BIMM 143 Class 14

Xaler Lu (A17388454)

Table of contents

Background

Previously, we have used DESeq to interpret RNA-seq data. We also annotated our data and used pathway analysis to map genes to known biological pathways. Here, we will work on a mini-project that will use the same methods.

(1) Differential Expression Analysis

```r
library(DESeq2)
```

```
Warning: package 'DESeq2' was built under R version 4.3.3
```

```
Warning: package 'S4Vectors' was built under R version 4.3.2
```

```
Warning: package 'GenomeInfoDb' was built under R version 4.3.3
```

```
Warning: package 'SummarizedExperiment' was built under R version 4.3.2
```

```
Warning: package 'matrixStats' was built under R version 4.3.3
```

Download both the count data and meta data (also called column data).

```
metaFile <- "GSE37704_metadata.csv"
countFile <- "GSE37704_featurecounts.csv"

metaData = read.csv(metaFile, row.names = 1)
head(metaData)
```

```
              condition
SRR493366 control_sirna
SRR493367 control_sirna
SRR493368 control_sirna
SRR493369      hoxa1_kd
SRR493370      hoxa1_kd
SRR493371      hoxa1_kd
```

```
countsA = read.csv(countFile, row.names = 1)
head(countsA)
```

```
                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
```

> Q. Complete the code below to remove the troublesome first column from counts

Now, we need to match the count data and meta data with a 1:1 correspondence, but the first column of the count data is just the length and needs to be removed.

```
counts <- as.matrix(countsA[,-1])
```

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

```
counts <- counts[rowSums(counts) != 0,]
head(counts)
```

```
                SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
ENSG00000279457        23        28        29        29        28        46
ENSG00000187634       124       123       205       207       212       258
ENSG00000188976      1637      1831      2383      1226      1326      1504
ENSG00000187961       120       153       180       236       255       357
ENSG00000187583        24        48        65        44        48        64
ENSG00000187642         4         9        16        14        16        16
```

DESeq

We will run DESeq2 with `DESeqDataSetFromMatrix()` with three required arguments: `counts`, `metaData`, and `design`. `design` is the name of the column in `metaData`

```
dds <- DESeqDataSetFromMatrix(countData = counts,
                              colData = metaData,
                              design = ~condition)
```

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

With `dds`, we will run it with `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(2): condition sizeFactor
```

> 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.

Here are the results. 4349 upregulated genes below 0.1 p-value, and 4396 downregulated genes below 0.1 p-value.

```
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
```

Volcano Plot

```
library(ggplot2)
```

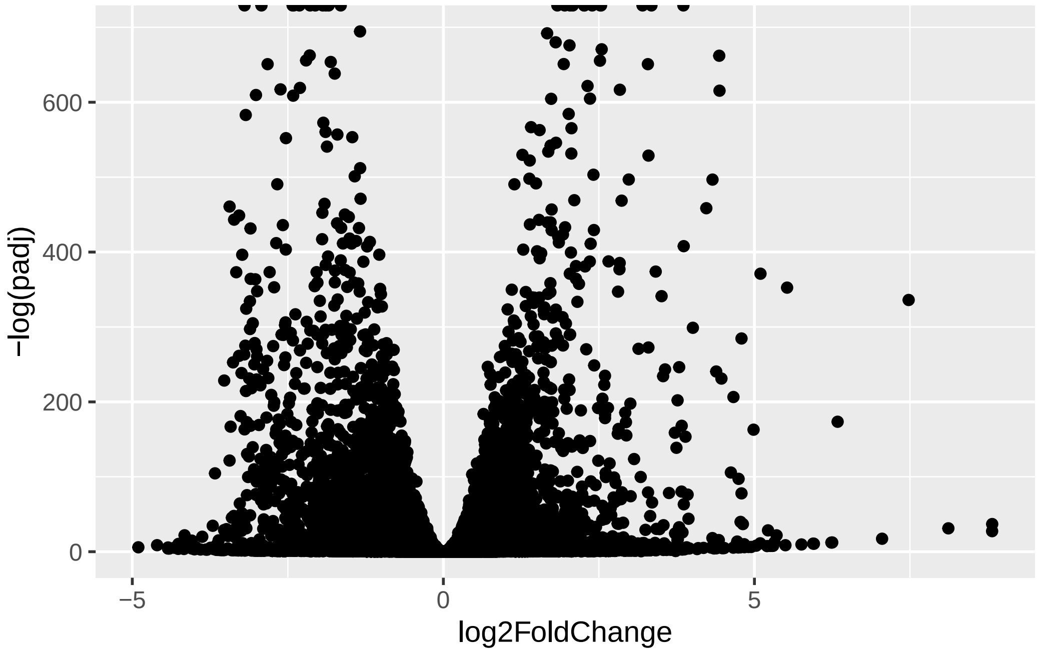
```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
head(res$log2FoldChange)
```

```
[1]  0.17925708  0.42645712 -0.69272046  0.72975561  0.04057653  0.54281049
```

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

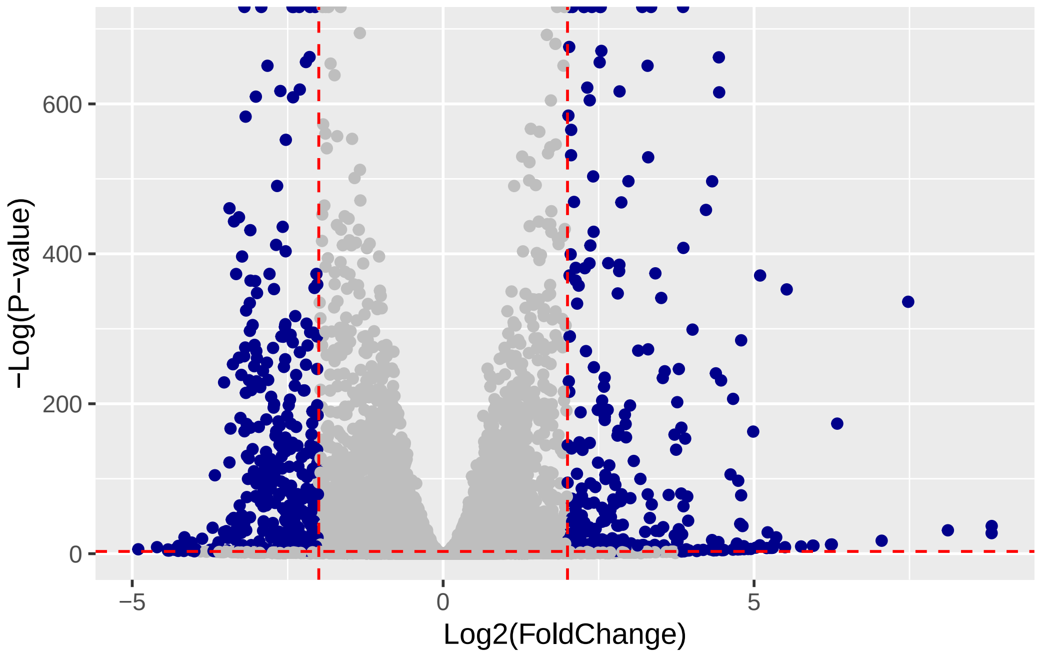
```
Warning: Removed 1237 rows containing missing values or values outside the scale range
(`geom_point()`).
```



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

```
# Make a color vector for all genes
mycols <- rep("gray", nrow(res) )

# Color blue the genes with fold change above 2
mycols[ abs(res$log2FoldChange) > 2 ] <- "blue4"

# Color gray those with adjusted p-value more than 0.01
mycols[ res$padj > 0.05 ] <- "gray"

ggplot(res) +
  aes(log2FoldChange, -log(padj)) +
  geom_point(col = mycols) +
  xlab("Log2(FoldChange)") +
  ylab("-Log(P-value)") +
  geom_vline(xintercept = c(-2,2), col = "red", lty = 2) +
  geom_hline(yintercept = -log(0.05), col = "red", lty = 2)
```

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



Gene Annotation

We want to use pathway analysis using the KEGG pathway. Let's first annotate with ENTREZID.

> Q. 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")
```

```
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"     
```

Essentially, we want to use `mapIds()` to create new columns with symbol using `SYMBOL`, entrez using `ENTREZID`, and gene name using `GENENAME`. The keytype is `ENSEMBLE`

```
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   29.913579      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 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 ..
```

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")
```

(2) Pathway Analysis

We will use gage and the **KEGG** database, specifically kegg.sets.hs. We can also use others like go.sets.hs or sigmet.idx.hs.

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

```r
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"  
```

With the data, we will use `gage()` which would require a vector of ENTREZID values because we are using **KEGG*

```r
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 
```

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

```r
attributes(keggres)
```

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

Here are the top six downregulated pathways

```r
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
```

Here is the pathway of the Cell Cycle pathway

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

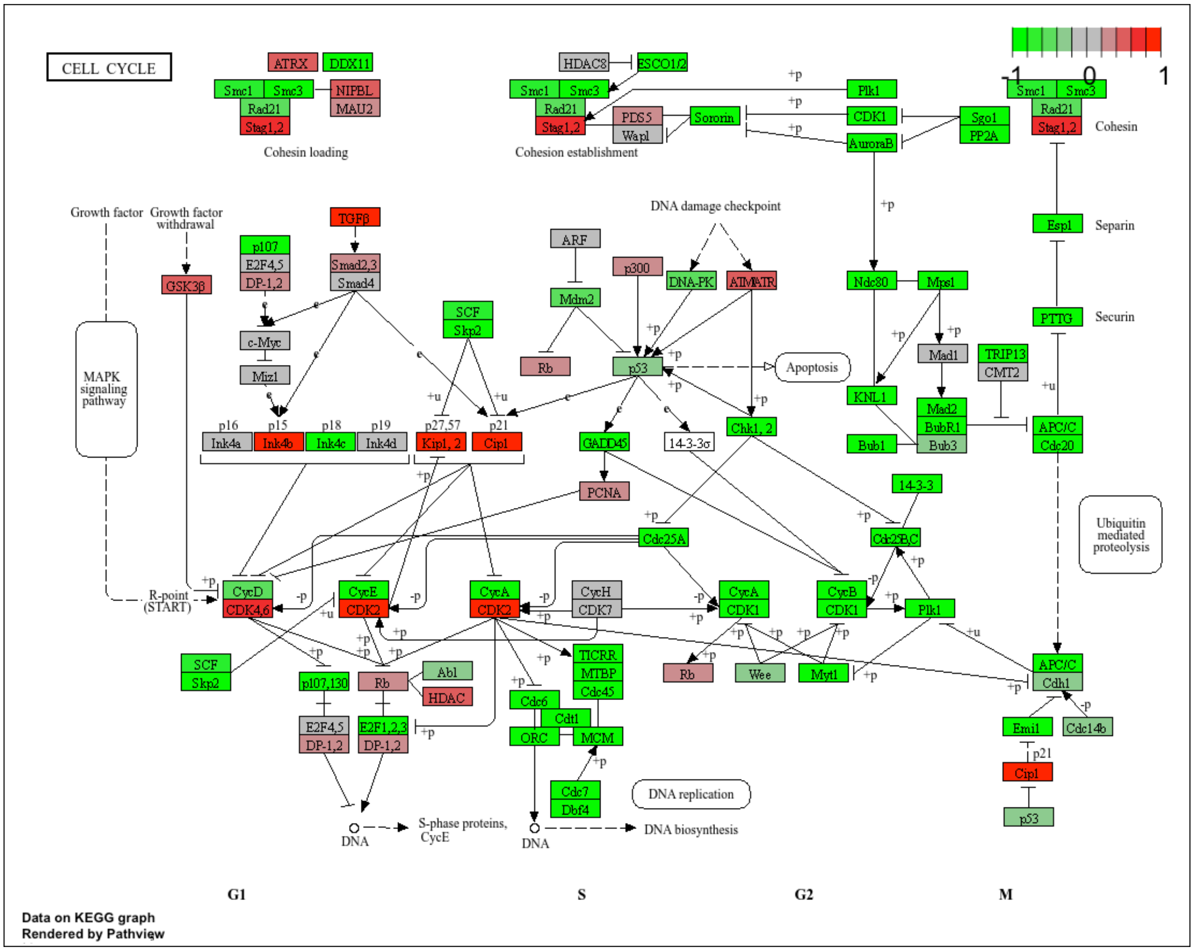
```
'select()' returned 1:1 mapping between keys and columns
```

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

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

```
Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14
```

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



Below is a demo of creating multiple pathviews of the top five upregulated pathways at once.

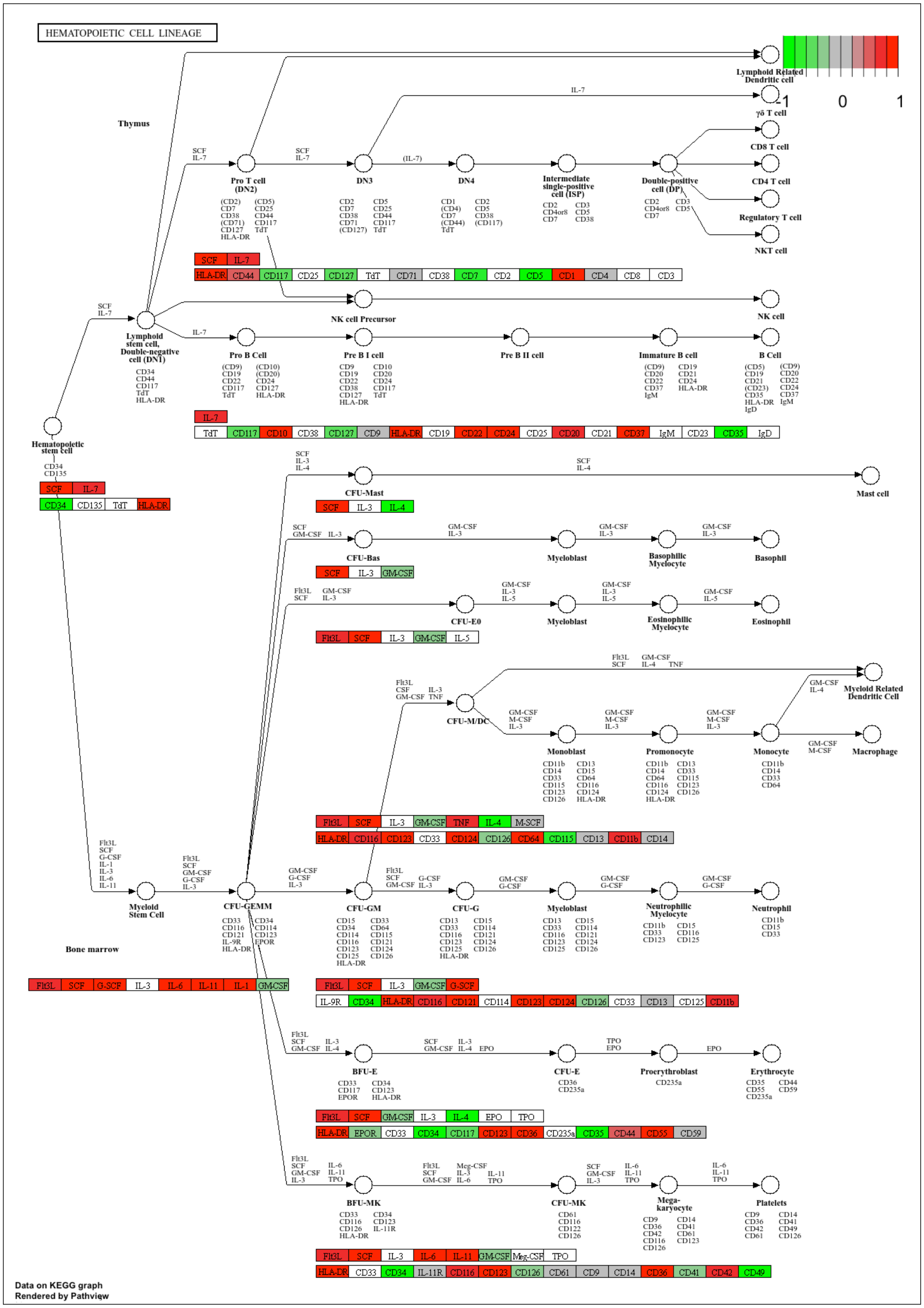
```
keggrespathways <- rownames(keggres$greater)[1:5]

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

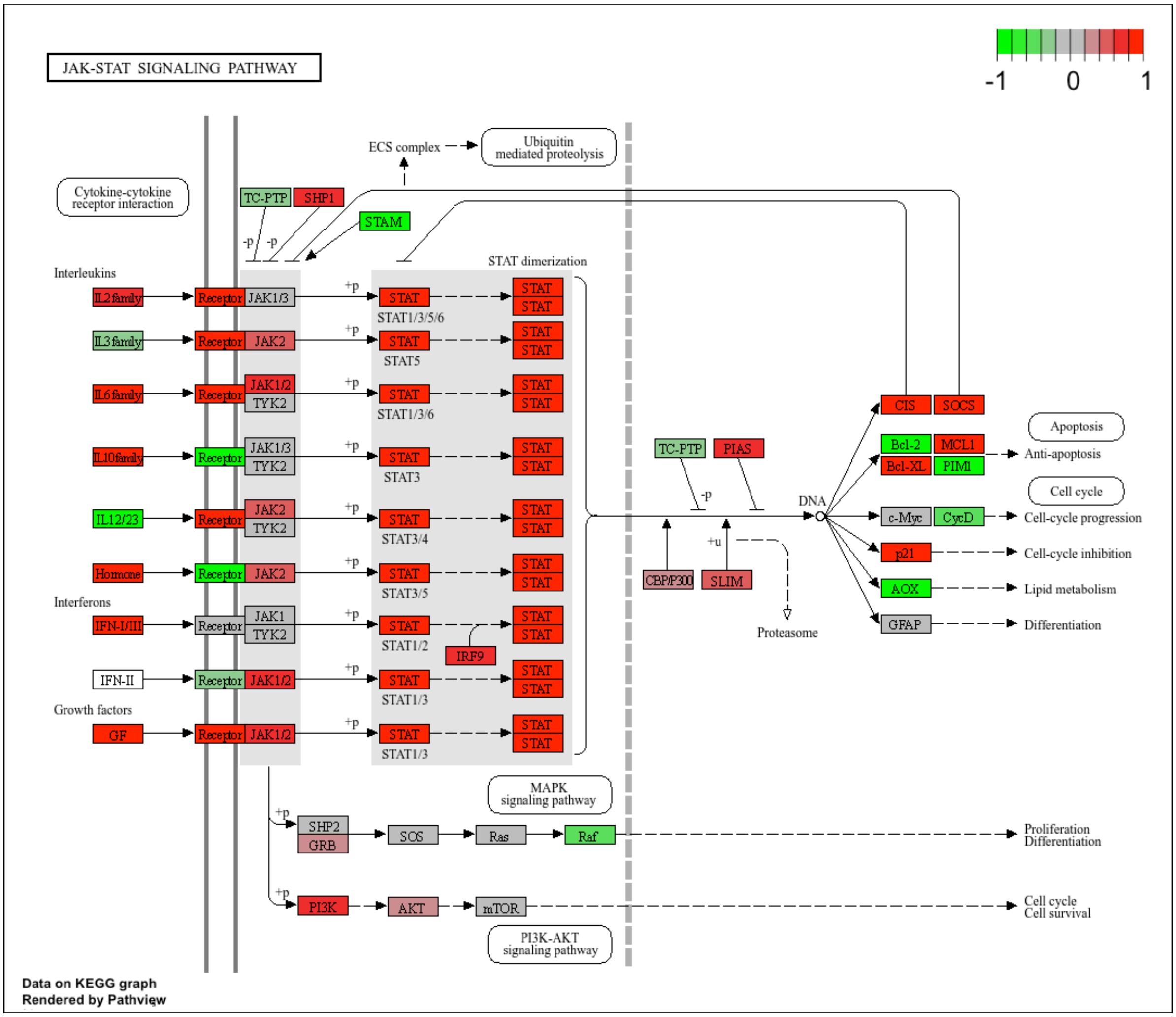
```
[1] "hsa04640" "hsa04630" "hsa00140" "hsa04142" "hsa04330"
```

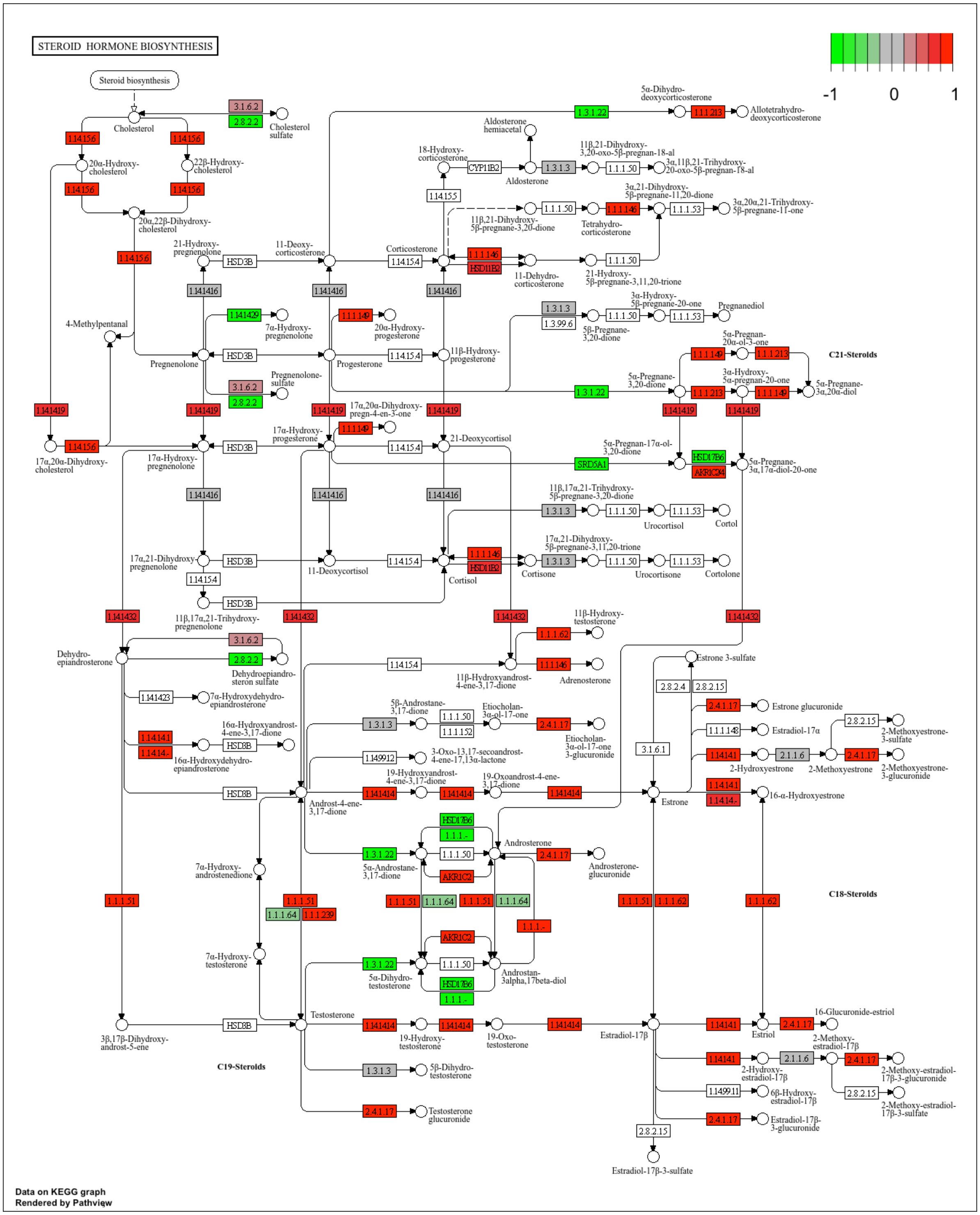
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

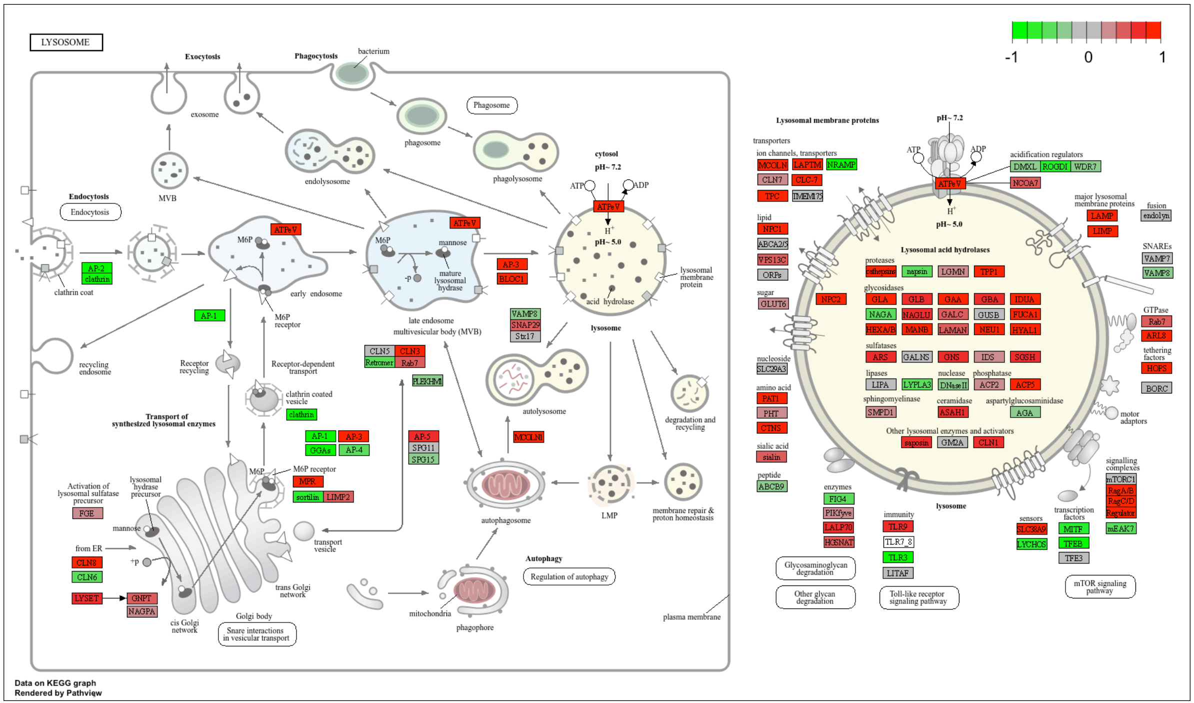
```
'select()' returned 1:1 mapping between keys and columns
```

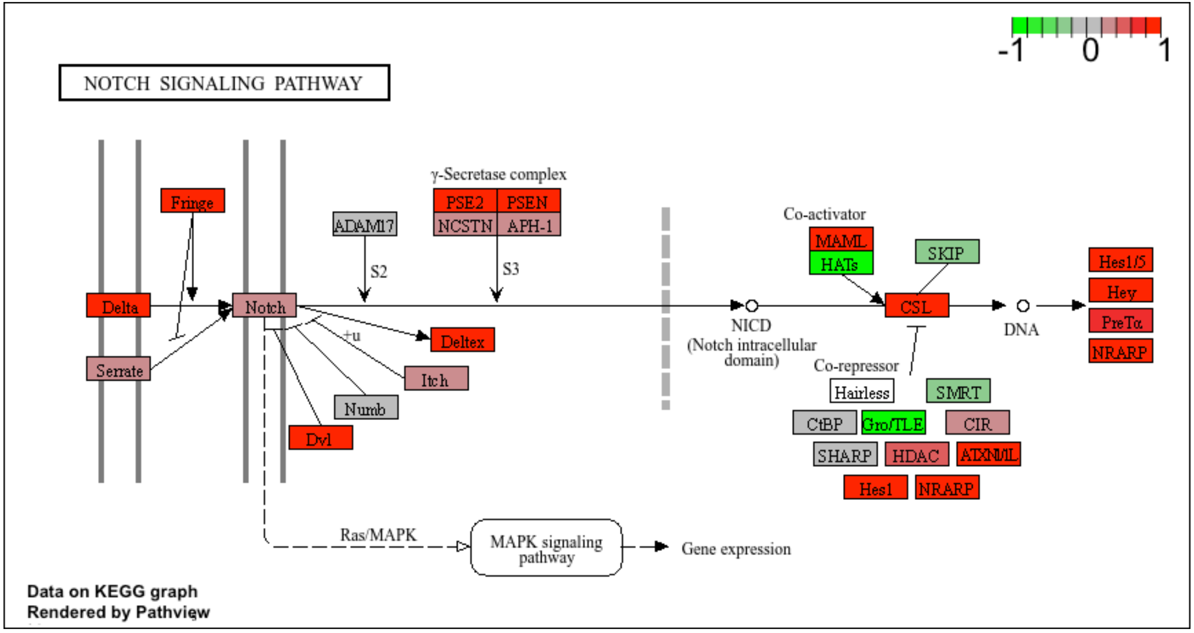
```
Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa04142.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

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


HEMATOPOIETIC CELL LINEAGE
-1
0
1
Thymus
Lymphoid Related Dendritic cell
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF
IL-7
Pro T cell (DN2)
DN3
(IL-7)
DN4
Intermediate single-positive cell (ISP)
Double-positive cell (DP)
(CD2) CD7 CD38 (CD71) CD127 HLA-DR
(CD5) CD25 CD44 CD117 TdT
CD2 CD7 CD38 CD71 (CD127)
CD5 CD25 CD44 CD117 TdT
CD1 (CD4) CD7 (CD44) TdT
CD2 CD5 CD38 (CD117)
CD2 CD4or8 CD7
CD3 CD5 CD38
CD2 CD4or8 CD7
CD3 CD5
SCF
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)
CD34 CD44 CD117 TdT HLA-DR
IL-7
Pro B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
(CD9) CD19 CD22 CD117 TdT
(CD10) (CD20) CD24 CD127 HLA-DR
CD9 CD19 CD22 CD38 CD127 HLA-DR
CD10 CD20 CD24 CD117 TdT
(CD9) CD20 CD22 CD37 IgM
CD19 CD21 CD24 HLA-DR
(CD5) CD19 CD21 (CD23) CD35 HLA-DR IgD
(CD9) CD20 CD22 CD24 CD37 IgM
IL-7
TdT
CD117
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
SCF IL-3 IL-4
CFU-Mast
SCF IL-4
Mast cell
SCF
IL-3
IL-4
SCF GM-CSF IL-3
CFU-Bas
GM-CSF IL-3
Myeloblast
GM-CSF IL-3
Basophilic Myelocyte
GM-CSF IL-3
Basophil
SCF
IL-3
GM-CSF
Flt3L SCF
GM-CSF IL-3
CFU-E0
GM-CSF IL-3 IL-5
Myeloblast
GM-CSF IL-3 IL-5
Eosinophilic Myelocyte
GM-CSF IL-5
Eosinophil
Flt3L
SCF
IL-3
GM-CSF
IL-5
Flt3L SCF
GM-CSF IL-4 TNF
Flt3L CSF GM-CSF
IL-3 TNF
CFU-M/DC
GM-CSF IL-4
Myeloid Related Dendritic Cell
GM-CSF M-CSF IL-3
Monoblast
GM-CSF M-CSF IL-3
Promonocyte
GM-CSF M-CSF IL-3
Monocyte
GM-CSF M-CSF
Macrophage
CD11b CD14 CD33 CD115 CD123 CD126
CD13 CD15 CD64 CD116 CD124 HLA-DR
CD11b CD14 CD64 CD116 CD124 HLA-DR
CD13 CD33 CD115 CD123 CD126
CD11b CD14 CD33 CD64
Flt3L
SCF
IL-3
GM-CSF
TNF
IL-4
M-SCF
HLA-DR
CD116
CD123
CD33
CD124
CD126
CD64
CD115
CD13
CD11b
CD14
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Flt3L SCF GM-CSF G-CSF IL-3
GM-CSF G-CSF IL-3
Flt3L SCF GM-CSF
G-CSF IL-3
GM-CSF G-CSF
GM-CSF G-CSF
GM-CSF G-CSF
Myeloid Stem Cell
CFU-GEMM
CFU-GM
CFU-G
Myeloblast
Neutrophilic Myelocyte
Neutrophil
CD33 CD116 CD121 IL-9R HLA-DR
CD34 CD114 CD123 EPOR
CD15 CD34 CD114 CD116 CD123 CD125 HLA-DR
CD33 CD64 CD115 CD121 CD124 CD126
CD13 CD33 CD116 CD123 CD125 HLA-DR
CD15 CD114 CD121 CD124 CD126
CD13 CD33 CD116 CD123 CD125
CD15 CD114 CD121 CD124 CD126
CD11b CD33 CD123
CD15 CD116 CD125
CD11b CD15 CD33
Bone marrow
Flt3L
SCF
G-SCF
IL-3
IL-6
IL-11
IL-1
GM-CSF
Flt3L
SCF
IL-3
GM-CSF
G-SCF
IL-9R
CD34
HLA-DR
CD116
CD121
CD114
CD123
CD124
CD126
CD33
CD13
CD125
CD11b
Flt3L SCF GM-CSF
IL-3 IL-4
BFU-E
SCF GM-CSF
IL-3 IL-4
EPO
CFU-E
TPO EPO
Proerythroblast
EPO
Erythrocyte
CD33 CD117 EPOR
CD34 CD123 HLA-DR
CD36 CD235a
CD235a
CD35 CD55 CD235a
CD44 CD59
Flt3L
SCF
GM-CSF
IL-3
IL-4
EPO
TPO
HLA-DR
EPOR
CD33
CD34
CD117
CD123
CD36
CD235a
CD35
CD44
CD55
CD59
Flt3L SCF GM-CSF IL-3
IL-6 IL-11 TPO
Flt3L SCF GM-CSF
Meg-CSF IL-3 IL-6
IL-11 TPO
SCF GM-CSF IL-3
IL-6 IL-11 TPO
IL-6 IL-11 TPO
BFU-MK
CFU-MK
Mega-karyocyte
Platelets
CD33 CD116 CD126 HLA-DR
CD34 CD123 IL-11R
CD61 CD116 CD122 CD126
CD9 CD36 CD42 CD116 CD126
CD14 CD41 CD61 CD123
CD9 CD36 CD42 CD61
CD14 CD41 CD49 CD126
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


JAK-STAT SIGNALING PATHWAY
-1
0
1
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
TC-PTP
SHP1
STAM
-p
-p
STAT dimerization
Interleukins
IL2 family
Receptor
JAK1/3
+p
STAT
STAT
STAT
STAT1/3/5/6
IL3 family
JAK2
STAT5
IL6 family
JAK1/2
TYK2
STAT1/3/6
IL10family
STAT3
IL12/23
STAT3/4
Hormone
STAT3/5
Interferons
IFN-I/III
JAK1
STAT1/2
IRF9
IFN-II
STAT1/3
Growth factors
GF
CIS
SOCS
Apoptosis
Bcl-2
MCL1
Bcl-XL
PIM1
Anti-apoptosis
Cell cycle
TC-PTP
PIAS
-p
DNA
c-Myc
CycD
Cell-cycle progression
+u
p21
Cell-cycle inhibition
CBP/P300
SLIM
AOX
Lipid metabolism
Proteasome
GFAP
Differentiation
MAPK signaling pathway
+p
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


STEROID HORMONE BIOSYNTHESIS
-1
0
1
Steroid biosynthesis
Cholesterol
3.1.6.2
2.8.2.2
Cholesterol sulfate
1.14.15.6
20α-Hydroxy-cholesterol
22β-Hydroxy-cholesterol
20α,22β-Dihydroxy-cholesterol
4-Methylpentanal
21-Hydroxy-pregnenolone
HSD3B
11-Deoxy-corticosterone
Corticosterone
1.14.15.4
1.14.15.5
18-Hydroxy-corticosterone
CYP11B2
Aldosterone hemiacetal
Aldosterone
1.3.1.3
11β,21-Dihydroxy-3,20-oxo-5β-pregnan-18-al
1.1.1.50
3α,11β,21-Trihydroxy-20-oxo-5β-pregnan-18-al
1.3.1.22
5α-Dihydro-deoxycorticosterone
1.1.1.213
Allotetrahydro-deoxycorticosterone
11β,21-Dihydroxy-5β-pregnane-3,20-dione
Tetrahydro-corticosterone
1.1.1.146
3α,21-Dihydroxy-5β-pregnane-11,20-dione
1.1.1.53
3α,20α,21-Trihydroxy-5β-pregnane-11-one
HSD11B2
11-Dehydro-corticosterone
21-Hydroxy-5β-pregnane-3,11,20-trione
1.14.14.16
1.14.14.29
7α-Hydroxy-pregnenolone
1.1.1.149
20α-Hydroxy-progesterone
Pregnenolone
Progesterone
11β-Hydroxy-progesterone
1.3.99.6
5β-Pregnane-3,20-dione
3α-Hydroxy-5β-pregnane-20-one
Pregnanediol
Pregnenolone-sulfate
5α-Pregnan-20α-ol-3-one
C21-Steroids
5α-Pregnane-3,20-dione
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
1.14.14.19
17α,20α-Dihydroxy-pregn-4-en-3-one
17α,20α-Dihydroxy-cholesterol
17α-Hydroxy-pregnenolone
17α-Hydroxy-progesterone
21-Deoxycortisol
SRD5A1
5α-Pregnan-17α-ol-3,20-dione
HSD17B6
AKR1C24
5α-Pregnane-3α,17α-diol-20-one
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
Urocortisol
Cortol
17α,21-Dihydroxy-pregnenolone
11-Deoxycortisol
Cortisol
Cortisone
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Urocortisone
Cortolone
11β,17α,21-Trihydroxy-pregnenolone
1.14.14.32
11β-Hydroxy-testosterone
1.1.1.62
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
Dehydro-epiandrosterone
Dehydroepiandro-steron sulfate
1.14.14.23
7α-Hydroxydehydro-epiandrosterone
1.14.14.1
1.14.14.-
16α-Hydroxyandrost-4-ene-3,17-dione
16α-Hydroxydehydro-epiandrosterone
5β-Androstane-3,17-dione
1.1.1.152
Etiocholan-3α-ol-17-one
2.4.1.17
Etiocholan-3α-ol-17-one 3-glucuronide
1.14.99.12
3-Oxo-13,17-secoandrost-4-ene-17,13α-lactone
19-Hydroxyandrost-4-ene-3,17-dione
1.14.14.14
19-Oxoandrost-4-ene-3,17-dione
Androst-4-ene-3,17-dione
Estrone 3-sulfate
2.8.2.4
2.8.2.15
Estrone glucuronide
1.1.1.148
Estradiol-17α
2-Methoxyestrone-3-sulfate
3.1.6.1
2-Hydroxyestrone
2.1.1.6
2-Methoxyestrone
2-Methoxyestrone-3-glucuronide
Estrone
16-α-Hydroxyestrone
1.1.1.-
Androsterone
5α-Androstane-3,17-dione
AKR1C2
Androsterone-glucuronide
7α-Hydroxy-androstenedione
1.1.1.51
1.1.1.64
1.1.1.239
C18-Steroids
7α-Hydroxy-testosterone
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
Testosterone
19-Hydroxy-testosterone
19-Oxo-testosterone
Estradiol-17β
Estriol
16-Glucuronide-estriol
2-Methoxy-estradiol-17β
3β,17β-Dihydroxy-androst-5-ene
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β-3-glucuronide
C19-Steroids
5β-Dihydro-testosterone
1.14.99.11
6β-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β-3-sulfate
Testosterone glucuronide
Estradiol-17β-3-glucuronide
Estradiol-17β-3-sulfate
Data on KEGG graph
Rendered by Pathview





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

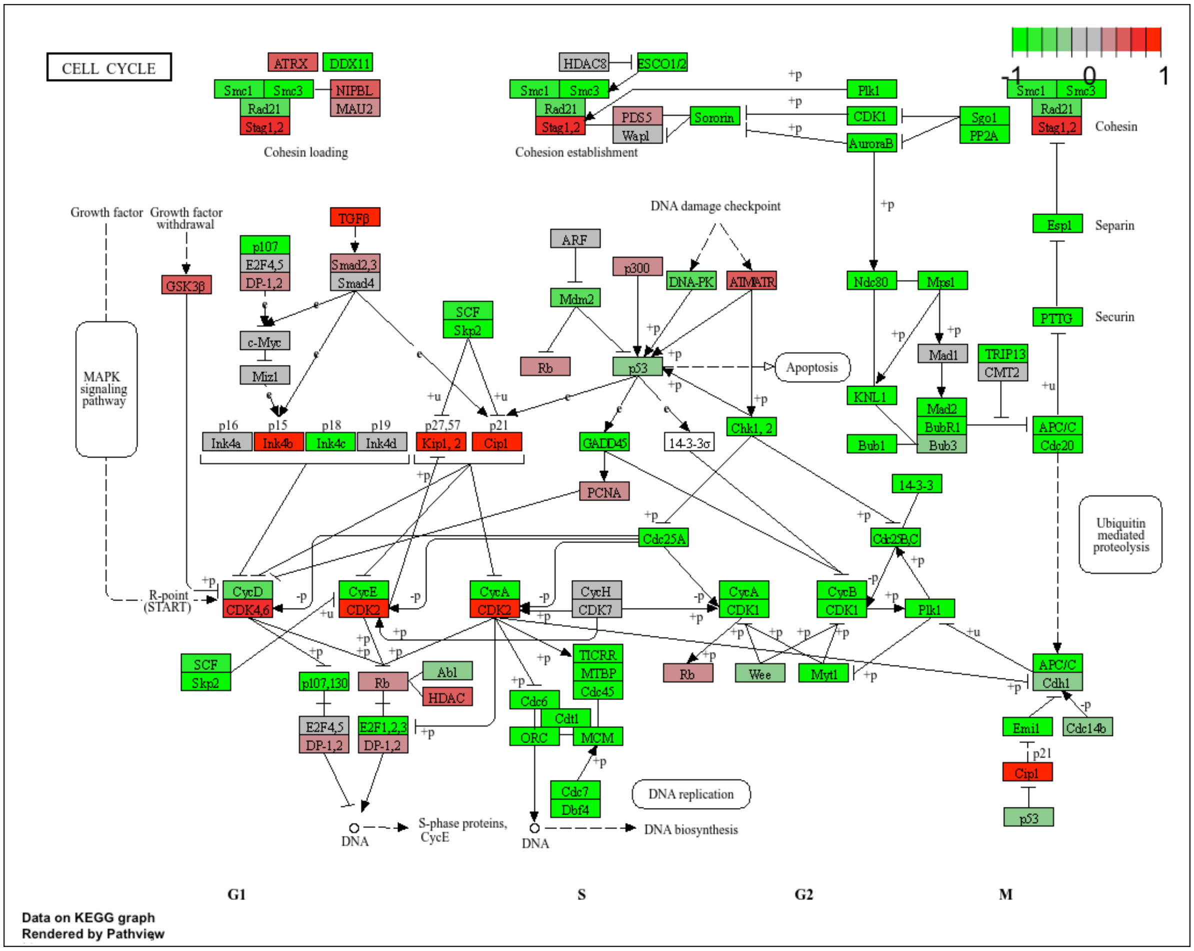
```
keggrespathways.down <- row.names(keggres$less)[1:5]
```

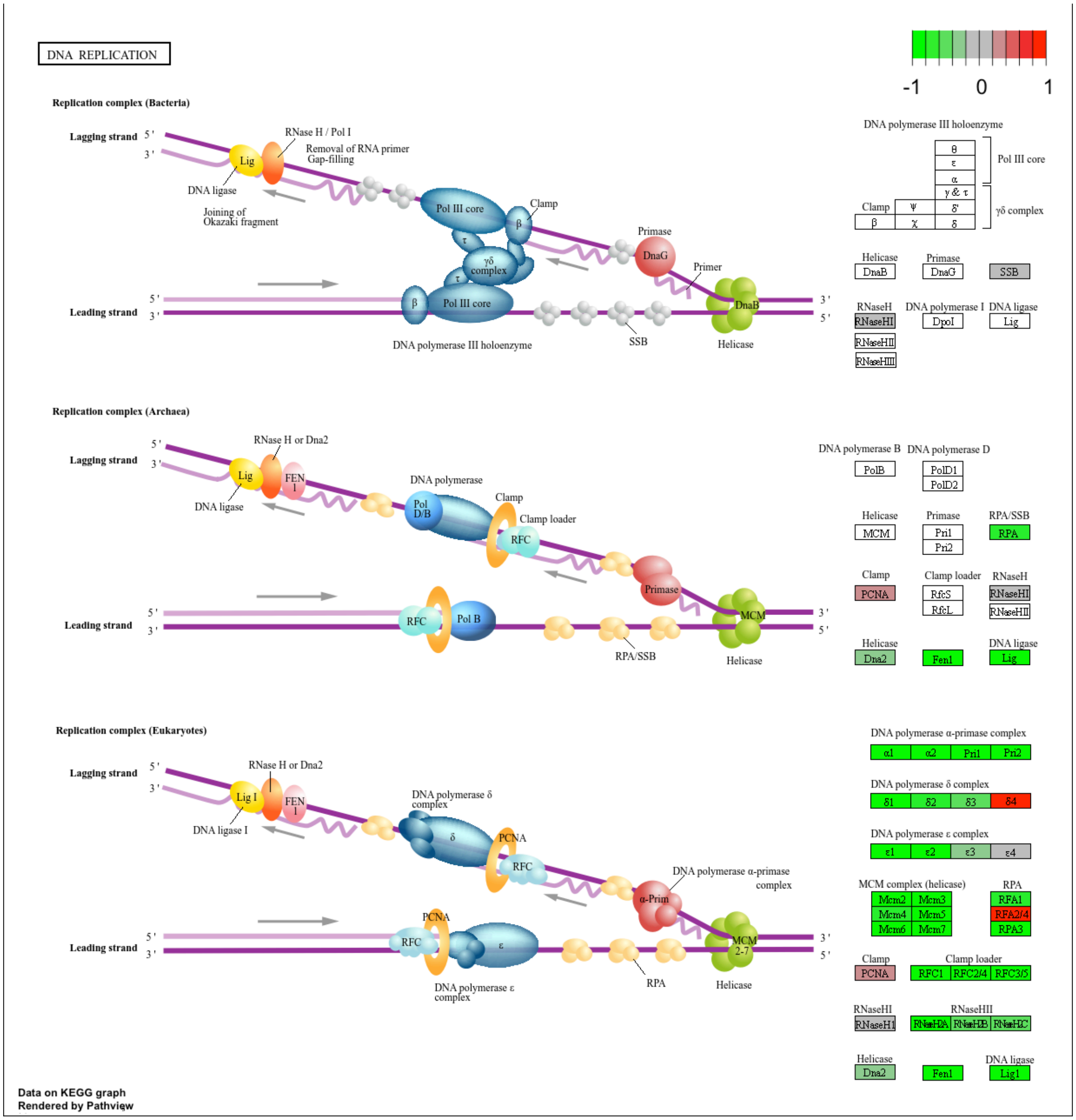
```r
keggresids.down <- substr(keggrespathways.down, start = 1, stop = 8)
keggresids.down
```

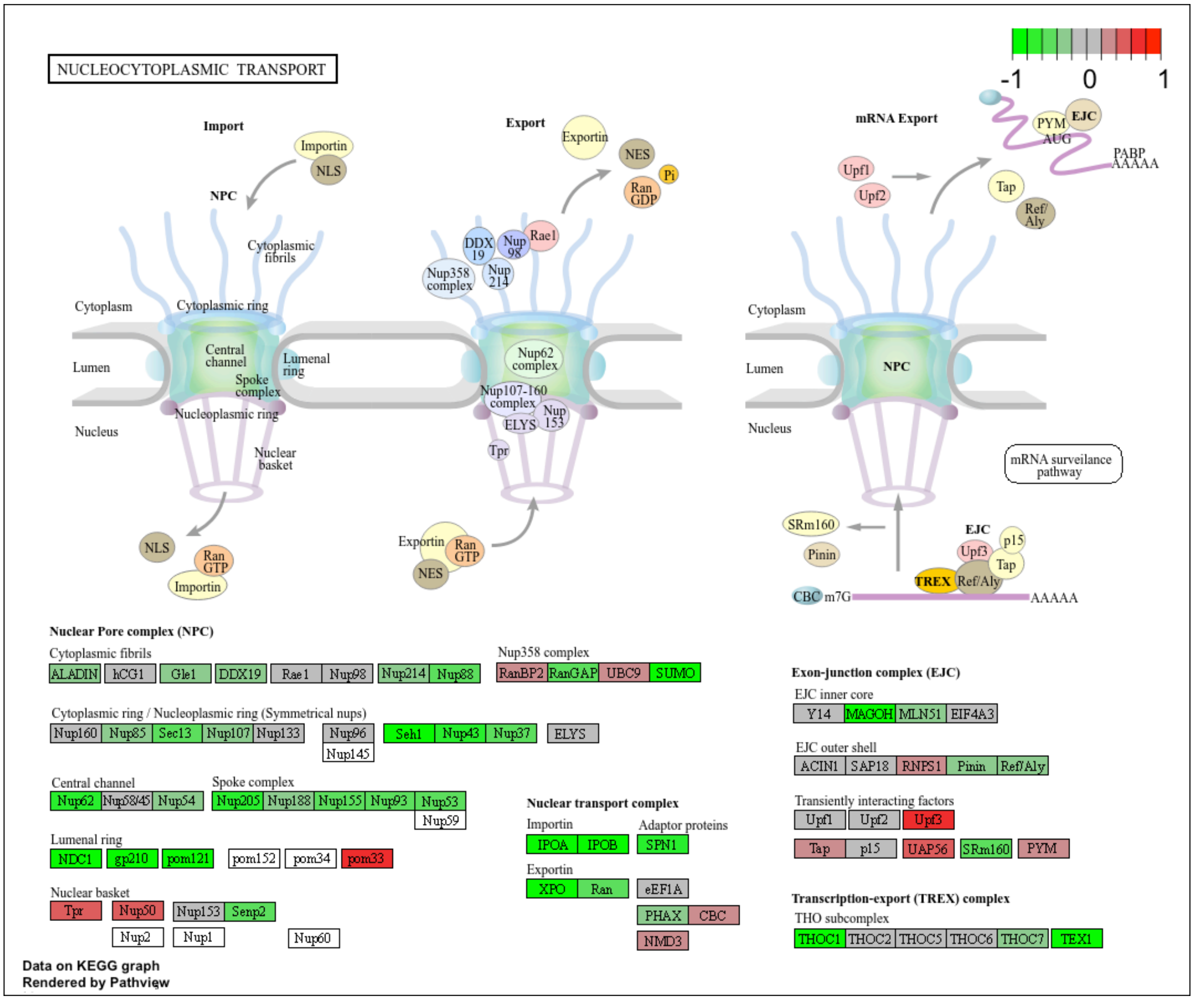
```
[1] "hsa04110" "hsa03030" "hsa03013" "hsa03440" "hsa04114"
```

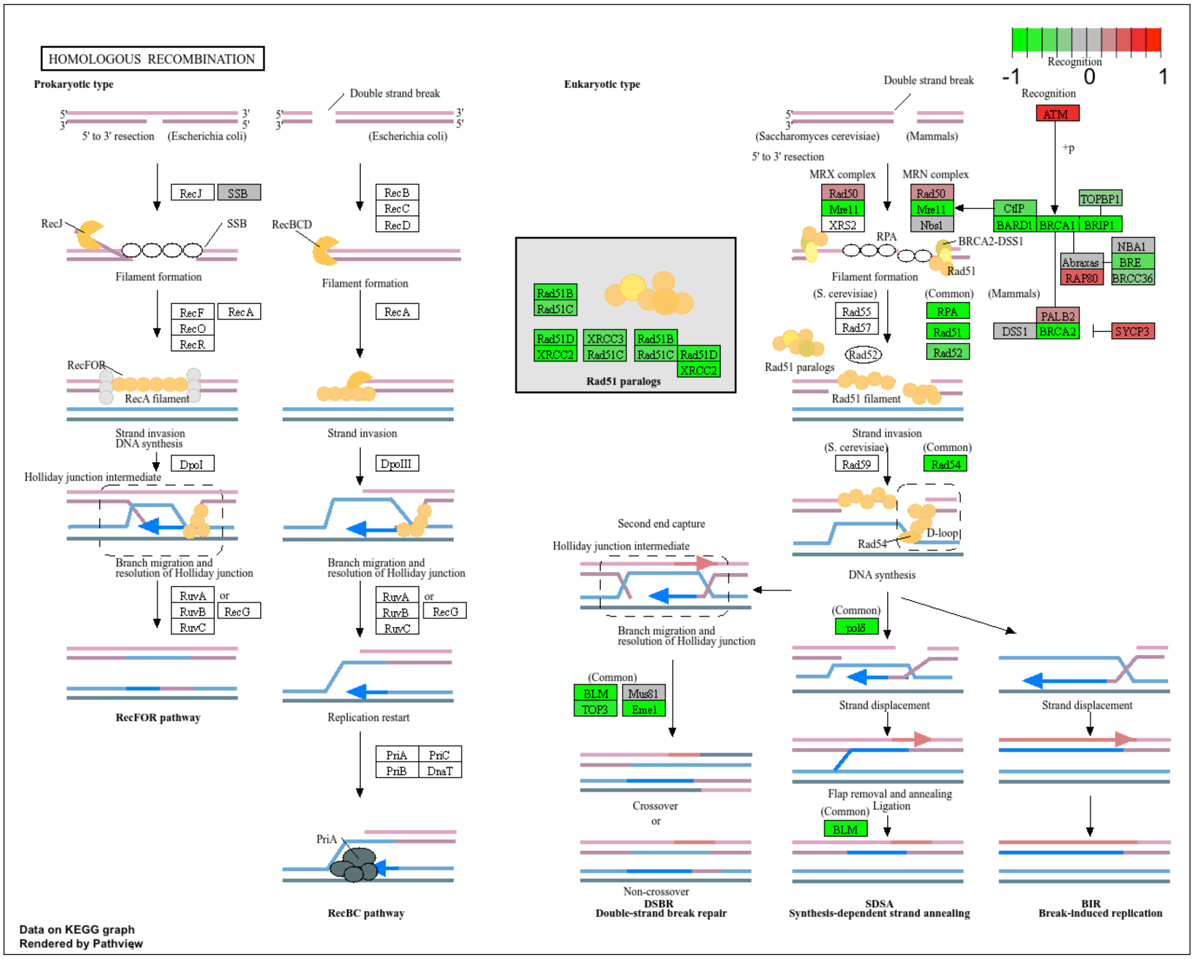
```r
pathview(gene.data=foldchanges, pathway.id=keggresids.down, species="hsa")
```

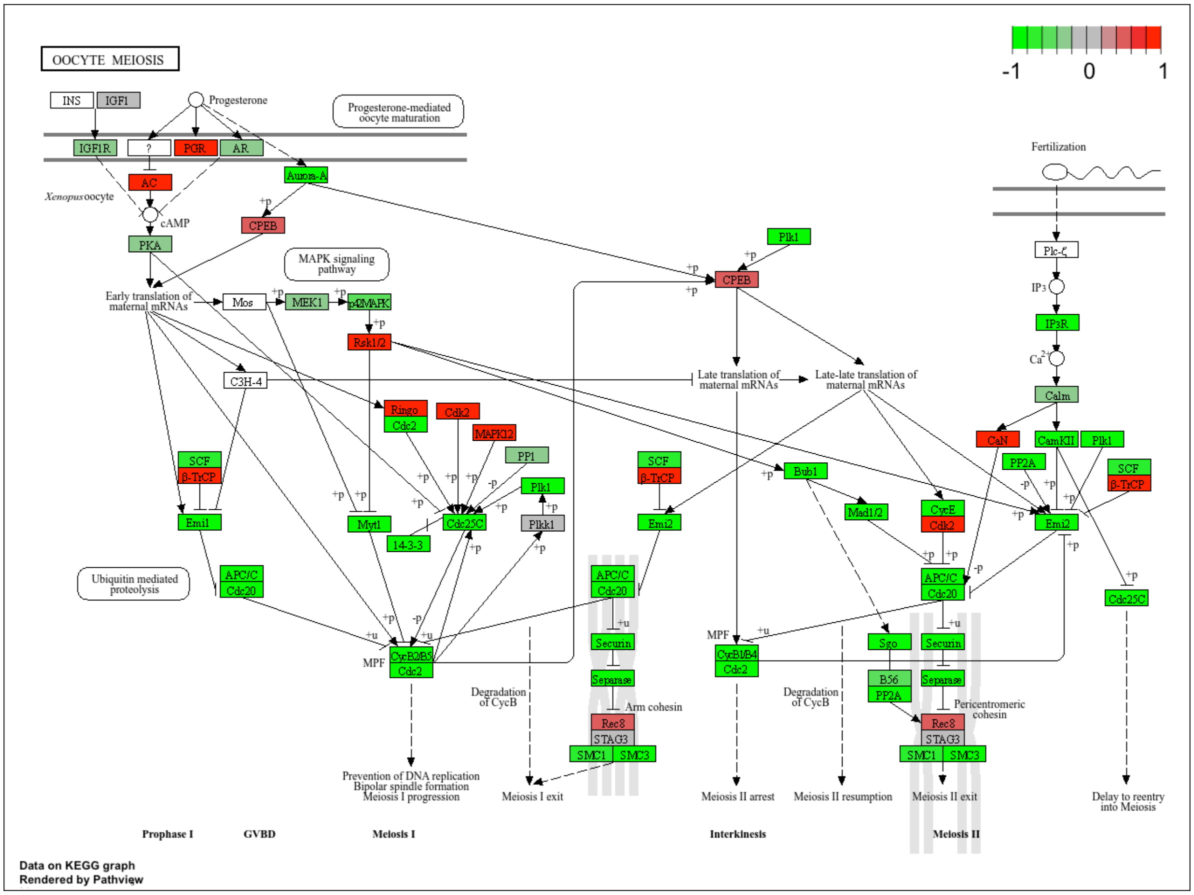
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa04110.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

Info: Writing image file hsa03440.pathview.png

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

Info: Working in directory /Users/xaler/Desktop/BIMM 143 Files/BIMM143Class14

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


CELL CYCLE
-1 0 1
ATRX
DDX11
Smc1
Smc3
Rad21
Stag1,2
NIPBL
MAU2
Cohesin loading
HDAC8
ESCO1/2
PDS5
Wapl
Sororin
Cohesion establishment
Plk1
CDK1
AuroraB
Sgo1
PP2A
Cohesin
+p
Growth factor
Growth factor withdrawal
GSK3β
MAPK signaling pathway
TGFβ
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
c-Myc
Miz1
p16
p15
p18
p19
Ink4a
Ink4b
Ink4c
Ink4d
SCF
Skp2
p27,57
p21
Kip1, 2
Cip1
ARF
Mdm2
Rb
p300
DNA damage checkpoint
DNA-PK
ATM/ATR
p53
Apoptosis
GADD45
14-3-3σ
Chk1, 2
PCNA
Ndc80
Mps1
KNL1
Mad1
TRIP13
CMT2
Mad2
BubR1
Bub3
Bub1
APC/C
Cdc20
Espl
Separin
PTTG
Securin
14-3-3
Cdc25B,C
Cdc25A
Ubiquitin mediated proteolysis
R-point (START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
+p
-p
+u
e
SCF
Skp2
p107,130
E2F4,5
DP-1,2
Rb
Abl
HDAC
E2F1,2,3
DP-1,2
TICRR
MTBP
Cdc45
Cdc6
Cdt1
ORC
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
APC/C
Cdh1
Emi1
Cdc14b
p21
Cip1
p53
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'
Lig
RNase H / Pol I
Removal of RNA primer
Gap-filling
DNA ligase
Joining of
Okazaki fragment
Pol III core
Clamp
β
τ
γδ
complex
τ
Primase
DnaG
Primer
Leading strand
5'
3'
β
Pol III core
DnaB
3'
5'
DNA polymerase III holoenzyme
SSB
Helicase
DNA polymerase III holoenzyme
θ
ε
α
γ & τ
Pol III core
Clamp
ψ
δ'
γδ complex
β
χ
δ
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
RNaseHI
RNaseHII
RNaseHIII
DNA polymerase I
DpoI
DNA ligase
Lig
Replication complex (Archaea)
Lagging strand
5'
3'
RNase H or Dna2
Lig
FEN
1
DNA ligase
DNA polymerase
Pol
D/B
Clamp
Clamp loader
RFC
Primase
Leading strand
5'
3'
RFC
Pol B
MCM
3'
5'
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)
Lagging strand
5'
3'
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
Leading strand
5'
3'
PCNA
RFC
ε
MCM
2-7
3'
5'
DNA polymerase ε
complex
RPA
Helicase
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
RPA1
RPA2/4
RPA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHI
RNaseH1
RNaseHII
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
Nup 153
ELYS
Tpr
Exportin
Ran GTP
NES
mRNA Export
PYM
EJC
AUG
PABP
AAAAA
Upf1
Upf2
Tap
Ref/ Aly
Cytoplasm
Lumen
NPC
Nucleus
mRNA surveilance pathway
SRm160
Pinin
EJC
p15
Upf3
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
Eukaryotic type
-1
0
1
Recognition
Double strand break
5'
3'
5' to 3' resection
(Escherichia coli)
RecJ
SSB
RecB
RecC
RecD
RecBCD
Filament formation
RecF
RecO
RecR
RecA
RecFOR
RecA filament
Strand invasion
DNA synthesis
DpoI
DpoIII
Holliday junction intermediate
Branch migration and resolution of Holliday junction
RuvA
RuvB
RuvC
RecG
or
RecFOR pathway
Replication restart
PriA
PriB
PriC
DnaT
RecBC pathway
Rad51B
Rad51C
Rad51D
XRCC2
XRCC3
Rad51 paralogs
(Saccharomyces cerevisiae)
(Mammals)
MRX complex
MRN complex
Rad50
Mre11
XRS2
Nbs1
CtIP
BARD1
BRCA1
BRIP1
TOPBP1
ATM
+p
NBA1
Abraxas
BRE
RAP80
BRCC36
BRCA2-DSS1
RPA
Rad51
(S. cerevisiae)
(Common)
Rad55
Rad57
Rad52
PALB2
DSS1
BRCA2
SYCP3
Rad51 filament
Rad59
Rad54
D-loop
Second end capture
DNA synthesis
pol δ
BLM
TOP3
Mus81
Eme1
Crossover
Non-crossover
Strand displacement
Flap removal and annealing
Ligation
DSBR
Double-strand break repair
SDSA
Synthesis-dependent strand annealing
BIR
Break-induced replication
Data on KEGG graph
Rendered by Pathview



(3) Gene Ontology

We will do a similar procedure with Gene Ontology using `go.sets.hs` that us all GO terms. `go.subs.hs` is a named list containing BP, CC, and MF ontologies.

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

```
gobpsets = go.sets.hs[go.subs.hs$BP]

gobpres = gage(foldchanges, gsets=gobpsets)

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       0.1952430      113 8.519724e-05
GO:0002009 morphogenesis of an epithelium 0.1952430      339 1.396681e-04
GO:0048729 tissue morphogenesis           0.1952430      424 1.432451e-04
GO:0007610 behavior                       0.1968058      426 1.925222e-04
GO:0060562 epithelial tube morphogenesis  0.3566193      257 5.932837e-04
GO:0035295 tube development               0.3566193      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             5.843127e-12      376 1.536227e-15
GO:0000280 nuclear division              5.843127e-12      352 4.286961e-15
GO:0007067 mitosis                       5.843127e-12      352 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.195965e-11      362 1.169934e-14
GO:0007059 chromosome segregation        1.659009e-08      142 2.028624e-11
GO:0000236 mitotic prometaphase          1.178690e-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
```

(4) Reactome Analysis

Reactome is a database consisting of biological molecules and their relation to pathways and processes. Let's conduct over-representation enrichment analysis and pathway-topology analysis. https://bioconductor.org/packages/release/bioc/html/ReactomePA.html and https://reactome.org/ Don't forget to install `BiocManager::install("ReactomePA")` if you want to do this in R, but otherwise, do this on the web page.

```
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"
```

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

> Q. What pathway has the most significant "Entities p-value"? Do the most significant pathways listed match your previous KEGG results? What factors could cause differences between the two methods?

The pathway with the most significance is the mitotic cell cycle pathway with a P-value of 2.02E-5. The cell cycle in KEGG is also the most significant. The difference between KEGG and Reactome is that KEGG shows the cell cycle at one layer, but Reactome shows the cell cycle at various levels.