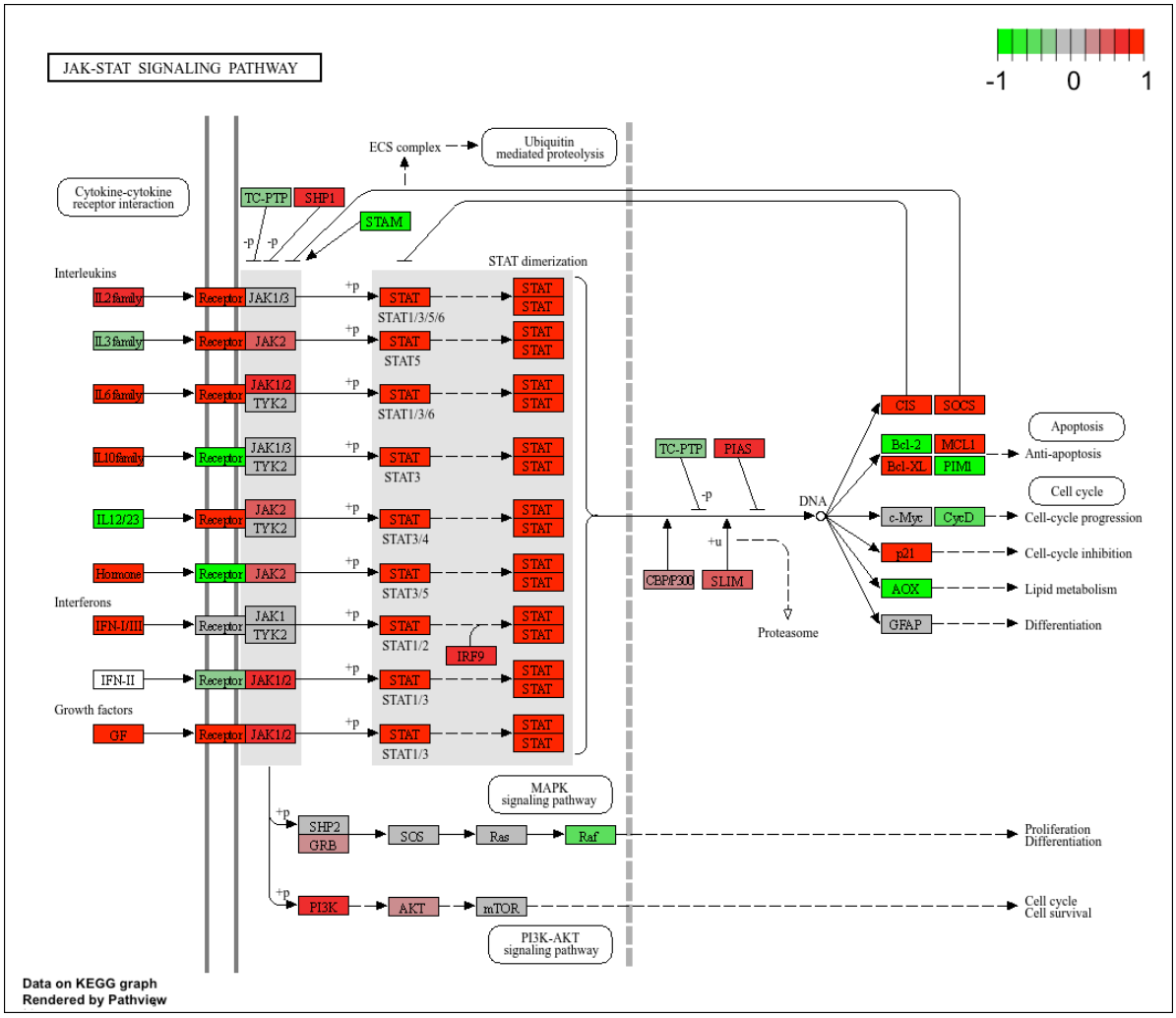
```
[1] "hsa04640" "hsa04630" "hsa00140" "hsa04142" "hsa04330"
```

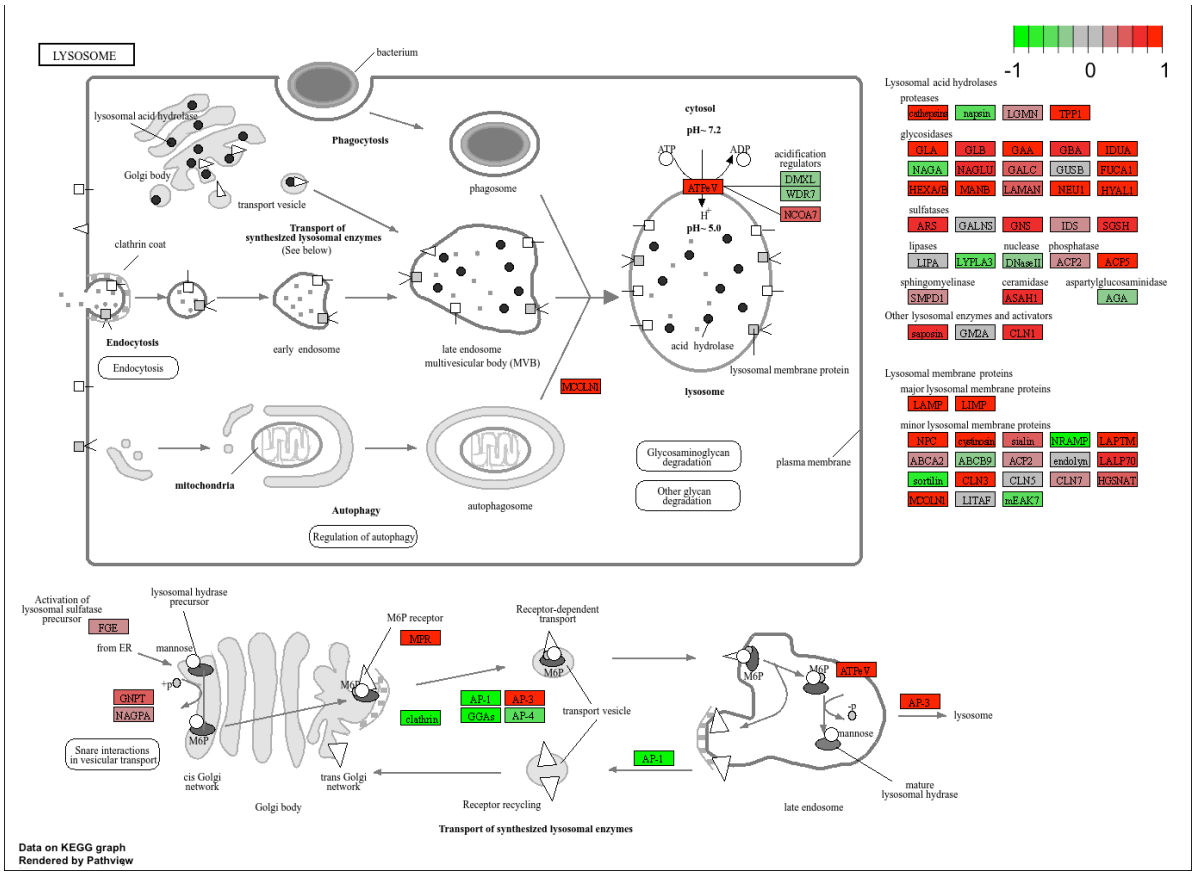
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

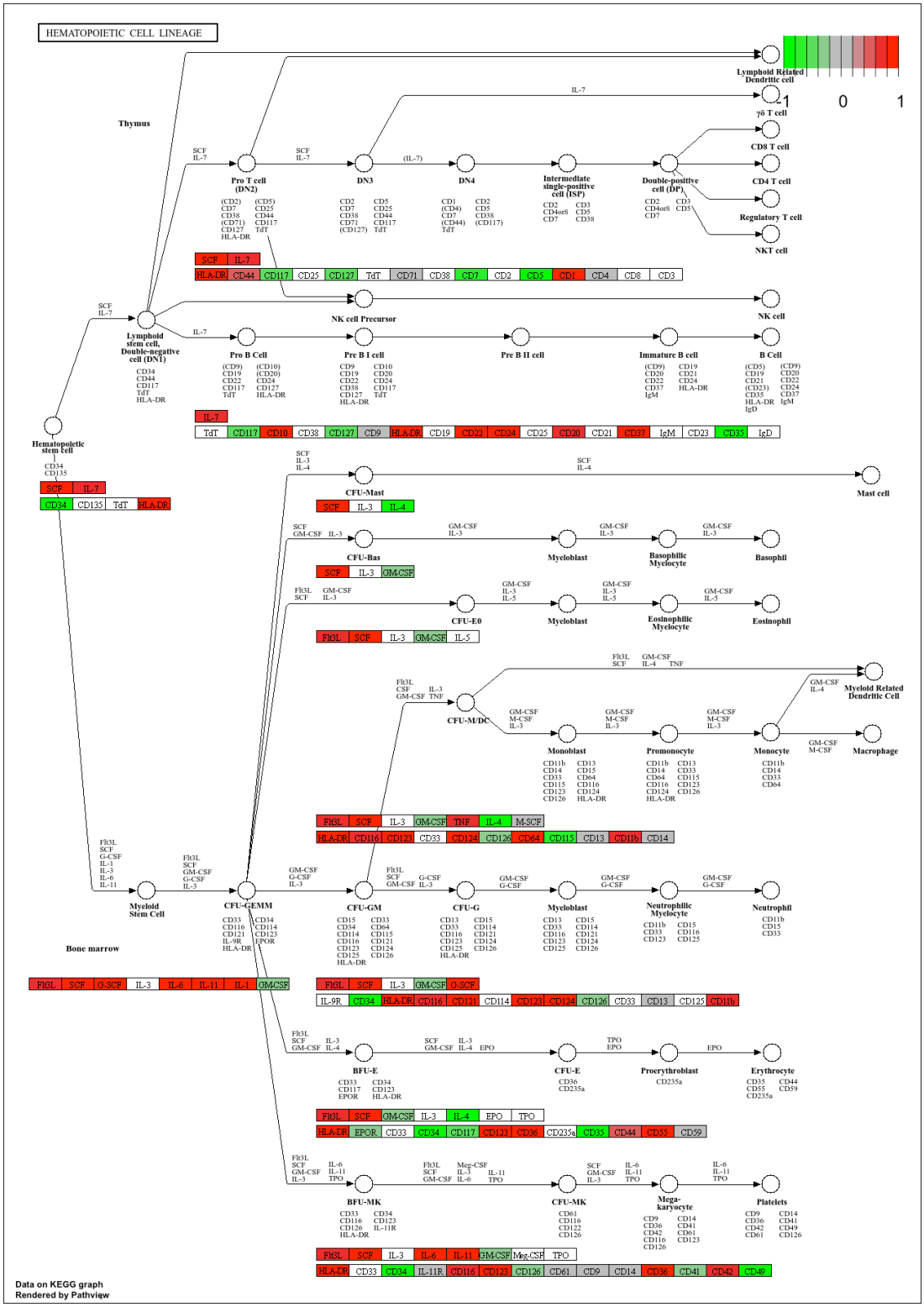
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa04142.pathview.png
```

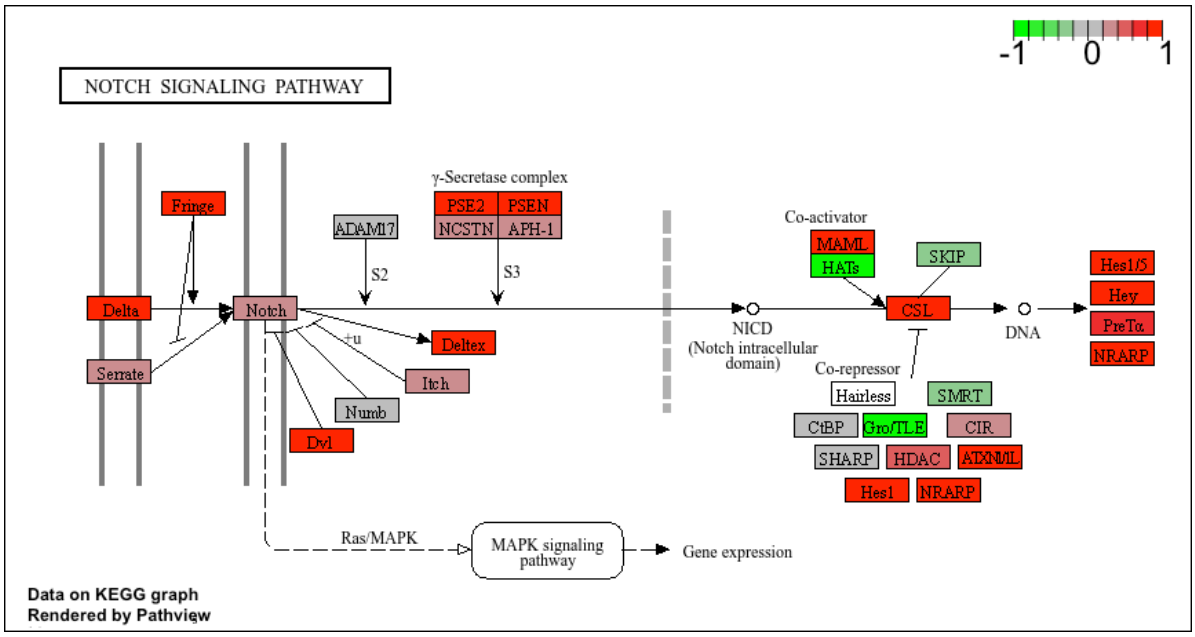
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

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


STEROID HORMONE BIOSYNTHESIS
-1
0
1
Steroid biosynthesis
Cholesterol
Cholesterol sulfate
20α-Hydroxy-cholesterol
22β-Hydroxy-cholesterol
20α,22β-Dihydroxy-cholesterol
21-Hydroxy-pregnenolone
11-Deoxy-corticosterone
Corticosterone
11-Dehydro-corticosterone
Aldosterone
Aldosterone hemiacetal
18-Hydroxy-corticosterone
11β,21-Dihydroxy-3,20-oxo-5β-pregnan-18-al
3α,11β,21-Trihydroxy-20-oxo-5β-pregnan-18-al
3α,21-Dihydroxy-5β-pregnane-11,20-dione
3α,20α,21-Trihydroxy-5β-pregnane-11-one
11β,21-Dihydroxy-5β-pregnane-3,20-dione
Tetrahydro-corticosterone
21-Hydroxy-5β-pregnane-3,11,20-trione
5α-Dihydro-deoxycorticosterone
Allotetrahydro-deoxycorticosterone
4-Methylpentanal
Pregnenolone
7α-Hydroxy-pregnenolone
Progesterone
11α-Hydroxy-progesterone
20α-Hydroxy-progesterone
11β-Hydroxy-progesterone
Pregnenolone-sulfate
17α,20α-Dihydroxy-pregn-4-en-3-one
17α,20α-Dihydroxy-cholesterol
17α-Hydroxy-pregnenolone
17α-Hydroxy-progesterone
21-Deoxycortisol
11-Deoxycortisol
17α,21-Dihydroxy-pregnenolone
Cortisol
Cortisone
11β,17α,21-Trihydroxy-pregnenolone
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Urocortisol
Cortol
Urocortisone
Cortolone
3α-Hydroxy-5β-pregnane-20-one
Pregnanediol
5β-Pregnane-3,20-dione
5α-Pregnan-20α-ol-3-one
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
5α-Pregnane-3,20-dione
5α-Pregnan-17α-ol-3,20-dione
5α-Pregnane-3α,17α-diol-20-one
C21-Steroids
Dehydro-epiandrosterone
Dehydroepiandro-steron sulfate
7α-Hydroxydehydro-epiandrosterone
16α-Hydroxyandrost-4-ene-3,17-dione
16α-Hydroxydehydro-epiandrosterone
3-Oxo-13,17-secoandrost-4-ene-17,13α-lactone
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
11β-Hydroxy-testosterone
5β-Androstane-3,17-dione
Etiocholan-3α-ol-17-one
Etiocholan-3α-ol-17-one 3-glucuronide
19-Oxoandrost-4-ene-3,17-dione
Androst-4-ene-3,17-dione
19-Hydroxyandrost-4-ene-3,17-dione
Androsterone
Androsterone-glucuronide
5α-Androstane-3,17-dione
7α-Hydroxy-androstenedione
Estrone 3-sulfate
Estrone glucuronide
Estradiol-17α
2-Hydroxyestrone
2-Methoxyestrone
2-Methoxyestrone-3-sulfate
2-Methoxyestrone-3-glucuronide
Estrone
16-α-Hydroxyestrone
C18-Steroids
7α-Hydroxy-testosterone
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
Testosterone
19-Hydroxy-testosterone
19-Oxo-testosterone
Estradiol-17β
3β,17β-Dihydroxy-androst-5-ene
5β-Dihydro-testosterone
Testosterone glucuronide
C19-Steroids
16-Glucuronide-estriol
Estriol
2-Methoxy-estradiol-17β
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β-3-glucuronide
6β-Hydroxy-estradiol-17β
Estradiol-17β-3-glucuronide
2-Methoxy-estradiol-17β-3-sulfate
Estradiol-17β-3-sulfate
Data on KEGG graph
Rendered by Pathview


JAK-STAT SIGNALING PATHWAY
-1
0
1
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
TC-PTP
SHP1
STAM
STAT dimerization
Interleukins
IL2 family
IL3 family
IL6 family
IL10 family
IL12/23
Hormone
Interferons
IFN-I/III
IFN-II
Growth factors
GF
Receptor
JAK1/3
JAK2
JAK1/2
TYK2
JAK1
STAT
STAT1/3/5/6
STAT5
STAT1/3/6
STAT3
STAT3/4
STAT3/5
STAT1/2
STAT1/3
IRF9
TC-PTP
PIAS
CBP/P300
SLIM
Proteasome
DNA
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
Apoptosis
Anti-apoptosis
Cell cycle
c-Myc
CycD
Cell-cycle progression
p21
Cell-cycle inhibition
AOX
Lipid metabolism
GFAP
Differentiation
MAPK signaling pathway
SHP2
GRB
SOS
Ras
Raf
Proliferation
Differentiation
PI3K
AKT
mTOR
Cell cycle
Cell survival
PI3K-AKT signaling pathway
Data on KEGG graph
Rendered by Pathview


LYSOSOME
bacterium
lysosomal acid hydrolase
Phagocytosis
Golgi body
transport vesicle
phagosome
cytosol
pH~ 7.2
ATP
ADP
acidification regulators
ATPeV
DMXL
WDR7
NCOA7
Transport of synthesized lysosomal enzymes
(See below)
clathrin coat
Endocytosis
Endocytosis
early endosome
late endosome
multivesicular body (MVB)
MCOLN1
pH~ 5.0
acid hydrolase
lysosomal membrane protein
lysosome
mitochondria
Autophagy
autophagosome
Regulation of autophagy
Glycosaminoglycan degradation
Other glycan degradation
plasma membrane
-1
0
1
Lysosomal acid hydrolases
proteases
cathepsins
saposin
LGMN
TPP1
glycosidases
GLA
GLB
GAA
GBA
IDUA
NAGA
NAGLU
GALC
GUSB
FUCA1
HEXA/B
MANB
LAMAN
NEU1
HYAL1
sulfatases
ARS
GALNS
GNS
IDS
SGSH
lipases
nuclease
phosphatase
LIPA
LYPLA3
DNaseII
ACP2
ACP5
sphingomyelinase
ceramidase
aspartylglucosaminidase
SMPD1
ASAH1
AGA
Other lysosomal enzymes and activators
saposin
GM2A
CLN1
Lysosomal membrane proteins
major lysosomal membrane proteins
LAMP
LIMP
minor lysosomal membrane proteins
NPC
cystinosin
sialin
VRAMP
LAPTM
ABCA2
ABCB9
ACP2
endolyn
LALP70
sortilin
CLN3
CLN5
CLN7
HGSNAT
MCOLN1
LITAF
mEAK7
Activation of lysosomal sulfatase precursor
FGE
lysosomal hydrase precursor
from ER
mannose
GNPT
NAGPA
M6P
M6P receptor
MPR
Receptor-dependent transport
M6P
clathrin
AP-1
GGAs
AP-3
AP-4
transport vesicle
Snare interactions in vesicular transport
cis Golgi network
trans Golgi network
Golgi body
Receptor recycling
AP-1
M6P
ATPeV
AP-3
lysosome
mannose
mature lysosomal hydrase
late endosome
Transport of synthesized lysosomal enzymes
Data on KEGG graph
Rendered by Pathview


HEMATOPOIETIC CELL LINEAGE
Lymphoid Related Dendritic cell
-1
0
1
Thymus
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF
IL-7
Pro T cell (DN2)
DN3
DN4
Intermediate single-positive cell (ISP)
Double-positive cell (DP)
(IL-7)
HLA-DR CD44 CD117 CD25 CD127 TdT CD71 CD38 CD7 CD2 CD5 CD1 CD4 CD8 CD3
NK cell Precursor
NK cell
Lymphoid stem cell, Double-negative cell (DN1)
Pro B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
TdT CD10 CD38 CD127 CD9 HLA-DR CD19 CD22 CD24 CD25 CD20 CD21 CD37 IgM CD23 CD35 IgD
Hematopoietic stem cell
CD34 CD135
SCF IL-7 CD34 CD135 TdT HLA-DR
CFU-Mast
Mast cell
SCF IL-3 IL-4
CFU-Bas
Myeloblast
Basophilic Myelocyte
Basophil
SCF IL-3 GM-CSF
CFU-E0
Eosinophilic Myelocyte
Eosinophil
Flt3L SCF IL-3 GM-CSF IL-5
CFU-M/DC
Myeloid Related Dendritic Cell
Monoblast
Promonocyte
Monocyte
Macrophage
Flt3L SCF IL-3 GM-CSF TNF IL-4 M-SCF
HLA-DR CD116 CD123 CD33 CD124 CD126 CD64 CD115 CD13 CD11b CD14
Myeloid Stem Cell
CFU-GEMM
CFU-GM
CFU-G
Myeloblast
Neutrophilic Myelocyte
Neutrophil
Bone marrow
Flt3L SCF G-CSF IL-3 IL-6 IL-11 IL-1 GM-CSF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD11b
BFU-E
CFU-E
Procrythroblast
Erythrocyte
Flt3L SCF GM-CSF IL-3 EPO TPO
HLA-DR EPOR CD33 CD34 CD117 CD123 CD36 CD235a CD35 CD44 CD55 CD59
BFU-MK
CFU-MK
Mega-karyocyte
Platelets
Flt3L SCF IL-3 IL-6 IL-11 GM-CSF Meg-CSF TPO
HLA-DR CD33 CD34 IL-11R CD116 CD123 CD126 CD61 CD9 CD14 CD36 CD41 CD42 CD49
Data on KEGG graph
Rendered by Pathview




Q7. Can you do the same procedure as above to plot the pathview figures for the top 5 down-reguled pathways?

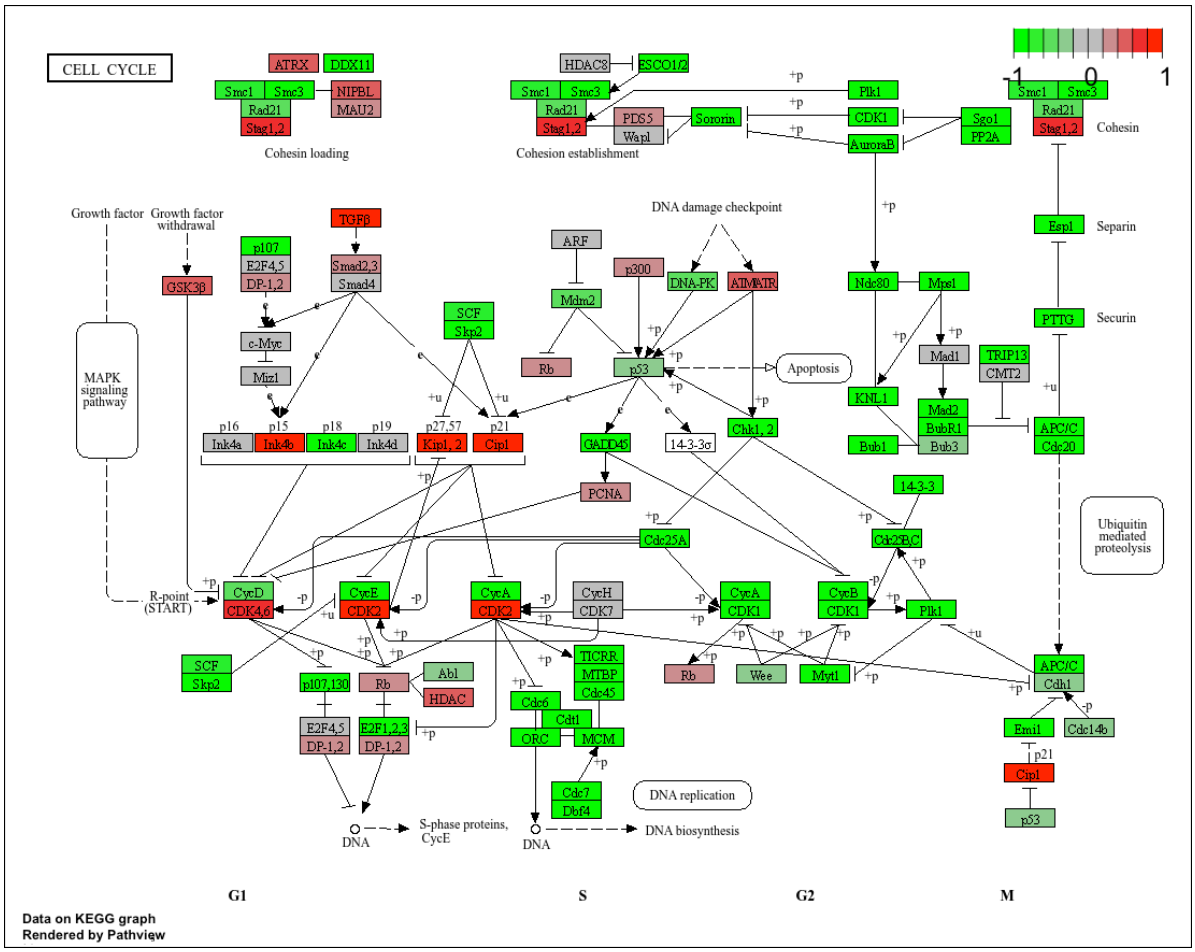
```
# Get top 5 DOWN-regulated pathways
keggrespathways_down <- rownames(keggres$less)[1:5]

# Extract only the KEGG ID part (first 8 characters)
keggresids_down <- substr(keggrespathways_down, start = 1, stop = 8)

keggresids_down
```

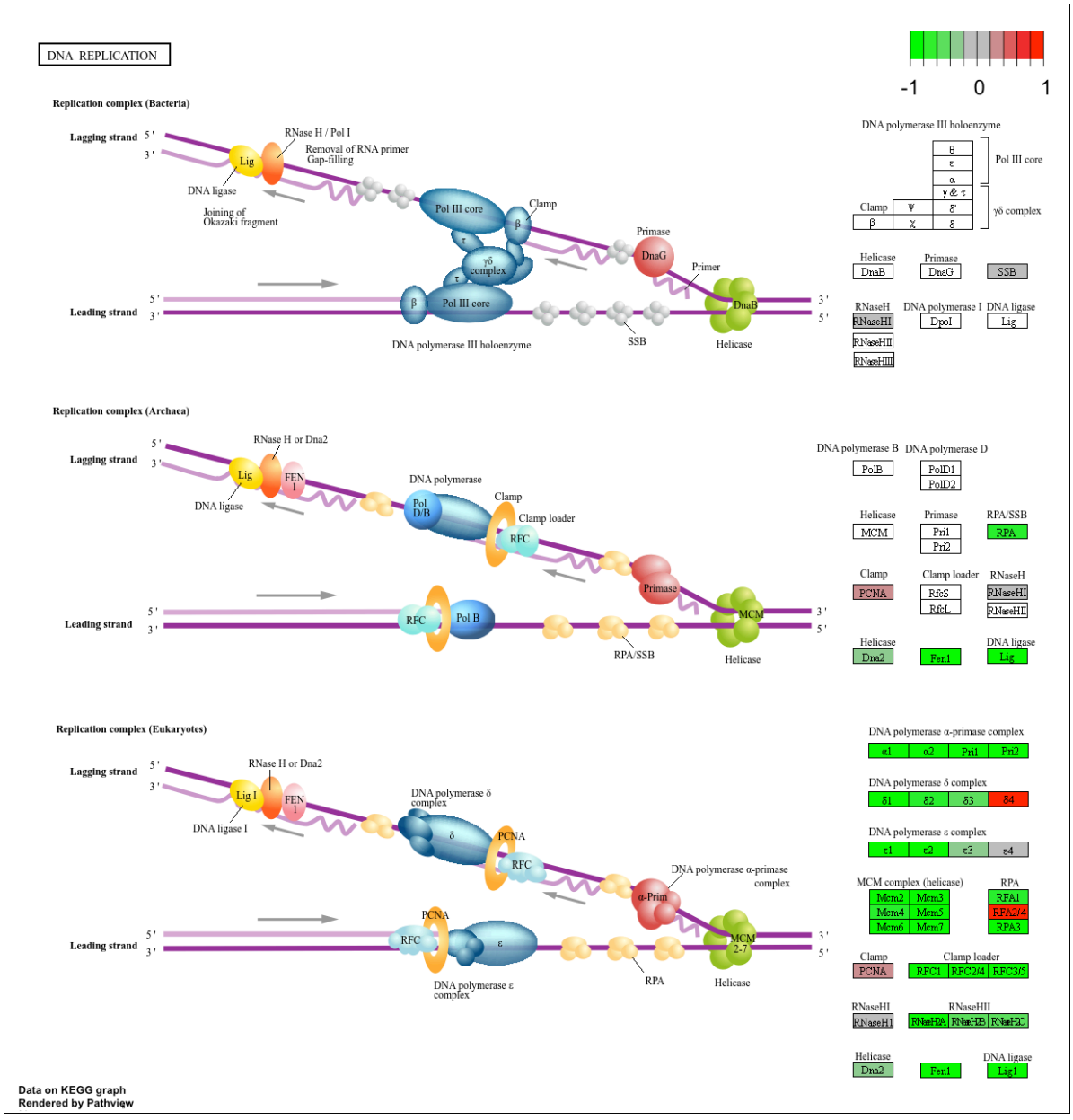
```
[1] "hsa04110" "hsa03030" "hsa03013" "hsa03440" "hsa04114"
```

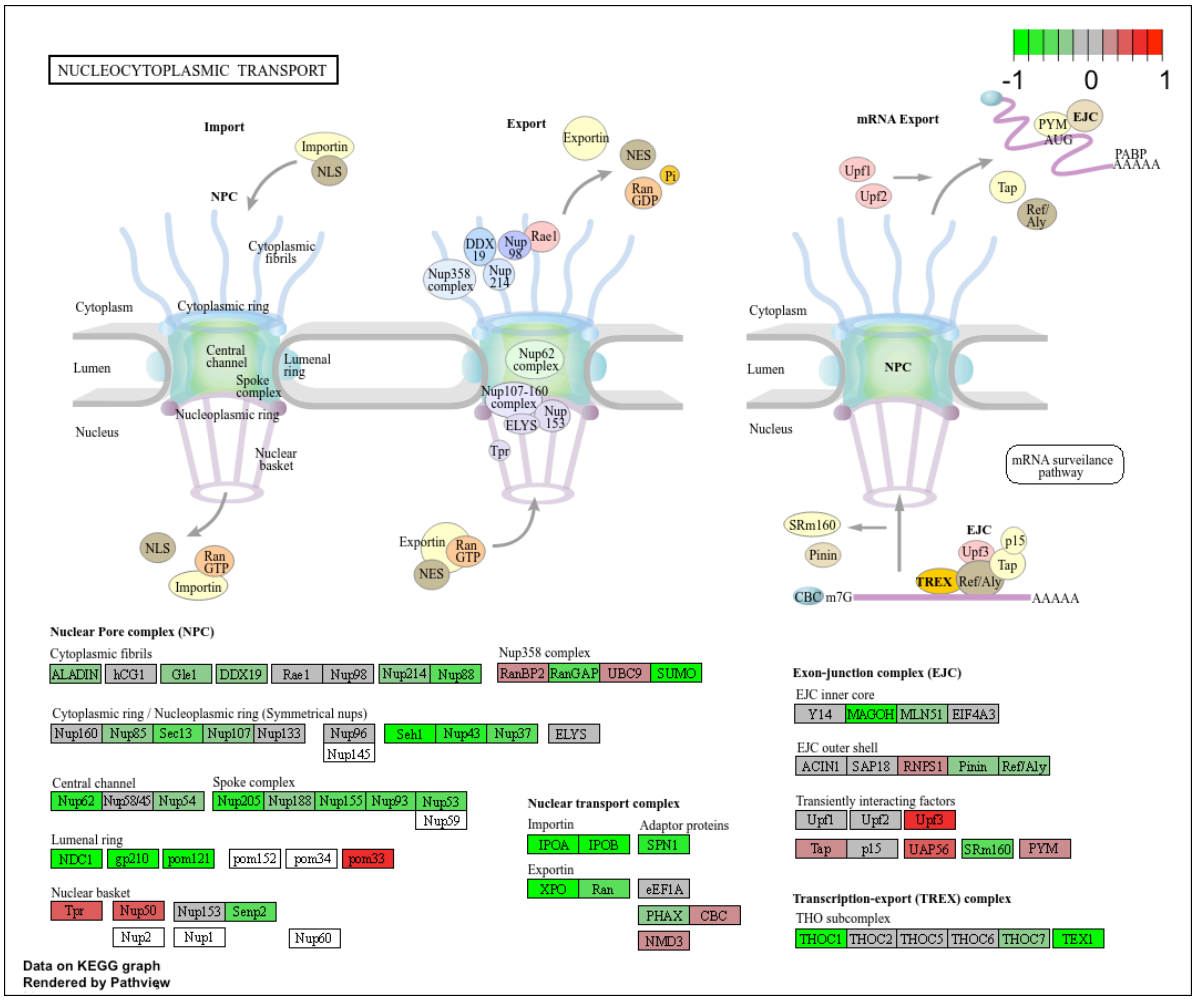
```
pathview(gene.data = foldchanges,
         pathway.id = keggresids_down,
         species = "hsa")
```

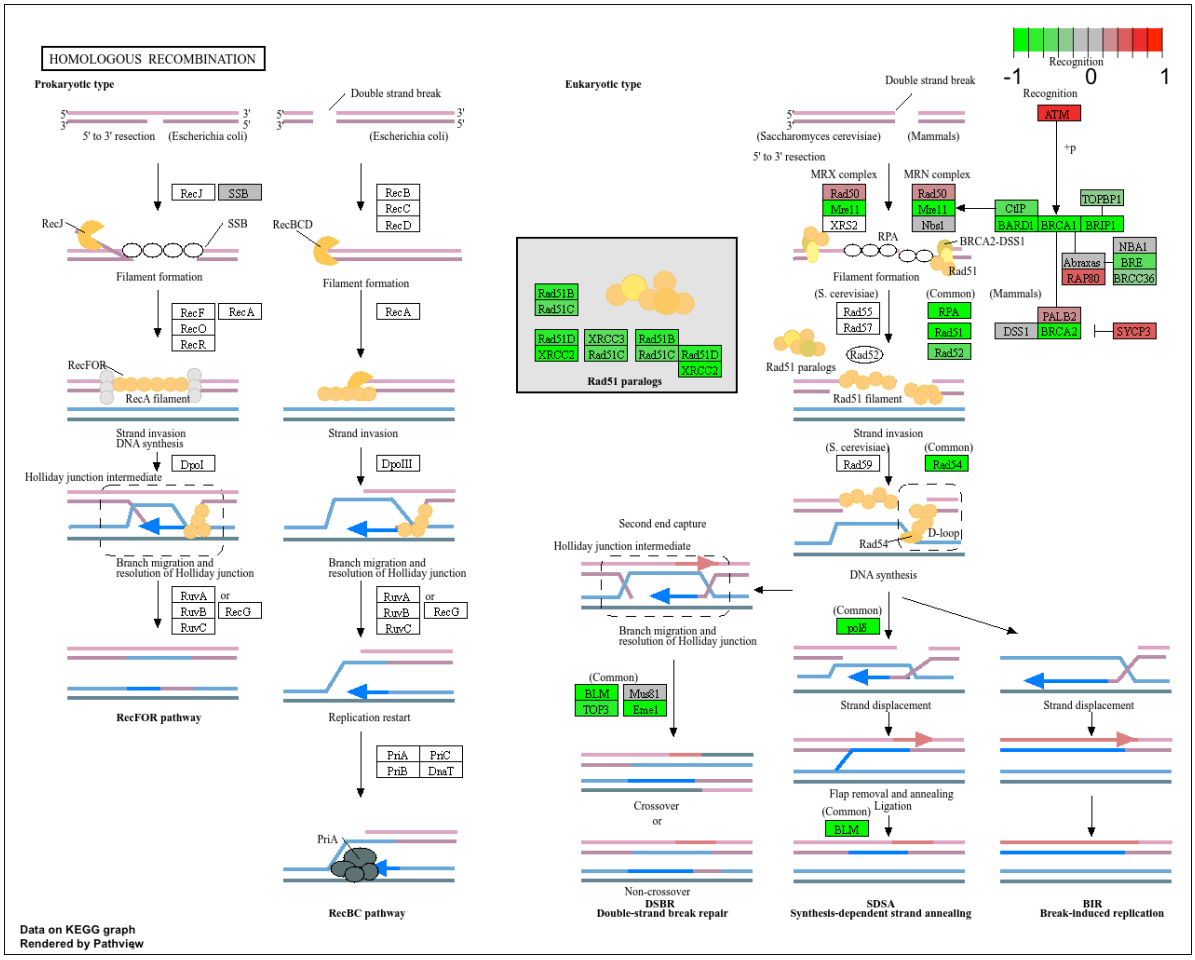
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

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

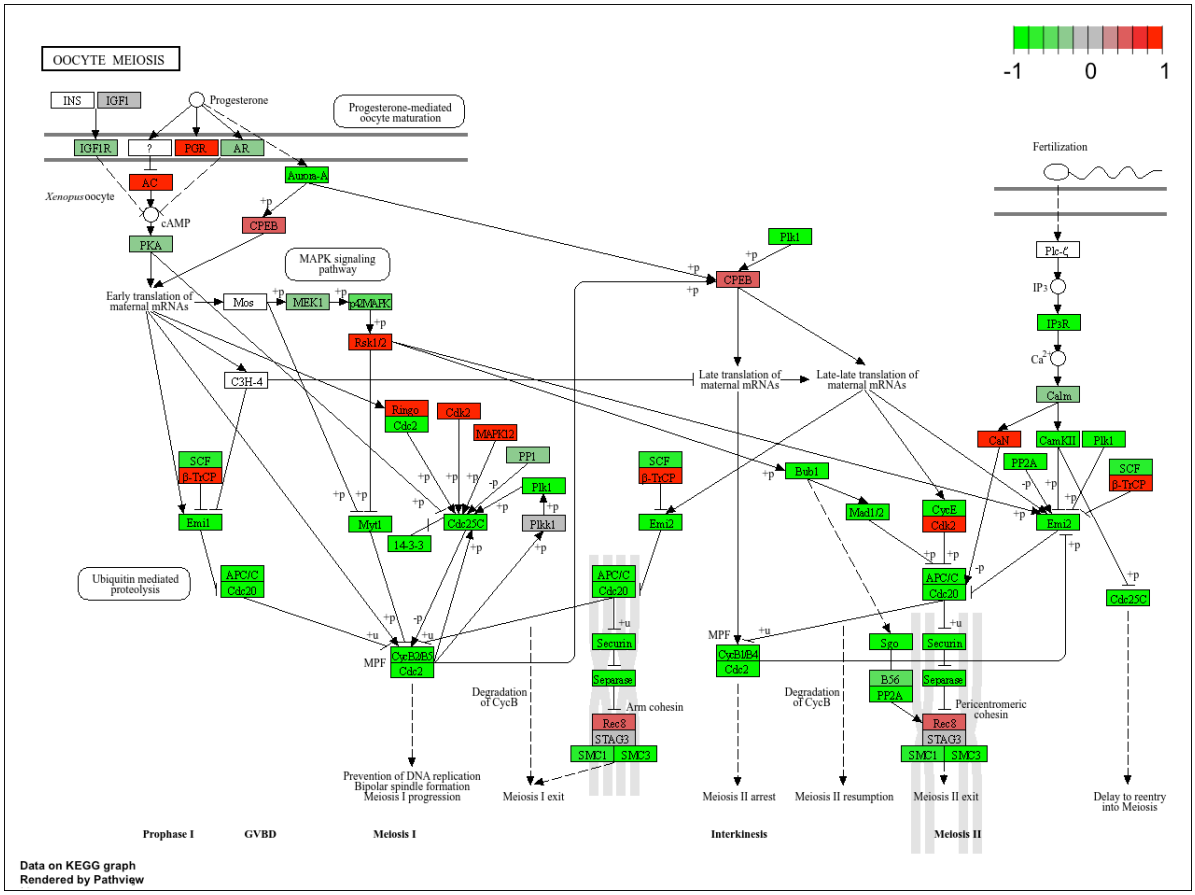
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

Info: Writing image file hsa03440.pathview.png

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

Info: Working in directory /Users/sarah/Downloads/BIMM 143/Class14

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


CELL CYCLE
ATRX
DDX11
Smc1
Smc3
NIPBL
MAU2
Rad21
Stag1,2
Cohesin loading
HDAC8
ESCO1/2
Smc1
Smc3
Rad21
Stag1,2
PDS5
Wapl
Sororin
Cohesion establishment
+p
Plk1
CDK1
AuroraB
Sgo1
PP2A
-1
0
1
Smc1
Smc3
Rad21
Stag1,2
Cohesin
Esp1
Separin
PTTG
Securin
Growth factor
Growth factor withdrawal
TGFβ
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
GSK3β
MAPK signaling pathway
c-Myc
Miz1
SCF
Skp2
ARF
Mdm2
p300
DNA-PK
ATM&ATR
DNA damage checkpoint
Rb
p53
Apoptosis
Ndc80
Mps1
KNL1
Mad1
Mad2
BubR1
Bub1
Bub3
TRIP13
CMT2
APC/C
Cdc20
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
GADD45
14-3-3σ
Chk1, 2
PCNA
14-3-3
Cdc25A
Cdk2BC
Ubiquitin mediated proteolysis
R-point
(START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
TICRR
MTBP
Cdc45
Cdc6
Cdt1
ORC
MCM
Rb
Wee
Myt1
APC/C
Cdh1
Emi1
Cdc14b
p21
Cip1
p53
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
Cdc7
Dbf4
DNA replication
DNA
S-phase proteins, CycE
DNA
DNA biosynthesis
G1
S
G2
M
Data on KEGG graph
Rendered by Pathview


DNA REPLICATION
-1
0
1
Replication complex (Bacteria)
Lagging strand
5 '
3 '
RNase H / Pol I
Removal of RNA primer
Gap-filling
Lig
DNA ligase
Joining of
Okazaki fragment
Pol III core
Clamp
β
τ
γδ
complex
Primase
DnaG
Primer
DnaB
Leading strand
Pol III core
β
DNA polymerase III holoenzyme
SSB
Helicase
DNA polymerase III holoenzyme
θ
ε
α
γ & τ
Pol III core
Clamp
ψ
χ
δ'
δ
γδ complex
β
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
RNaseHI
RNaseHII
RNaseHIII
DNA polymerase I
DpoI
DNA ligase
Lig
Replication complex (Archaea)
RNase H or Dna2
Lig
FEN
1
DNA ligase
DNA polymerase
Pol
D/B
Clamp
Clamp loader
RFC
Primase
MCM
RPA/SSB
Helicase
DNA polymerase B
PolB
DNA polymerase D
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
RfcS
RfcL
RNaseH
RNaseHI
RNaseHII
Helicase
Dna2
Fen1
DNA ligase
Lig
Replication complex (Eukaryotes)
RNase H or Dna2
Lig I
FEN
1
DNA ligase I
DNA polymerase δ
complex
δ
PCNA
RFC
DNA polymerase α-primase
complex
α-Prim
MCM
2-7
RPA
Helicase
PCNA
RFC
ε
DNA polymerase ε
complex
DNA polymerase α-primase complex
α1
α2
Pri1
Pri2
DNA polymerase δ complex
δ1
δ2
δ3
δ4
DNA polymerase ε complex
ε1
ε2
ε3
ε4
MCM complex (helicase)
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
RPA
RFA1
RFA2/4
RFA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHI
RNaseH1
RNaseHII
RNaseHIII
RNaseH2A
RNaseH2B
RNaseH2C
Helicase
Dna2
Fen1
DNA ligase
Lig1
Data on KEGG graph
Rendered by Pathview


NUCLEOCYTOPLASMIC TRANSPORT
-1
0
1
Import
Importin
NLS
NPC
Cytoplasmic fibrils
Cytoplasm
Cytoplasmic ring
Central channel
Lumenal ring
Lumen
Spoke complex
Nucleoplasmic ring
Nucleus
Nuclear basket
NLS
Ran GTP
Importin
Export
Exportin
NES
Pi
Ran GDP
DDX 19
Nup 98
Rae1
Nup358 complex
Nup 214
Nup62 complex
Nup107-160 complex
ELYS
Nup 153
Tpr
Exportin
Ran GTP
NES
mRNA Export
EJC
PYM
AUG
PABP
AAAAA
Upf1
Upf2
Tap
Ref/ Aly
Cytoplasm
Lumen
NPC
Nucleus
mRNA surveilance pathway
SRm160
Pinin
EJC
Upf3
p15
Tap
TREX
Ref/Aly
CBC
m7G
AAAAA
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 nups)
Nup160
Nup85
Sec13
Nup107
Nup133
Nup96
Nup145
Seh1
Nup43
Nup37
ELYS
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
MAGOH
MLN51
EIF4A3
EJC outer shell
ACIN1
SAP18
RNPS1
Pinin
Ref/Aly
Transiently interacting factors
Upf1
Upf2
Upf3
Tap
p15
UAP56
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
Double strand break
5' to 3' resection
(Escherichia coli)
(Escherichia coli)
RecJ
SSB
RecJ
SSB
RecBCD
RecB
RecC
RecD
Filament formation
Filament formation
RecF
RecO
RecR
RecA
RecA
RecFOR
RecA filament
Strand invasion
DNA synthesis
Strand invasion
DpoI
DpoIII
Holliday junction intermediate
Branch migration and
resolution of Holliday junction
Branch migration and
resolution of Holliday junction
RuvA
RuvB
RuvC
or
RecG
RuvA
RuvB
RuvC
or
RecG
RecFOR pathway
Replication restart
PriA
PriB
PriC
DnaT
PriA
RecBC pathway
Eukaryotic type
Recognition
-1
0
1
Double strand break
Recognition
ATM
(Saccharomyces cerevisiae)
(Mammals)
+p
5' to 3' resection
MRX complex
MRN complex
Rad50
Mre11
XRS2
Rad50
Mre11
Nbs1
CtIP
BARD1
BRCA1
BRIP1
TOPBP1
RPA
BRCA2-DSS1
NBA1
BRE
BRCC36
Abraxas
RAP80
Rad51
Filament formation
(S. cerevisiae)
(Common)
(Mammals)
Rad55
Rad57
RPA
Rad51
Rad52
PALB2
BRCA2
DSS1
SYCP3
Rad51 paralogs
(Rad52)
Rad51 filament
Rad51B
Rad51C
Rad51D
XRCC2
XRCC3
Rad51C
Rad51B
Rad51C
Rad51D
XRCC2
Rad51 paralogs
Strand invasion
(S. cerevisiae)
(Common)
Rad59
Rad54
D-loop
Rad54
DNA synthesis
Second end capture
Holliday junction intermediate
(Common)
pol δ
Branch migration and
resolution of Holliday junction
(Common)
BLM
TOP3
Mus81
Eme1
Strand displacement
Strand displacement
Crossover
or
Flap removal and annealing
(Common) Ligation
BLM
Non-crossover
DSBR
Double-strand break repair
SDSA
Synthesis-dependent strand annealing
BIR
Break-induced replication
Data on KEGG graph
Rendered by Pathview




## GO terms

Same analysis but using GO terms rather than KEGG.

```
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)
```

```
head(gobpres$less, 4)
```

```
                                      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
                                               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
```

## Reactome

Lots of folks like the reactome web interface. You can also run this as an R function but lets look at the website first. < https://reactome.org/>

The website wants a text file with one gene symbol per line of the genes you want to map to pathways.

```
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj),] $symbol
head(sig_genes)
```

```
ENSG00000117519 ENSG00000183508 ENSG00000159176 ENSG00000116016 ENSG00000164251
         "CNN3"        "TENT5C"         "CSRP1"         "EPAS1"         "F2RL1"
ENSG00000124766
         "SOX4"
```

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

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

And write out to a file.

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

> Q8. 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 most significant pathway is Cell Cycle (lowest entities p-value of 2.65x10^-5). KEGG did not have the exact same pathway but they both analyze cell cycle biology. The factors that could cause the differences between the two methods are that they use different pathway definitions, gene annotations and statistical methods. Reactome uses hierarchical pathway structures with uses only significant genes while KEGG uses broader metabolic and signaling categories with fold-change values.

### Save our results

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