Class16: RNASeq Mini Project

Wai Lam Adele Hong A15999023

11/18/2021

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

# Import metadata and take a peek
colData = read.csv(metaFile, row.names=1)
head(colData)
```

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

```
# Import countdata
countData = read.csv(countFile, row.names=1)
head(countData)
```

```
##                 length SRR493366 SRR493367 SRR493368 SRR493369 SRR493370
## ENSG00000186092    918         0         0         0         0         0
## ENSG00000279928    718         0         0         0         0         0
## ENSG00000279457   1982        23        28        29        29        28
## ENSG00000278566    939         0         0         0         0         0
## ENSG00000273547    939         0         0         0         0         0
## ENSG00000187634   3214       124       123       205       207       212
##                 SRR493371
## ENSG00000186092         0
## ENSG00000279928         0
## ENSG00000279457        46
## ENSG00000278566         0
## ENSG00000273547         0
## ENSG00000187634       258
```

Q1. Complete the code below to remove the troublesome first column from countData. We need to get rid of this funny first column.

```
countData <- 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
```

```
#We should have 6 rows and 6 columns! If we run the [, -1] code again and again, each time it will get
```

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).

```
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
```

```
countsnozero <- countData[rowSums(countData) !=0,]
```

```
# BiocManager::install("DESeq2")
library(DESeq2)
```

```
## Loading required package: S4Vectors

## Loading required package: stats4

## Loading required package: BiocGenerics

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which.max, which.min
```

```
##
## Attaching package: 'S4Vectors'

## The following objects are masked from 'package:base':
##
##     expand.grid, I, unname

## Loading required package: IRanges

## Loading required package: GenomicRanges

## Warning: package 'GenomicRanges' was built under R version 4.1.2

## Loading required package: GenomeInfoDb

## Loading required package: SummarizedExperiment

## Loading required package: MatrixGenerics

## Loading required package: matrixStats

##
## Attaching package: 'MatrixGenerics'

## The following objects are masked from 'package:matrixStats':
##
##     colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
##     colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
##     colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
##     colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
##     colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
##     colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
##     colWeightedMeans, colWeightedMedians, colWeightedSds,
##     colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
##     rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
##     rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
##     rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
##     rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
##     rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
##     rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
##     rowWeightedSds, rowWeightedVars

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
##
## Attaching package: 'Biobase'

## The following object is masked from 'package:MatrixGenerics':
##
##     rowMedians

## The following objects are masked from 'package:matrixStats':
##
##     anyMissing, rowMedians
```

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

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

```
dds = DESeq(dds)
```

```
## estimating size factors

## estimating dispersions

## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing
```

```
dds
```

```
## class: DESeqDataSet
## dim: 19808 6
## metadata(1): version
## assays(4): counts mu H cooks
## rownames(19808): ENSG00000186092 ENSG00000279928 ... ENSG00000277475
##   ENSG00000268674
## rowData names(22): baseMean baseVar ... deviance maxCooks
## colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
## colData names(2): condition sizeFactor
```

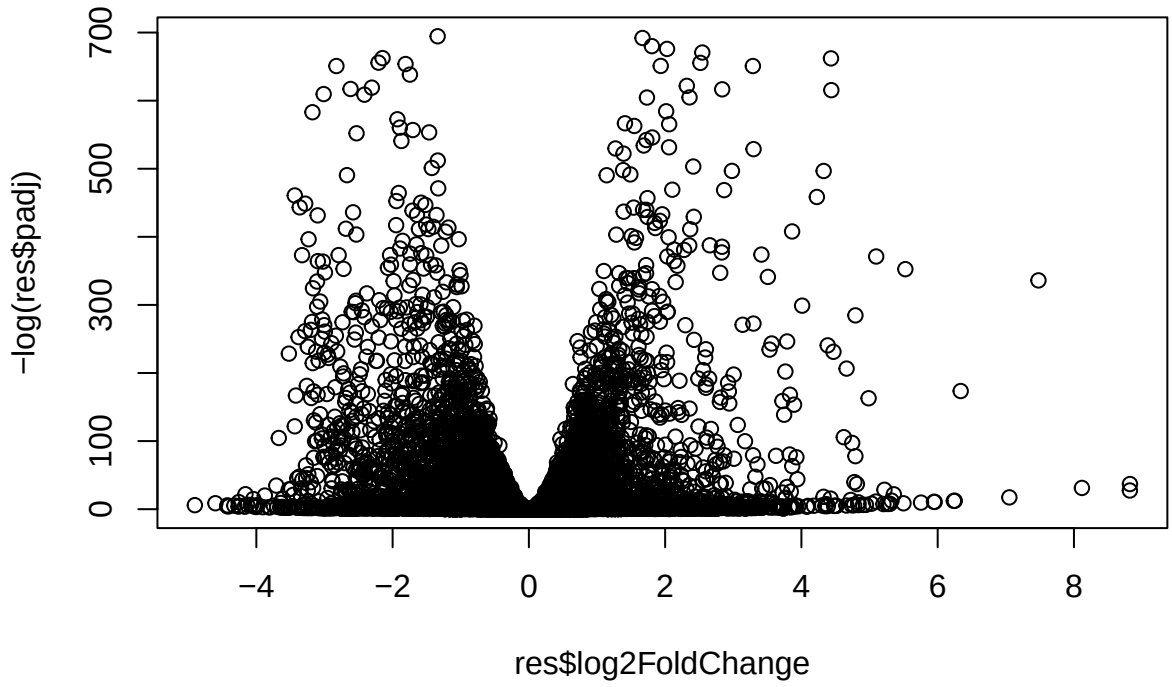
Q3. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

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

summary(res)
```

```
##
## out of 15975 with nonzero total read count
## adjusted p-value < 0.1
## LFC > 0 (up)       : 4349, 27%
## LFC < 0 (down)     : 4393, 27%
## outliers [1]       : 0, 0%
## low counts [2]     : 1221, 7.6%
## (mean count < 0)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

##Volcano Plot

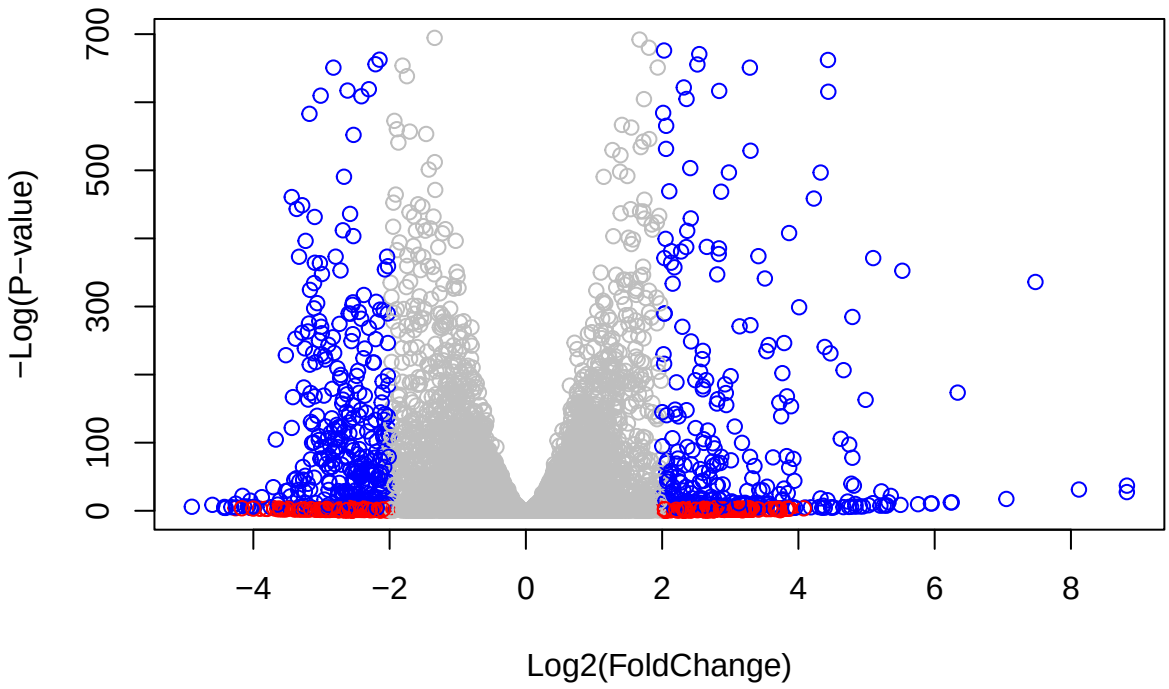
```
plot( res$log2FoldChange, -log(res$padj) )
```



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$pvalue < 0.01) & (abs(res$log2FoldChange) > 2 )
```

```
mycols[ inds ] <- "blue"

plot( res$log2FoldChange, -log(res$padj), col=mycols, xlab="Log2(FoldChange)", ylab="-Log(P-value)" )
```




```
library("AnnotationDbi")
```

```
## Warning: package 'AnnotationDbi' was built under R version 4.1.2
```

```
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"
```

```
res$symbol <- 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$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
```

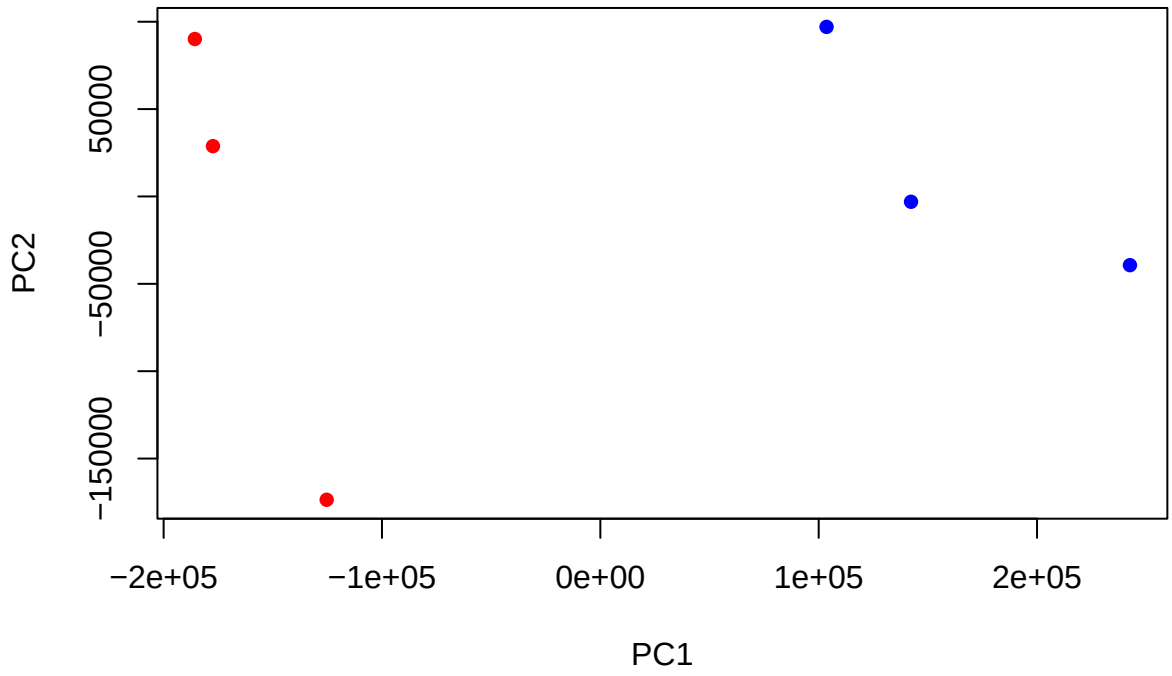
```
res$name =   mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="ENTREZID",
                    multiVals="first")
```

```
## '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>
## ENSG00000186092    0.0000             NA        NA         NA          NA
## ENSG00000279928    0.0000             NA        NA         NA          NA
## ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
## ENSG00000278566    0.0000             NA        NA         NA          NA
## ENSG00000273547    0.0000             NA        NA         NA          NA
## ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43990e-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 0.5215598   1.040744 2.97994e-01
##                       padj      symbol      entrez        name
##                  <numeric> <character> <character> <character>
## ENSG00000186092          NA       79501       79501       79501
## ENSG00000279928          NA          NA          NA          NA
## ENSG00000279457 6.87080e-01   102723897   102723897   102723897
## ENSG00000278566          NA          NA          NA          NA
## ENSG00000273547          NA          NA          NA          NA
## ENSG00000187634 5.16278e-03      148398      148398      148398
## ENSG00000188976 1.76741e-35       26155       26155       26155
## ENSG00000187961 1.13536e-07      339451      339451      339451
## ENSG00000187583 9.18988e-01       84069       84069       84069
## ENSG00000187642 4.03817e-01       84808       84808       84808
```

```
pca <- prcomp(t(countsnozero))
mycols <- rep(c("red", "blue"), each=3)
plot(pca$x[,1:2], col=mycols, pch=16)
```




```
# Run in your R console (i.e. not your Rmarkdown doc!)
# BiocManager::install( c("pathview", "gage", "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).
## ##############################################################################
```

```
library(gage)
```

```
##
```

```
library(gageData)

# Focus on datasubset of KEGG
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"
```

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

```
##     79501      <NA> 102723897      <NA>      <NA>    148398
##        NA        NA 0.1792571        NA        NA 0.4264571
```

Let's run the gage pathway analysis.

```
# Get the results
keggres = gage(foldchanges, gsets=kegg.sets.hs)

attributes(keggres)
```

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

Let's look at the first few down-regulat4ed (less) pathways.

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

```
##                                       p.geomean stat.mean        p.val
## hsa04110 Cell cycle                   7.077982e-06 -4.432593 7.077982e-06
## hsa03030 DNA replication              9.424076e-05 -3.951803 9.424076e-05
## hsa03013 RNA transport                1.012277e-03 -3.122555 1.012277e-03
## hsa04114 Oocyte meiosis               2.563806e-03 -2.827297 2.563806e-03
## hsa03440 Homologous recombination     3.066756e-03 -2.852899 3.066756e-03
## hsa00010 Glycolysis / Gluconeogenesis 4.360092e-03 -2.663825 4.360092e-03
##                                            q.val set.size         exp1
## hsa04110 Cell cycle                   0.001160789      124 7.077982e-06
## hsa03030 DNA replication              0.007727742       36 9.424076e-05
## hsa03013 RNA transport                0.055337821      150 1.012277e-03
## hsa04114 Oocyte meiosis               0.100589607      112 2.563806e-03
## hsa03440 Homologous recombination     0.100589607       28 3.066756e-03
## hsa00010 Glycolysis / Gluconeogenesis 0.119175854       65 4.360092e-03
```

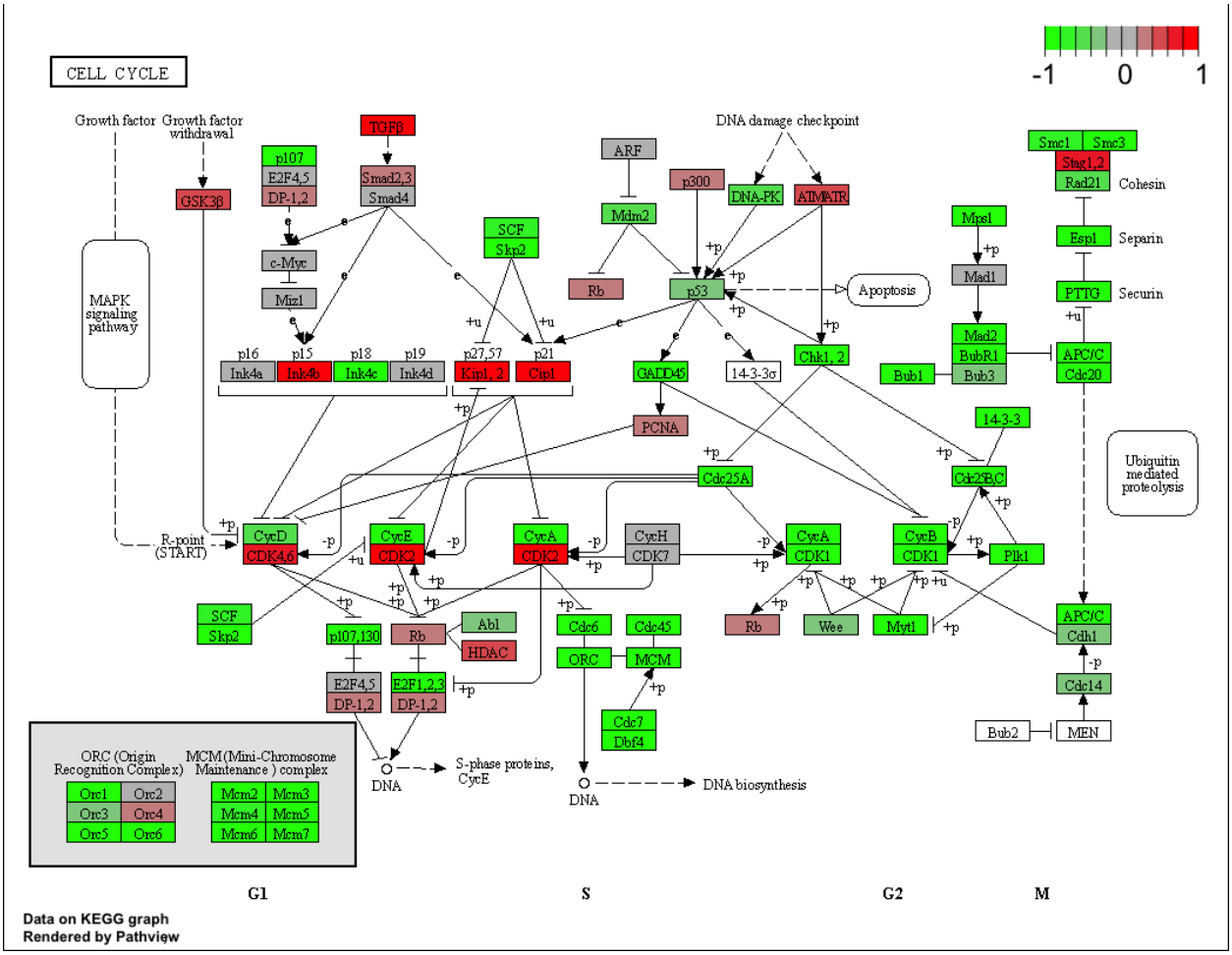
Using the **pathview()** function, we will make pathway plot for the RNASeq. expression results.

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

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

```
## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16
```

```
## 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
```

```
## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16
```

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

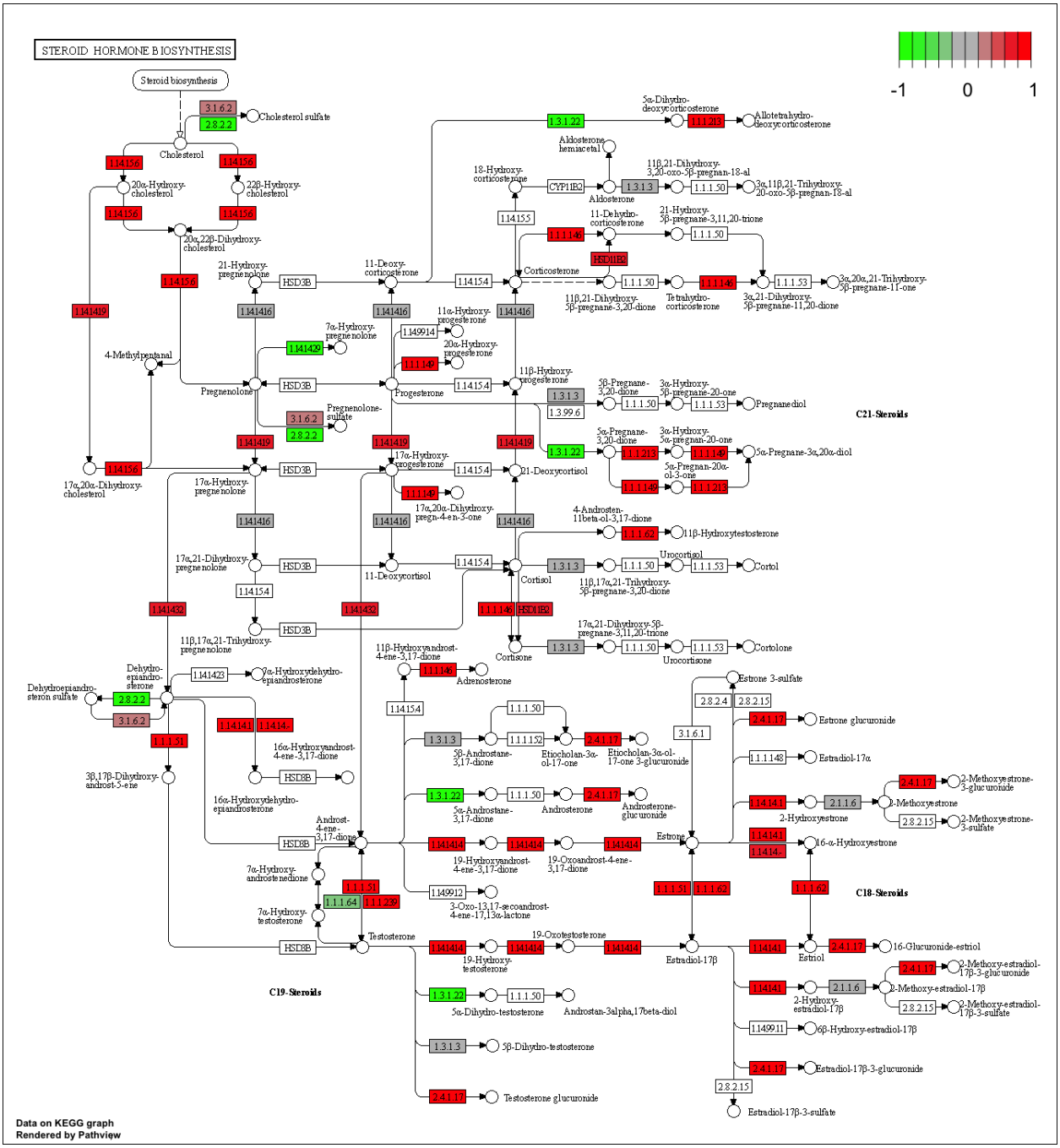
```
## Focus on top 5 upregulated pathways here for demo purposes only
keggrespathwaysup <- rownames(keggres$greater)[1:5]

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

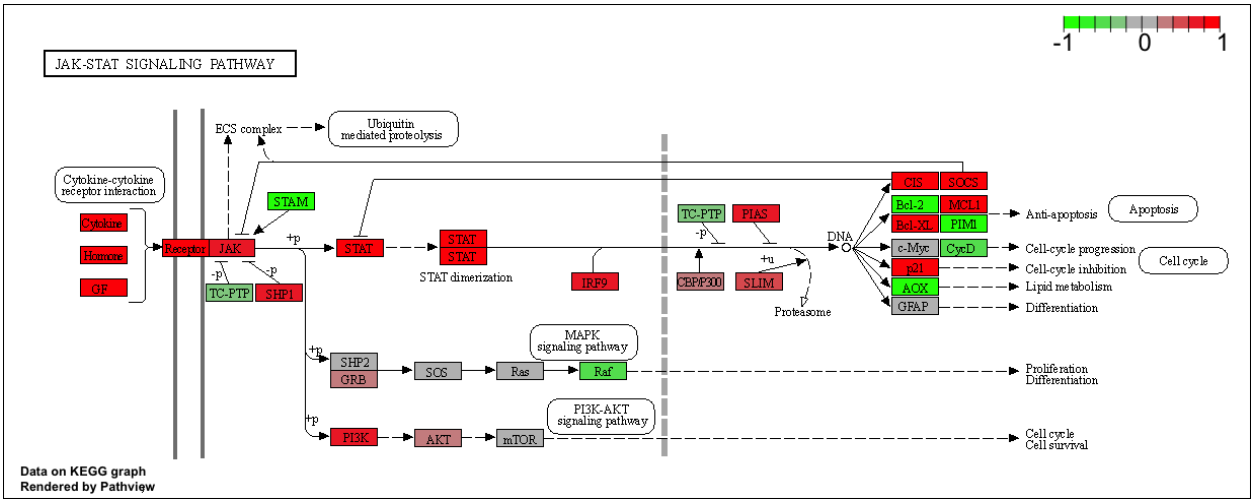
```
## [1] "hsa04740" "hsa04640" "hsa00140" "hsa04630" "hsa04976"
```

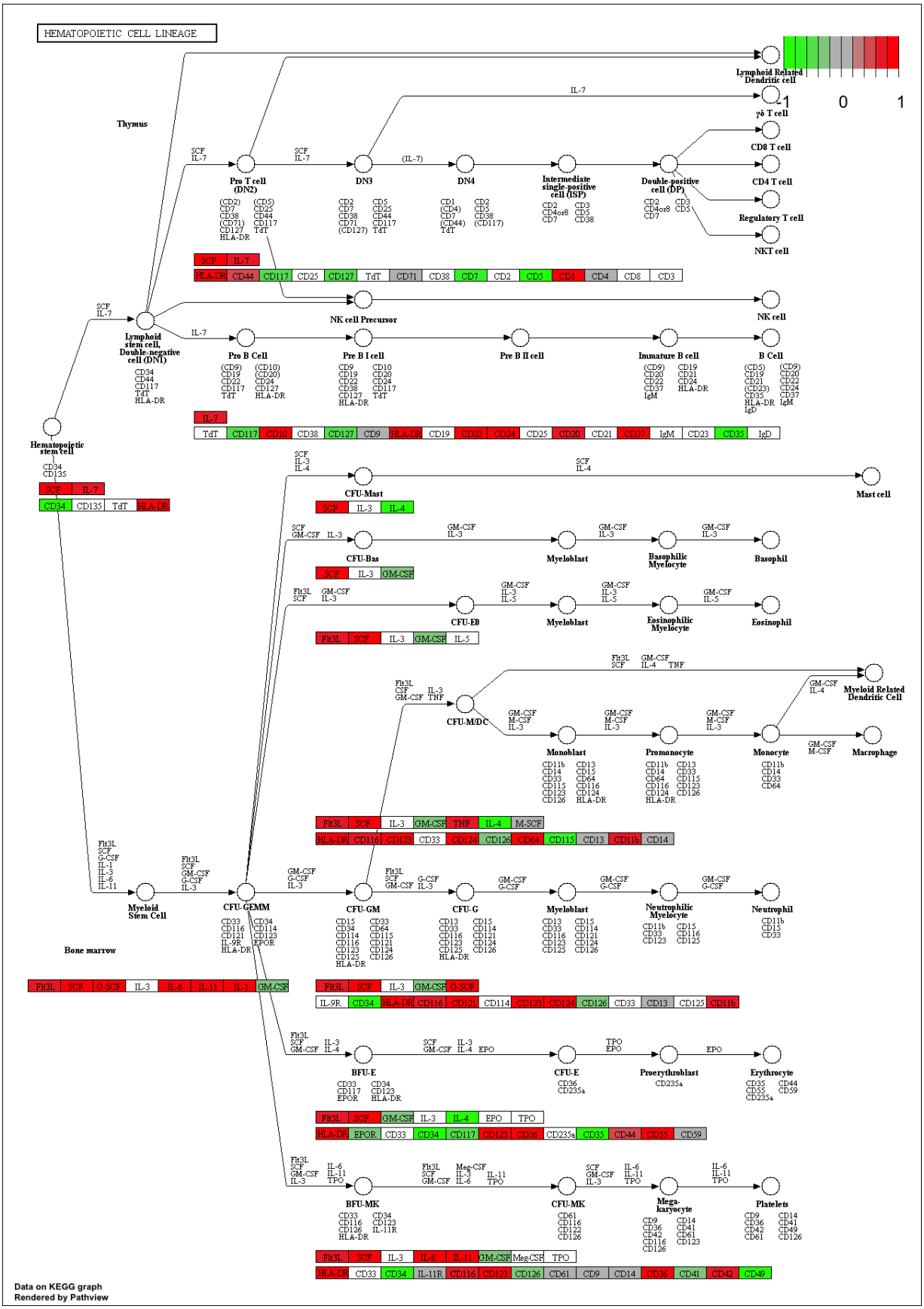
Let's use **pathview()** for all top 5 up-regulated pathways.

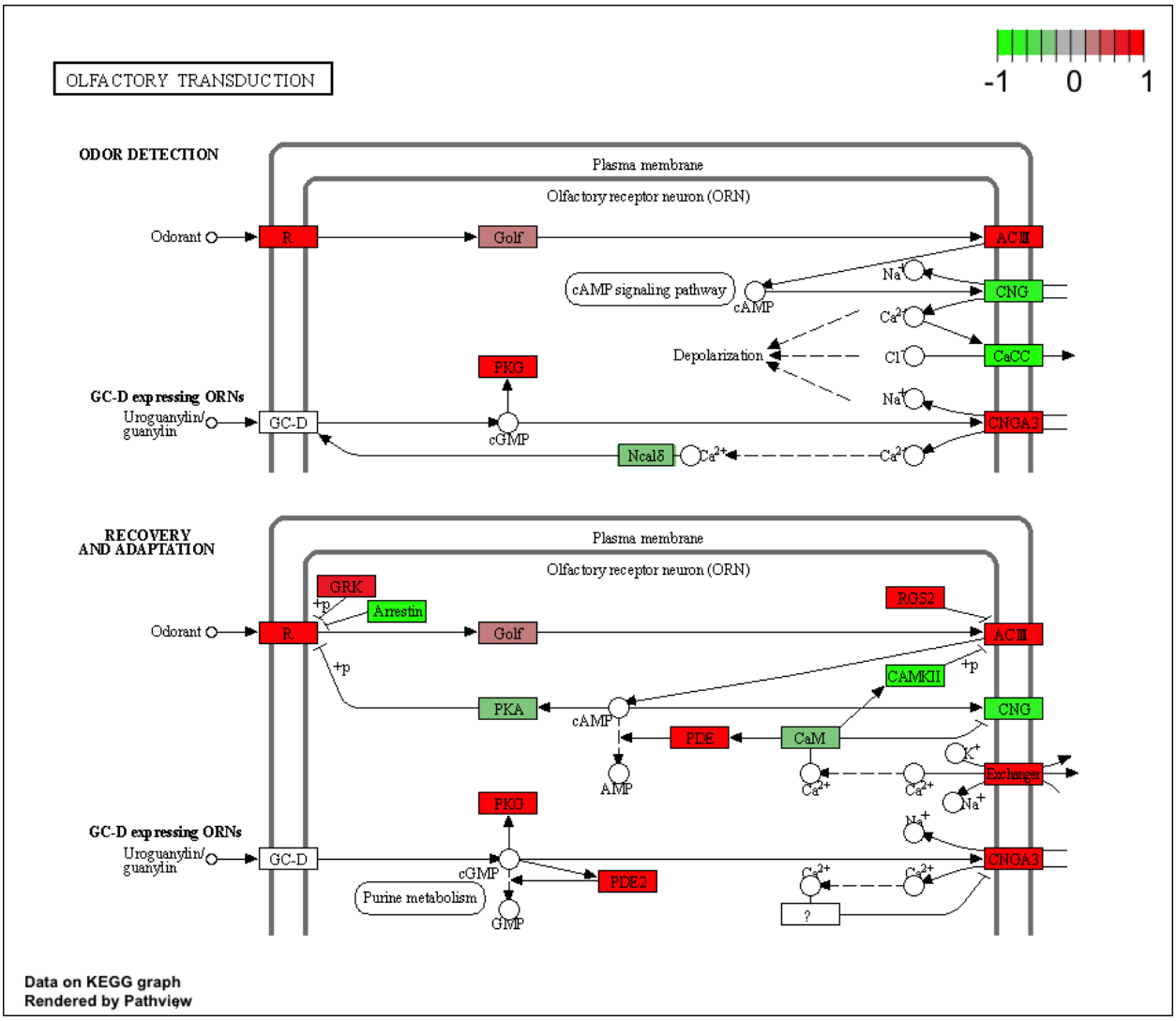
```
pathview(gene.data=foldchanges, pathway.id=keggresidsup, species="hsa")
```

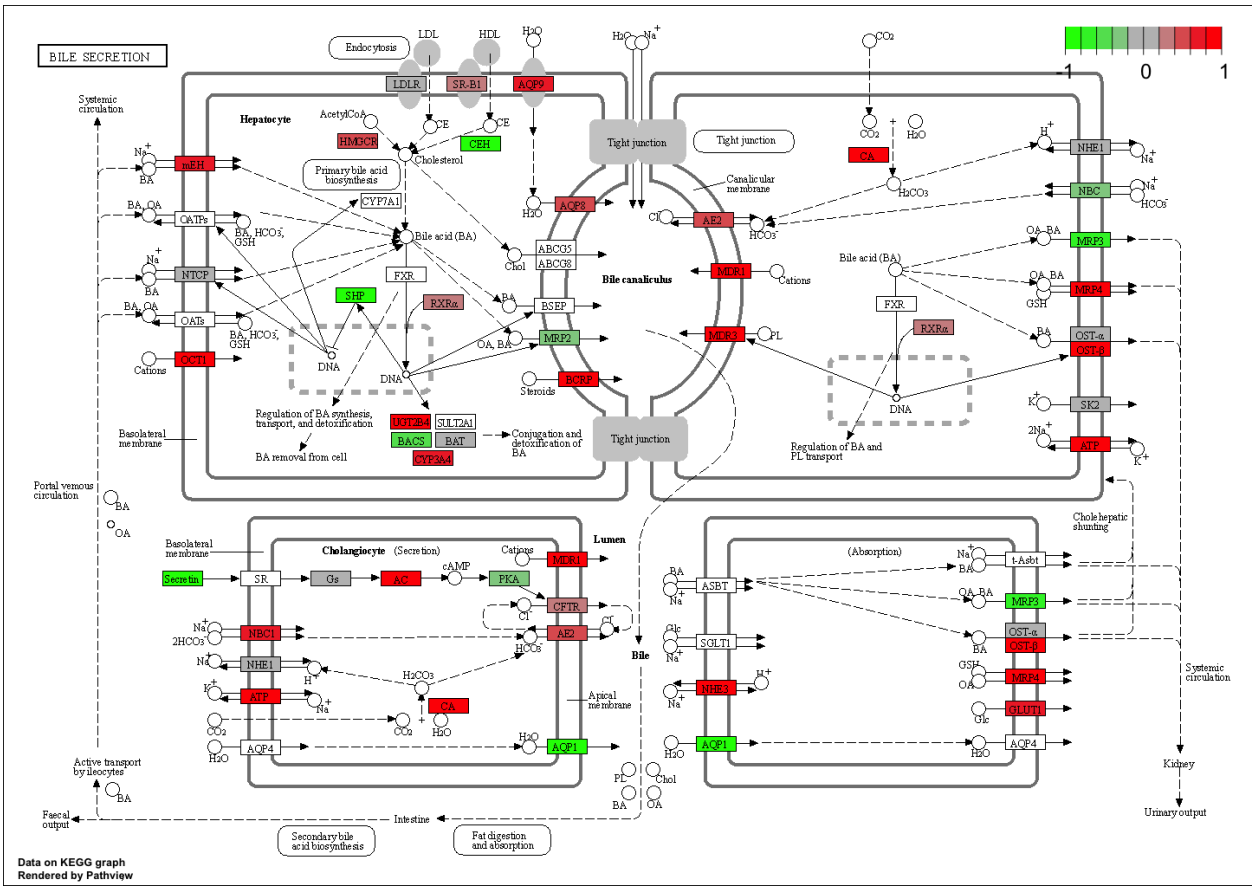
```
## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

## Info: Writing image file hsa04740.pathview.png

## Info: some node width is different from others, and hence adjusted!

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

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

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

## Info: Writing image file hsa00140.pathview.png

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

## Info: Writing image file hsa04630.pathview.png

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

## Info: Writing image file hsa04976.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
11β,21-Dihydroxy-5β-pregnane-3,20-dione
Tetrahydro-corticosterone
3α,21-Dihydroxy-5β-pregnane-11,20-dione
3α,20α,21-Trihydroxy-5β-pregnane-11-one
5α-Dihydro-deoxycorticosterone
Allotetrahydro-deoxycorticosterone
Aldosterone hemiacetal
18-Hydroxy-corticosterone
Aldosterone
11β,21-Dihydroxy-3,20-oxo-5β-pregnan-18-al
3α,11β,21-Trihydroxy-20-oxo-5β-pregnan-18-al
11-Dehydro-corticosterone
21-Hydroxy-5β-pregnane-3,11,20-trione
HSD11B2
7α-Hydroxy-pregnenolone
11α-Hydroxy-progesterone
20α-Hydroxy-progesterone
4-Methylpentanal
Pregnenolone
Progesterone
11β-Hydroxy-progesterone
5β-Pregnane-3,20-dione
3α-Hydroxy-5β-pregnane-20-one
Pregnanediol
Pregnenolone-sulfate
17α-Hydroxy-progesterone
21-Deoxycortisol
5α-Pregnane-3,20-dione
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
5α-Pregnane-20α-ol-3-one
17α,20α-Dihydroxy-cholesterol
17α-Hydroxy-pregnenolone
17α,20α-Dihydroxy-pregn-4-en-3-one
4-Androsten-11beta-ol-3,17-dione
11β-Hydroxytestosterone
17α,21-Dihydroxy-pregnenolone
11-Deoxycortisol
Cortisol
Urocortisol
Cortol
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
11β,17α,21-Trihydroxy-pregnenolone
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Cortisone
Urocortisone
Cortolone
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
Dehydroepiandro-sterone sulfate
Dehydro-epiandro-sterone
7α-Hydroxydehydro-epiandrosterone
16α-Hydroxyandrost-4-ene-3,17-dione
3β,17β-Dihydroxy-androst-5-ene
16α-Hydroxydehydro-epiandrosterone
5β-Androstane-3,17-dione
Etiocholan-3α-ol-17-one
Etiocholan-3α-ol-17-one 3-glucuronide
5α-Androstane-3,17-dione
Androsterone
Androsterone-glucuronide
Androst-4-ene-3,17-dione
19-Hydroxyandrost-4-ene-3,17-dione
19-Oxoandrost-4-ene-3,17-dione
Estrone
Estrone 3-sulfate
Estrone glucuronide
Estradiol-17α
2-Methoxyestrone-3-glucuronide
2-Hydroxyestrone
2-Methoxyestrone
2-Methoxyestrone-3-sulfate
16α-Hydroxyestrone
7α-Hydroxy-androstenedione
3-Oxo-13,17-secoandrost-4-ene-17,13α-lactone
7α-Hydroxy-testosterone
Testosterone
19-Oxotestosterone
19-Hydroxy-testosterone
Estradiol-17β
Estriol
16-Glucuronide-estriol
2-Methoxy-estradiol-17β-3-glucuronide
2-Methoxy-estradiol-17β
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β-3-sulfate
6β-Hydroxy-estradiol-17β
Estradiol-17β-3-glucuronide
Estradiol-17β-3-sulfate
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
5β-Dihydro-testosterone
Testosterone glucuronide
C21-Steroids
C18-Steroids
C19-Steroids
Data on KEGG graph
Rendered by Pathview


-1
0
1
JAK-STAT SIGNALING PATHWAY
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
Cytokine
Hormone
GF
Receptor
JAK
STAM
+p
-p
TC-PTP
SHP1
STAT
STAT
STAT
STAT dimerization
IRF9
TC-PTP
PIAS
-p
+u
CBP/P300
SLIM
Proteasome
DNA
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
c-Myc
CycD
p21
ACOX
GFAP
Anti-apoptosis
Apoptosis
Cell-cycle progression
Cell-cycle inhibition
Lipid metabolism
Differentiation
Cell cycle
MAPK signaling pathway
SHP2
GRB
SOS
Ras
Raf
Proliferation
Differentiation
PI3K-AKT signaling pathway
PI3K
AKT
mTOR
Cell cycle
Cell survival
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)
(CD2) (CD5) CD7 CD25 CD38 CD44 (CD71) CD117 CD127 T4T HLA-DR
CD2 CD5 CD7 CD25 CD38 CD44 CD71 CD117 (CD127) T4T
CD1 CD2 (CD4) CD5 CD7 CD38 (CD44) (CD117) T4T
CD2 CD3 CD4or8 CD5 CD7 CD38
CD2 CD3 CD4or8 CD5 CD7
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)
Pre B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
CD34 CD44 CD117 T4T HLA-DR
(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
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
CFU-Mast
Mast cell
SCF IL-3 IL-4
CFU-Bas
Myeloblast
Basophilic Myelocyte
Basophil
SCF IL-3 GM-CSF
CFU-Eo
Eosinophilic Myelocyte
Eosinophil
Flt3L SCF IL-3 GM-CSF IL-5
CFU-M/DC
Monoblast
Promonocyte
Monocyte
Myeloid Related Dendritic Cell
Macrophage
Flt3L SCF IL-4 TNF
Flt3L SCF GM-CSF IL-3 TNF
CD11b CD13 CD14 CD15 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
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
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Flt3L SCF GM-CSF G-CSF IL-3
CD33 CD34 CD116 CD114 CD121 CD123 IL-9R EPOR HLA-DR
CD15 CD33 CD34 CD64 CD114 CD115 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126
CD11b CD15 CD33 CD116 CD123 CD125
CD11b CD15 CD14 CD33
Bone marrow
Flt3L SCF G-CSF IL-3 IL-6 IL-11 IL-1 GM-CSF
Flt3L SCF IL-3 GM-CSF G-CSF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD11b
BFU-E
CFU-E
Proerythroblast
Erythrocyte
Flt3L SCF GM-CSF IL-3 IL-4
SCF GM-CSF IL-3 IL-4 EPO
TPO EPO
EPO
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
BFU-MK
CFU-MK
Mega-karyocyte
Platelets
Flt3L SCF GM-CSF IL-3
IL-6 IL-11 TPO
Flt3L SCF GM-CSF IL-3 IL-6 IL-11 TPO
Meg-CSF
SCF GM-CSF IL-3
IL-6 IL-11 TPO
CD33 CD116 CD126 HLA-DR
CD34 CD123 IL-11R
CD61 CD116 CD123 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


OLFACTORY TRANSDUCTION
-1
0
1
ODOR DETECTION
Plasma membrane
Olfactory receptor neuron (ORN)
Odorant
R
Golf
ACIII
cAMP signaling pathway
cAMP
Na+
CNG
Ca2+
Depolarization
Cl-
CaCC
PKG
GC-D expressing ORNs
Uroguanylin/
guanylin
GC-D
cGMP
Na+
CNGA3
Ncald
Ca2+
Ca2+
RECOVERY
AND ADAPTATION
Plasma membrane
Olfactory receptor neuron (ORN)
GRK
+p
Arrestin
RGS2
Odorant
R
Golf
ACIII
+p
CAMKII
+p
PKA
cAMP
CNG
PDE
CaM
K+
Exchanger
AMP
Ca2+
Ca2+
Na+
PKG
Na+
GC-D expressing ORNs
Uroguanylin/
guanylin
GC-D
CNGA3
cGMP
PDE2
Ca2+
Ca2+
Purine metabolism
GMP
?
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?

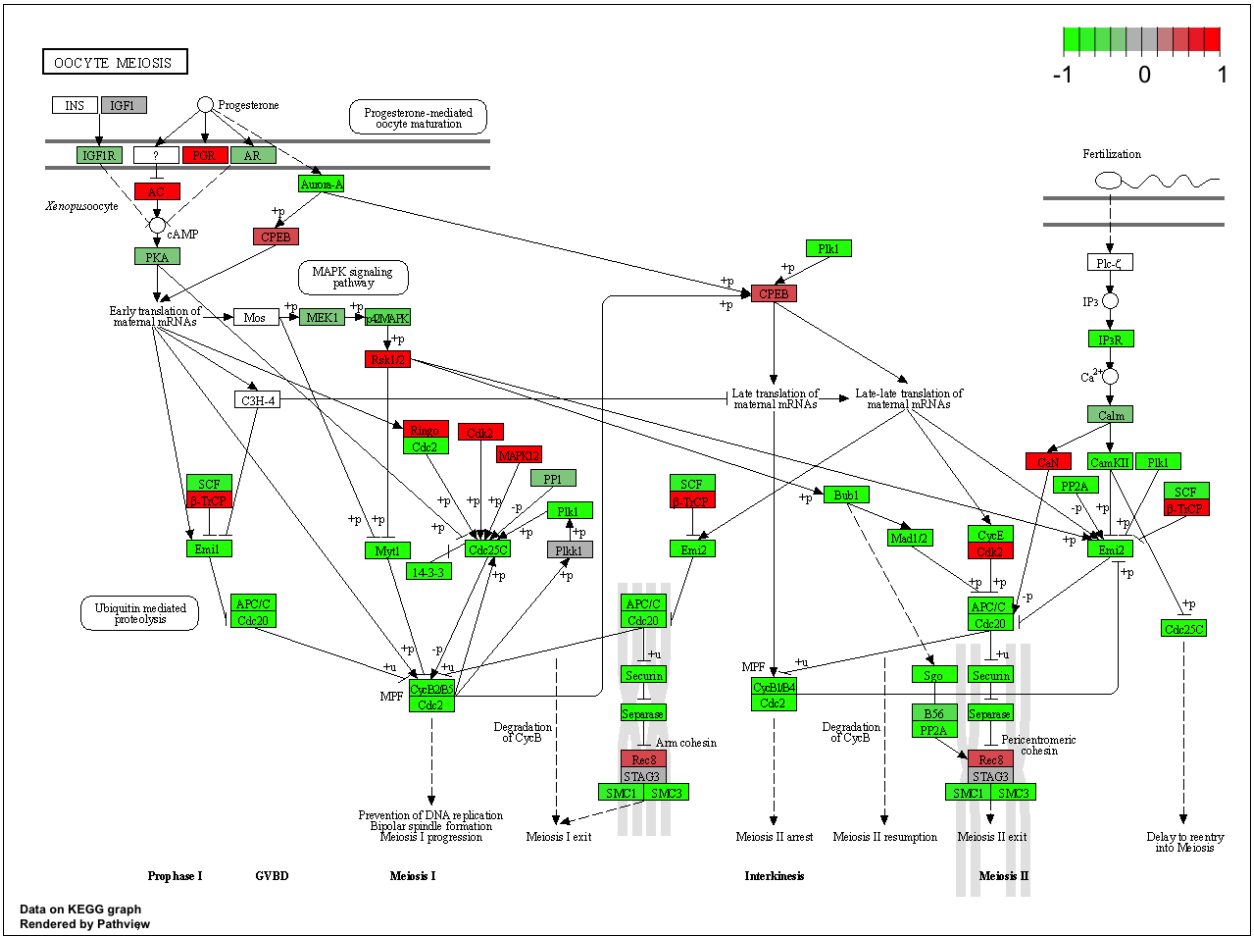
```
## Focus on top 5 down-regulated pathways here for demo purposes only
keggrespathwaysdown <- rownames(keggres$less)[1:5]

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

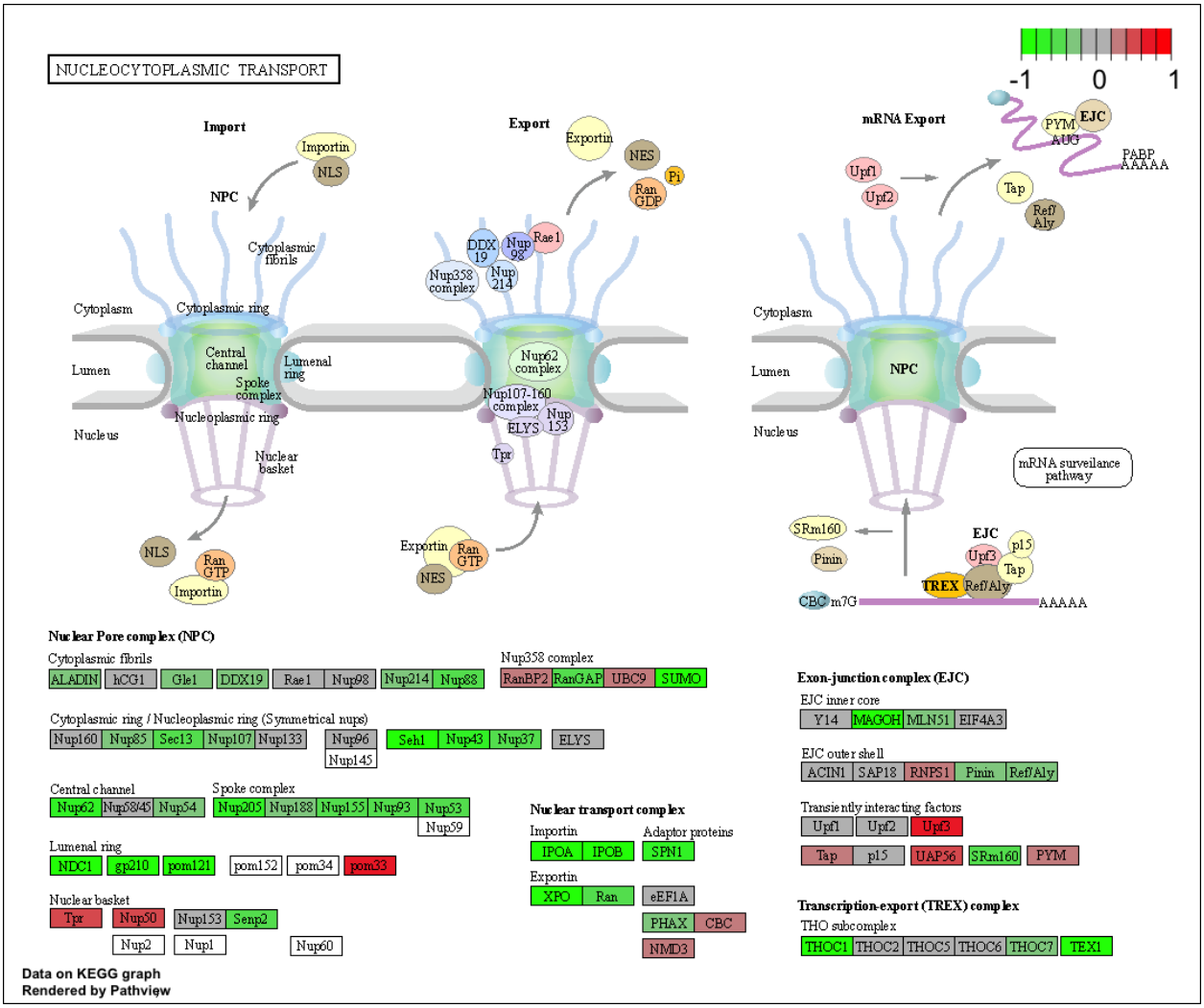
```
## [1] "hsa04110" "hsa03030" "hsa03013" "hsa04114" "hsa03440"
```

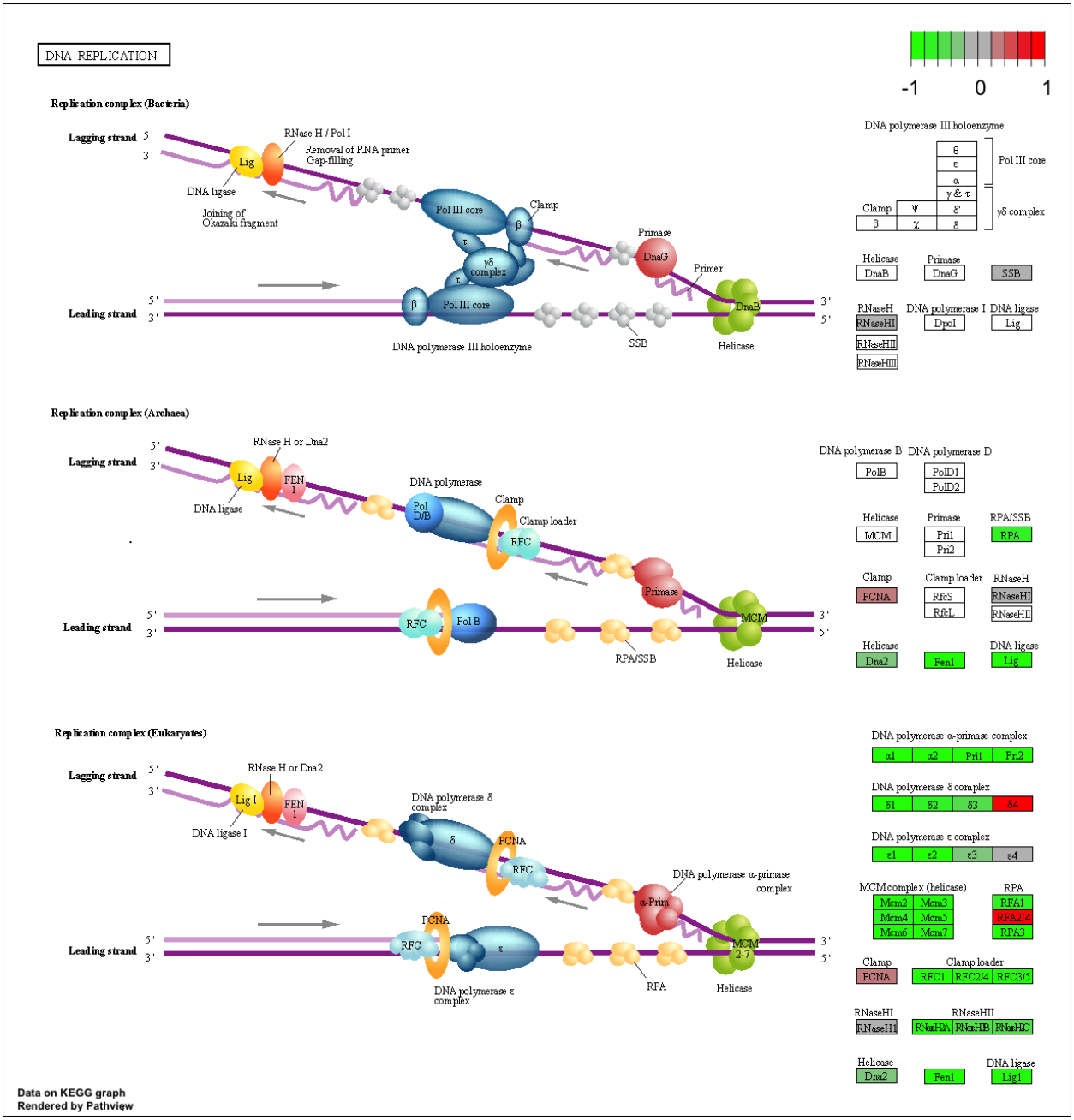
```
pathview(gene.data=foldchanges, pathway.id=keggresidsdown, species="hsa")
```

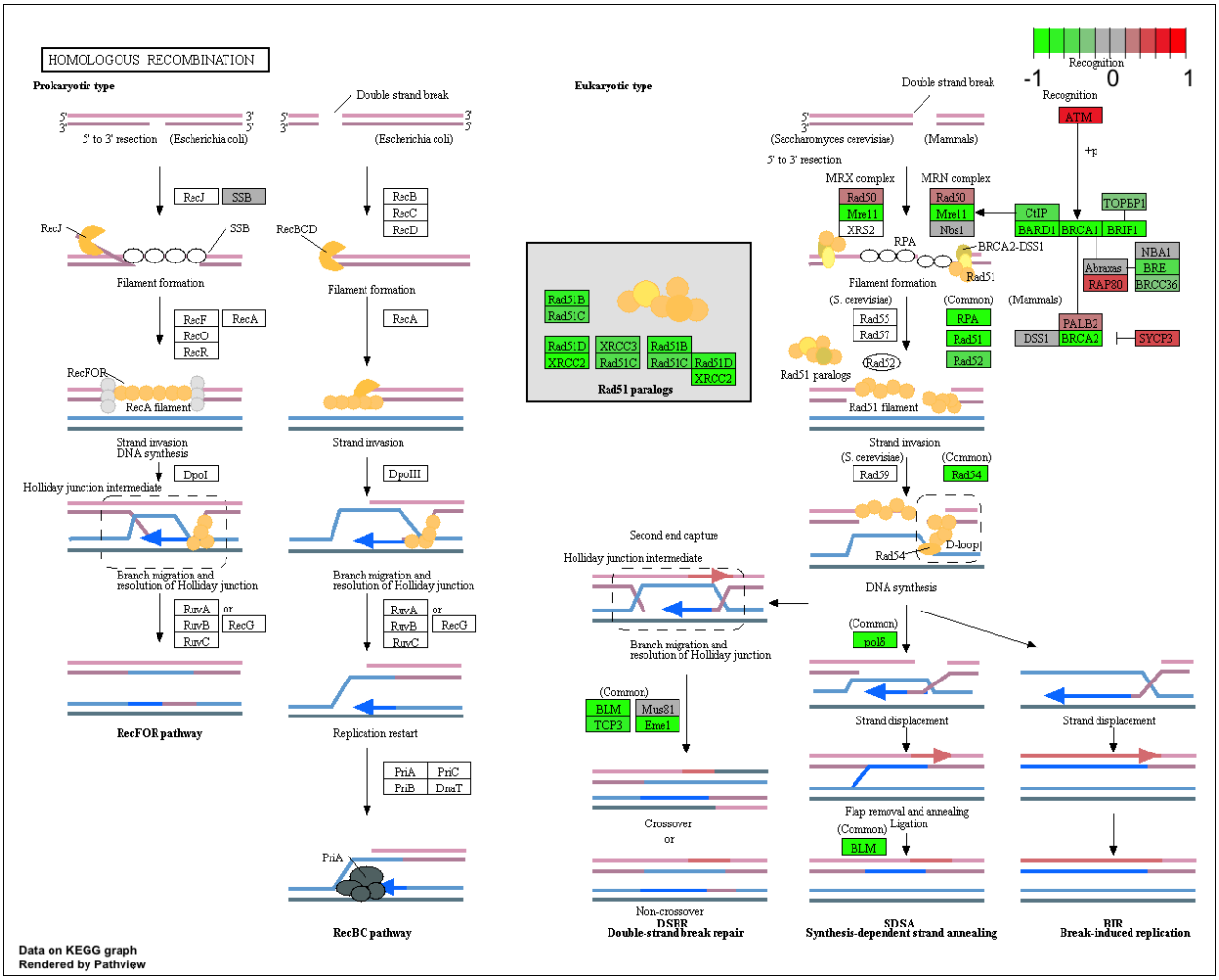
```
## 'select()' returned 1:1 mapping between keys and columns

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

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

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16
```

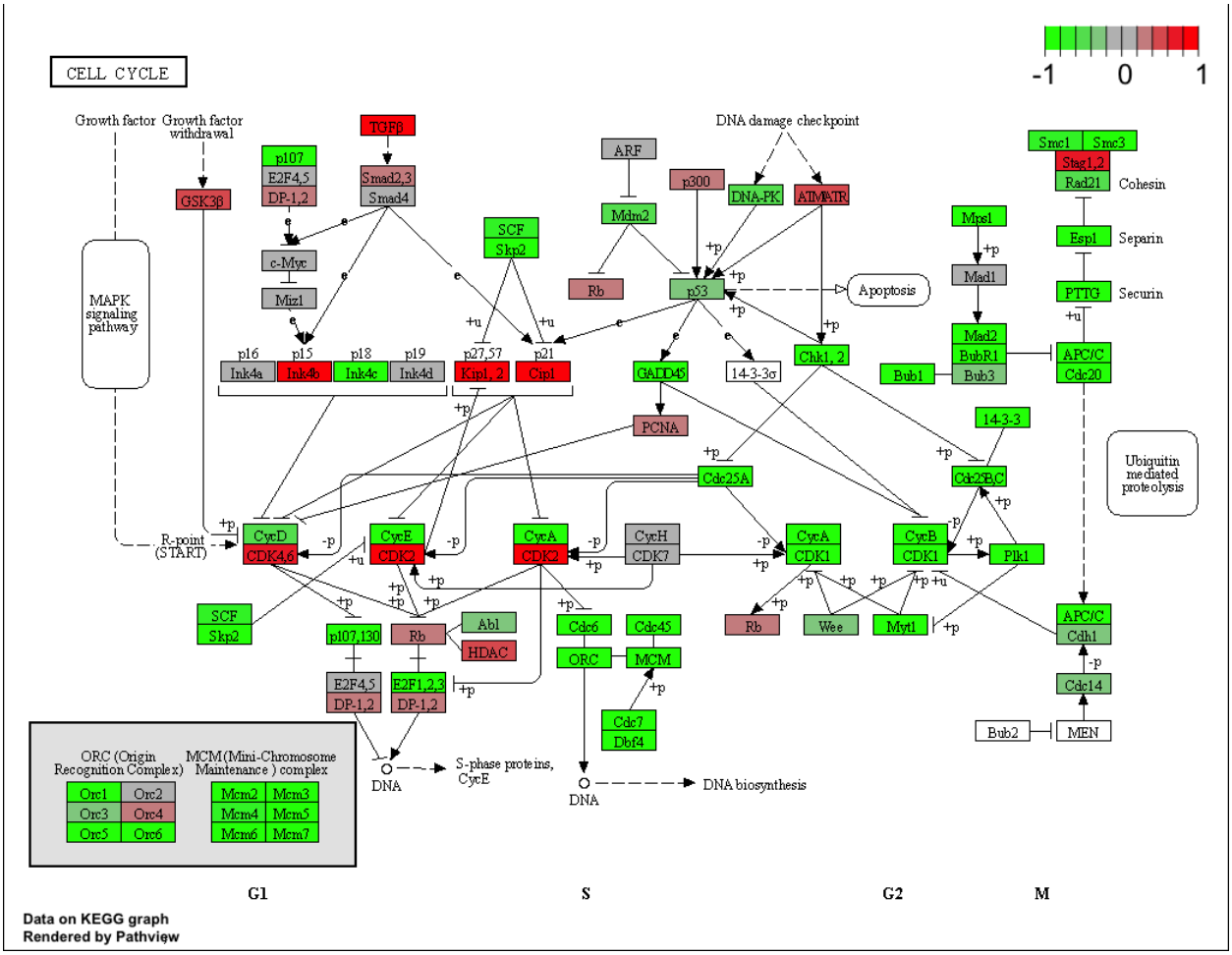
```
## Info: Writing image file hsa03030.pathview.png

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

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

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

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

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

## Info: Working in directory /Users/adelehong/BIMM 143/bimm143_github/Class16

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




NUCLEOCYTOPLASMIC TRANSPORT
-1
0
1
Import
Importin
NLS
NPC
Cytoplasmic fibrils
Cytoplasm
Cytoplasmic ring
Central channel
Spoke complex
Lumenal ring
Lumen
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
Adaptor proteins
IPOA
IPOB
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


DNA REPLICATION
-1
0
1
Replication complex (Bacteria)
Lagging strand
Leading strand
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
SSB
Helicase
DNA polymerase III holoenzyme
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
RNaseHII
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
Pol B
RPA/SSB
Helicase
DNA polymerase B
DNA polymerase D
PolB
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
RfcS
RfcL
RNaseH
RNaseHII
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
Helicase
RPA
DNA polymerase ε
complex
ε
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
RNaseHIIA
RNaseHIIB
RNaseHIIC
Helicase
Dna2
Fen1
DNA ligase
Lig1
Data on KEGG graph
Rendered by Pathview


HOMOLOGOUS RECOMBINATION
Prokaryotic type
5' to 3' resection
(Escherichia coli)
Double strand break
(Escherichia coli)
RecJ
SSB
RecJ
SSB
Filament formation
RecF
RecO
RecR
RecA
RecFOR
RecA filament
Strand invasion
DNA synthesis
DpoI
Holliday junction intermediate
Branch migration and
resolution of Holliday junction
RuvA
RuvB
RuvC
or
RecG
RecFOR pathway
RecB
RecC
RecD
RecBCD
Filament formation
RecA
Strand invasion
DpoIII
Branch migration and
resolution of Holliday junction
RuvA
RuvB
RuvC
or
RecG
Replication restart
PriA
PriB
PriC
DnaT
PriA
RecBC pathway
Eukaryotic type
Double strand break
(Saccharomyces cerevisiae)
(Mammals)
5' to 3' resection
Recognition
-1
0
1
Recognition
ATM
+p
MRX complex
MRN complex
Rad50
Mre11
XRS2
Rad50
Mre11
Nbs1
CtIP
BARD1
BRCA1
BRIP1
TOPBP1
RPA
BRCA2-DSS1
Rad51
Filament formation
NBA1
BRE
BRCC36
Abraxas
RAP80
(S. cerevisiae)
(Common)
(Mammals)
Rad55
Rad57
RPA
Rad51
Rad52
PALB2
BRCA2
DSS1
SYCP3
Rad51 paralogs
Rad51 filament
Rad51B
Rad51C
Rad51D
XRCC2
XRCC3
Rad51B
Rad51C
Rad51D
XRCC2
Rad51 paralogs
Strand invasion
(S. cerevisiae)
(Common)
Rad59
Rad54
D-loop
Rad54
DNA synthesis
Second end capture
Holliday junction intermediate
Branch migration and
resolution of Holliday junction
(Common)
pol δ
Strand displacement
Strand displacement
(Common)
BLM
Mus81
TOP3
Eme1
Crossover
or
Non-crossover
Flap removal and annealing
(Common)
Ligation
BLM
DSBR
Double-strand break repair
SDSA
Synthesis-dependent strand annealing
BIR
Break-induced replication
Data on KEGG graph
Rendered by Pathview



Gene Ontology

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

# Focus on Biological Process subset of GO
gobpsets = go.sets.hs[go.subs.hs$BP]

gobpres = gage(foldchanges, gsets=gobpsets, same.dir=TRUE)

lapply(gobpres, head)
```

```
## $greater
##                                             p.geomean stat.mean        p.val
## GO:0007156 homophilic cell adhesion       1.624062e-05  4.226117 1.624062e-05
## GO:0048729 tissue morphogenesis           5.407952e-05  3.888470 5.407952e-05
## GO:0002009 morphogenesis of an epithelium 5.727599e-05  3.878706 5.727599e-05
## GO:0030855 epithelial cell differentiation 2.053700e-04  3.554776 2.053700e-04
## GO:0060562 epithelial tube morphogenesis  2.927804e-04  3.458463 2.927804e-04
## GO:0048598 embryonic morphogenesis        2.959270e-04  3.446527 2.959270e-04
```

```
##                                              q.val set.size         exp1
## GO:0007156 homophilic cell adhesion        0.07103646      138 1.624062e-05
## GO:0048729 tissue morphogenesis            0.08350839      483 5.407952e-05
## GO:0002009 morphogenesis of an epithelium  0.08350839      382 5.727599e-05
## GO:0030855 epithelial cell differentiation 0.15370245      299 2.053700e-04
## GO:0060562 epithelial tube morphogenesis   0.15370245      289 2.927804e-04
## GO:0048598 embryonic morphogenesis         0.15370245      498 2.959270e-04
##
## $less
##                                               p.geomean stat.mean        p.val
## GO:0048285 organelle fission              6.386337e-16 -8.175381 6.386337e-16
## GO:0000280 nuclear division               1.726380e-15 -8.056666 1.726380e-15
## GO:0007067 mitosis                        1.726380e-15 -8.056666 1.726380e-15
## GO:0000087 M phase of mitotic cell cycle 4.593581e-15 -7.919909 4.593581e-15
## GO:0007059 chromosome segregation         9.576332e-12 -6.994852 9.576332e-12
## GO:0051301 cell division                  8.718528e-11 -6.455491 8.718528e-11
##                                              q.val set.size         exp1
## GO:0048285 organelle fission              2.517062e-12      386 6.386337e-16
## GO:0000280 nuclear division               2.517062e-12      362 1.726380e-15
## GO:0007067 mitosis                        2.517062e-12      362 1.726380e-15
## GO:0000087 M phase of mitotic cell cycle 5.023080e-12      373 4.593581e-15
## GO:0007059 chromosome segregation         8.377375e-09      146 9.576332e-12
## GO:0051301 cell division                  6.355807e-08      479 8.718528e-11
##
## $stats
##                                            stat.mean     exp1
## GO:0007156 homophilic cell adhesion         4.226117 4.226117
## GO:0048729 tissue morphogenesis             3.888470 3.888470
## GO:0002009 morphogenesis of an epithelium   3.878706 3.878706
## GO:0030855 epithelial cell differentiation  3.554776 3.554776
## GO:0060562 epithelial tube morphogenesis    3.458463 3.458463
## GO:0048598 embryonic morphogenesis          3.446527 3.446527
```

Reactome Analysis

We can also do a similar procedure with gene ontology. Similar to above, go.sets.hs has all GO terms. go.subs.hs is a named list containing indexes for the BP, CC, and MF ontologies. Let's focus on BP (a.k.a Biological Process) here.

Reactome Analysis

Reactome is database consisting of biological molecules and their relation to pathways and processes.

```
# output the list of significant genes at the 0.05 level as a plain text file:
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: 8146"
```

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

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?

```
# Performed significant gene analysis on https://reactome.org/PathwayBrowser/#TOOL=AT website. Download
mostsignificant <- read.csv(file="result.csv")
head(mostsignificant)
```

```
##   Pathway.identifier
## 1      R-HSA-9716542
## 2      R-HSA-9012999
## 3        R-HSA-69618
## 4       R-HSA-141424
## 5       R-HSA-141444
## 6       R-HSA-194315
##                                                                        Pathway.name
## 1                                  Signaling by Rho GTPases, Miro GTPases and RHOBTB3
## 2                                                                      RHO GTPase cycle
## 3                                                           Mitotic Spindle Checkpoint
## 4                                        Amplification of signal from the kinetochores
## 5 Amplification  of signal from unattached  kinetochores via a MAD2  inhibitory signal
## 6                                                            Signaling by Rho GTPases
##   X.Entities.found X.Entities.total Entities.ratio Entities.pValue Entities.FDR
## 1              495              725    0.050802326     0.006251003    0.7958572
## 2              323              460    0.032233200     0.006517206    0.7958572
## 3               89              111    0.007778011     0.007440536    0.7958572
## 4               77               94    0.006586784     0.007543102    0.7958572
## 5               77               94    0.006586784     0.007543102    0.7958572
## 6              483              709    0.049681172     0.007827996    0.7958572
##   X.Reactions.found X.Reactions.total Reactions.ratio Species.identifier
## 1               204               212    0.0156146424               9606
## 2                84                91    0.0067025116               9606
## 3                 7                 7    0.0005155778               9606
## 4                 4                 4    0.0002946159               9606
## 5                 4                 4    0.0002946159               9606
## 6               195               203    0.0149517566               9606
##   Species.name
## 1 Homo sapiens
## 2 Homo sapiens
## 3 Homo sapiens
## 4 Homo sapiens
## 5 Homo sapiens
## 6 Homo sapiens
##
## 1 908;25904;65124;2551;1460;9181;9184;25900;2669;79658;253959;5930;91526;3998;23616;81839;115703;201
## 2
## 3
## 4
## 5
## 6                                                                                                 908;25904;65124;2551;1460;9181;9
##   Mapped.entities
## 1              NA
## 2              NA
## 3              NA
## 4              NA
```

```
## 5               NA
## 6               NA
##
## 1 R-HSA-3858489;R-HSA-3858495;R-HSA-5668947;R-HSA-8981637;R-HSA-9014424;R-HSA-5668932;R-HSA-9017488;
## 2
## 3
## 4
## 5
## 6
```