Class14: RNA-Seq Mini Project

Yu (Ericsson) Cao (PID: A16421048)

Run a complete RNASeq analysis workflow from counts to enriched genesets...

Data import

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

Data exploration

```
head(counts, 3)
```

```
                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
                SRR493371
ENSG00000186092         0
ENSG00000279928         0
ENSG00000279457        46
```

```
head(metadata)
```

```
         id     condition
1 SRR493366 control_sirna
2 SRR493367 control_sirna
3 SRR493368 control_sirna
4 SRR493369      hoxa1_kd
```

```
5 SRR493370      hoxa1_kd
6 SRR493371      hoxa1_kd
```

Check if the colData and countData match up

```
metadata$id
```

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

```
colnames(counts)
```

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

```
#RM first column
countData <- counts[,-1]
```

Now check if the metadata and count data columns match

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

```
[1] TRUE
```

We need to remove all the zero count genes.

```
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
```

To identify these zero count genes we can sum across the rows and check if the sum is more than zero

```
to.keep.inds <- rowSums(countData) > 0
nonzero.counts <- countData[to.keep.inds,]
```

DESeq setup and analysis

```
library(DESeq2)
```

```
dds <- DESeqDataSetFromMatrix(countData = nonzero.counts,
                              colData = metadata,
                              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
```

Result extraction

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

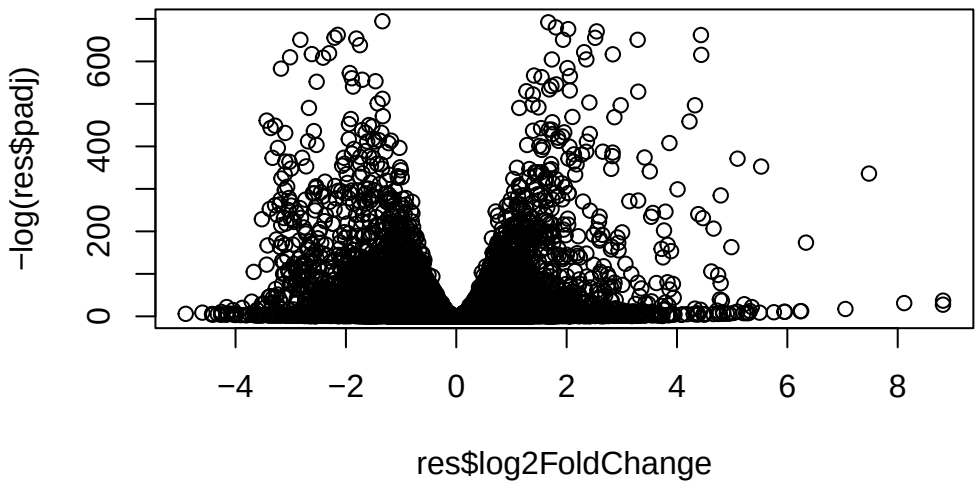
```
log2 fold change (MLE): condition hoxa1 kd vs control sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 6 rows and 6 columns
                 baseMean log2FoldChange     lfcSE       stat      pvalue
                <numeric>      <numeric> <numeric>  <numeric>   <numeric>
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
```

```
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
```

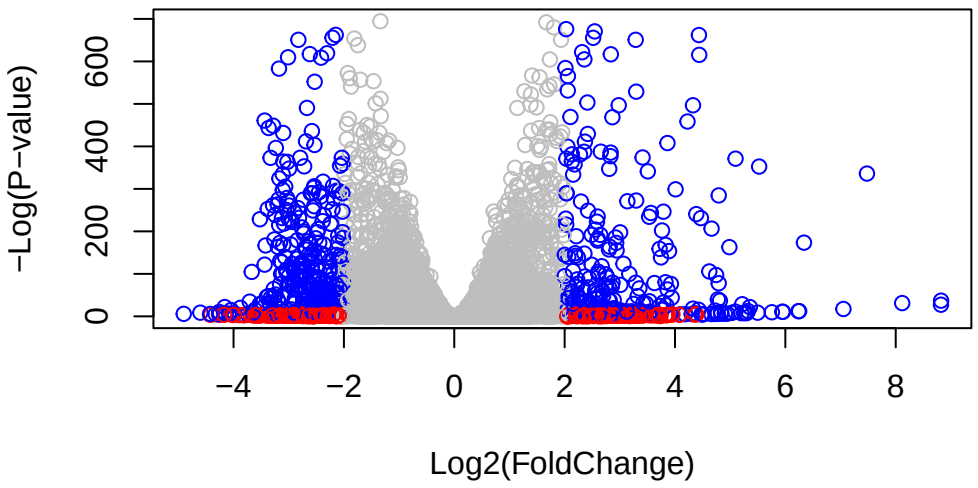
Volcano Plot

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



Setup a wee color vector for this plot

```
mycols <- rep("gray", nrow(res) )


mycols[ abs(res$log2FoldChange) > 2 ] <- "red"


inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"

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



Gene annotation

```
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(x=org.Hs.eg.db,
                     keys=rownames(nonzero.counts),
                     keytype="ENSEMBL",
```

```
                    column="SYMBOL")
```

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

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

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

```
res$name <- mapIds(x=org.Hs.eg.db,
                     keys=rownames(nonzero.counts),
                     keytype="ENSEMBL",
                     column="GENENAME")
```

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

```
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 9 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      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..
```

Pathway analysis

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

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

kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]

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)
```

```
      <NA>      148398       26155      339451       84069       84808
0.17925708  0.42645712 -0.69272046  0.72975561  0.04057653  0.54281049
```

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

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

Now, let's look at one of these pathways: hsa04110

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

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

Info: Working in directory /Users/yucao515/Desktop/BIMM 143/Class 14

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

```
# A different PDF based output of the same data
pathview(gene.data=foldchanges, pathway.id="hsa04110", kegg.native=FALSE)
```

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

Warning: reconcile groups sharing member nodes!
```

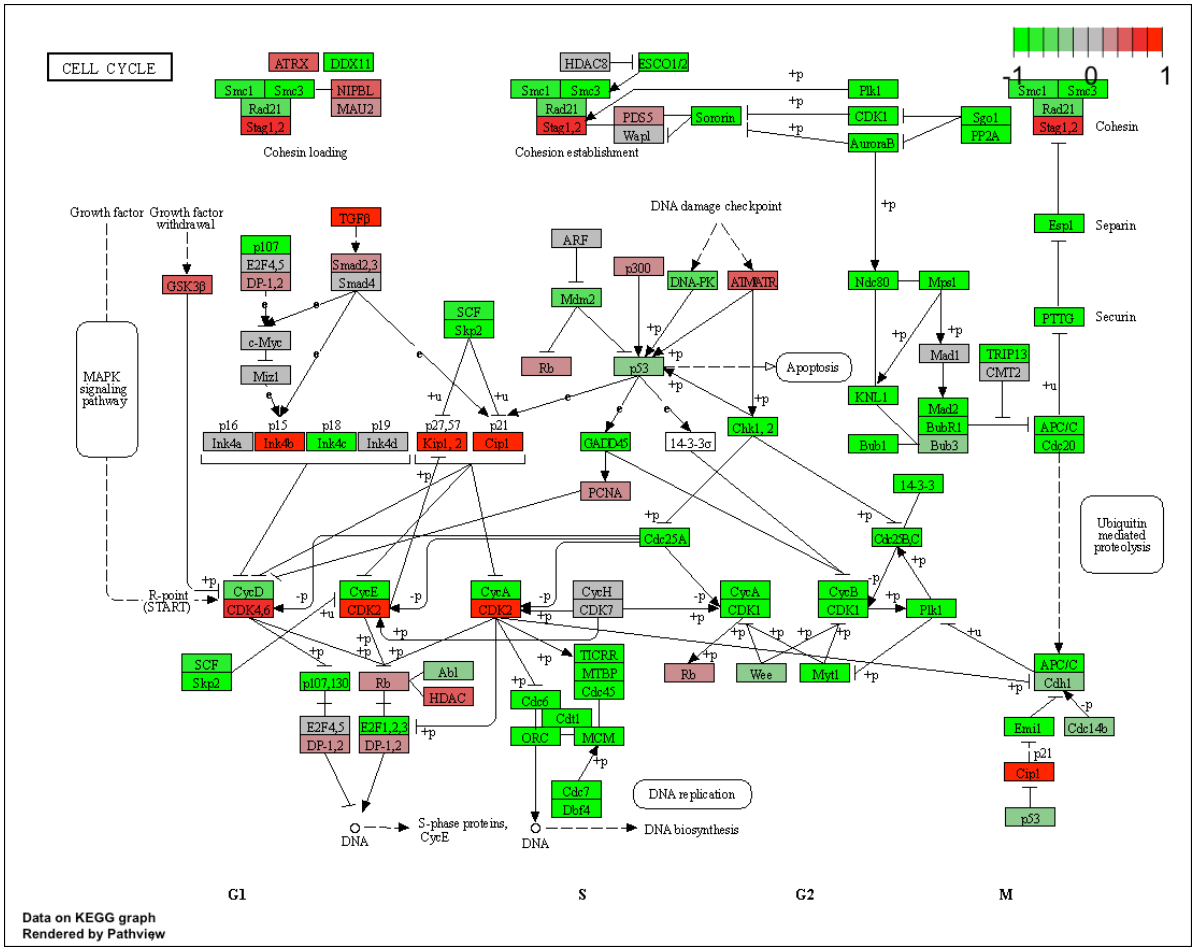


Figure 1: The KEGG Cell Cycle Pathway

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

```
Info: Working in directory /Users/yucao515/Desktop/BIMM 143/Class 14
```

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

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

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

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

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

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

```
Info: Working in directory /Users/yucao515/Desktop/BIMM 143/Class 14
```

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

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

```
Info: Working in directory /Users/yucao515/Desktop/BIMM 143/Class 14
```

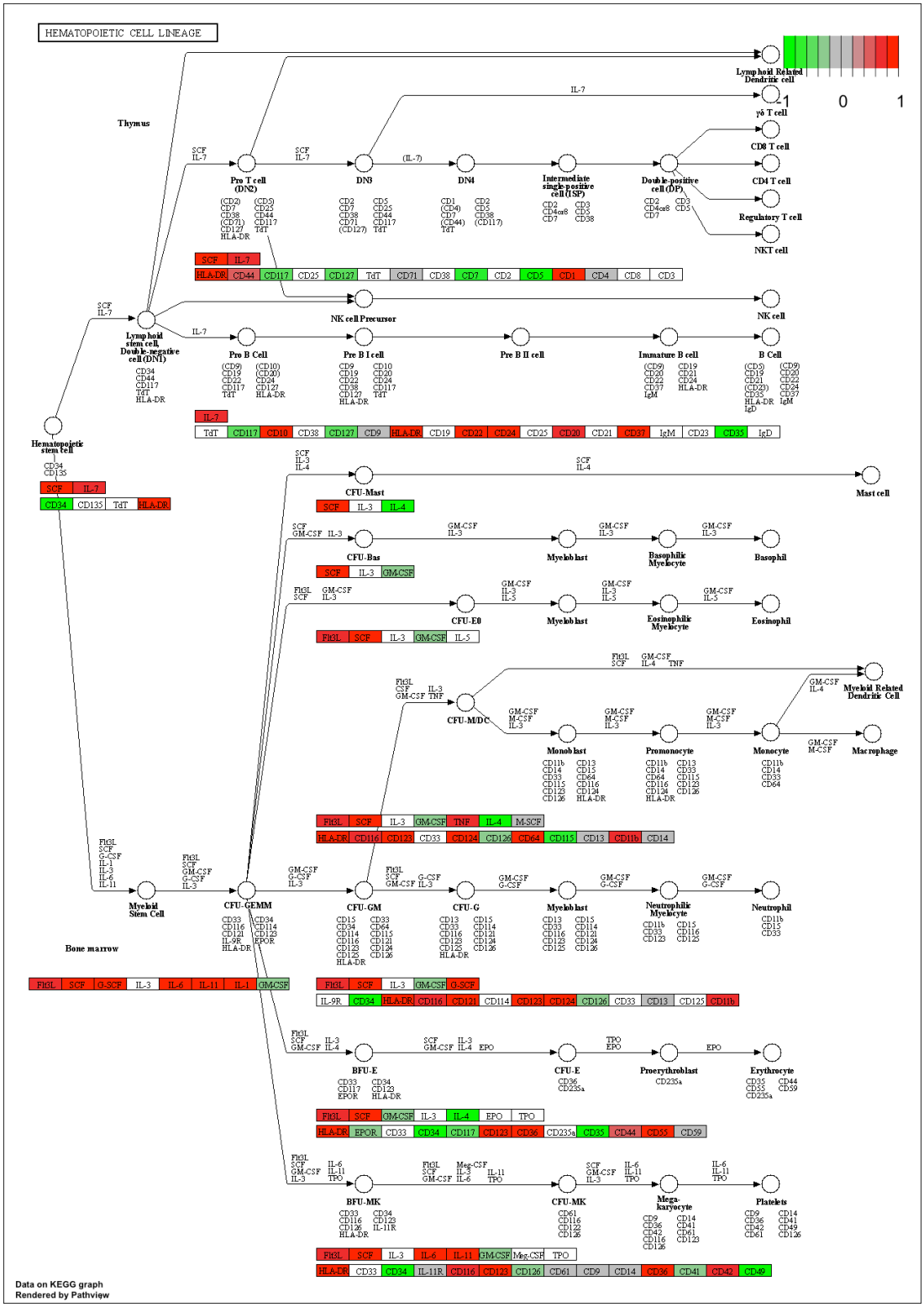
```
Info: Writing image file hsa04630.pathview.png
```

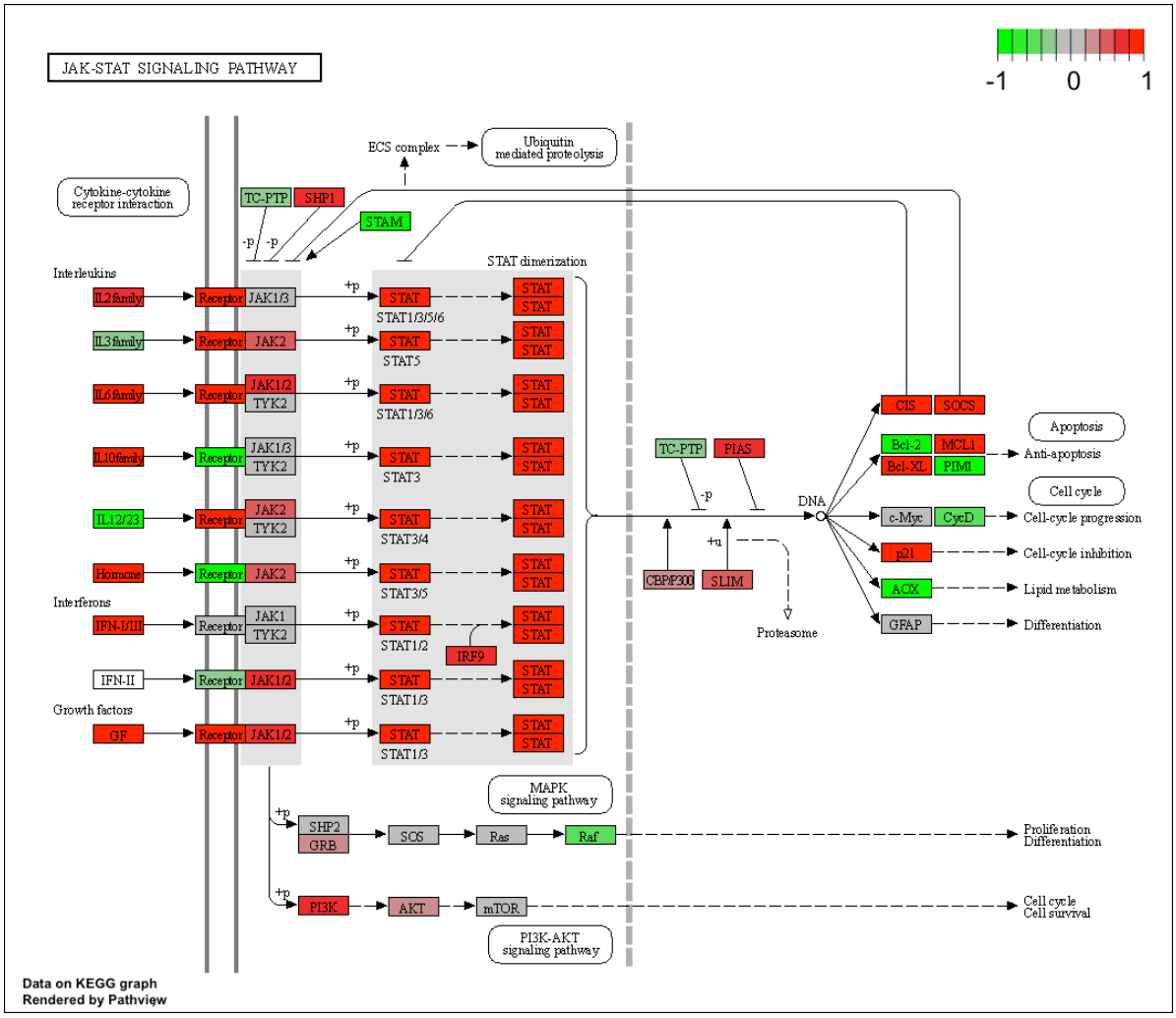
```
'select()' returned 1:1 mapping between keys and columns
```

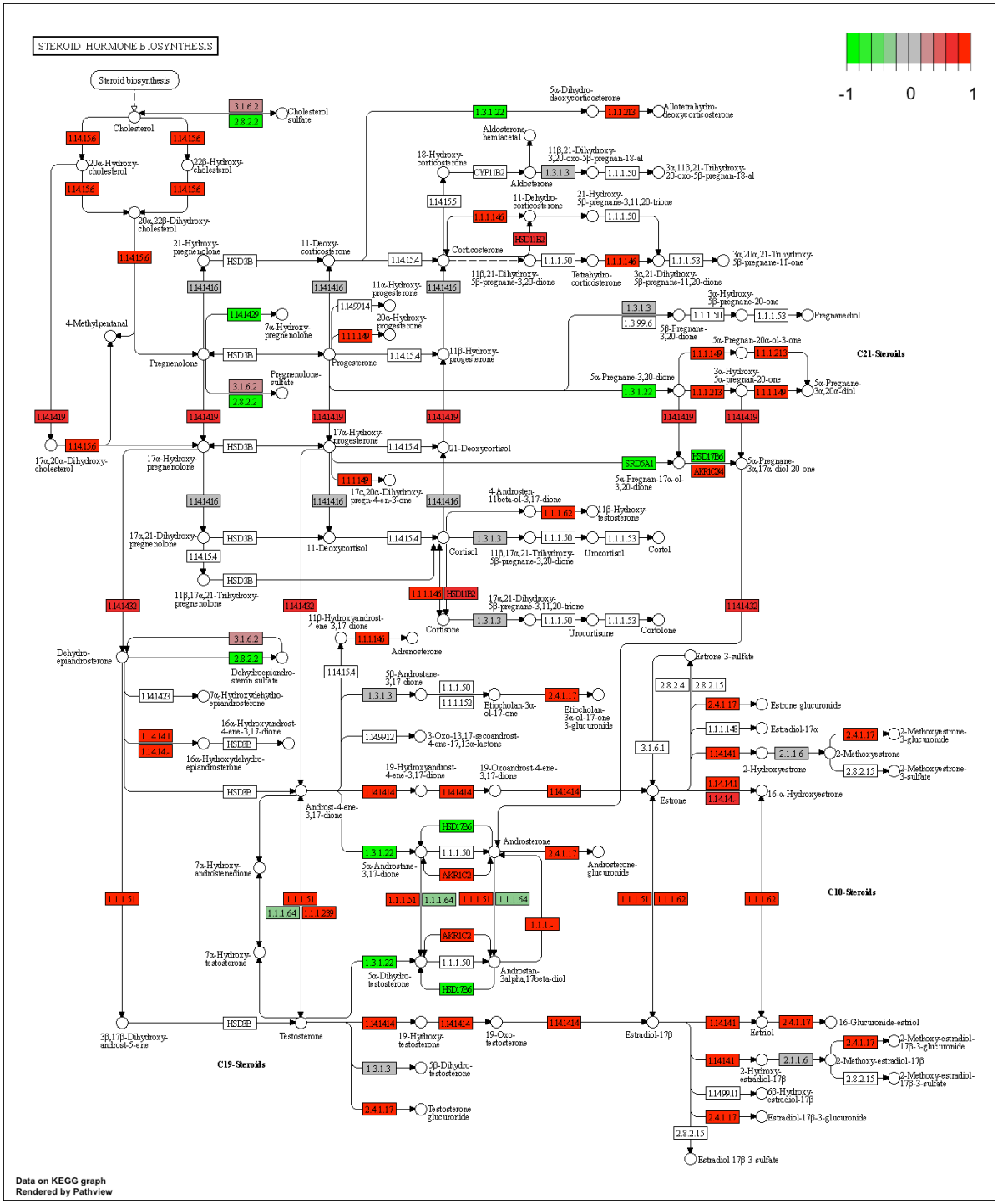
```
Info: Working in directory /Users/yucao515/Desktop/BIMM 143/Class 14
```

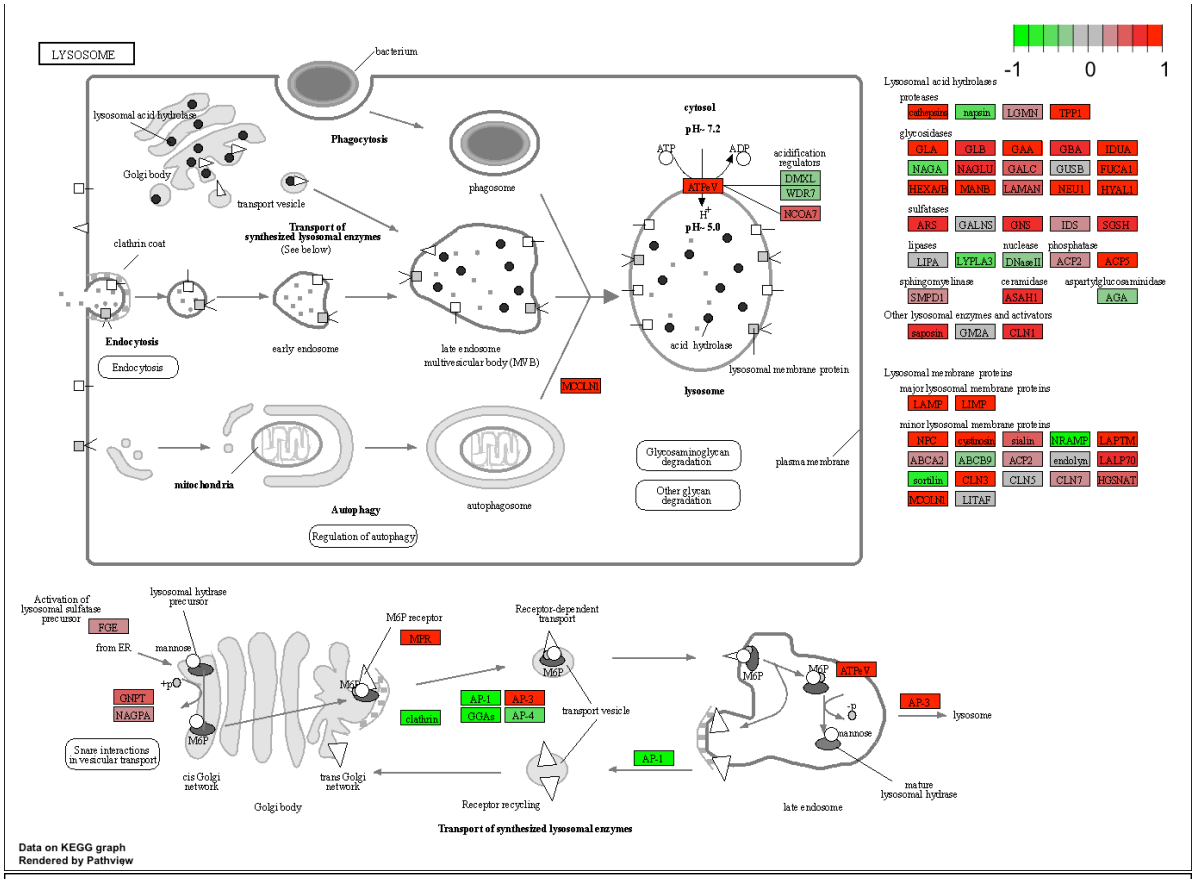
```
Info: Writing image file hsa00140.pathview.png
```

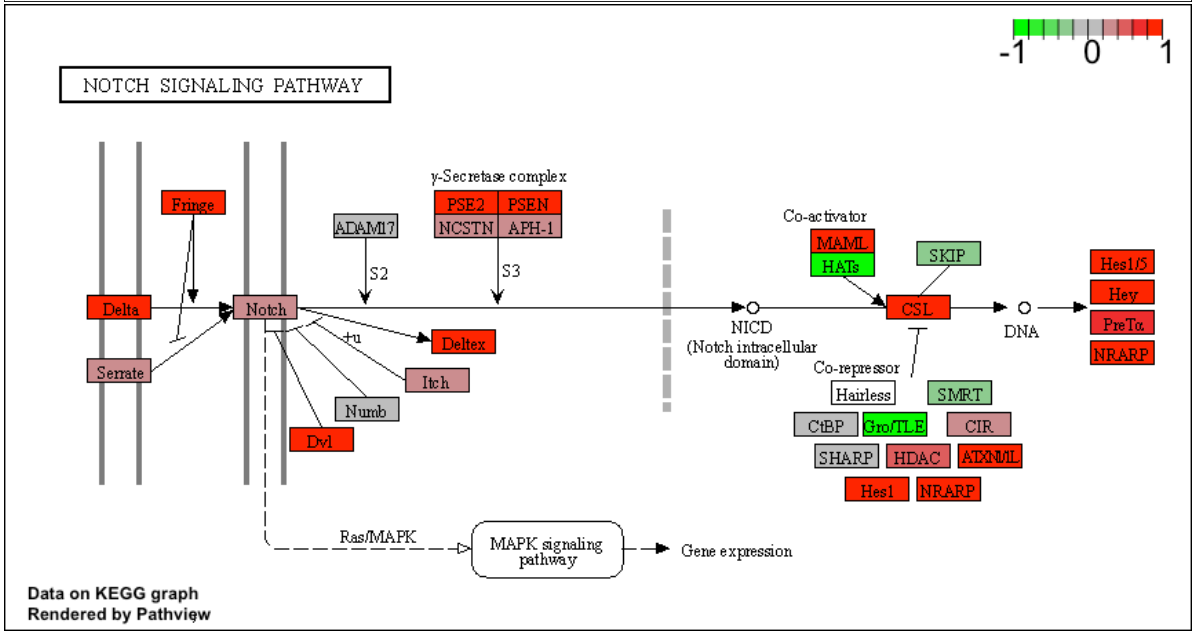
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/yucao515/Desktop/BIMM 143/Class 14

Info: Writing image file hsa04142.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/yucao515/Desktop/BIMM 143/Class 14

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


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
Pre T cell (DN2)
SCF
IL-7
DN3
(IL-7)
DN4
Intermediate single-positive cell (ISP)
Double-positive cell (DP)
(CD2) (CD5)
CD7 CD25
CD38 CD44
(CD71) CD117
CD127 TdT
HLA-DR
CD2 CD5
CD7 CD25
CD38 CD44
CD71 CD117
(CD127) TdT
CD1 CD2
(CD4) CD5
CD7 CD38
(CD44) (CD117)
TdT
CD2 CD3
CD4cor8 CD5
CD38
CD2 CD3
CD4cor8 CD5
CD7
SCF IL-7
HLA-DR CD44 CD117 CD25 CD127 TdT CD71 CD38 CD7 CD2 CD5 CD1 CD4 CD8 CD3
NK cell Precursor
NK cell
SCF
IL-7
Lymphoid stem cell, Double-negative cell (DN1)
IL-7
Pro B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
CD34
CD44
CD117
TdT
HLA-DR
(CD9) (CD10)
CD19 (CD20)
CD22 CD24
CD117 CD127
TdT HLA-DR
CD9 CD10
CD19 CD20
CD22 CD24
CD38 CD117
CD127 TdT
HLA-DR
(CD9) CD19
CD20 CD21
CD22 CD24
CD37 HLA-DR
IgM
(CD5) (CD9)
CD19 CD20
CD21 CD22
(CD22) CD24
CD25 CD37
HLA-DR IgM
IgD
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
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
FltGL
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
FltGL SCF IL-3 GM-CSF IL-5
FltGL
SCF
GM-CSF
TNF
CFU-M/DC
FltGL
SCF
IL-4
TNF
Myeloid Related Dendritic Cell
GM-CSF
IL-4
GM-CSF
M-CSF
IL-3
Monoblast
Promonocyte
Monocyte
Macrophage
CD11b CD13
CD14 CD15
CD33 CD64
CD115 CD116
CD123 CD124
CD126 HLA-DR
CD11b CD13
CD14 CD33
CD64 CD115
CD116 CD123
CD124 CD126
HLA-DR
CD11b CD13
CD14 CD33
CD64
FltGL SCF IL-3 GM-CSF TNF IL-4 M-SCF
HLA-DR CD116 CD123 CD33 CD124 CD126 CD64 CD115 CD13 CD11b CD14
FltGL
SCF
G-CSF
IL-1
IL-3
IL-6
IL-11
Myeloid Stem Cell
FltGL
SCF
GM-CSF
G-CSF
IL-3
CFU-GEMM
GM-CSF
G-CSF
IL-3
CFU-GM
FltGL
SCF
GM-CSF
G-CSF
IL-3
CFU-G
GM-CSF
G-CSF
Myeloblast
GM-CSF
G-CSF
Neutrophilic Myelocyte
GM-CSF
G-CSF
Neutrophil
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
CD33
Bone marrow
FltGL SCF G-CSF IL-3 IL-6 IL-11 IL-1 GM-CSF
FltGL SCF IL-3 GM-CSF G-CSF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD11b
FltGL
SCF
GM-CSF
IL-4
IL-3
BFU-E
SCF
GM-CSF
IL-4
IL-3
EPO
CFU-E
TPO
EPO
Proerythroblast
EPO
Erythrocyte
CD33 CD117 EPOR
CD34 CD123 HLA-DR
CD36 CD235a
CD235a
CD35 CD44 CD55 CD59 CD235a
FltGL SCF GM-CSF IL-3 IL-4 EPO TPO
HLA-DR EPOR CD33 CD34 CD117 CD123 CD36 CD235a CD35 CD44 CD55 CD59
FltGL
SCF
GM-CSF
IL-3
IL-6
IL-11
TPO
BFU-MK
Meg-CSF
IL-3
IL-6
IL-11
TPO
CFU-MK
SCF
GM-CSF
IL-3
IL-6
IL-11
TPO
Mega-karyocyte
IL-6
IL-11
TPO
Platelets
CD33 CD34
CD116 CD123
CD126 IL-11R
HLA-DR
CD61 CD14
CD116 CD41
CD122 CD61
CD126 CD123
CD9 CD14
CD36 CD41
CD42 CD49
CD116 CD61
CD126
CD9 CD14
CD36 CD41
CD42 CD49
CD61 CD126
FltGL 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
STAT dimerization
Interleukins
IL2family
IL3family
IL6family
IL10family
IL12/23
Hormone
Interferons
IFN-I/III
IFN-II
Growth factors
GF
Receptor
JAK1/3
JAK2
JAK1/2
TYK2
JAK1
STAT
STAT1/3/5/6
STAT5
STAT1/3/6
STAT3
STAT3/4
STAT3/5
STAT1/2
STAT1/3
IRF9
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
Apoptosis
Anti-apoptosis
Cell cycle
c-Myc
CycD
Cell-cycle progression
p21
Cell-cycle inhibition
AOX
Lipid metabolism
GFAP
Differentiation
PIAS
DNA
CBP/P300
SLIM
Proteasome
MAPK signaling pathway
SHP2
GRB
SOS
Ras
Raf
Proliferation
Differentiation
PI3K
AKT
mTOR
Cell cycle
Cell survival
PI3K-AKT signaling pathway
Data on KEGG graph
Rendered by Pathview


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


LYSOSOME
bacterium
lysosomal acid hydrolase
Golgi body
transport vesicle
Phagocytosis
phagosome
cytosol
pH~ 7.2
ATP
ADP
acidification regulators
DMXL
WDR7
NCOA7
ATPeV
H+
pH~ 5.0
Transport of synthesized lysosomal enzymes
(See below)
clathrin coat
Endocytosis
Endocytosis
early endosome
late endosome
multivesicular body (MVB)
MCOLN1
acid hydrolase
lysosomal membrane protein
lysosome
mitochondria
Autophagy
autophagosome
Regulation of autophagy
Glycosaminoglycan degradation
Other glycan degradation
plasma membrane
-1
0
1
Lysosomal acid hydrolases
proteases
cathepsin
napsin
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
LALP70
sortilin
CLN3
CLN5
CLN7
HGSNAT
MCOLN1
LITAF
Activation of lysosomal sulfatase precursor
FGE
lysosomal hydrase precursor
from ER
mannose
+pO
GNPT
NAGPA
M6P
M6P receptor
MPR
Receptor-dependent transport
M6P
clathrin
AP-1
GGAs
AP-3
AP-4
transport vesicle
Snare interactions in vesicular transport
cis Golgi network
trans Golgi network
Golgi body
Receptor recycling
AP-1
M6P
M6P
ATPeV
-p
mannose
AP-3
lysosome
mature lysosomal hydrase
late endosome
Transport of synthesized lysosomal enzymes
Data on KEGG graph
Rendered by Pathview


-1
0
1
NOTCH SIGNALING PATHWAY
Fringe
Delta
Serrate
Notch
γ-Secretase complex
PSE2
PSEN
NCSTN
APH-1
ADAM17
S2
S3
+u
Deltex
Itch
Numb
Dvl
Co-activator
MAML
HATs
SKIP
CSL
DNA
Hes1/5
Hey
PreTα
NRARP
NICD
(Notch intracellular domain)
Co-repressor
Hairless
SMRT
CtBP
Gro/TLE
CIR
SHARP
HDAC
ATXN1L
Hes1
NRARP
Ras/MAPK
MAPK signaling pathway
Gene expression
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)
```

```
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             5.843127e-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
```

Reactome Analysis

We can use reactome via an R package or use their relatively new website interface. Let's use the latter.

It wants a list of our most interesting (i.e. significant) genes in gene SYMBOL format.

```
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"
```

```
#sig_genes
```

We will write these our to a wee file so we can use them on the website:

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

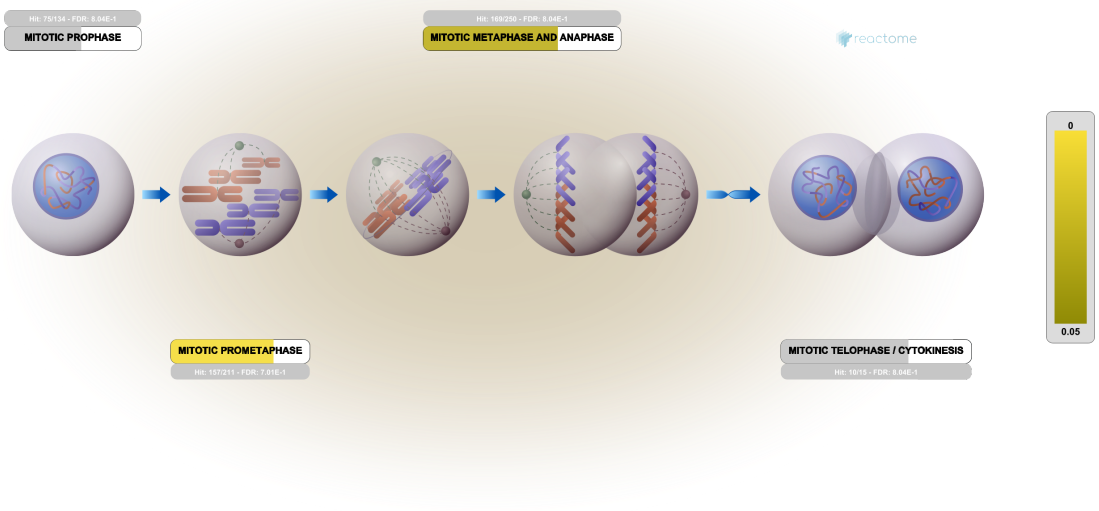


Figure 2: Example figure from reactome online - mitosis

> 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?

A: Cell Cycle has the most significant entities P-value. Yes, the most significant pathways listed do match my previous KEGG results. They analyze pathways from different perspectives.