RNA-Seq Analysis Mini-Project

Nhi To (A18053310)

2025-02-20

Table of contents

Background

The data for this 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.

Section 1. Differential Expression Analysis

Data Import

```
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, aperm, 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, saveRDS, setdiff,
    table, tapply, union, unique, unsplit, which.max, which.min
```

```

Attaching package: 'S4Vectors'
```

```
The following object is masked from 'package:utils':

    findMatches
```

```
The following objects are masked from 'package:base':

    expand.grid, I, unname

Loading required package: IRanges

Loading required package: GenomicRanges

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
```

```
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? > No, it does exactly match because there is an extra column, hence we need to fix that

```
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
```

```
head(colData)
```

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

```
colData$id
```

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

```
colnames(counts)
```

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

Q1. Complete the code below to remove the troublesome first column from countData

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

ANSWER: 19808 genes

```
nrow(countData)
```

```
[1] 19808
```

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).How many genes are left?

ANSWER: 15975 are the amount of genes left after the filtered countData to exclude genes where we have 0 read count across all samples.

Tip: What will rowSums() of countData return and how could you use it in this context?

Answer: rowSums(countData) shows a table of the values for each row and column. We should use this to be able to filter to filter out the values that are zero, and keep the values that are above zero.

```
head(rowSums(countData))
```

```
ENSG00000186092 ENSG00000279928 ENSG00000279457 ENSG00000278566 ENSG00000273547
              0               0             183               0               0
ENSG00000187634
           1129
```

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

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

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

```
[1] 15975
```

Setup for DESeq

```
#/ message: false
library(DESeq2)
```

Setup input object for DEseq

```
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

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

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

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

```
head(dds)
```

```
class: DESeqDataSet
dim: 6 6
metadata(1): version
assays(4): counts mu H cooks
rownames(6): ENSG00000279457 ENSG00000187634 ... ENSG00000187583
  ENSG00000187642
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(3): id condition sizeFactor
```

```
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
```

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.

ANSWER: 4349 (27%) genes are up regulated, while 4396 (28%) are down-regulated at the default 0.1 p-value cutoff.

```
summary(res)
```

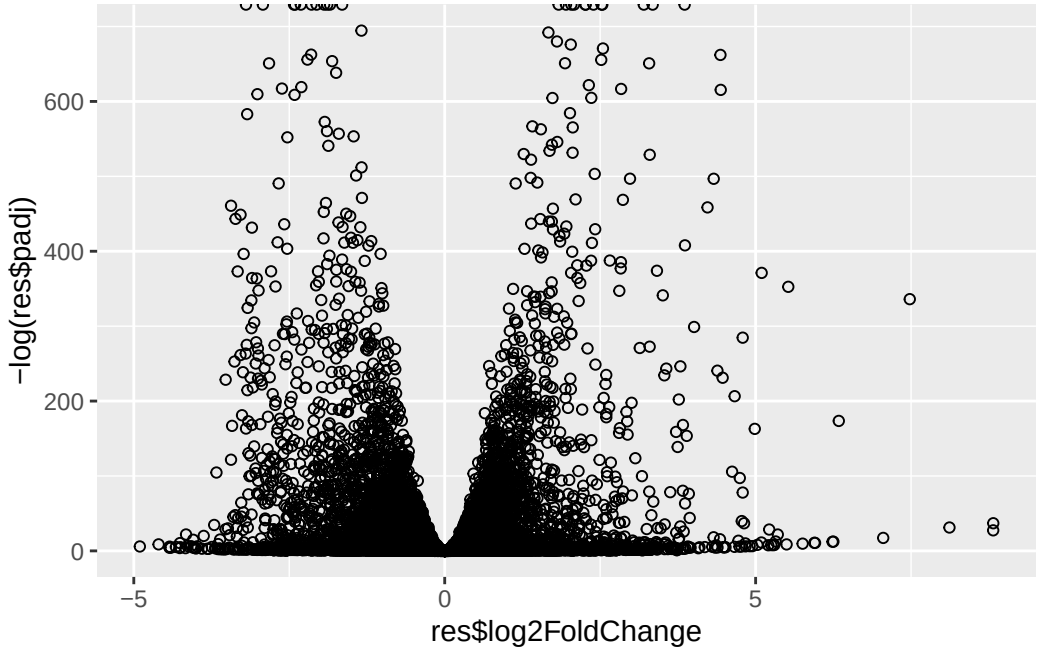
```
out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4349, 27%
LFC < 0 (down)     : 4396, 28%
outliers [1]       : 0, 0%
low counts [2]     : 1237, 7.7%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

Volcano plot of results

```
library(ggplot2)
```

```
ggplot(res) +
  aes(res$log2FoldChange, -log(res$padj)) +
  geom_point(shape=1)
```

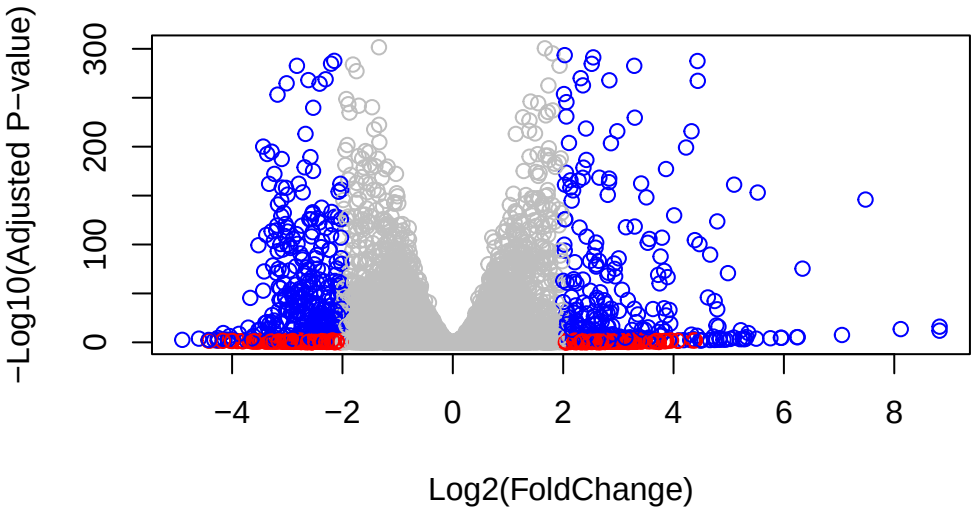
```
Warning: Removed 1237 rows containing missing values or values outside the scale range
(`geom_point()`).
```



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

ANSWER:

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

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

# Color blue those with adjusted p-value less than 0.01
# and absolute fold change more than 2
inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2)
mycols[ inds ] <- "blue"

# Create the volcano plot with colors and axis labels
plot( res$log2FoldChange, -log10(res$padj), col=mycols,
      xlab="Log2(FoldChange)", ylab="-Log10(Adjusted P-value)" )
```



Gene annotation

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

ANSWER:

```
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"
```

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

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

```
res$entrez = mapIds(org.Hs.eg.db,
                    keys=rownames(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 ..
```

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

ANSWER:

```
res <- res[order(res$pvalue),]
write.csv(res, file="deseq_results.csv", row.names=TRUE)
```

Section 2. Pathway Analysis

```
library(gage)
```

```
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()'

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

```
data(kegg.sets.hs)
data(sigmet.idx.hs)
```

```
# Focus on signaling and metabolic pathways only
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]
```

Runpathway analysis in Keggres

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

```
head(keggres$less, 3)
```

```
                          p.geomean stat.mean        p.val       q.val
hsa04110 Cell cycle      8.995727e-06 -4.378644 8.995727e-06 0.001448312
hsa03030 DNA replication 9.424076e-05 -3.951803 9.424076e-05 0.007586381
hsa03013 RNA transport   1.375901e-03 -3.028500 1.375901e-03 0.073840037
                         set.size         exp1
hsa04110 Cell cycle           121 8.995727e-06
hsa03030 DNA replication       36 9.424076e-05
hsa03013 RNA transport        144 1.375901e-03
```

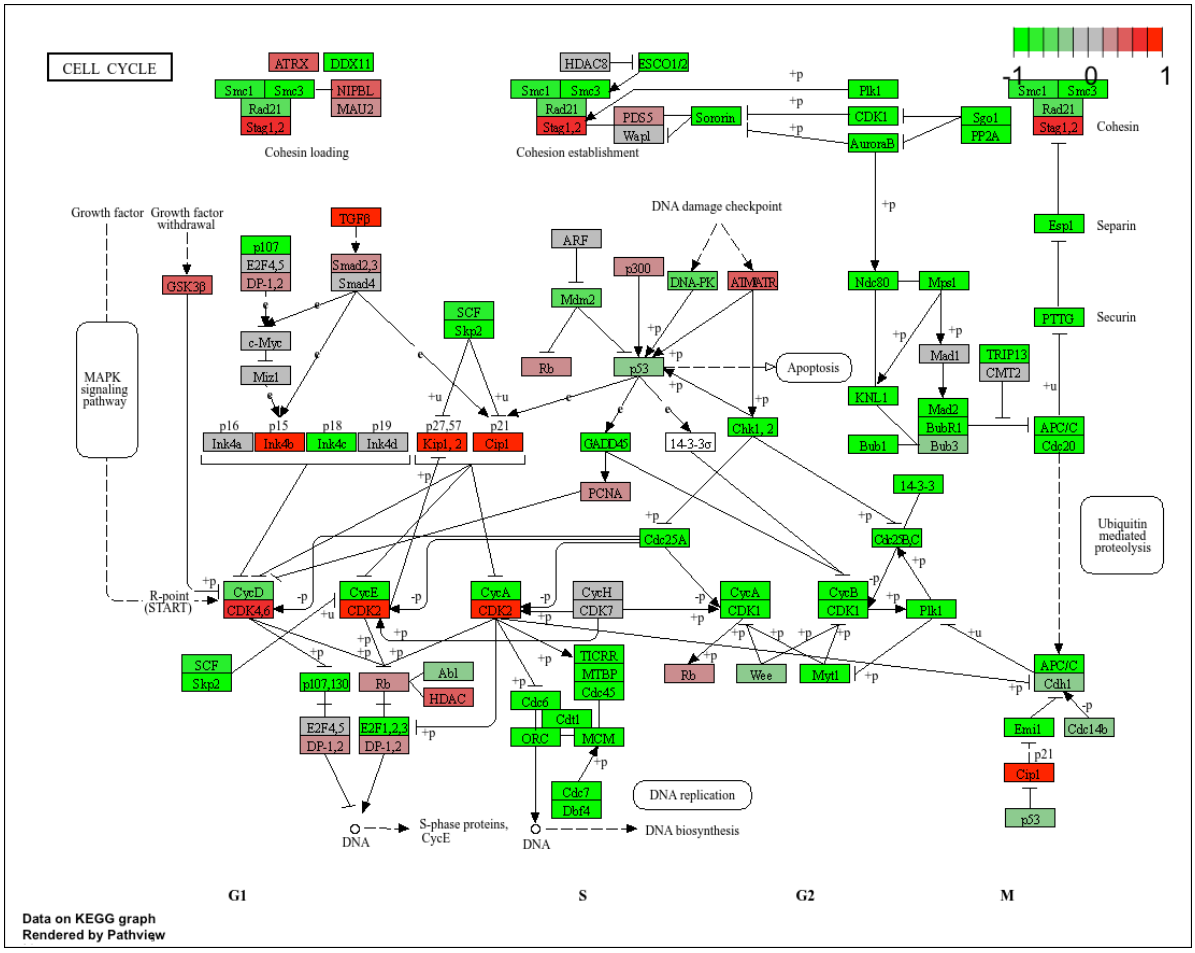
Cell cycle figure

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

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

```
Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM
```

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



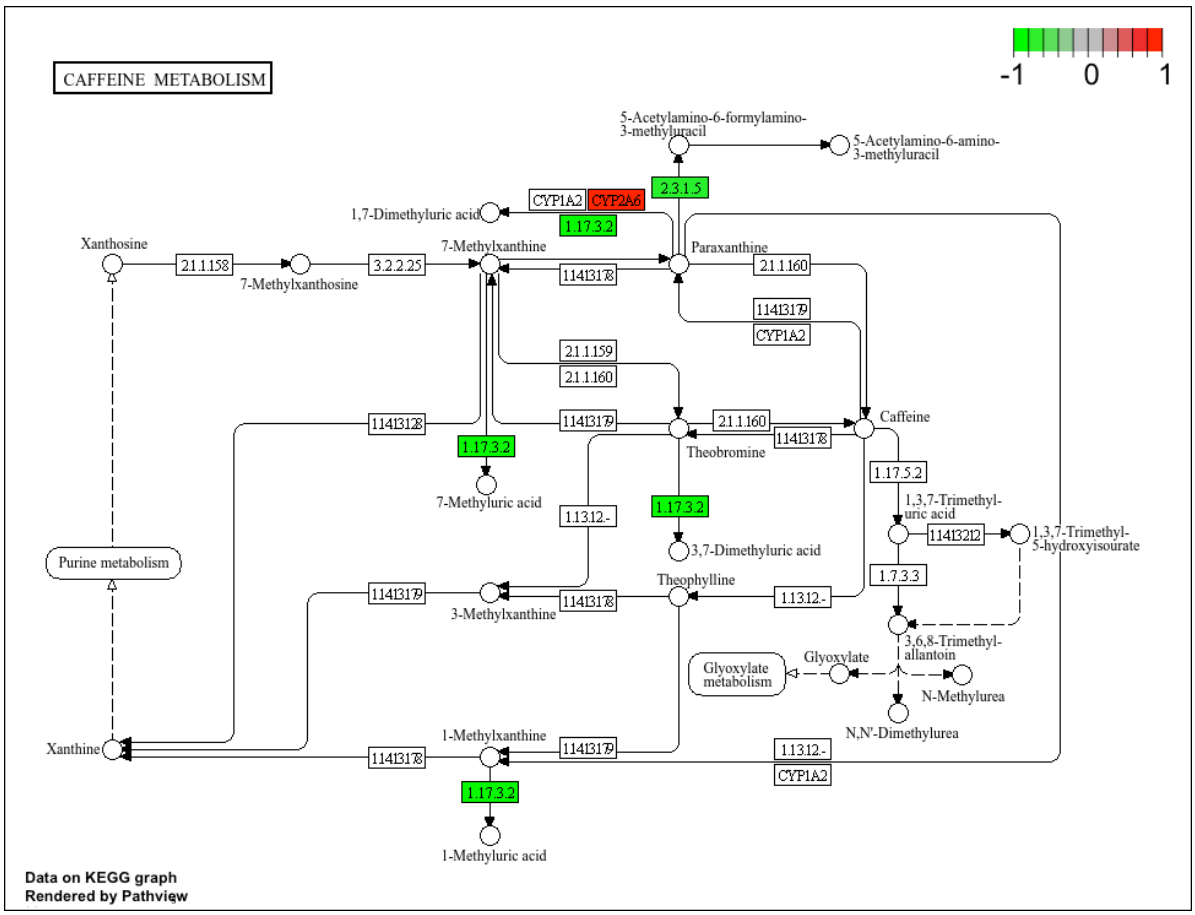
Caffeine Metabolism figure

```
pathview (foldchanges, pathway.id= "hsa00232")
```

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

```
Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM
```

```
Info: Writing image file hsa00232.pathview.png
```



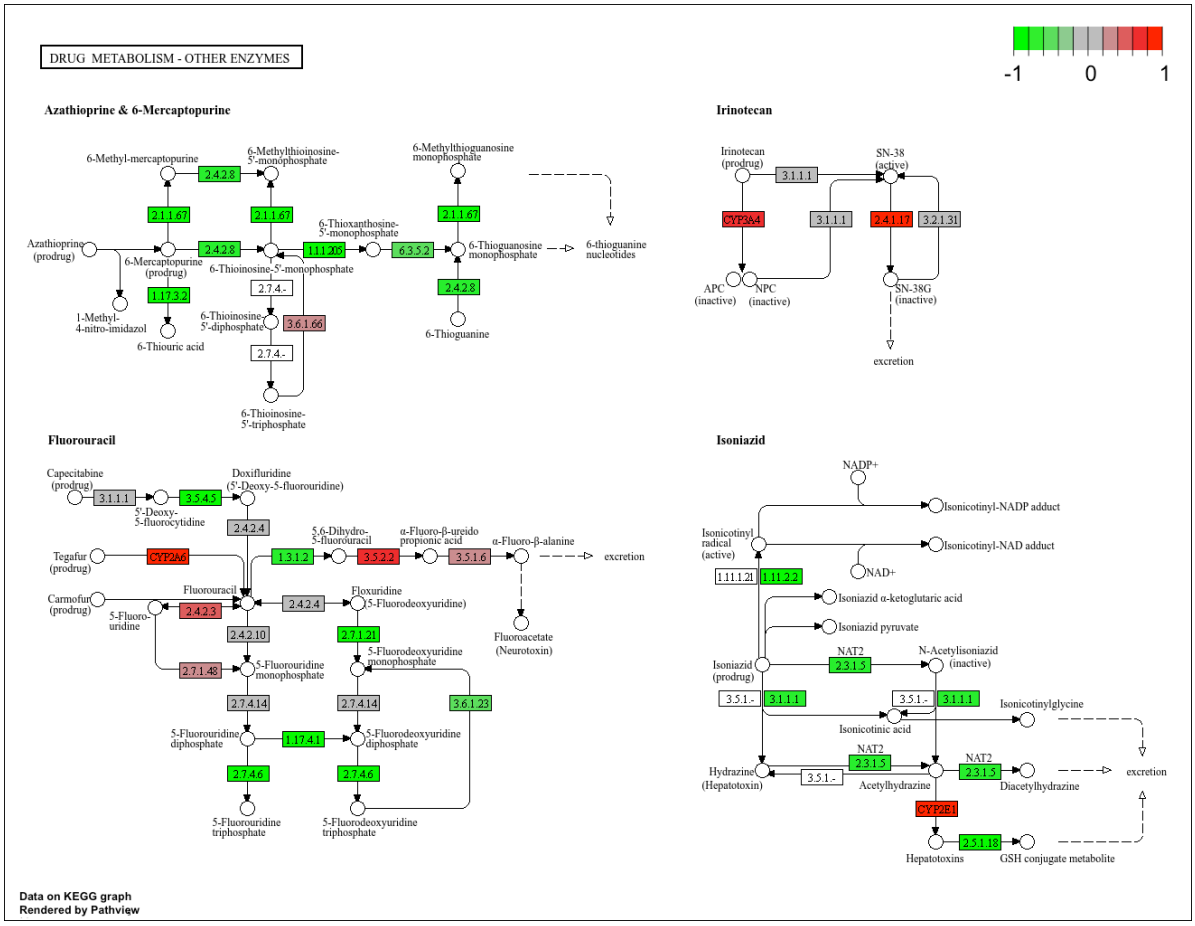
Drug metabolism - other enzymes Figure

```
pathview (foldchanges, pathway.id= "hsa00983")
```

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

```
Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM
```

```
Info: Writing image file hsa00983.pathview.png
```




```
names(foldchanges)= res$entrez
# Get the results
keggres = gage(foldchanges, gsets=kegg.sets.hs)
attributes(keggres)
```

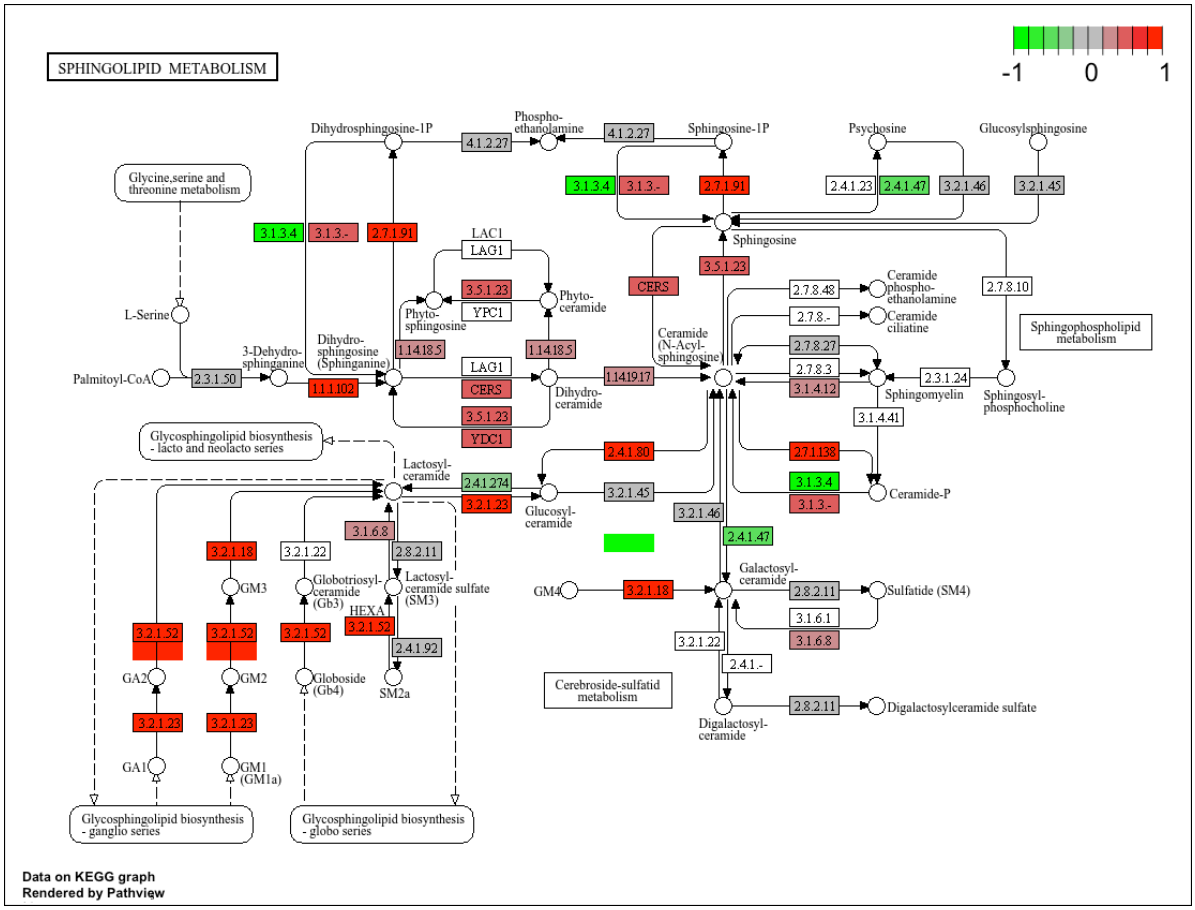
```
$names
[1] "greater" "less"    "stats"
```

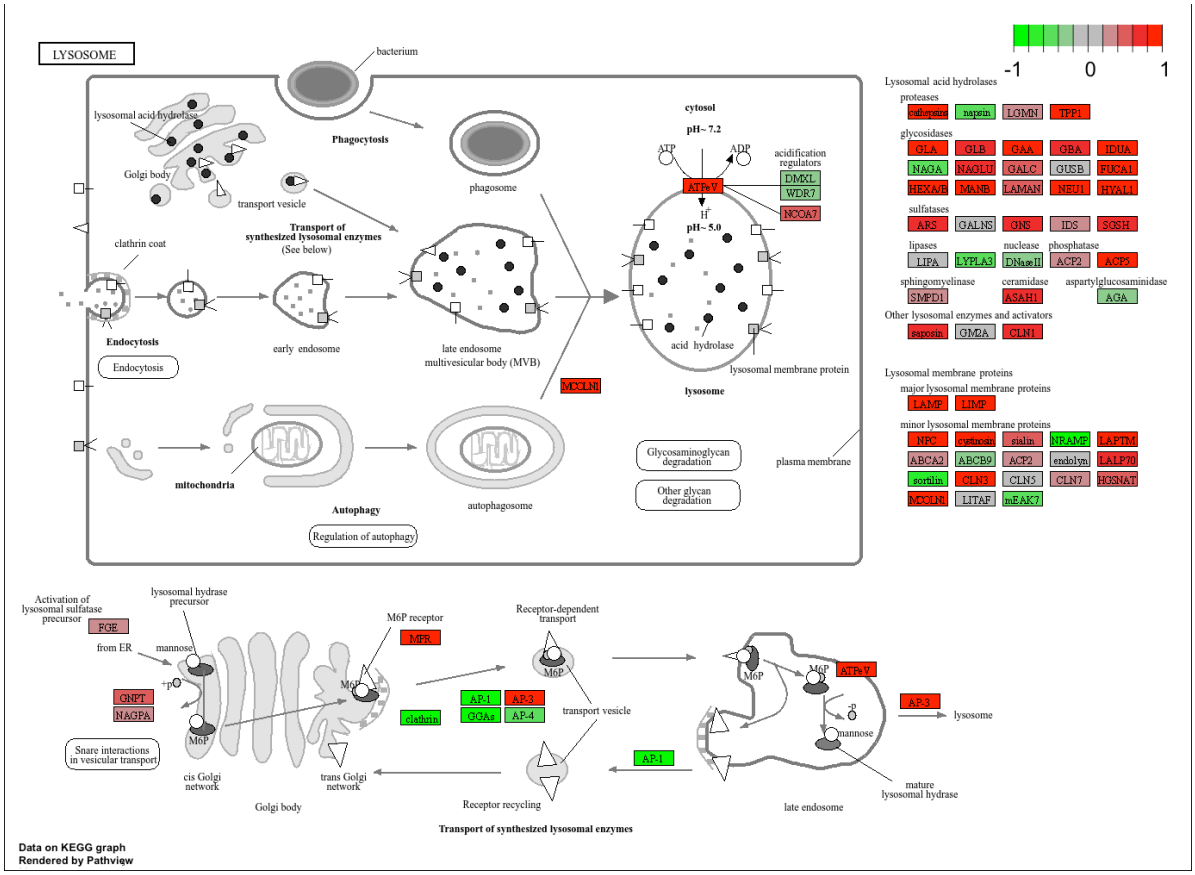
```
## Focus on top 5 upregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$greater)[1:5]

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

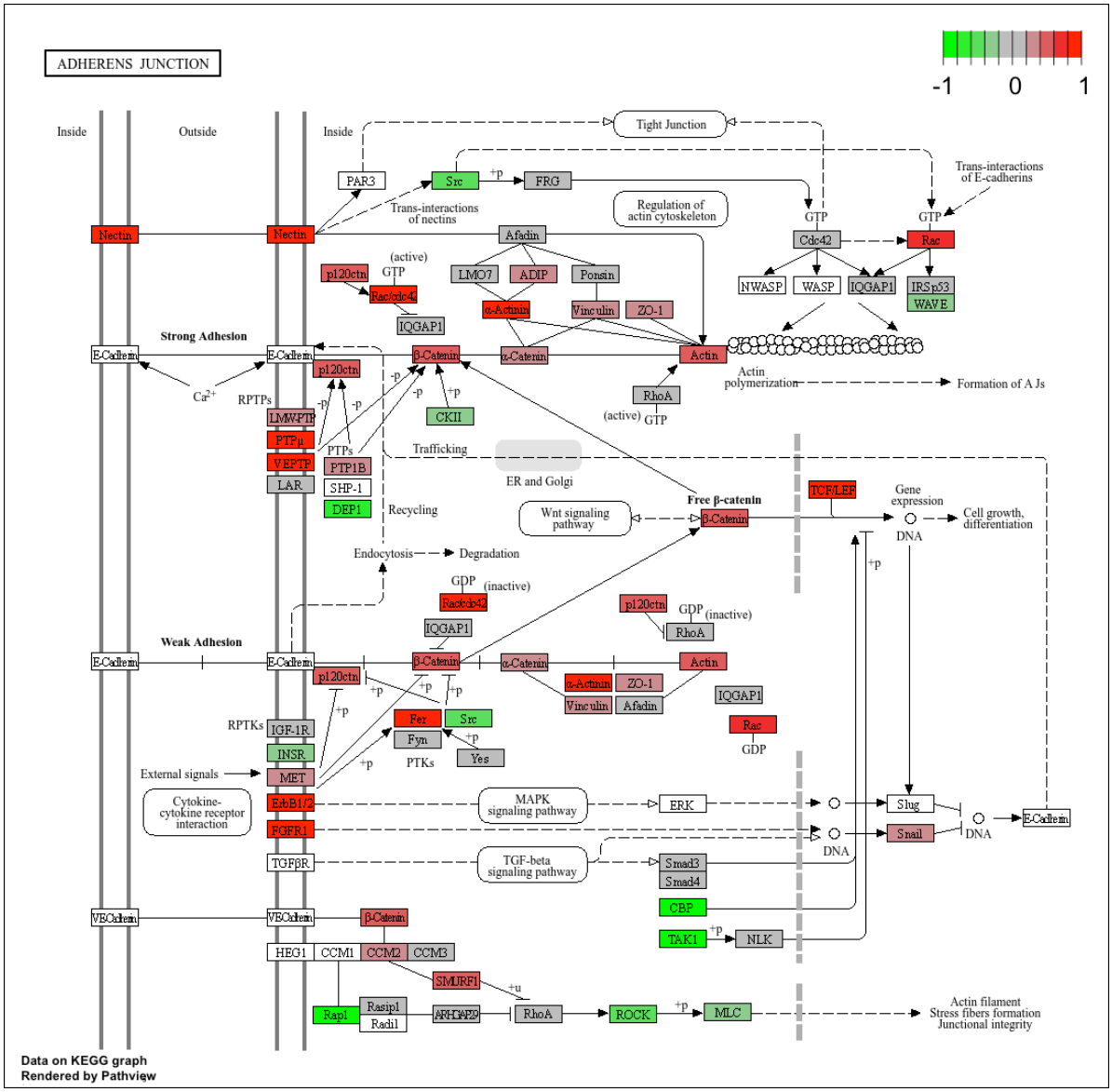
```
[1] "hsa04640" "hsa04630" "hsa00140" "hsa04142" "hsa04330"
```

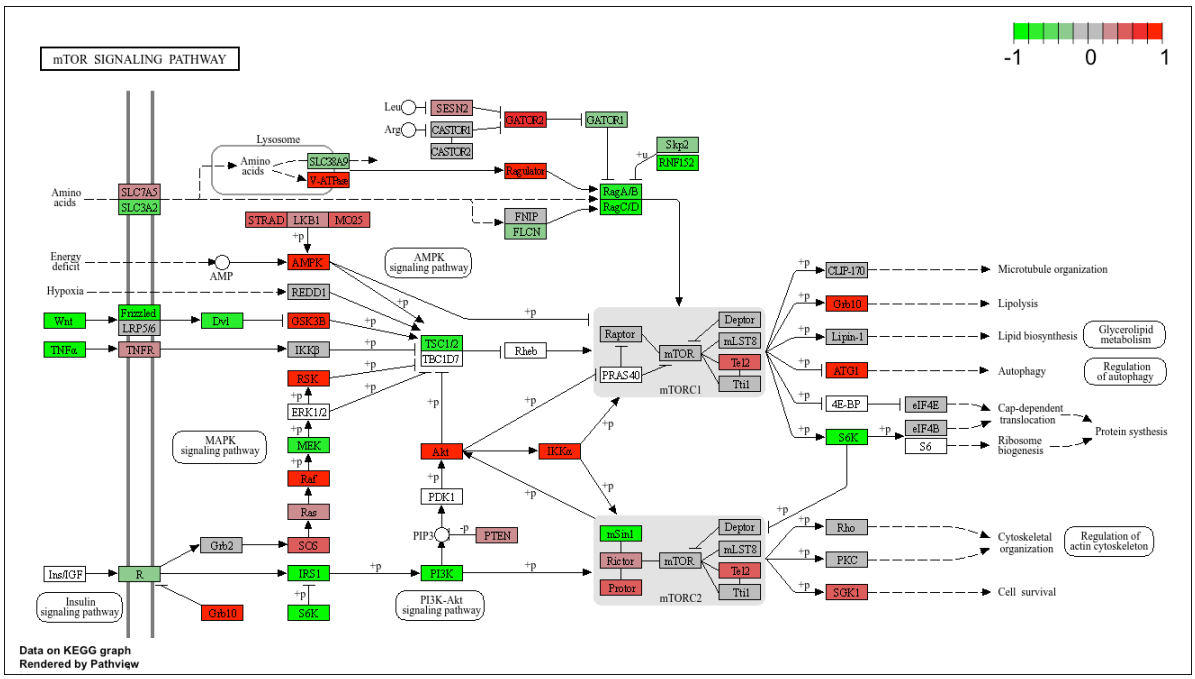
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

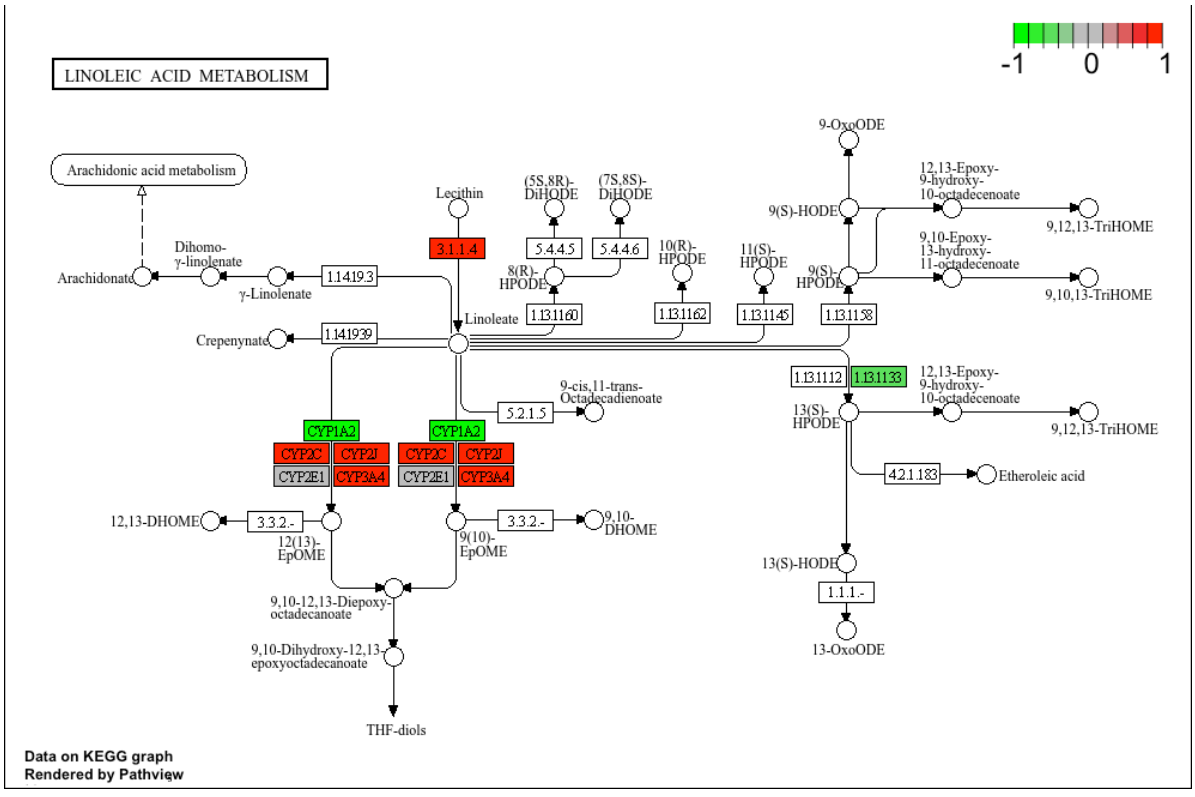
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa04142.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

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


SPHINGOLIPID METABOLISM
-1
0
1
Dihydrosphingosine-1P
Phospho-ethanolamine
Sphingosine-1P
Psychosine
Glucosylsphingosine
Glycine,serine and threonine metabolism
Sphingosine
L-Serine
Phyto-sphingosine
Phyto-ceramide
Ceramide (N-Acyl-sphingosine)
Ceramide phospho-ethanolamine
Ceramide ciliatine
Sphingophospholipid metabolism
3-Dehydro-sphinganine
Dihydro-sphingosine (Sphinganine)
Palmitoyl-CoA
Dihydro-ceramide
Sphingomyelin
Sphingosyl-phosphocholine
Glycosphingolipid biosynthesis - lacto and neolacto series
Lactosyl-ceramide
Glucosyl-ceramide
Ceramide-P
Galactosyl-ceramide
Sulfatide (SM4)
GM4
GM3
Globotriosyl-ceramide (Gb3)
Lactosyl-ceramide sulfate (SM3)
HEXA
GA2
GM2
Globoside (Gb4)
SM2a
Cerebroside-sulfatid metabolism
Digalactosyl-ceramide
Digalactosylceramide sulfate
GA1
GM1 (GM1a)
Glycosphingolipid biosynthesis - ganglio series
Glycosphingolipid biosynthesis - globo series
Data on KEGG graph
Rendered by Pathview


LYSOSOME
bacterium
lysosomal acid hydrolase
Phagocytosis
Golgi body
transport vesicle
phagosome
cytosol
pH~ 7.2
ATP
ADP
acidification regulators
DMXL
WDR7
ATPeV
NCOA7
Transport of synthesized lysosomal enzymes
(See below)
clathrin coat
Endocytosis
Endocytosis
early endosome
late endosome
multivesicular body (MVB)
pH~ 5.0
acid hydrolase
lysosomal membrane protein
lysosome
MCOLN1
mitochondria
Autophagy
autophagosome
Regulation of autophagy
Glycosaminoglycan degradation
Other glycan degradation
plasma membrane
-1
0
1
Lysosomal acid hydrolases
proteases
cathepsins
saposin
LGMN
TPP1
glycosidases
GLA
GLB
GAA
GBA
IDUA
NAGA
NAGLU
GALC
GUSB
FUCA1
HEXA/B
MANB
LAMAN
NEU1
HYAL1
sulfatases
ARS
GALNS
GNS
IDS
SGSH
lipases
nuclease
phosphatase
LIPA
LYPLA3
DNaseII
ACP2
ACP5
sphingomyelinase
ceramidase
aspartylglucosaminidase
SMPD1
ASAH1
AGA
Other lysosomal enzymes and activators
saposin
GM2A
CLN1
Lysosomal membrane proteins
major lysosomal membrane proteins
LAMP
LIMP
minor lysosomal membrane proteins
NPC
cystinosin
sialin
NRAMP
LAPTM
ABCA2
ABCB9
ACP2
endolyn
LAL P70
sortilin
CLN3
CLN5
CLN7
HGSNAT
MCOLN1
LITAF
mEAK7
Activation of lysosomal sulfatase precursor
FGE
lysosomal hydrase precursor
from ER
mannose
GNPT
NAGPA
M6P
Snare interactions in vesicular transport
cis Golgi network
trans Golgi network
Golgi body
M6P receptor
MPR
clathrin
AP-1
GGAs
AP-3
AP-4
Receptor-dependent transport
transport vesicle
Receptor recycling
AP-1
late endosome
ATPeV
AP-3
lysosome
mannose
mature lysosomal hydrase
Transport of synthesized lysosomal enzymes
Data on KEGG graph
Rendered by Pathview


ADHERENS JUNCTION
-1
0
1
Inside
Outside
Inside
Tight Junction
PAR3
Src
+p
FRG
Trans-interactions of E-cadherins
Trans-interactions of nectins
Regulation of actin cytoskeleton
GTP
GTP
Nectin
Nectin
Afadin
Cdc42
Rac
(active)
GTP
p120ctn
Rac/cdc42
LMO7
ADIP
Ponsin
NWASP
WASP
IQGAP1
IRSp53
WAVE
α-Actinin
Vinculin
ZO-1
IQGAP1
Strong Adhesion
E-Cadherin
E-Cadherin
β-Catenin
α-Catenin
Actin
p120ctn
Ca2+
RPTPs
-p
-p
-p
-p
+p
CKII
RhoA
(active)
GTP
Actin polymerization
Formation of A Js
LMW-PTP
PTPμ
VEPTP
LAR
PTPs
PTP1B
SHP-1
DEP1
Trafficking
ER and Golgi
Recycling
Wnt signaling pathway
Free β-catenin
β-Catenin
TCF/LEF
Gene expression
Cell growth, differentiation
DNA
Endocytosis
Degradation
+p
GDP
(inactive)
Rac/cdc42
p120ctn
GDP
(inactive)
RhoA
IQGAP1
Weak Adhesion
E-Cadherin
E-Cadherin
β-Catenin
α-Catenin
Actin
p120ctn
+p
+p
+p
+p
α-Actinin
ZO-1
Vinculin
Afadin
IQGAP1
Rac
GDP
RPTKs
IGF-1R
INSR
Fer
Src
Fyn
Yes
+p
PTKs
External signals
MET
Cytokine-cytokine receptor interaction
ErbB1/2
FGFR1
TGFβR
MAPK signaling pathway
ERK
Slug
Snail
DNA
DNA
E-Cadherin
TGF-beta signaling pathway
Smad3
Smad4
CBP
TAK1
+p
NLK
VECadherin
VECadherin
β-Catenin
HEG1
CCM1
CCM2
CCM3
SMURF1
+u
Rap1
Rasip1
Radil
ARHGAP29
RhoA
ROCK
+p
MLC
Actin filament
Stress fibers formation
Junctional integrity
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?

ANSWER:

```
keggrespathways_down <- rownames(keggres$less)[1:5]
```

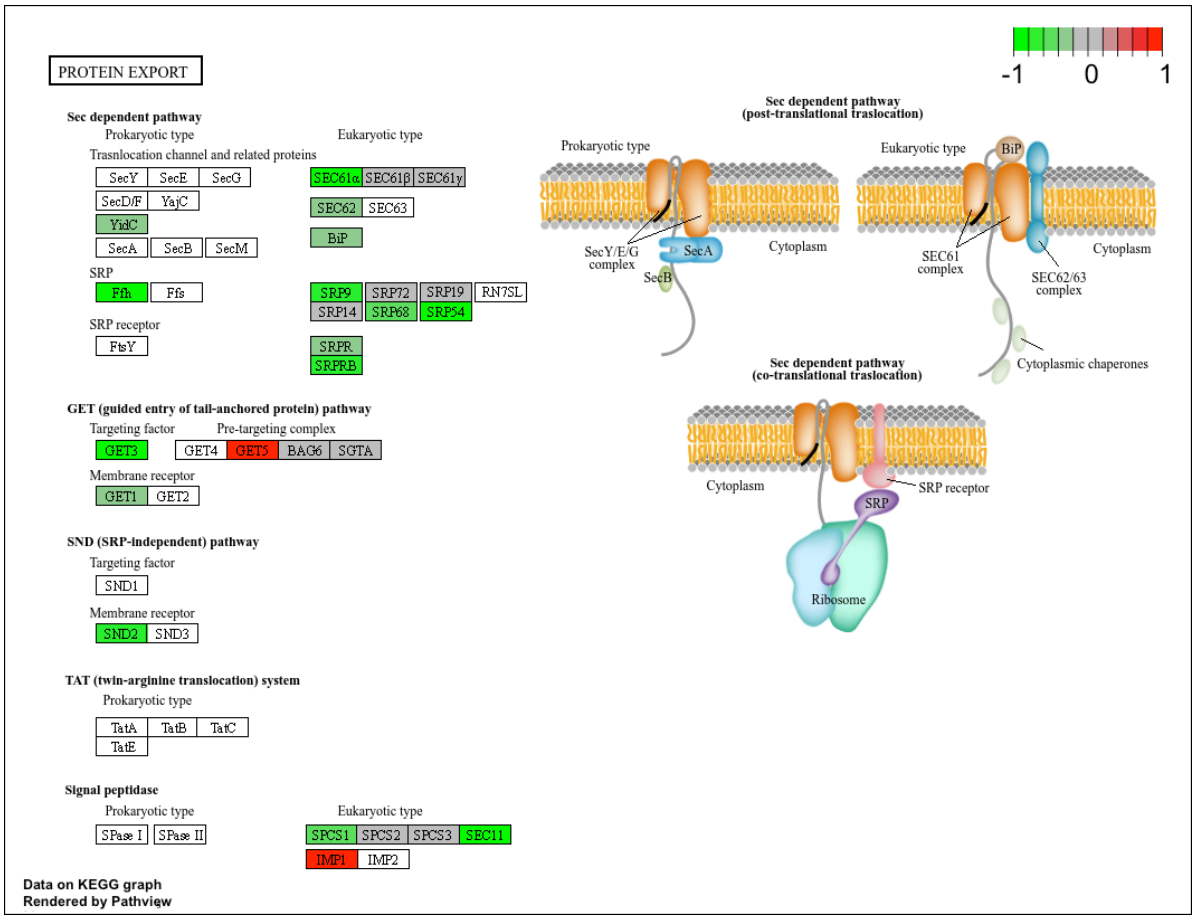
```
keggresids_down <- substr(keggrespathways_down, start=1, stop=8)
keggresids_down
```

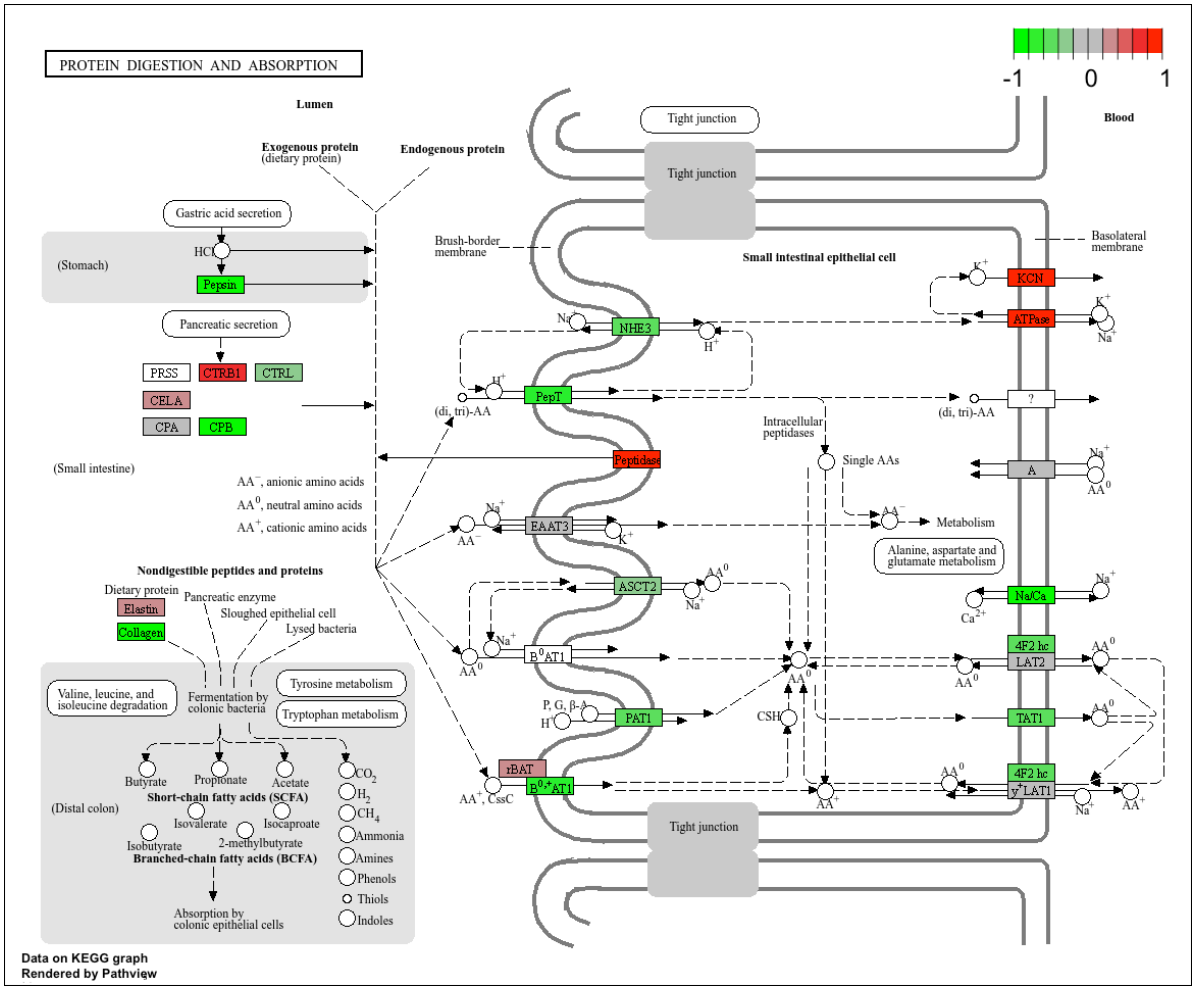
```
[1] "hsa04110" "hsa03030" "hsa03013" "hsa03440" "hsa04114"
```

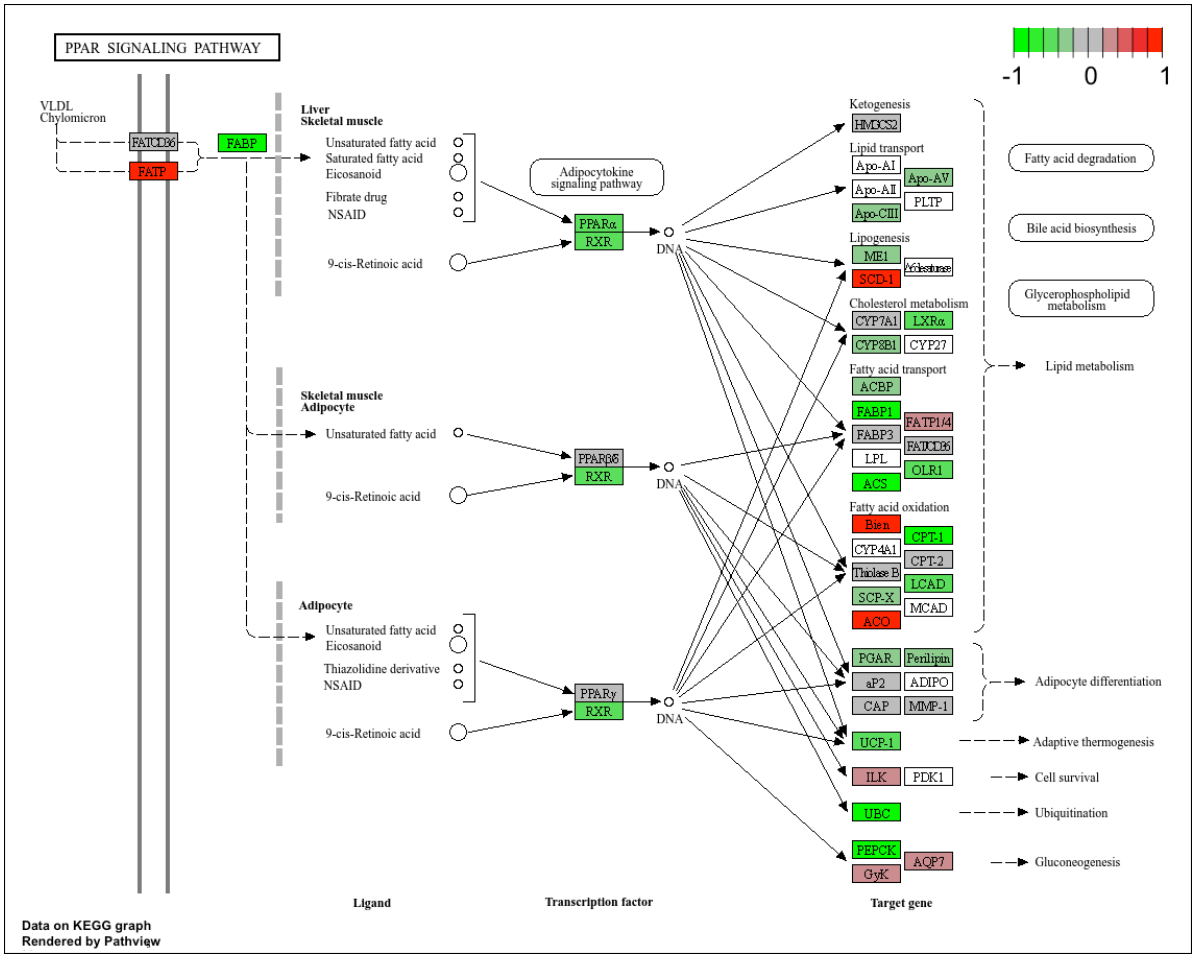
```
pathview(gene.data=foldchanges, pathway.id=keggresids_down, species="hsa")
```

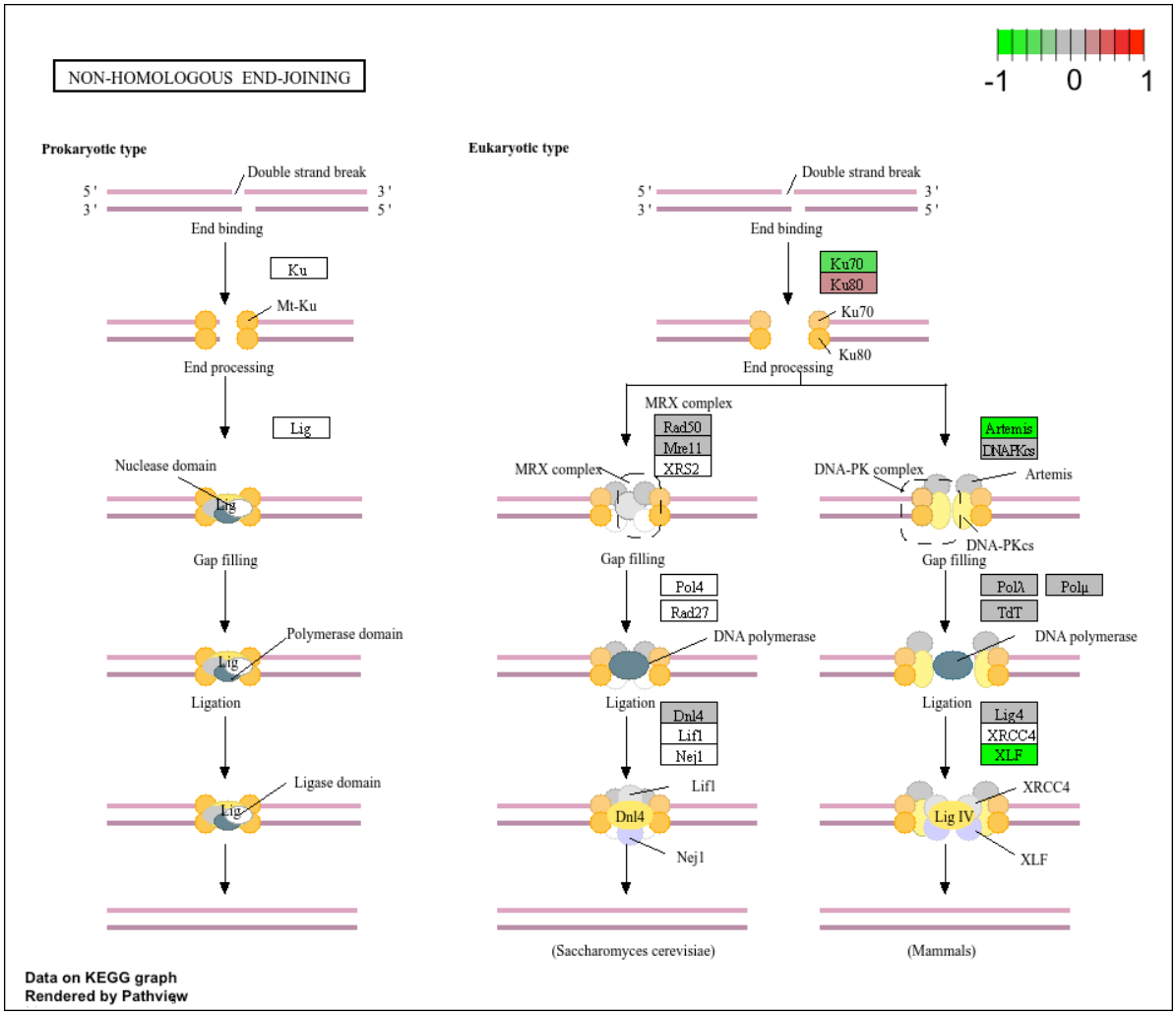
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa04110.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

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

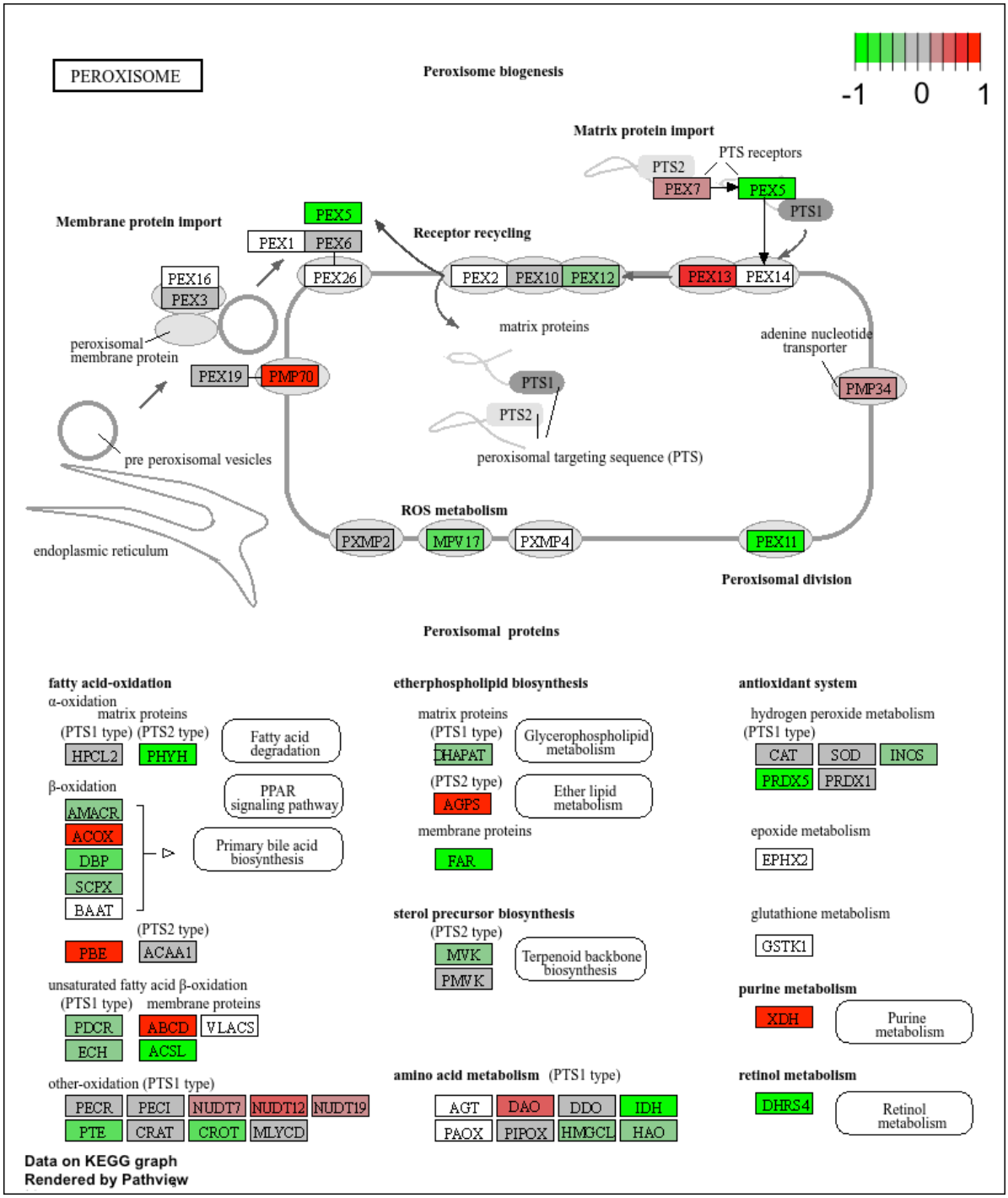
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/nhito/Desktop/Desktop - Nhi's MacBook Air/School/UCSD/BIMM

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




PROTEIN DIGESTION AND ABSORPTION
-1
0
1
Lumen
Blood
Exogenous protein
(dietary protein)
Endogenous protein
Tight junction
Tight junction
Gastric acid secretion
(Stomach)
HCl
Pepsin
Brush-border membrane
Small intestinal epithelial cell
Basolateral membrane
KCN
ATPase
NHE3
Pancreatic secretion
PRSS
CTRB1
CTRL
CELA
CPA
CPB
PepT
(di, tri)-AA
?
(di, tri)-AA
Intracellular peptidases
Peptidase
Single AAs
A
(Small intestine)
AA−, anionic amino acids
AA0, neutral amino acids
AA+, cationic amino acids
EAAT3
Metabolism
Alanine, aspartate and glutamate metabolism
Nondigestible peptides and proteins
Dietary protein
Elastin
Collagen
Pancreatic enzyme
Sloughed epithelial cell
Lysed bacteria
ASCT2
Na/Ca
B0AT1
4F2 hc
LAT2
Valine, leucine, and isoleucine degradation
Fermentation by colonic bacteria
Tyrosine metabolism
Tryptophan metabolism
PAT1
TAT1
rBAT
B0,+AT1
y+LAT1
Butyrate
Propionate
Acetate
Short-chain fatty acids (SCFA)
(Distal colon)
Isovalerate
Isocaproate
Isobutyrate
2-methylbutyrate
Branched-chain fatty acids (BCFA)
Absorption by colonic epithelial cells
CO2
H2
CH4
Ammonia
Amines
Phenols
Thiols
Indoles
Tight junction
Data on KEGG graph
Rendered by Pathview


PPAR SIGNALING PATHWAY
-1
0
1
VLDL
Chylomicron
FATCD36
FATP
FABP
Liver
Skeletal muscle
Unsaturated fatty acid
Saturated fatty acid
Eicosanoid
Fibrate drug
NSAID
9-cis-Retinoic acid
Adipocytokine signaling pathway
PPARα
RXR
DNA
Skeletal muscle
Adipocyte
Unsaturated fatty acid
9-cis-Retinoic acid
PPARβδ
RXR
DNA
Adipocyte
Unsaturated fatty acid
Eicosanoid
Thiazolidine derivative
NSAID
9-cis-Retinoic acid
PPARγ
RXR
DNA
Ligand
Transcription factor
Target gene
Ketogenesis
HMGCS2
Lipid transport
Apo-AI
Apo-AV
Apo-AII
PLTP
Apo-CIII
Lipogenesis
ME1
SCD-1
Cholesterol metabolism
CYP7A1
LXRα
CYP8B1
CYP27
Fatty acid transport
ACBP
FABP1
FATP1/4
FABP3
FATCD36
LPL
OLR1
ACS
Fatty acid oxidation
Bien
CPT-1
CYP4A1
CPT-2
Thiolase B
LCAD
SCP-X
MCAD
ACO
PGAR
Perilipin
aP2
ADIPO
CAP
MMP-1
UCP-1
ILK
PDK1
UBC
PEPCK
AQP7
GyK
Fatty acid degradation
Bile acid biosynthesis
Glycerophospholipid metabolism
Lipid metabolism
Adipocyte differentiation
Adaptive thermogenesis
Cell survival
Ubiquitination
Gluconeogenesis
Data on KEGG graph
Rendered by Pathview


NON-HOMOLOGOUS END-JOINING
-1
0
1
Prokaryotic type
Double strand break
5'
3'
3'
5'
End binding
Ku
Mt-Ku
End processing
Lig
Nuclease domain
Lig
Gap filling
Polymerase domain
Lig
Ligation
Ligase domain
Lig
Eukaryotic type
Double strand break
5'
3'
3'
5'
End binding
Ku70
Ku80
Ku70
Ku80
End processing
MRX complex
Rad50
Mre11
XRS2
MRX complex
Gap filling
Pol4
Rad27
DNA polymerase
Ligation
Dnl4
Lif1
Nej1
Lif1
Dnl4
Nej1
(Saccharomyces cerevisiae)
Artemis
DNAPKcs
DNA-PK complex
Artemis
DNA-PKcs
Gap filling
Polλ
Polμ
TdT
DNA polymerase
Ligation
Lig4
XRCC4
XLF
XRCC4
Lig IV
XLF
(Mammals)
Data on KEGG graph
Rendered by Pathview


PEROXISOME
Peroxisome biogenesis
-1
0
1
Matrix protein import
PTS receptors
PTS2
PEX7
PEX5
PTS1
Membrane protein import
PEX5
PEX1
PEX6
Receptor recycling
PEX16
PEX3
PEX26
PEX2
PEX10
PEX12
PEX13
PEX14
matrix proteins
peroxisomal membrane protein
adenine nucleotide transporter
PEX19
PMP70
PTS1
PMP34
PTS2
peroxisomal targeting sequence (PTS)
pre peroxisomal vesicles
ROS metabolism
endoplasmic reticulum
PXMP2
MPV17
PXMP4
PEX11
Peroxisomal division
Peroxisomal proteins
fatty acid-oxidation
α-oxidation
matrix proteins
(PTS1 type) (PTS2 type)
HPCL2
PHYH
Fatty acid degradation
β-oxidation
PPAR signaling pathway
AMACR
ACOX
DBP
SCPX
BAAT
Primary bile acid biosynthesis
(PTS2 type)
PBE
ACAA1
unsaturated fatty acid β-oxidation
(PTS1 type)
membrane proteins
PDCR
ECH
ABCD
ACSL
VLACS
other-oxidation (PTS1 type)
PECR
PECI
NUDT7
NUDT12
NUDT19
PTE
CRAT
CROT
MLYCD
etherphospholipid biosynthesis
matrix proteins
(PTS1 type)
DHAPAT
Glycerophospholipid metabolism
(PTS2 type)
AGPS
Ether lipid metabolism
membrane proteins
FAR
sterol precursor biosynthesis
(PTS2 type)
MVK
PMVK
Terpenoid backbone biosynthesis
amino acid metabolism (PTS1 type)
AGT
DAO
DDO
IDH
PAOX
PIPOX
HMGCL
HAO
antioxidant system
hydrogen peroxide metabolism
(PTS1 type)
CAT
SOD
INOS
PRDX5
PRDX1
epoxide metabolism
EPHX2
glutathione metabolism
GSTK1
purine metabolism
XDH
Purine metabolism
retinol metabolism
DHRS4
Retinol metabolism
Data on KEGG graph
Rendered by Pathview

Section 3. Gene Ontology (GO) Analysis

Run pathway analysis with GO

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

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

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

lapply(gobpres, head)
```

```
$greater
                                           p.geomean stat.mean        p.val
GO:0007156 homophilic cell adhesion       8.519724e-05  3.824205 8.519724e-05
GO:0002009 morphogenesis of an epithelium 1.396681e-04  3.653886 1.396681e-04
GO:0048729 tissue morphogenesis           1.432451e-04  3.643242 1.432451e-04
GO:0007610 behavior                       1.925222e-04  3.565432 1.925222e-04
GO:0060562 epithelial tube morphogenesis  5.932837e-04  3.261376 5.932837e-04
GO:0035295 tube development               5.953254e-04  3.253665 5.953254e-04
                                               q.val set.size         exp1
GO:0007156 homophilic cell adhesion       0.1951953      113 8.519724e-05
GO:0002009 morphogenesis of an epithelium 0.1951953      339 1.396681e-04
GO:0048729 tissue morphogenesis           0.1951953      424 1.432451e-04
GO:0007610 behavior                       0.1967577      426 1.925222e-04
GO:0060562 epithelial tube morphogenesis  0.3565320      257 5.932837e-04
GO:0035295 tube development               0.3565320      391 5.953254e-04

$less
                                             p.geomean stat.mean        p.val
GO:0048285 organelle fission              1.536227e-15 -8.063910 1.536227e-15
GO:0000280 nuclear division               4.286961e-15 -7.939217 4.286961e-15
GO:0007067 mitosis                        4.286961e-15 -7.939217 4.286961e-15
GO:0000087 M phase of mitotic cell cycle  1.169934e-14 -7.797496 1.169934e-14
GO:0007059 chromosome segregation         2.028624e-11 -6.878340 2.028624e-11
GO:0000236 mitotic prometaphase           1.729553e-10 -6.695966 1.729553e-10
                                                 q.val set.size         exp1
GO:0048285 organelle fission              5.841698e-12      376 1.536227e-15
GO:0000280 nuclear division               5.841698e-12      352 4.286961e-15
GO:0007067 mitosis                        5.841698e-12      352 4.286961e-15
```

```
GO:0000087 M phase of mitotic cell cycle 1.195672e-11      362 1.169934e-14
GO:0007059 chromosome segregation        1.658603e-08      142 2.028624e-11
GO:0000236 mitotic prometaphase          1.178402e-07       84 1.729553e-10

$stats
                                          stat.mean     exp1
GO:0007156 homophilic cell adhesion        3.824205 3.824205
GO:0002009 morphogenesis of an epithelium  3.653886 3.653886
GO:0048729 tissue morphogenesis            3.643242 3.643242
GO:0007610 behavior                        3.565432 3.565432
GO:0060562 epithelial tube morphogenesis   3.261376 3.261376
GO:0035295 tube development                3.253665 3.253665
```

```
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
```

Section 4. Reactome Analysis

```
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, quote=
```

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?

ANSWER: Signaling by PDGF has the most significant "entities p-value" in Reactome because it has the smallest value at 8.41 E-5. No, Signaling by PDGF, the most significant pathway does not patch the pervious KEGG results for neither top 5 upregulated nor downregulated KEGG pathways identified earlier. Some factors that could be causing the differences between the two methods is that Reactome details specific singaling, such as PDGF signaling, however, KEGG connects signaling pathways with each other. Therefore, PDGF may not be considered a separate pathway in KEGG.

Section 5. GO online (OPTIONAL)

Q9: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?

ANSWER: