# Class 14: RNA-Seq analysis mini-project

Ruth Barnes: A16747659

## Table of contents

## Background

The data for hands-on session comes from GEO entry: GSE37704, which is associated with the following publication:

> Trapnell C, Hendrickson DG, Sauvageau M, Goff L et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". Nat Biotechnol 2013 Jan;31(1):46-53. PMID: 23222703

The authors report on differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOXA1. Their results and others indicate that HOXA1 is required for lung fibroblast and HeLa cell cycle progression. In particular their analysis show that "loss of HOXA1 results in significant expression level changes in thousands of individual transcripts, along with isoform switching events in key regulators of the cell cycle". For our session we have used their Sailfish gene-level estimated counts and hence are restricted to protein-coding genes only.

## Data Import

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

## Inspect and Tidy Data

Does the `counts` columns match the `colData` rows?

```
head(counts)
```

```
                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
```

```
colData$id
```

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

```
colnames(counts)
```

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

**Q. Complete the code below to remove the troublesome first column from countData**

The fix here looks to be removing the first "length" column from counts:

```
countData <- counts[,-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
```

Check for matching countData and colData:

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

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

**Q. How many genes in total?**

```
nrow(countData)
```

```
[1] 19808
```

**Q. Filter to remove zero count genes (rows where there are zero counts in all columns). How many genes are left** and **Q. 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).**

```r
to.keep.inds <- rowSums(countData) > 0
```

```r
new.counts <- countData[to.keep.inds,]
```

```r
nrow(new.counts)
```

```
[1] 15975
```

## Setup for DESeq

```r
library(DESeq2)
```

Setup input object for DESeq:

```r
dds <- DESeqDataSetFromMatrix(countData = new.counts,
                              colData = colData,
                              design = ~condition)
```

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

## Run DESeq

```r
dds <- DESeq(dds)
```

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates

fitting model and testing
```

```r
res <- results(dds)
```

```r
head(res)
```

```
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 1.43989e-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 0.5215599   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
```

**Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.**

```r
summary(res)
```

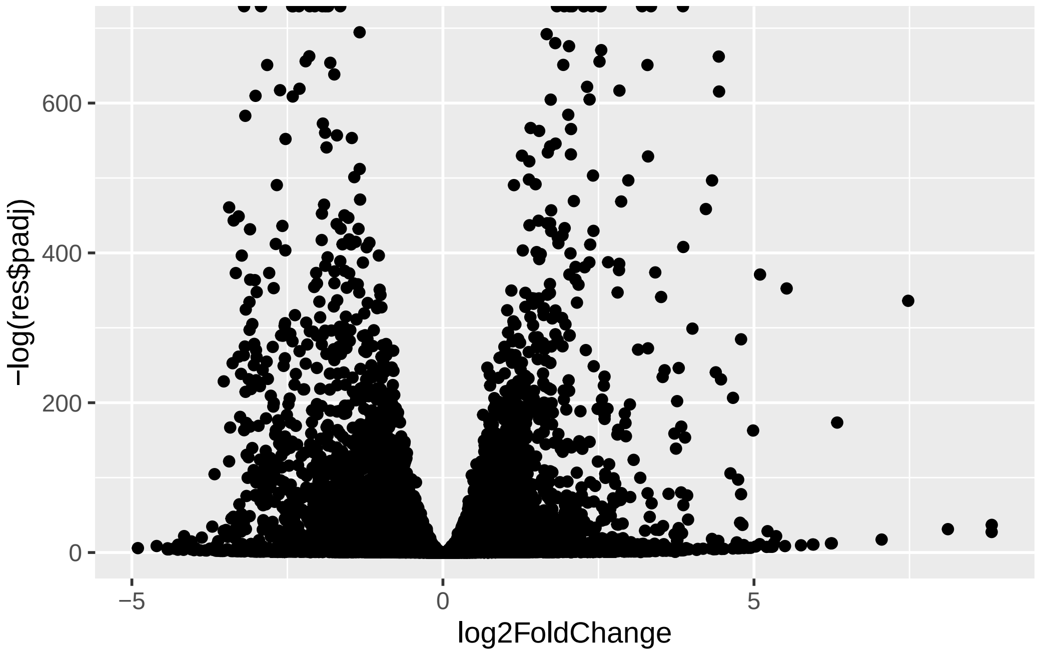
```

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
```

## Volcano Plot of Results

```
library(ggplot2)
```

```
ggplot(res) +
  aes(log2FoldChange, -log(res$padj)) +
  geom_point()
```

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

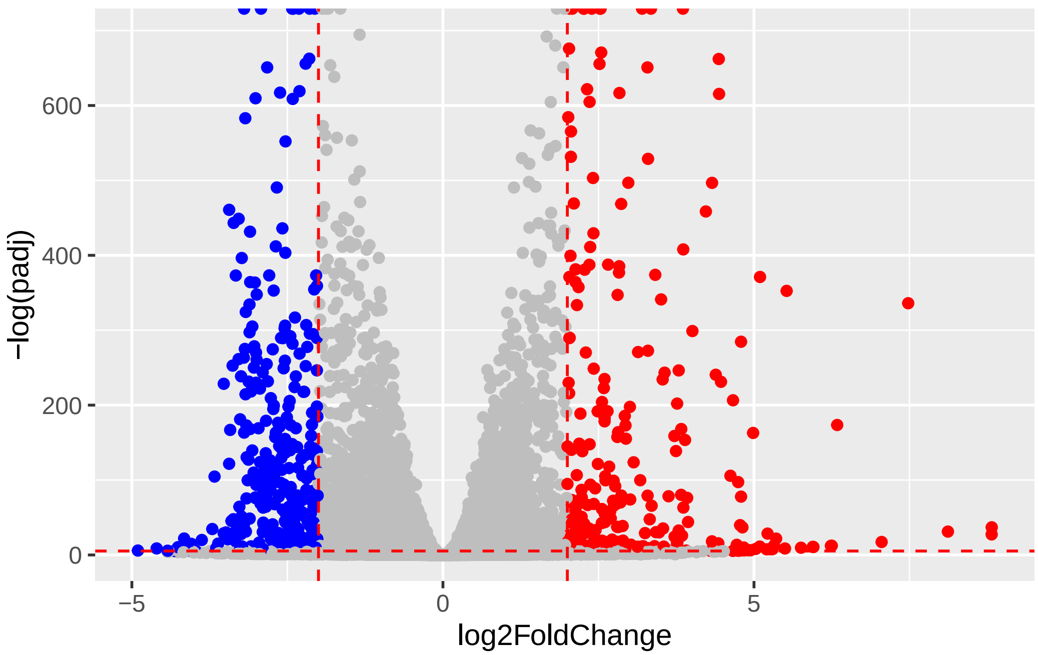


**Q. Improve this plot by completing the below code, which adds color and axis labels.**

```
mycols <- rep("grey", nrow(res))
mycols[res$log2FoldChange >= 2] <- "red"
mycols[res$log2FoldChange <= -2] <- "blue"
mycols[res$padj > 0.005] <- "gray"
```

```r
ggplot(res) +
  aes(log2FoldChange, -log(padj)) +
  geom_point(col=mycols) +
  geom_vline(xintercept = c(-2,2), col="red", linetype = "dashed") +
  geom_hline(yintercept = -log(0.005), col = "red", linetype = "dashed")
```

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



## Gene Annotation

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

```r
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"     
```

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

Add gene SYMBOL and ENTREZID

```
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"     
```

```
res$symbol = mapIds(org.Hs.eg.db,
                    keys=rownames(res),
                    keytype="ENSEMBL",
                    column="SYMBOL")
```

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

```
res$entrez = mapIds(org.Hs.eg.db,
                    keys=rownames(res),
                    keytype="ENSEMBL",
                    column="ENTREZID")
```

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

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

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

```r
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 1.43989e-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 0.5215599   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      148398
ENSG00000188976 1.76549e-35       NOC2L       26155       26155
ENSG00000187961 1.13413e-07      KLHL17      339451      339451
ENSG00000187583 9.19031e-01     PLEKHN1       84069       84069
ENSG00000187642 4.03379e-01       PERM1       84808       84808
ENSG00000188290 1.30538e-24        HES4       57801       57801
ENSG00000187608 2.37452e-02       ISG15        9636        9636
ENSG00000188157 4.21963e-16        AGRN      375790      375790
ENSG00000237330          NA      RNF223      401934      401934
```

**Q. 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.**

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

```r
head(read.csv("deseq_results.csv"))
```

```
                X baseMean log2FoldChange      lfcSE      stat pvalue padj
```

```
1 ENSG00000117519 4483.627      -2.422719 0.06000162 -40.37756      0    0
2 ENSG00000183508 2053.881       3.201955 0.07241720  44.21540      0    0
3 ENSG00000159176 5692.463      -2.313738 0.05755337 -40.20160      0    0
4 ENSG00000150938 7442.986      -2.059631 0.05384491 -38.25118      0    0
5 ENSG00000116016 4423.947      -1.888019 0.04316799 -43.73656      0    0
6 ENSG00000136068 3796.127      -1.649792 0.04393544 -37.55037      0    0
  symbol entrez  name
1   CNN3   1266  1266
2 TENT5C  54855 54855
3  CSRP1   1465  1465
4  CRIM1  51232 51232
5  EPAS1   2034  2034
6   FLNB   2317  2317
```

## Pathway Analysis

```r
library(gage)
```

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

```
##############################################################################
Pathview is an open source software package distributed under GNU General
Public License version 3 (GPLv3). Details of GPLv3 is available at
http://www.gnu.org/licenses/gpl-3.0.html. Particullary, users are required to
formally cite the original Pathview paper (not just mention it) in publications
or products. For details, do citation("pathview") within R.

The pathview downloads and uses KEGG data. Non-academic uses may require a KEGG
license agreement (details at http://www.kegg.jp/kegg/legal.html).
##############################################################################
```

Input vector for gage():

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

Load up the KEGG gene-sets:

```r
data(kegg.sets.hs)
```

Run pathway analysis:

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

```r
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                         1.375901e-03 -3.028500
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                         1.375901e-03 0.072234819
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 1.375901e-03
hsa03440 Homologous recombination                    28 3.066756e-03
hsa04114 Oocyte meiosis                             102 3.784520e-03
```

Cell cycle figure:

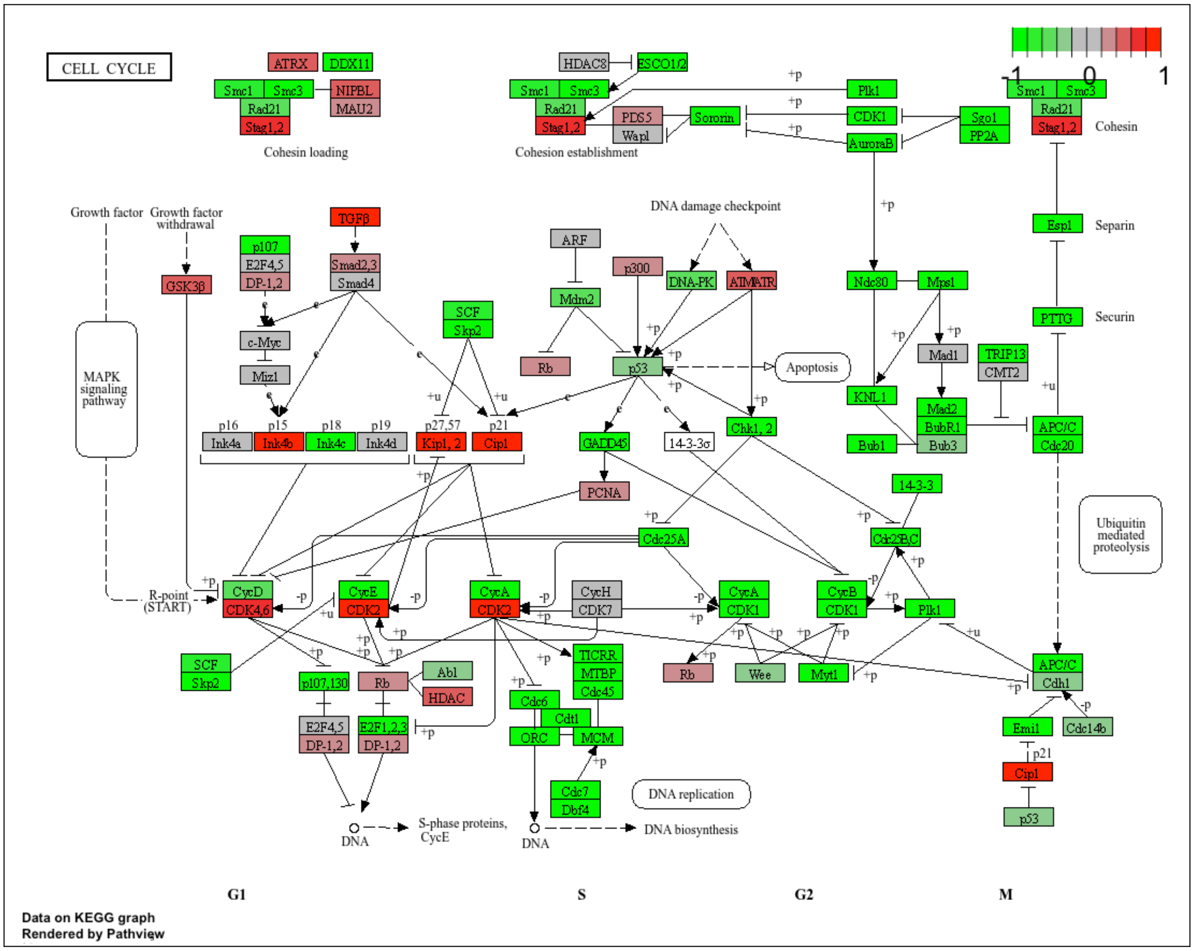
```r
pathview(foldchanges, pathway.id = "hsa04110")
```

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

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

Insert this figure:



Change the display in various ways including generating a PDF graph:

```r
# 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/ruthbarnes/Desktop/School/bimm143/Class 14
```

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

```r
## 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] "hsa04060" "hsa05323" "hsa05146" "hsa05332" "hsa04640"
```

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

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

```
Info: Writing image file hsa04060.pathview.png
```

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

```
Info: Writing image file hsa05323.pathview.png
```

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

```
Info: Writing image file hsa05146.pathview.png
```

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

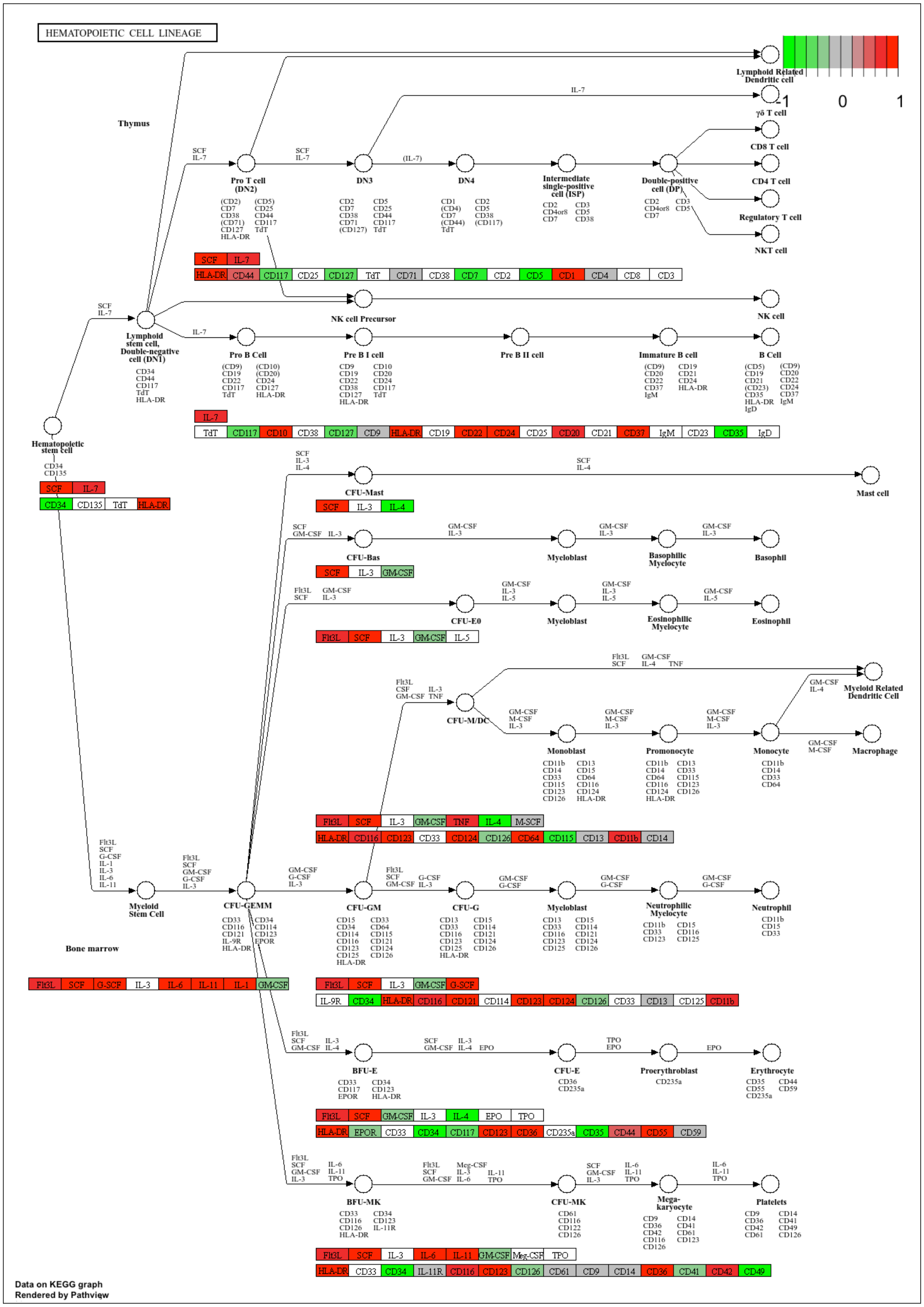
```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

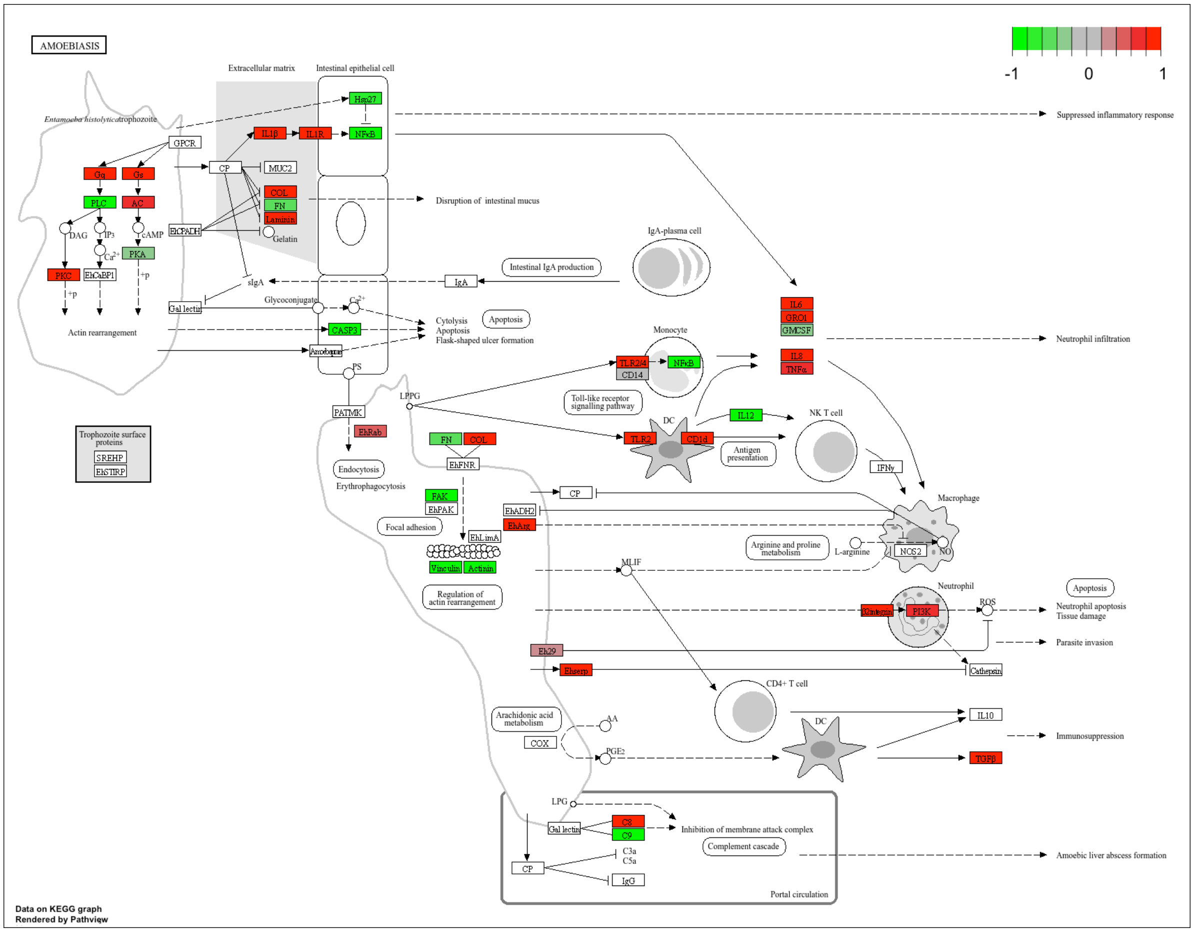
```
Info: Writing image file hsa05332.pathview.png
```

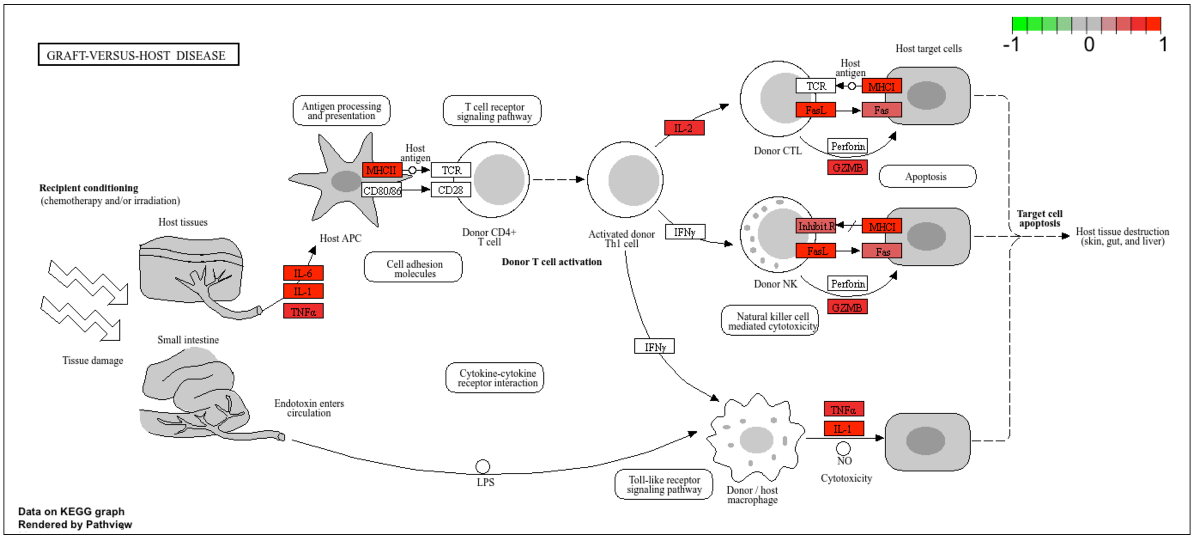
```
'select()' returned 1:1 mapping between keys and columns
```

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

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

Insert images: 

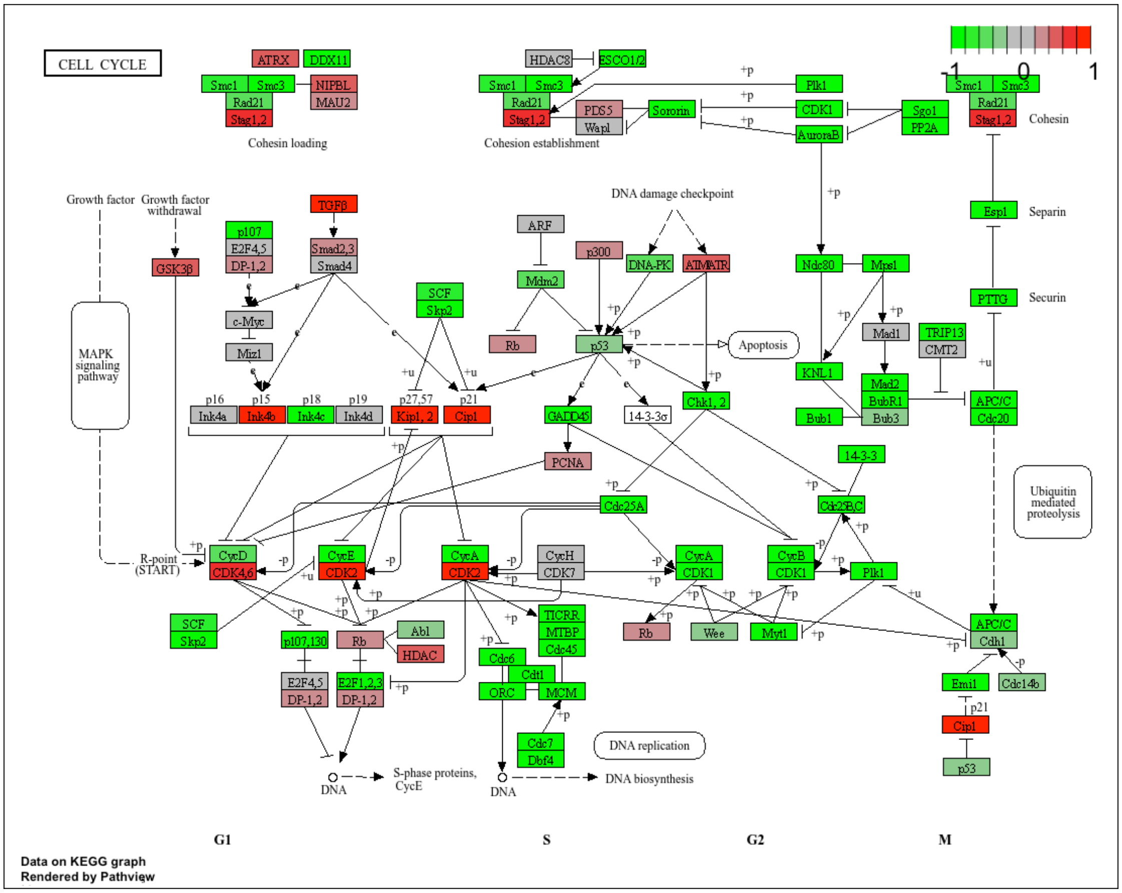




**Q. Can you do the same procedure as above to plot the pathview figures for the**

### top 5 down-regulated pathways?

Down-regulated pathways:

1. 

Data on KEGG graph
Rendered by Pathview

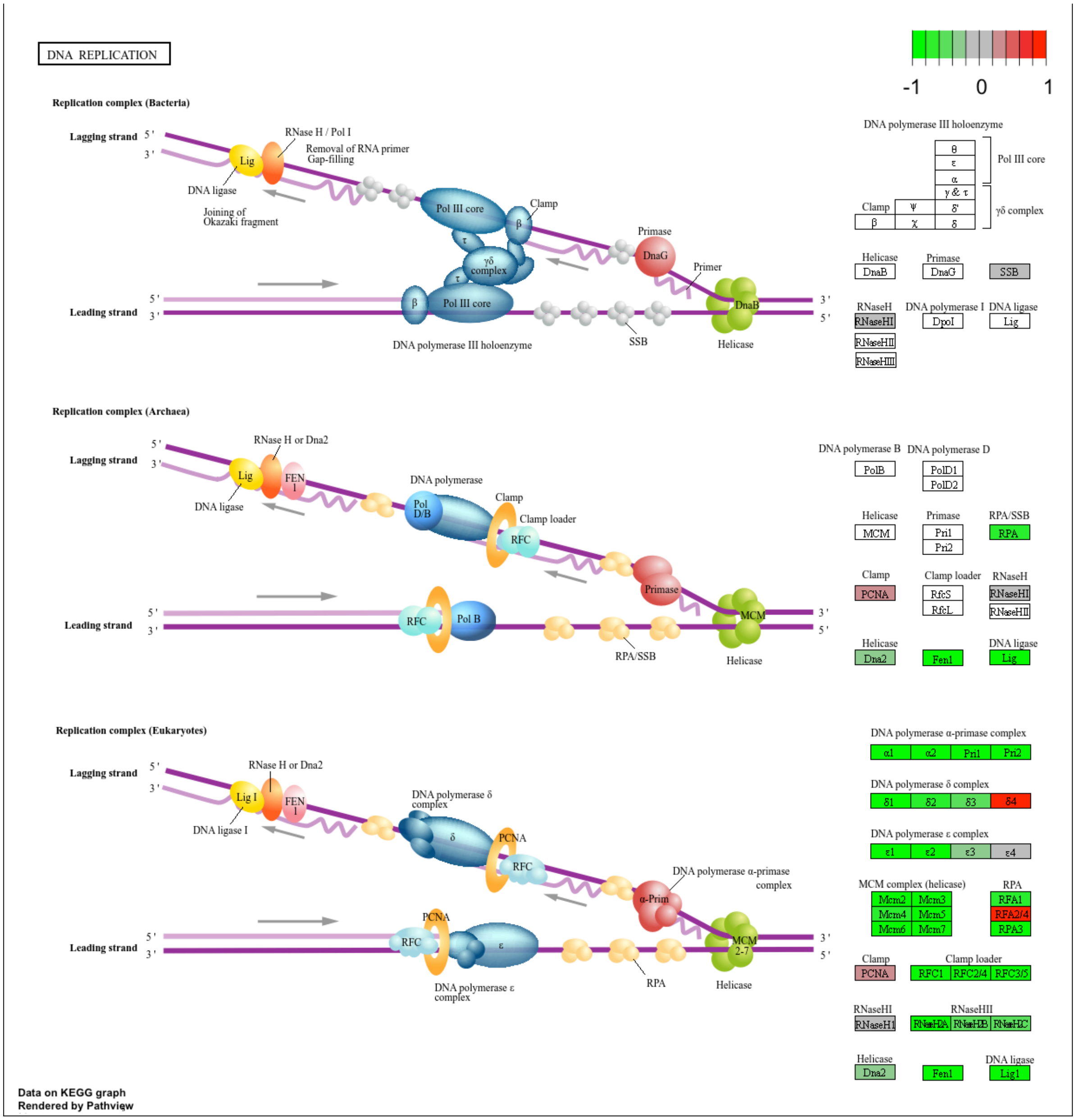
2.

```
pathview(foldchanges, pathway.id = "hsa03030")
```

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

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



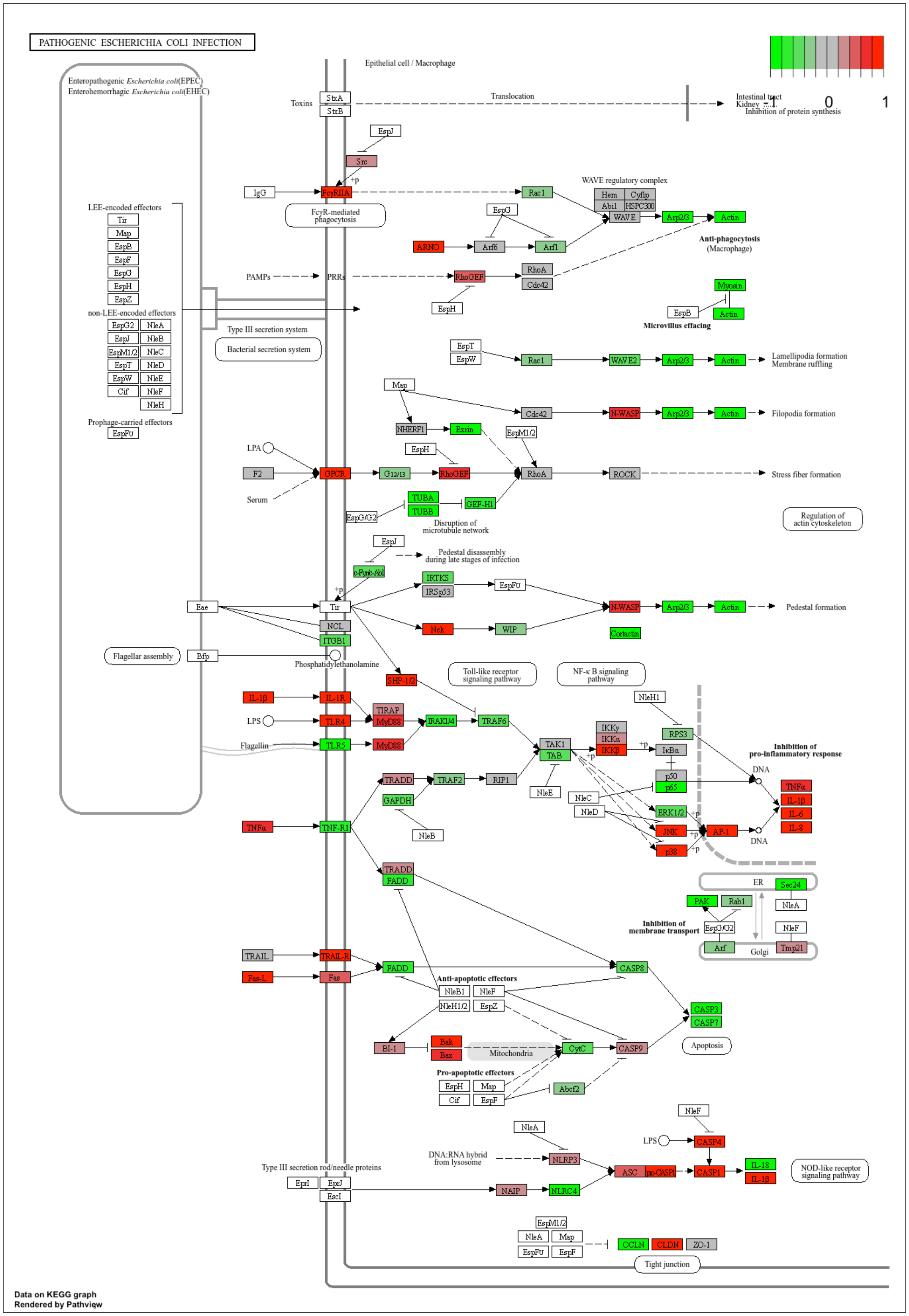
3.

```r
pathview(foldchanges, pathway.id = "hsa05130")
```

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

```
Info: Writing image file hsa05130.pathview.png
```


PATHOGENIC ESCHERICHIA COLI INFECTION
-1
0
1
Epithelial cell / Macrophage
Enteropathogenic Escherichia coli(EPEC)
Enterohemorrhagic Escherichia coli(EHEC)
Toxins
StxA
StxB
Translocation
Intestinal tract
Kidney
Inhibition of protein synthesis
EspJ
Src
+p
IgG
FcγRIIA
FcγR-mediated phagocytosis
Rac1
WAVE regulatory complex
Hem
Cyfip
Abi1
HSPC300
WAVE
Arp2/3
Actin
EspG
ARNO
Arf6
Arf1
Anti-phagocytosis
(Macrophage)
LEE-encoded effectors
Tir
Map
EspB
EspF
EspG
EspH
EspZ
non-LEE-encoded effectors
EspG2
NleA
EspJ
NleB
EspM1/2
NleC
EspT
NleD
EspW
NleE
Cif
NleF
NleH
Prophage-carried effectors
EspFu
PAMPs
PRRs
RhoGEF
RhoA
Cdc42
EspH
Myosin
EspB
Actin
Microvillus effacing
Type III secretion system
Bacterial secretion system
EspT
EspW
Rac1
WAVE2
Arp2/3
Actin
Lamellipodia formation
Membrane ruffling
Map
Cdc42
N-WASP
Arp2/3
Actin
Filopodia formation
NHERF1
Ezrin
EspM1/2
EspH
LPA
F2
GPCR
G12/13
RhoGEF
RhoA
ROCK
Stress fiber formation
Serum
TUBA
TUBB
GEF-H1
EspG/G2
Disruption of microtubule network
Regulation of actin cytoskeleton
EspJ
Pedestal disassembly during late stages of infection
c-Fyn/c-Abl
+p
IRTKS
IRSp53
EspFu
Eae
Tir
N-WASP
Arp2/3
Actin
Pedestal formation
NCL
Nck
WIP
ITGB1
Cortactin
Flagellar assembly
Bfp
Phosphatidylethanolamine
SHP-1/2
Toll-like receptor signaling pathway
NF-κ B signaling pathway
IL-1β
IL-1R
TIRAP
NleH1
LPS
TLR4
MyD88
IRAK1/4
TRAF6
IKKγ
RPS3
IKKα
TAK1
+p
Flagellin
TLR5
MyD88
TAB
IKKβ
IκBα
Inhibition of pro-inflammatory response
+p
DNA
TRADD
TRAF2
RIP1
NleE
p50
TNFα
p65
IL-1β
GAPDH
NleC
IL-6
NleD
ERK1/2
+p
IL-8
NleB
JNK
+p
AP-1
TNFα
TNF-R1
DNA
p38
+p
TRADD
FADD
ER
Sec24
PAK
Rab1
NleA
Inhibition of membrane transport
EspG/G2
NleF
Arf
Golgi
Tmp21
TRAIL
TRAIL-R
Fas-L
Fas
FADD
CASP8
Anti-apoptotic effectors
NleB1
NleF
NleH1/2
EspZ
CASP3
CASP7
Apoptosis
Bl-1
Bak
Bax
Mitochondria
CytC
CASP9
Pro-apoptotic effectors
EspH
Map
Cif
EspF
Abcf2
NleF
NleA
LPS
CASP4
DNA:RNA hybrid from lysosome
NLRP3
ASC
pro-CASP1
CASP1
IL-18
IL-1β
NOD-like receptor signaling pathway
Type III secretion rod/needle proteins
EprI
EprJ
EscI
NAIP
NLRC4
EspM1/2
NleA
Map
EspFu
EspF
OCLN
CLDN
ZO-1
Tight junction
Data on KEGG graph
Rendered by Pathview

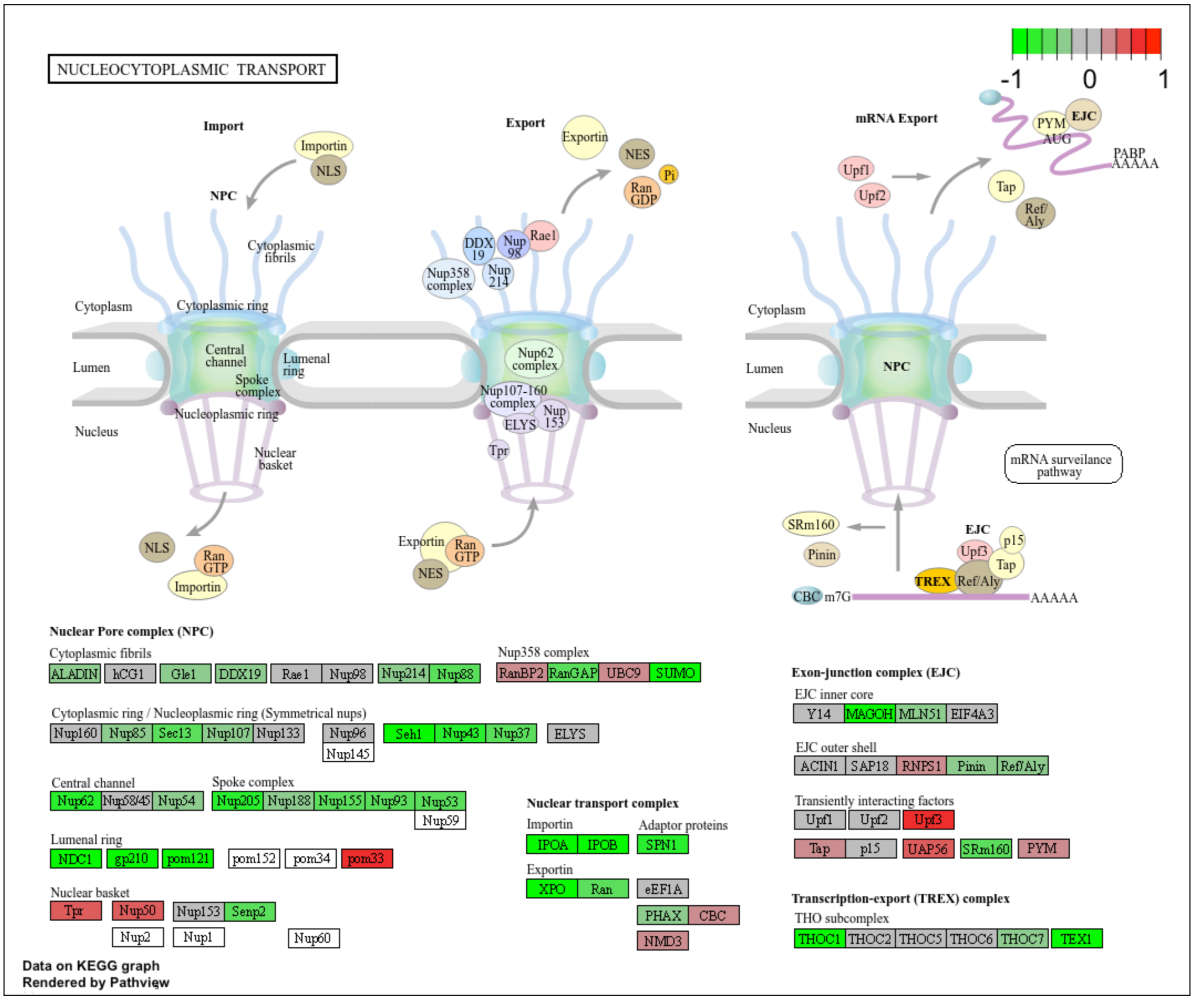
4.

```r
pathview(foldchanges, pathway.id = "hsa03013")
```

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

```
Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14
```

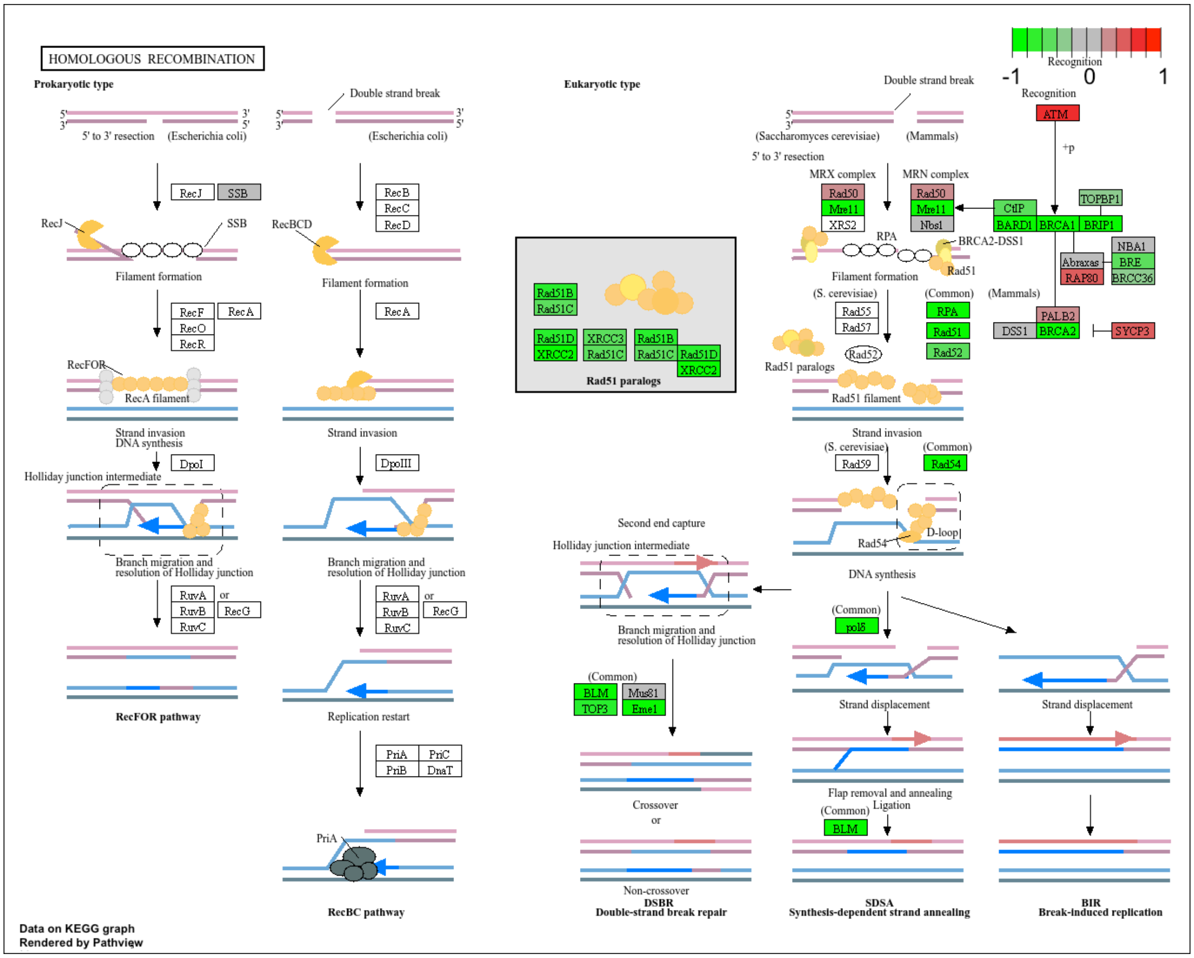
```
Info: Writing image file hsa03013.pathview.png
```



5.

```r
pathview(foldchanges, pathway.id = "hsa03440")
```

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

Info: Working in directory /Users/ruthbarnes/Desktop/School/bimm143/Class 14

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



## Gene Ontology Analysis

Run pathway analysis with GO

```r
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)
```

```
                                            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
```