Class 13: RNA-Seq Analysis Mini Project

Michelle Woo

1. Differential Expression Analysis

Loading in DESeq2 and our data files:

```r
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, 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


Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':

    windows

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
```

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

# Import metadata
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
```

```r
# 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`

```r
# Removing the odd first $length col
countData <- as.matrix(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
```

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).

```r
# Filter count data to have 0 read count across all samples
countData <- countData[-c(1,2,4,5),]

head(countData)
```

```
                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
```

2. Running DESeq2

```r
# Setting up DESeqDataSet
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
```

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

```
estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

```r
dds
```

```
class: DESeqDataSet
dim: 19804 6
metadata(1): version
assays(4): counts mu H cooks
rownames(19804): ENSG00000279457 ENSG00000187634 ... ENSG00000277475
  ENSG00000268674
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

```r
# HoxA1 knockdown vs control siRNA
res <- results(dds, contrast=c("condition", "hoxa1_kd", "control_sirna"))
```

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.

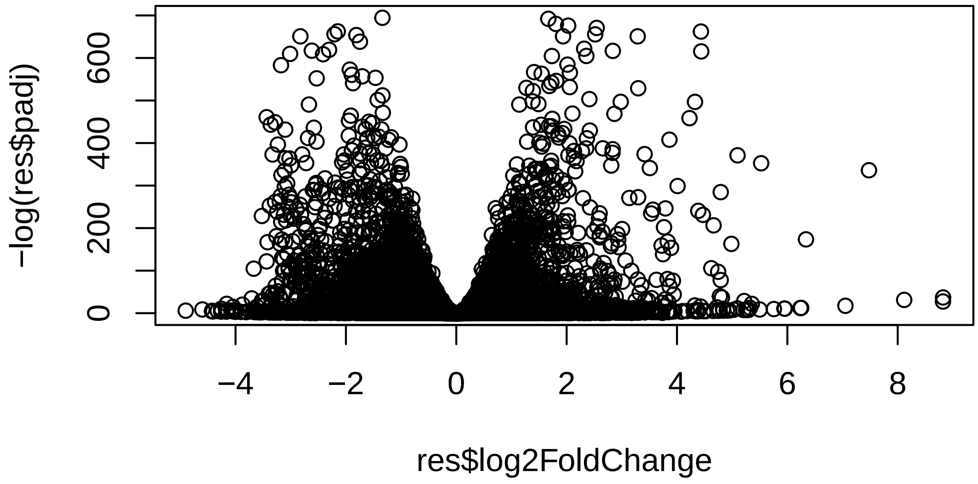
```r
summary(res)
```

```

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
```

3. Volcano Plot

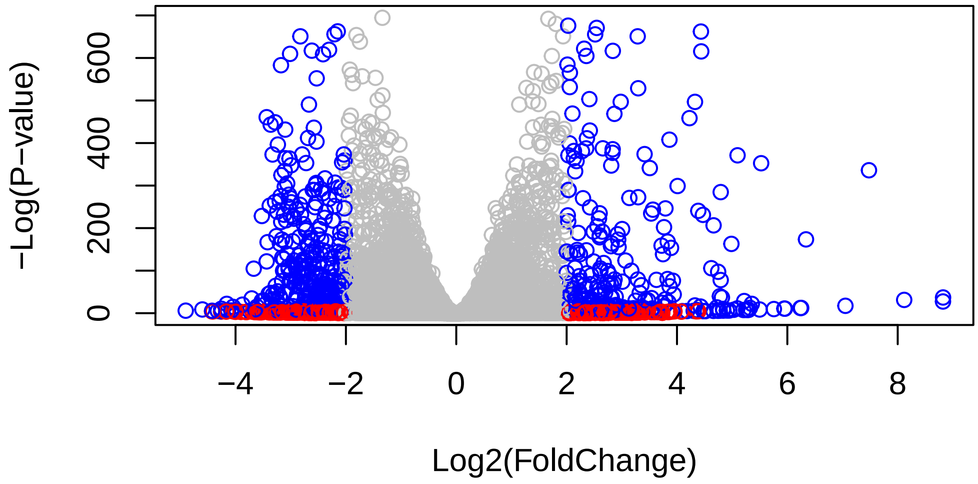
A plot of log2 fold change vs -log adjusted p-value

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

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



Adding gene annotation

The results only have information about Ensembl gene IDs since it was mapped and counted against the Ensembl annotation. However, the pathway analysis downstream will use KEGG pathways, and genes in KEGG pathways are annotated with Entrez gene IDs. So we will need to add this to our data.

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

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

```
# Checking the column names
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="SYMBOL",
                    multiVals="first")
```

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

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

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

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

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

```
head(res, 10)
```

```
log2 fold change (MLE): condition hoxa1_kd vs control_sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 10 rows and 9 columns
                 baseMean log2FoldChange     lfcSE      stat      pvalue
                <numeric>      <numeric> <numeric> <numeric>   <numeric>
ENSG00000279457 29.913579      0.1792571 0.3248216  0.551863 5.81042e-01
```

```
ENSG00000187634  183.229650      0.4264571 0.1402658   3.040350 2.36304e-03
ENSG00000188976 1651.188076     -0.6927205 0.0548465 -12.630158 1.43990e-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 0.5215598   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 6.87080e-01          NA          NA                     NA
ENSG00000187634 5.16278e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76741e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13536e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.18988e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03817e-01       PERM1       84808 PPARGC1 and ESRR ind..
ENSG00000188290 1.30680e-24        HES4       57801 hes family bHLH tran..
ENSG00000187608 2.37710e-02       ISG15        9636 ISG15 ubiquitin like..
ENSG00000188157 4.22421e-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.**

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

4. Pathway Analysis

First we will use the gage package for pathway analysis. And once we have a list of enriched pathways, we will use the pathview package to draw pathway diagrams with a focus on the up and down-regulation degrees.

KEGG Pathways

Installing the required bioconductor packages:

```
# BiocManager::install( c("pathview", "gage", "gageData") )
```

Now loading the packages and setting up the KEGG data-sets:

```
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)

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"
```

```r
foldchanges <- res$log2FoldChange

names(foldchanges) <- res$entrez

head(foldchanges)
```

```
     1266     54855      1465     51232      2034      2317
-2.422719  3.201955 -2.313738 -2.059631 -1.888019 -1.649792
```

Running the gage pathway analysis:

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

```
# object returned from gage()
attributes(keggres)
```

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

```
# Look at the first few down(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               1.121279e-03 -3.090949 1.121279e-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               0.061296597      150 1.121279e-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
```

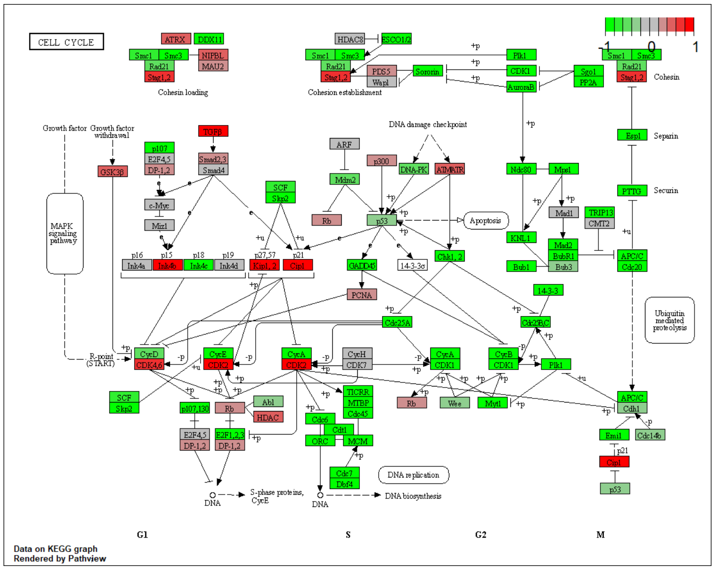
Making the pathway plot:

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

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

```
Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj
```

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



Data on KEGG graph
Rendered by Pathview

Now we can pull out the 5 unregulated pathways and get the pathway IDs:

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

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

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

Now we can get the IDs into `keggresids` to `pathview()`. This will draw plots for all the top 5 pathways:

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

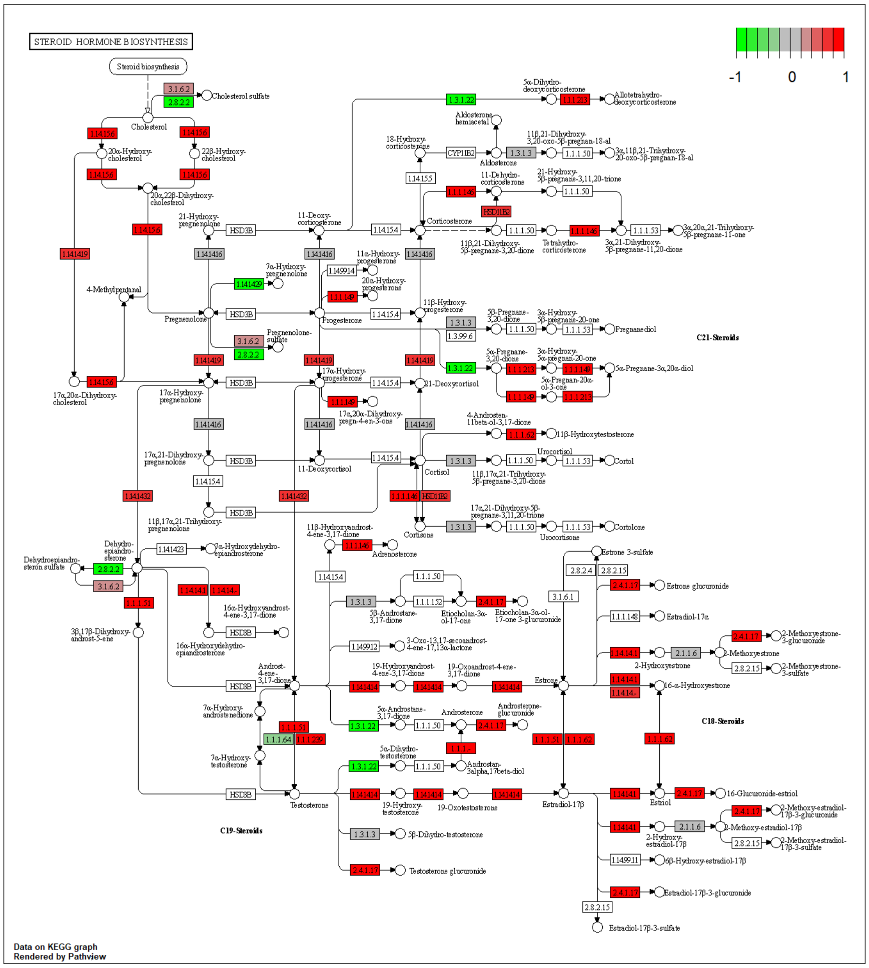
```
'select()' returned 1:1 mapping between keys and columns
```

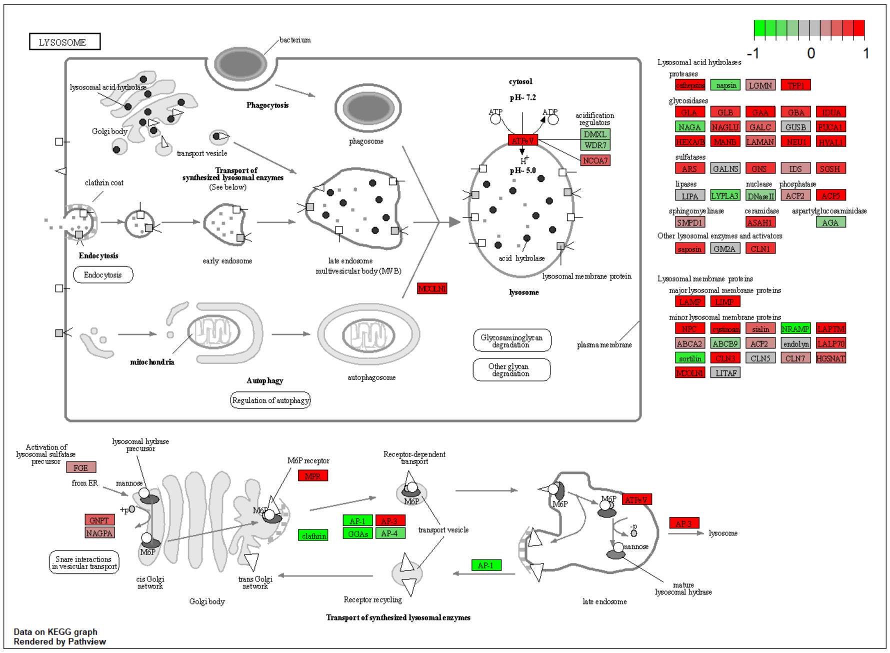
```
Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj
```

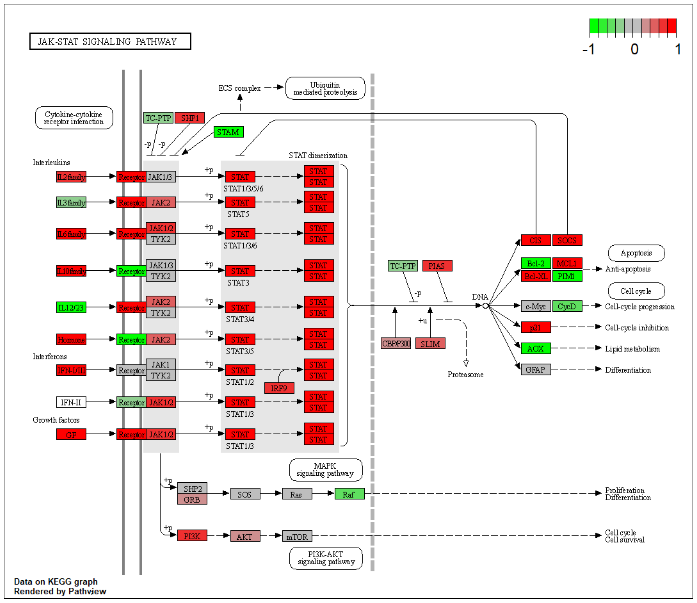
```
Info: Writing image file hsa04740.pathview.png
```

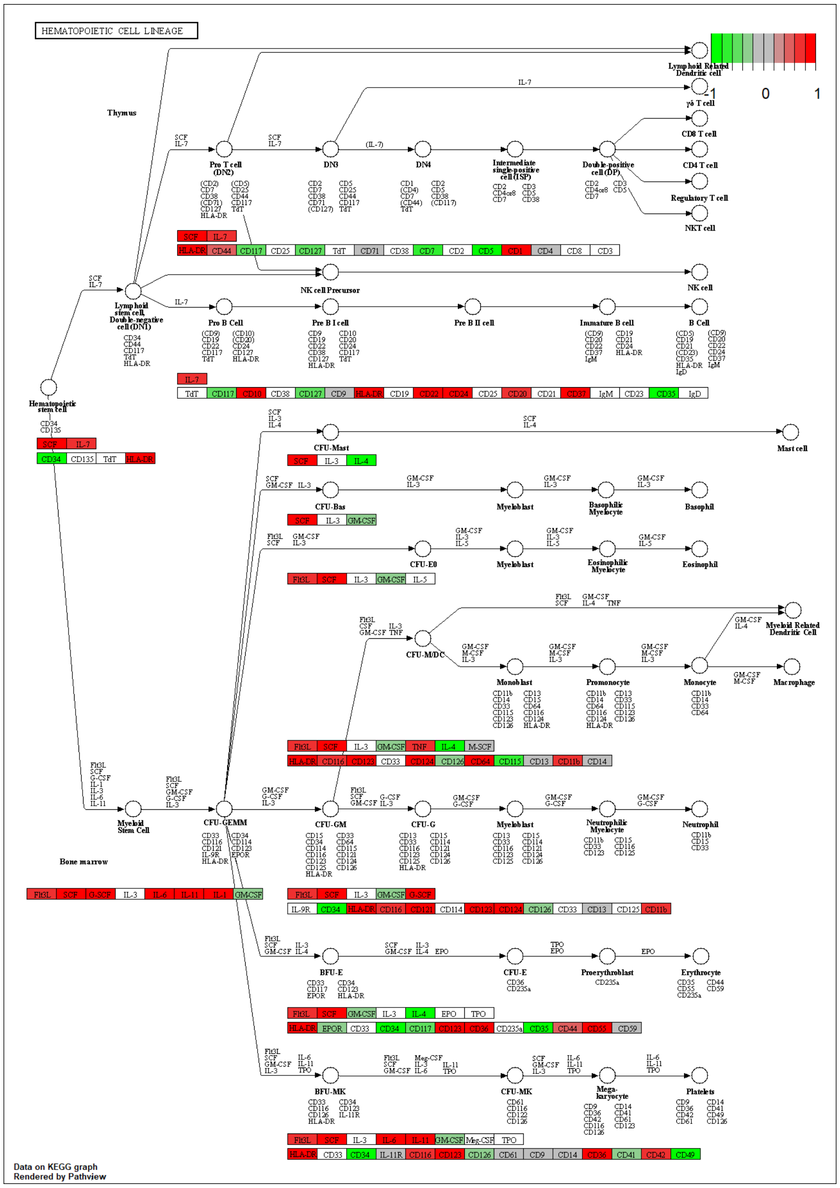
```
Info: some node width is different from others, and hence adjusted!
```

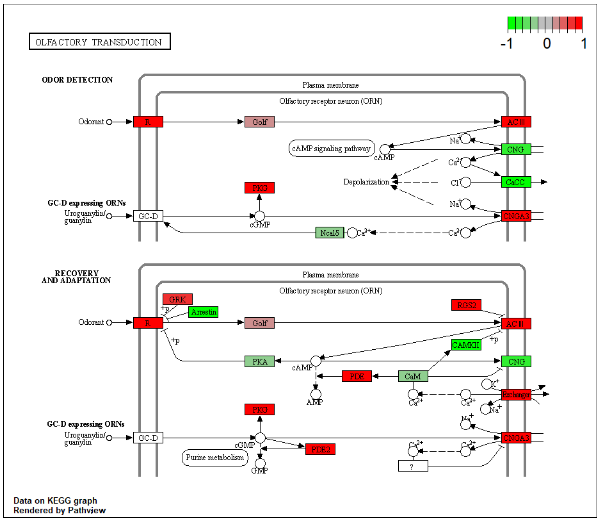
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

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
4-Methylpentanal
17α,20α-Dihydroxy-cholesterol
21-Hydroxy-pregnenolone
7α-Hydroxy-pregnenolone
Pregnenolone
Pregnenolone-sulfate
17α-Hydroxy-pregnenolone
17α,21-Dihydroxy-pregnenolone
11β,17α,21-Trihydroxy-pregnenolone
11-Deoxy-corticosterone
11α-Hydroxy-progesterone
20α-Hydroxy-progesterone
Progesterone
17α-Hydroxy-progesterone
17α,20α-Dihydroxy-pregn-4-en-3-one
11-Deoxycortisol
18-Hydroxy-corticosterone
Corticosterone
11β-Hydroxy-progesterone
21-Deoxycortisol
Cortisol
Cortisone
5α-Dihydro-deoxycorticosterone
Allotetrahydro-deoxycorticosterone
Aldosterone hemiacetal
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
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
5β-Pregnane-3,20-dione
3α-Hydroxy-5β-pregnane-20-one
Pregnanediol
5α-Pregnane-3,20-dione
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
5α-Pregnan-20α-ol-3-one
C21-Steroids
4-Androsten-11beta-ol-3,17-dione
11β-Hydroxytestosterone
Urocortisol
Cortol
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Urocortisone
Cortolone
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
Dehydro-epiandro-sterone
Dehydroepiandro-steron sulfate
7α-Hydroxydehydro-epiandrosterone
16α-Hydroxyandrost-4-ene-3,17-dione
16α-Hydroxydehydro-epiandrosterone
3β,17β-Dihydroxy-androst-5-ene
Androst-4-ene-3,17-dione
7α-Hydroxy-androstenedione
7α-Hydroxy-testosterone
Testosterone
C19-Steroids
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
19-Hydroxyandrost-4-ene-3,17-dione
19-Oxoandrost-4-ene-3,17-dione
5α-Androstane-3,17-dione
Androsterone
Androsterone-glucuronide
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
19-Hydroxy-testosterone
19-Oxotestosterone
5β-Dihydro-testosterone
Testosterone glucuronide
Estrone 3-sulfate
Estrone glucuronide
Estradiol-17α
2-Hydroxyestrone
2-Methoxyestrone
2-Methoxyestrone-3-glucuronide
2-Methoxyestrone-3-sulfate
Estrone
16-α-Hydroxyestrone
C18-Steroids
Estradiol-17β
Estriol
16-Glucuronide-estriol
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β
2-Methoxy-estradiol-17β-3-glucuronide
2-Methoxy-estradiol-17β-3-sulfate
6β-Hydroxy-estradiol-17β
Estradiol-17β-3-glucuronide
Estradiol-17β-3-sulfate
1.14.15.6
3.1.6.2
2.8.2.2
1.14.14.19
1.14.14.16
1.14.14.29
1.14.99.14
1.1.1.149
HSD3B
1.14.15.4
1.14.15.5
CYP11B2
1.3.1.3
1.1.1.50
1.1.1.146
HSD11B2
1.1.1.53
1.3.1.22
1.1.1.213
1.3.99.6
1.1.1.62
1.14.14.32
1.14.14.23
1.14.14.1
1.14.14.-
1.1.1.51
1.1.1.64
1.1.1.239
1.14.99.12
1.14.14.14
1.1.1.152
2.4.1.17
1.1.1.-
2.8.2.4
2.8.2.15
3.1.6.1
1.1.1.148
2.1.1.6
1.14.99.11
Data on KEGG graph
Rendered by Pathview


LYSOSOME
-1
0
1
bacterium
lysosomal acid hydrolase
Golgi body
Phagocytosis
phagosome
transport vesicle
Transport of synthesized lysosomal enzymes (See below)
clathrin coat
Endocytosis
Endocytosis
early endosome
late endosome multivesicular body (MVB)
cytosol
pH~ 7.2
ATP
ADP
ATPeV
acidification regulators
DMXL
WDR7
NCOA7
H+
pH~ 5.0
acid hydrolase
lysosome
lysosomal membrane protein
plasma membrane
MCOLN1
mitochondria
Autophagy
autophagosome
Regulation of autophagy
Glycosaminoglycan degradation
Other glycan degradation
Lysosomal acid hydrolases
proteases
cathepsins
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
LIPA
LYPLA3
nuclease
DNase II
phosphatase
ACP2
ACP5
sphingomyelinase
SMPD1
ceramidase
ASAH1
aspartylglucosaminidase
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
from ER
mannose
+p
GNPT
NAGPA
M6P
lysosomal hydrase precursor
M6P receptor
MPR
clathrin
AP-1
AP-3
GGAs
AP-4
Receptor-dependent transport
M6P
transport vesicle
AP-1
M6P
ATPeV
AP-3
lysosome
-p
mannose
mature lysosomal hydrase
late endosome
Snare interactions in vesicular transport
cis Golgi network
trans Golgi network
Golgi body
Receptor recycling
Transport of synthesized lysosomal enzymes
Data on KEGG graph
Rendered by Pathview


JAK-STAT SIGNALING PATHWAY
-1
0
1
Cytokine-cytokine receptor interaction
ECS complex
Ubiquitin mediated proteolysis
TC-PTP
SHP1
STAM
-p
-p
STAT dimerization
Interleukins
IL2 family
Receptor
JAK1/3
+p
STAT
STAT1/3/5/6
IL3 family
Receptor
JAK2
STAT5
IL6 family
Receptor
JAK1/2
TYK2
STAT1/3/6
IL10 family
Receptor
JAK1/3
TYK2
STAT3
IL12/23
Receptor
JAK2
TYK2
STAT3/4
Hormone
Receptor
JAK2
STAT3/5
Interferons
IFN-I/III
Receptor
JAK1
TYK2
STAT1/2
IRF9
IFN-II
Receptor
JAK1/2
STAT1/3
Growth factors
GF
Receptor
JAK1/2
STAT1/3
STAT
STAT
TC-PTP
PIAS
-p
CBP/P300
SLIM
+u
Proteasome
DNA
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
MAPK signaling pathway
+p
SHP2
GRB
SOS
Ras
Raf
Proliferation Differentiation
+p
PI3K
AKT
mTOR
Cell cycle Cell survival
PI3K-AKT signaling pathway
Data on KEGG graph
Rendered by Pathview


HEMATOPOIETIC CELL LINEAGE
-1
0
1
Thymus
Bone marrow
Hematopoietic stem cell
CD34 CD135
SCF IL-7
CD34 CD135 TdT HLA-DR
SCF IL-7
Lymphoid stem cell, Double-negative cell (DN1)
CD34 CD44 CD117 TdT HLA-DR
Pro T cell (DN2)
(CD2) CD7 CD38 (CD71) CD127 HLA-DR (CD5) CD25 CD44 CD117 TdT
DN3
CD2 CD7 CD38 CD71 (CD127) CD5 CD25 CD44 CD117 TdT
(IL-7)
DN4
CD1 (CD4) CD7 (CD44) TdT CD2 CD5 CD38 (CD117)
Intermediate single-positive cell (ISP)
CD2 CD4α8 CD7 CD3 CD5 CD38
Double-positive cell (DP)
CD2 CD4α8 CD7 CD3 CD5
Lymphoid Related Dendritic cell
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF IL-7
HLA-DR CD44 CD117 CD25 CD127 TdT CD71 CD38 CD7 CD2 CD5 CD1 CD4 CD8 CD3
NK cell Precursor
NK cell
IL-7
Pro B Cell
(CD9) CD19 CD22 CD117 TdT (CD10) (CD20) CD24 CD127 HLA-DR
Pre B I cell
CD9 CD19 CD22 CD38 CD127 HLA-DR CD10 CD20 CD24 CD117 TdT
Pre B II cell
Immature B cell
(CD9) CD20 CD22 CD37 IgM CD19 CD21 CD24 HLA-DR
B Cell
(CD5) CD19 CD21 (CD23) CD35 HLA-DR IgD (CD9) CD20 CD22 CD24 CD37 IgM
IL-7
TdT CD117 CD10 CD38 CD127 CD9 HLA-DR CD19 CD22 CD24 CD25 CD20 CD21 CD37 IgM CD23 CD35 IgD
SCF IL-3 IL-4
CFU-Mast
SCF IL-4
Mast cell
SCF IL-3 IL-4
SCF GM-CSF IL-3
CFU-Bas
GM-CSF IL-3
Myeloblast
Basophilic Myelocyte
Basophil
SCF IL-3 GM-CSF
Flt3L SCF GM-CSF IL-3
CFU-E0
GM-CSF IL-3 IL-5
Myeloblast
Eosinophilic Myelocyte
GM-CSF IL-5
Eosinophil
Flt3L SCF IL-3 GM-CSF IL-5
Flt3L CSF GM-CSF IL-3 TNF
CFU-M/DC
Flt3L SCF GM-CSF IL-4 TNF
Myeloid Related Dendritic Cell
GM-CSF IL-4
GM-CSF M-CSF IL-3
Monoblast
CD11b CD14 CD33 CD115 CD123 CD126 CD13 CD15 CD64 CD116 CD124 HLA-DR
Promonocyte
CD11b CD14 CD64 CD116 CD124 HLA-DR CD13 CD33 CD115 CD123 CD126
Monocyte
CD11b CD14 CD33 CD64
GM-CSF M-CSF
Macrophage
Flt3L SCF IL-3 GM-CSF TNF IL-4 M-SCF
HLA-DR CD116 CD123 CD33 CD124 CD126 CD64 CD115 CD13 CD11b CD14
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Myeloid Stem Cell
Flt3L SCF GM-CSF G-CSF IL-3
CFU-GEMM
CD33 CD116 CD121 IL-9R HLA-DR CD34 CD114 CD123 EPOR
Flt3L SCF G-CSF IL-3 IL-6 IL-11 IL-1 GM-CSF
GM-CSF G-CSF IL-3
CFU-GM
CD15 CD34 CD114 CD116 CD123 CD125 HLA-DR CD33 CD64 CD115 CD121 CD124 CD126
Flt3L SCF GM-CSF G-CSF IL-3
CFU-G
CD13 CD33 CD116 CD123 CD125 HLA-DR CD15 CD114 CD121 CD124 CD126
GM-CSF G-CSF
Myeloblast
CD13 CD33 CD116 CD123 CD125 CD15 CD114 CD121 CD124 CD126
Neutrophilic Myelocyte
CD11b CD33 CD123 CD15 CD116 CD125
Neutrophil
CD11b CD15 CD33
Flt3L SCF IL-3 GM-CSF G-SCF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD11b
Flt3L SCF GM-CSF IL-3 IL-4
BFU-E
CD33 CD117 EPOR CD34 CD123 HLA-DR
SCF GM-CSF IL-3 IL-4 EPO
CFU-E
CD36 CD235a
TPO EPO
Proerythroblast
CD235a
EPO
Erythrocyte
CD35 CD55 CD235a CD44 CD59
Flt3L SCF GM-CSF IL-3 IL-4 EPO TPO
HLA-DR EPOR CD33 CD34 CD117 CD123 CD36 CD235a CD35 CD44 CD55 CD59
Flt3L SCF GM-CSF IL-3 IL-6 IL-11 TPO
BFU-MK
CD33 CD116 CD126 HLA-DR CD34 CD123 IL-11R
Flt3L SCF GM-CSF Meg-CSF IL-3 IL-6 IL-11 TPO
CFU-MK
CD61 CD116 CD122 CD126
SCF GM-CSF IL-3 IL-6 IL-11 TPO
Mega-karyocyte
CD9 CD36 CD42 CD116 CD126 CD14 CD41 CD61 CD123
Platelets
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



Q7. Can you do the same procedure as above to plot the pathview figures for the top 5 down-reguled pathways?

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

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

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

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

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

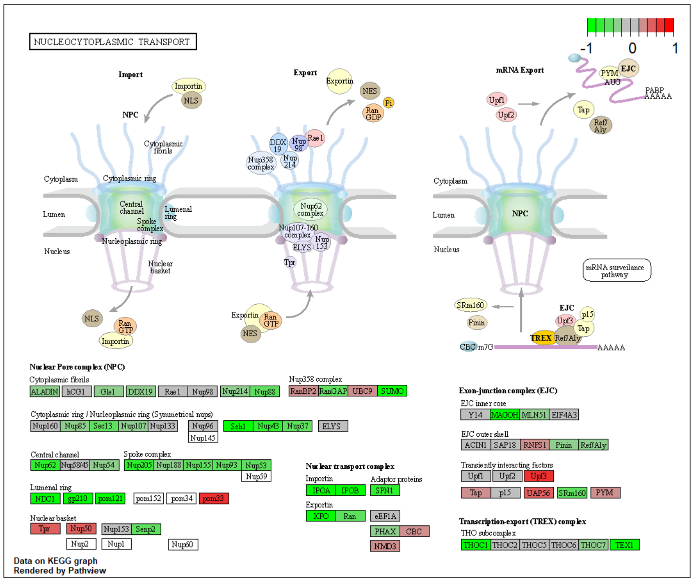
```
Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj
```

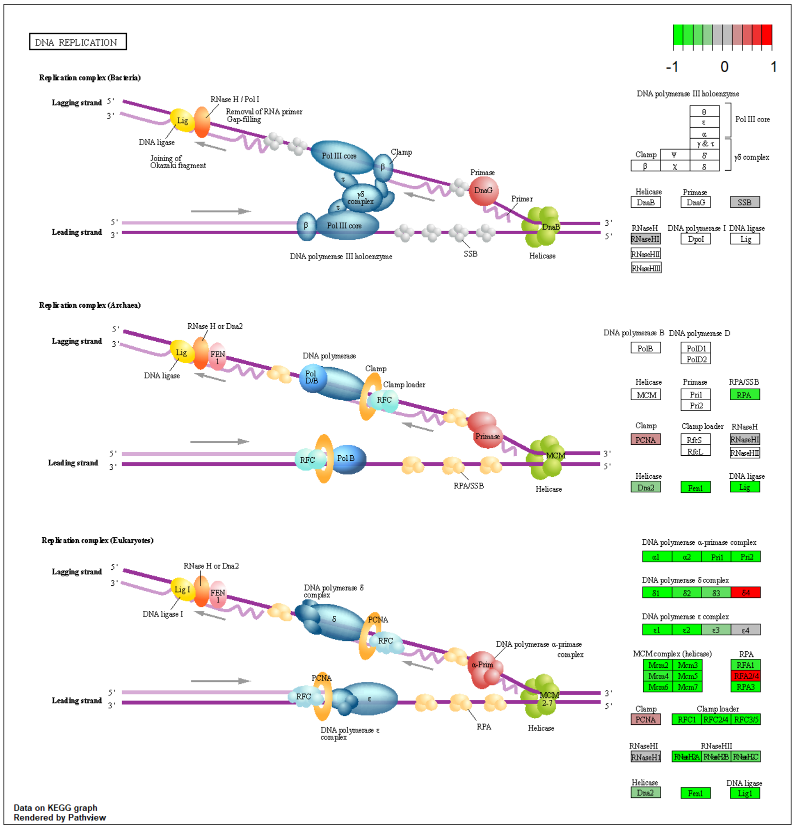
```
Info: Writing image file hsa04740.pathview.png
```

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

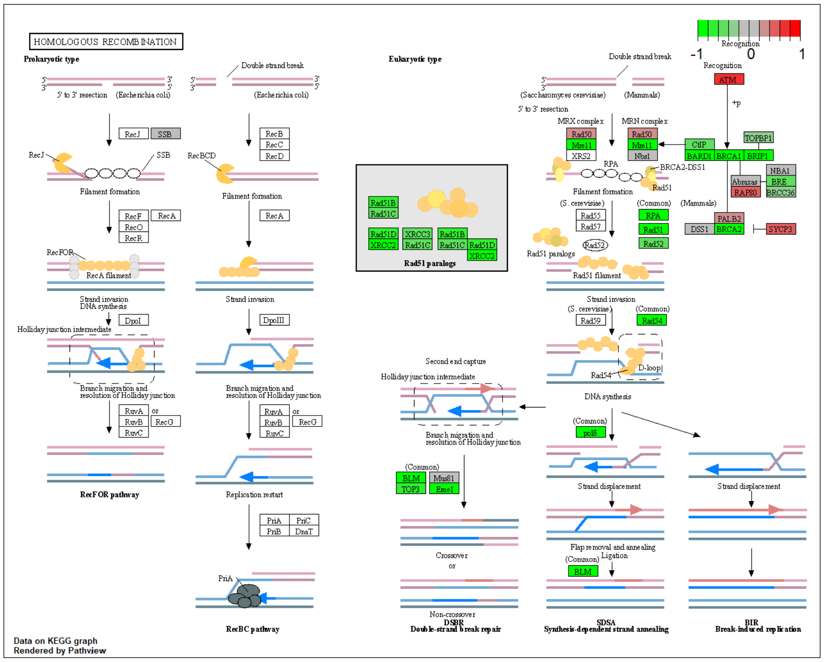
```
'select()' returned 1:1 mapping between keys and columns
```

```
Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/Users/miche/OneDrive/Desktop/class_13_miniproj

Info: Writing image file hsa04976.pathview.png
```

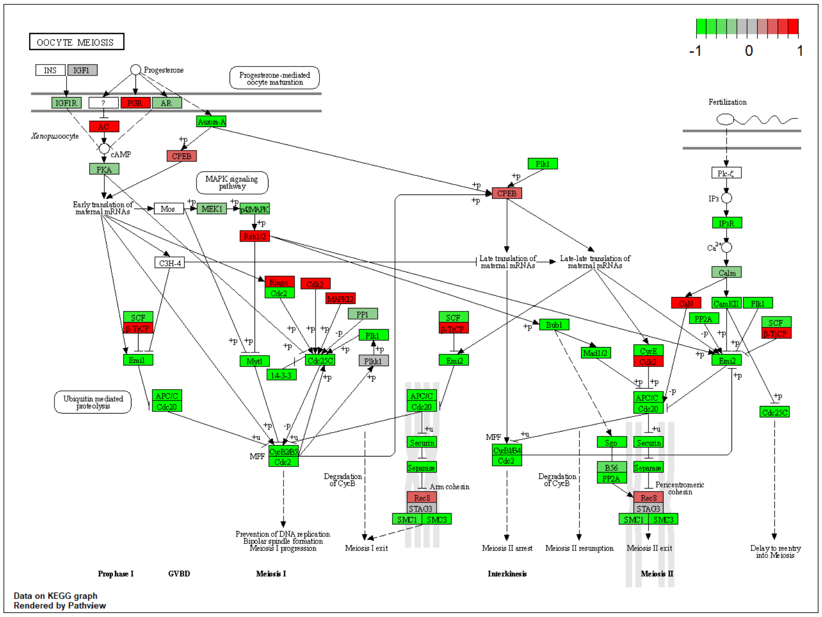


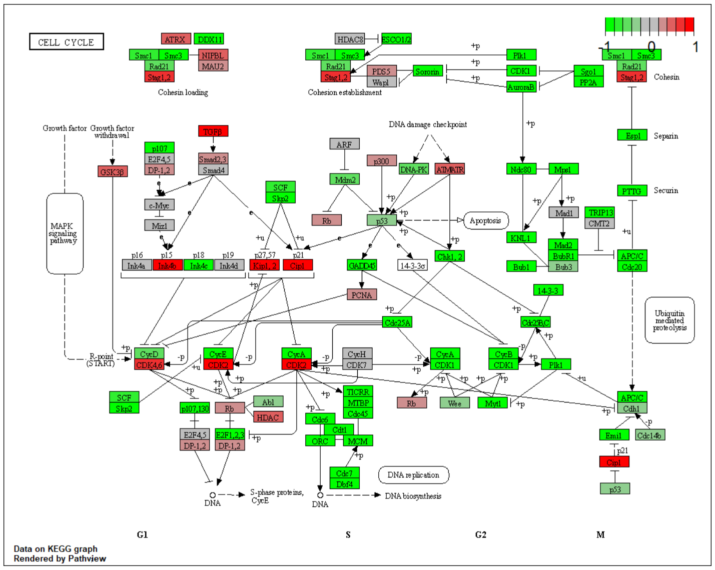


DNA REPLICATION
-1
0
1
Replication complex (Bacteria)
Lagging strand
Leading strand
RNase H / Pol I
Removal of RNA primer
Gap-filling
Lig
DNA ligase
Joining of Okazaki fragment
Pol III core
Clamp
β
τ
γδ complex
Primase
DnaG
Primer
DnaB
SSB
Helicase
DNA polymerase III holoenzyme
θ
ε
α
γ & τ
δ'
ψ
χ
δ
Pol III core
γδ complex
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
RNaseHI
RNaseHII
RNaseHIII
DNA polymerase I
DpoI
DNA ligase
Lig
Replication complex (Archaea)
RNase H or Dna2
FEN 1
DNA polymerase
Pol D/B
Clamp
Clamp loader
RFC
Primase
MCM
RPA/SSB
Pol B
DNA polymerase B
PolB
DNA polymerase D
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
RfcS
RfcL
RNaseH
RNaseHI
RNaseHII
Helicase
Dna2
Fen1
DNA ligase
Lig
Replication complex (Eukaryotes)
Lig I
DNA ligase I
DNA polymerase δ complex
δ
PCNA
RFC
DNA polymerase α-primase complex
α-Prim
MCM 2-7
ε
DNA polymerase ε complex
RPA
Helicase
DNA polymerase α-primase complex
α1
α2
Pri1
Pri2
DNA polymerase δ complex
δ1
δ2
δ3
δ4
DNA polymerase ε complex
ε1
ε2
ε3
ε4
MCM complex (helicase)
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
RPA
RFA1
RPA2/4
RPA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHI
RNaseHII
RNaseHI
RNaseHIA
RNaseHIB
RNaseHIC
Helicase
Dna2
Fen1
DNA ligase
Lig1
Data on KEGG graph
Rendered by Pathview



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





5. Gene Ontology (GO)

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. We will be focusing on BP (a.k.a Biological Process).

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

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

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

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             6.626774e-16 -8.170439 6.626774e-16
GO:0000280 nuclear division              1.797050e-15 -8.051200 1.797050e-15
GO:0007067 mitosis                       1.797050e-15 -8.051200 1.797050e-15
GO:0000087 M phase of mitotic cell cycle 4.757263e-15 -7.915080 4.757263e-15
GO:0007059 chromosome segregation        1.081862e-11 -6.974546 1.081862e-11
GO:0051301 cell division                 8.718528e-11 -6.455491 8.718528e-11
                                                q.val set.size         exp1
GO:0048285 organelle fission             2.620099e-12      386 6.626774e-16
GO:0000280 nuclear division              2.620099e-12      362 1.797050e-15
GO:0007067 mitosis                       2.620099e-12      362 1.797050e-15
GO:0000087 M phase of mitotic cell cycle 5.202068e-12      373 4.757263e-15
GO:0007059 chromosome segregation        9.464127e-09      146 1.081862e-11
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
```

6. Reactome Analysis

```
# Output list of signicant genes at 0.05 level
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)
```

Then we can perform a pathway analysis on the Reactome website

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?**

RHOBTB2 GTPase cycle has the most significant or least entities p-value of 1.83E-1. There are some similarities such as cell cycle and gene expression. The most significant pathways from Reactome is the signal transduction, a pathway from disease, one from gene expression, and from the cell cycle. Reactome is not very reliable or accurate given that the it a starting point in analyzing these pathways. The differences could be caused by experimental inaccuracies or even mistakes in the data file.