Lab 14 Thu 2.20 - RNA Seq Analysis Mini-Project

Jessica Le (PID: A17321021)

Table of contents

Background

The data for for 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.

Data Import

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

```
colData <- read.csv("GSE37704_metadata.csv")
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
```

Inspect and tidy data

Does the `counts` columns match the `colData` rows?

```
ncol(counts)
```

```
[1] 7
```

```
colData$id
```

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

```
nrow(colData)
```

```
[1] 6
```

```
names(colData)
```

```
[1] "id"        "condition"
```

No, the `counts` column does not match the `colData` rows. The fix here looks to be removing the first column from "counts".

Q1. Remove the troublesome first column from countData.

```
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
```

How many genes in total?

```
nrow(countData)
```

```
[1] 19808
```

Q2. Filter to remove zero count genes(rows where there are zero counts in all columns). How many genes are left?

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

```
[1] 15975
```

Setup for DESeq

```
#/ message: false
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
```

Setup input object for DESeq

```
library(DESeq2)
dds = DESeqDataSetFromMatrix(countData=countData,
                             colData=colData,
                             design=~condition)
```

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

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

```
estimating size factors

estimating dispersions
```

```
gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

Run DESeq

```
res=results(dds)
```

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.

```
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
```

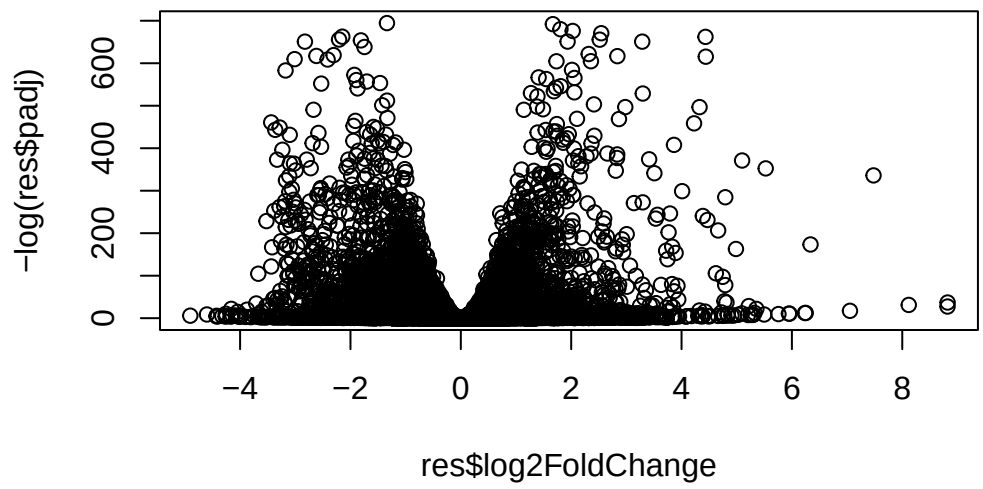
```
head(res)
```

```
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>
ENSG00000186092    0.0000             NA        NA        NA         NA
ENSG00000279928    0.0000             NA        NA        NA         NA
ENSG00000279457   29.9136       0.179257  0.324822  0.551863 0.58104205
ENSG00000278566    0.0000             NA        NA        NA         NA
```

```
ENSG00000273547    0.0000             NA        NA        NA          NA
ENSG00000187634  183.2296       0.426457  0.140266  3.040350 0.00236304
                      padj
                 <numeric>
ENSG00000186092         NA
ENSG00000279928         NA
ENSG00000279457 0.68707978
ENSG00000278566         NA
ENSG00000273547         NA
ENSG00000187634 0.00516278
```

Volcano plot of results

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

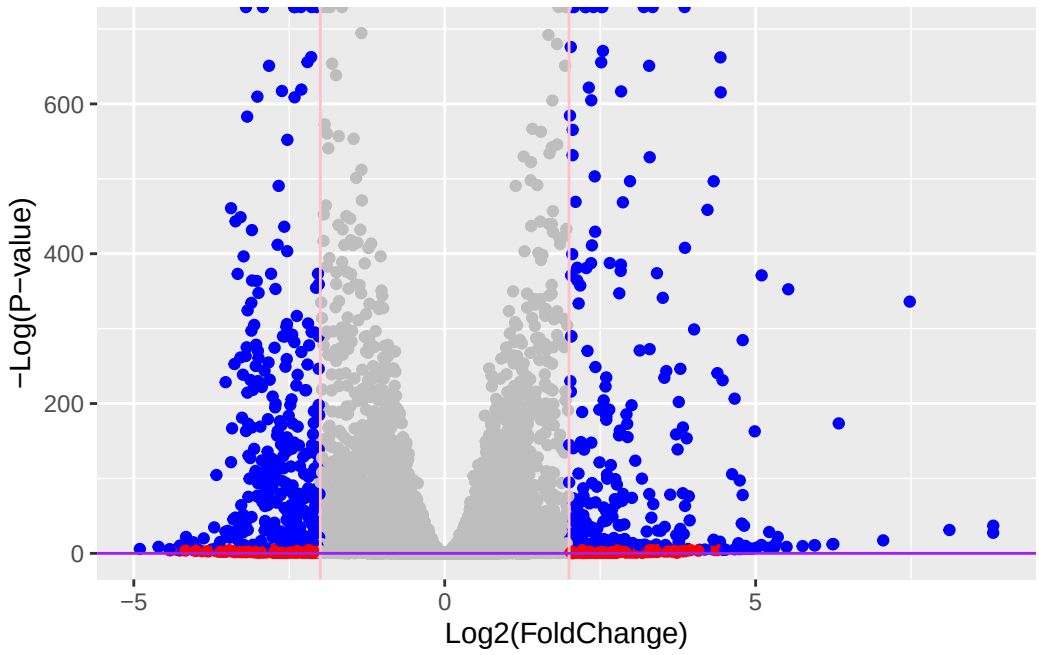


Q4. Improve this plot by adding color and axis labels.

```
mycols <- rep("gray", nrow(res))
mycols[abs(res$log2FoldChange) > 2] <- "red"
inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"
```

```
library(ggplot2)

ggplot(res) +
  aes(log2FoldChange, -log(padj)) +
  geom_point(col=mycols) +
  geom_vline(xintercept=c(-2,2), col="pink") +
  geom_hline(yintercept=0.01, col="purple") +
  labs(x="Log2(FoldChange)", y="-Log(P-value)")
```

```
Warning: Removed 5054 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Gene annotation

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

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

```
res$symbol <- mapIds(org.Hs.eg.db,
                     key=rownames(res),
                     keytype="ENSEMBL",
                     column="SYMBOL")
```

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

```
res$genename <- mapIds(org.Hs.eg.db,
                     key=rownames(res),
                     keytype="ENSEMBL",
                     column="GENENAME")
```

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

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

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

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

```
log2 fold change (MLE): condition hoxa1 kd vs control sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 10 rows and 9 columns
                 baseMean log2FoldChange     lfcSE       stat      pvalue
                <numeric>      <numeric> <numeric>  <numeric>   <numeric>
ENSG00000186092    0.0000             NA        NA         NA          NA
ENSG00000279928    0.0000             NA        NA         NA          NA
ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
ENSG00000278566    0.0000             NA        NA         NA          NA
ENSG00000273547    0.0000             NA        NA         NA          NA
ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43989e-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               genename      entrez
                <numeric> <character>            <character> <character>
ENSG00000186092          NA       OR4F5 olfactory receptor f..       79501
ENSG00000279928          NA          NA                      NA          NA
ENSG00000279457 6.87080e-01          NA                      NA          NA
ENSG00000278566          NA          NA                      NA          NA
ENSG00000273547          NA          NA                      NA          NA
ENSG00000187634 5.16278e-03      SAMD11 sterile alpha motif ..      148398
ENSG00000188976 1.76740e-35       NOC2L NOC2 like nucleolar ..       26155
ENSG00000187961 1.13536e-07      KLHL17 kelch like family me..      339451
ENSG00000187583 9.18988e-01     PLEKHN1 pleckstrin homology ..       84069
ENSG00000187642 4.03817e-01       PERM1 PPARGC1 and ESRR ind..       84808
```

Q7. 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.

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

Pathway Analysis

```
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)
```

```r
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"
```

```
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
```

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

Now lets look at the 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              1.160132e-03 -3.080629 1.160132e-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              0.063420543      149 1.160132e-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
```

Now, let's try out the pathview() function from the pathview package to make a pathway plot with our RNA-Seq expression results shown in color.

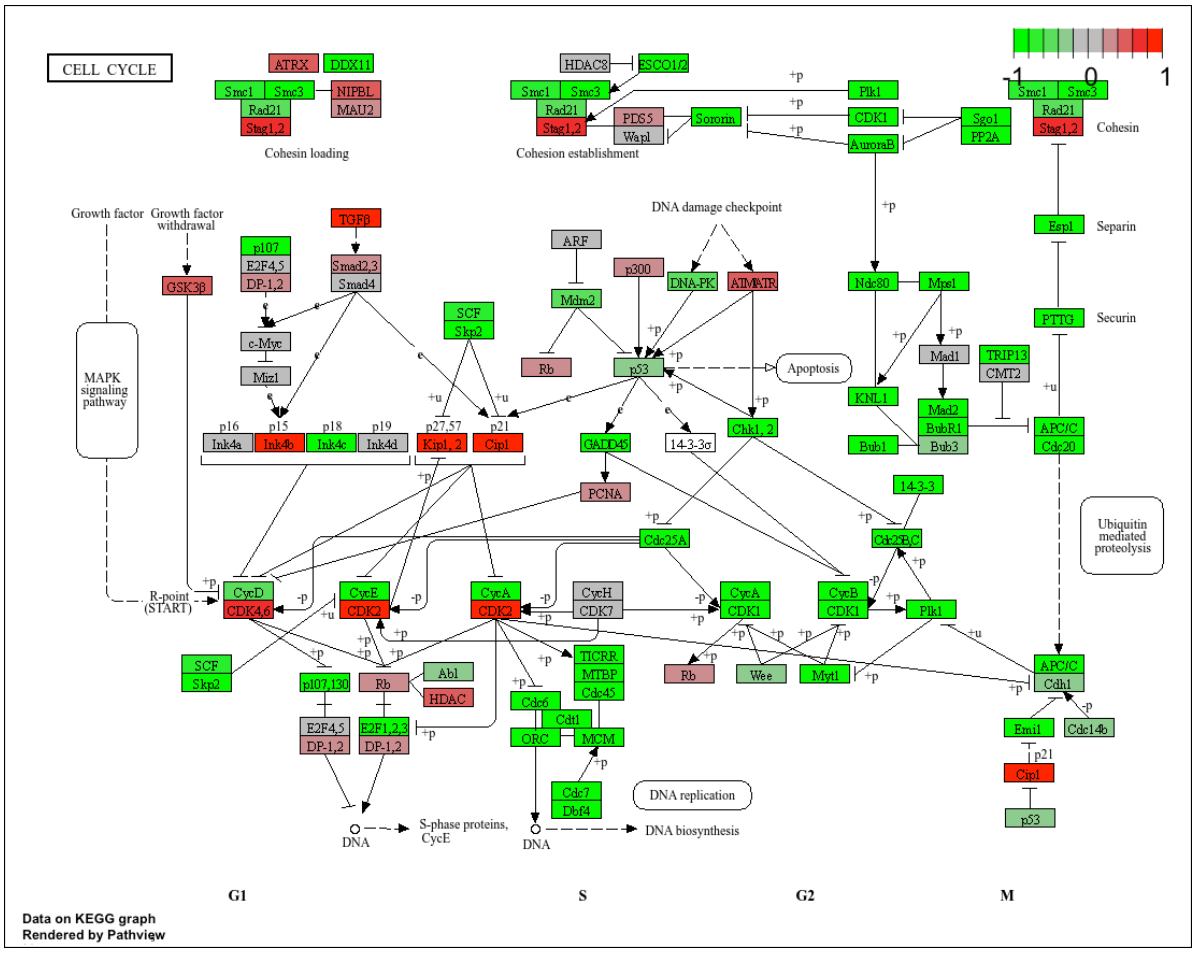
To begin with lets manually supply a pathway.id (namely the first part of the "hsa04110 Cell cycle") that we could see from the print out above.

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

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

```
Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20
```

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



You can play with the other input arguments to pathview() to change the display in various ways including generating a PDF graph. For example:

```
# 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!
```

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

```
Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20
```

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

Now, let's process our results a bit more to automagicaly pull out the top 5 upregulated pathways, then further process that just to get the pathway IDs needed by the pathview() function. We'll use these KEGG pathway IDs for pathview plotting below.

```
## 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"
```

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

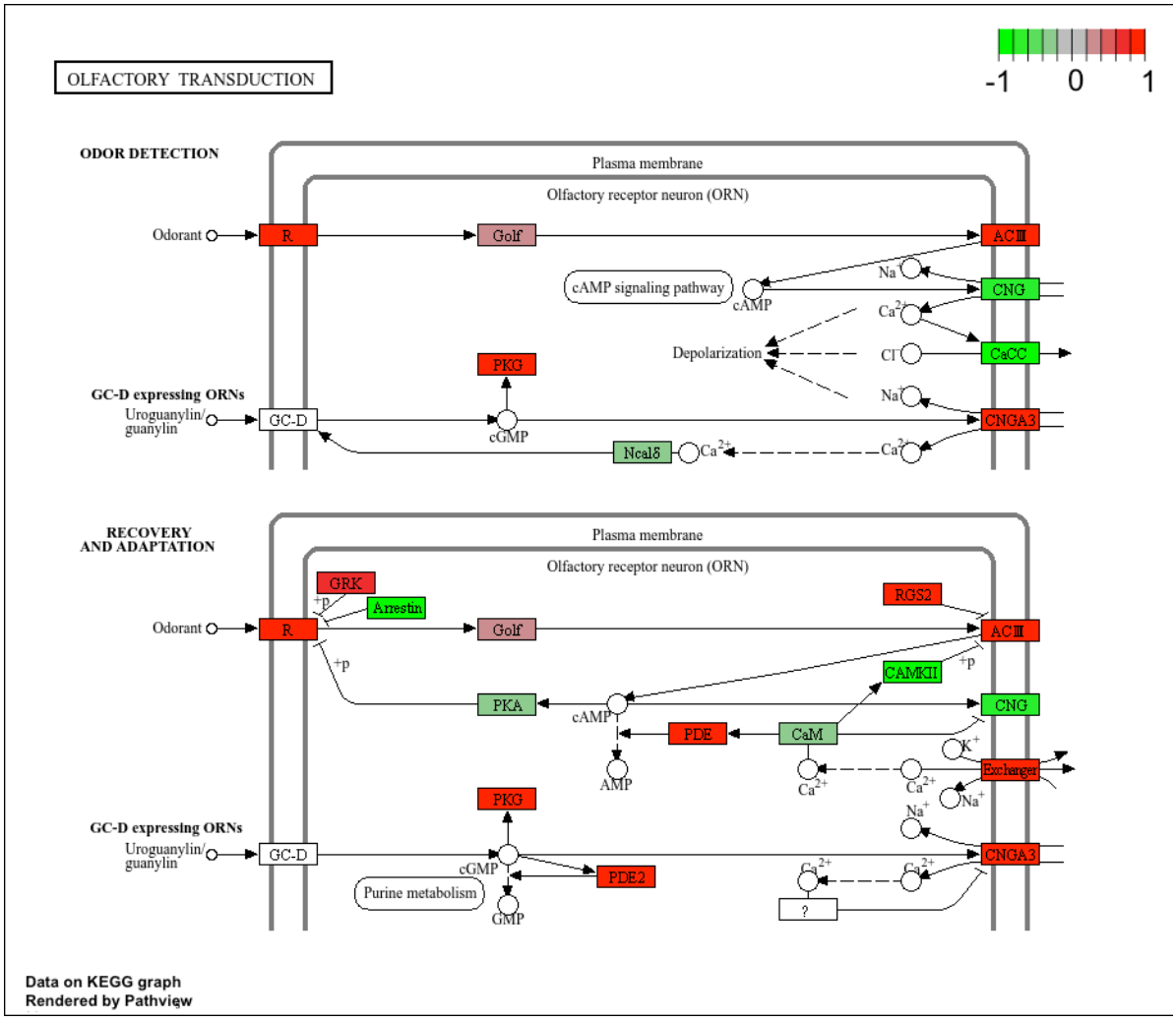
```
'select()' returned 1:1 mapping between keys and columns
```

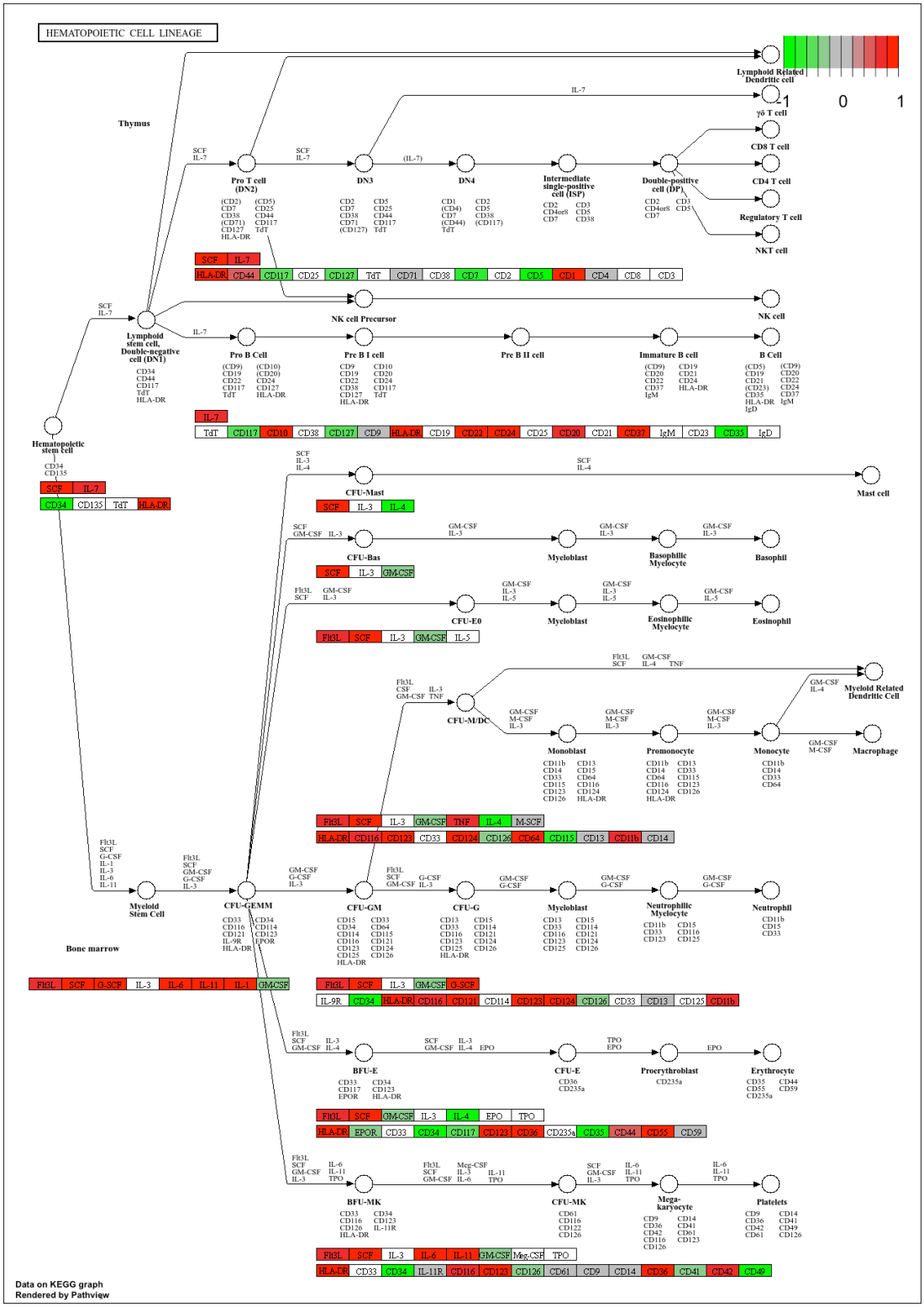
```
Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20
```

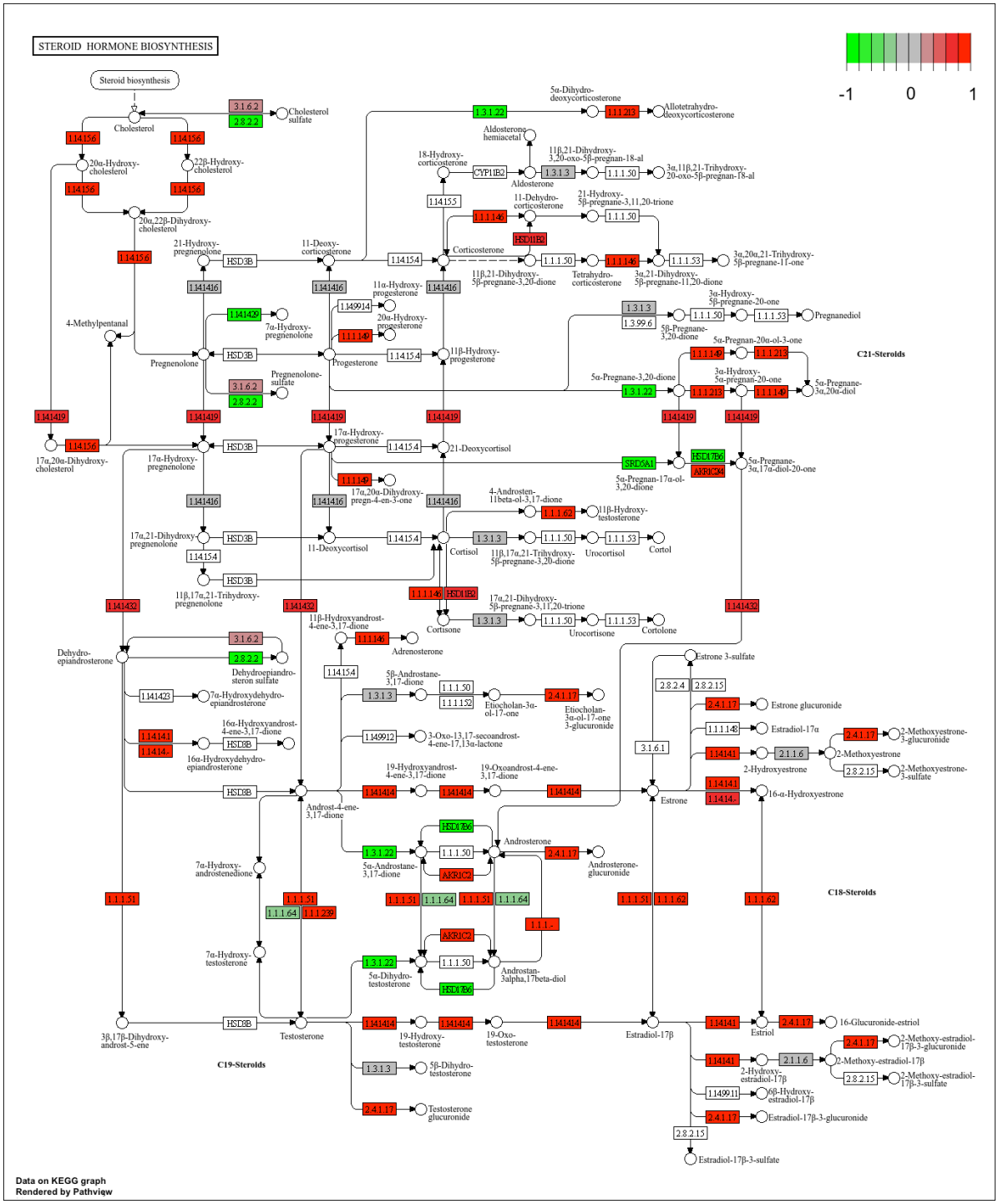
```
Info: Writing image file hsa04740.pathview.png
```

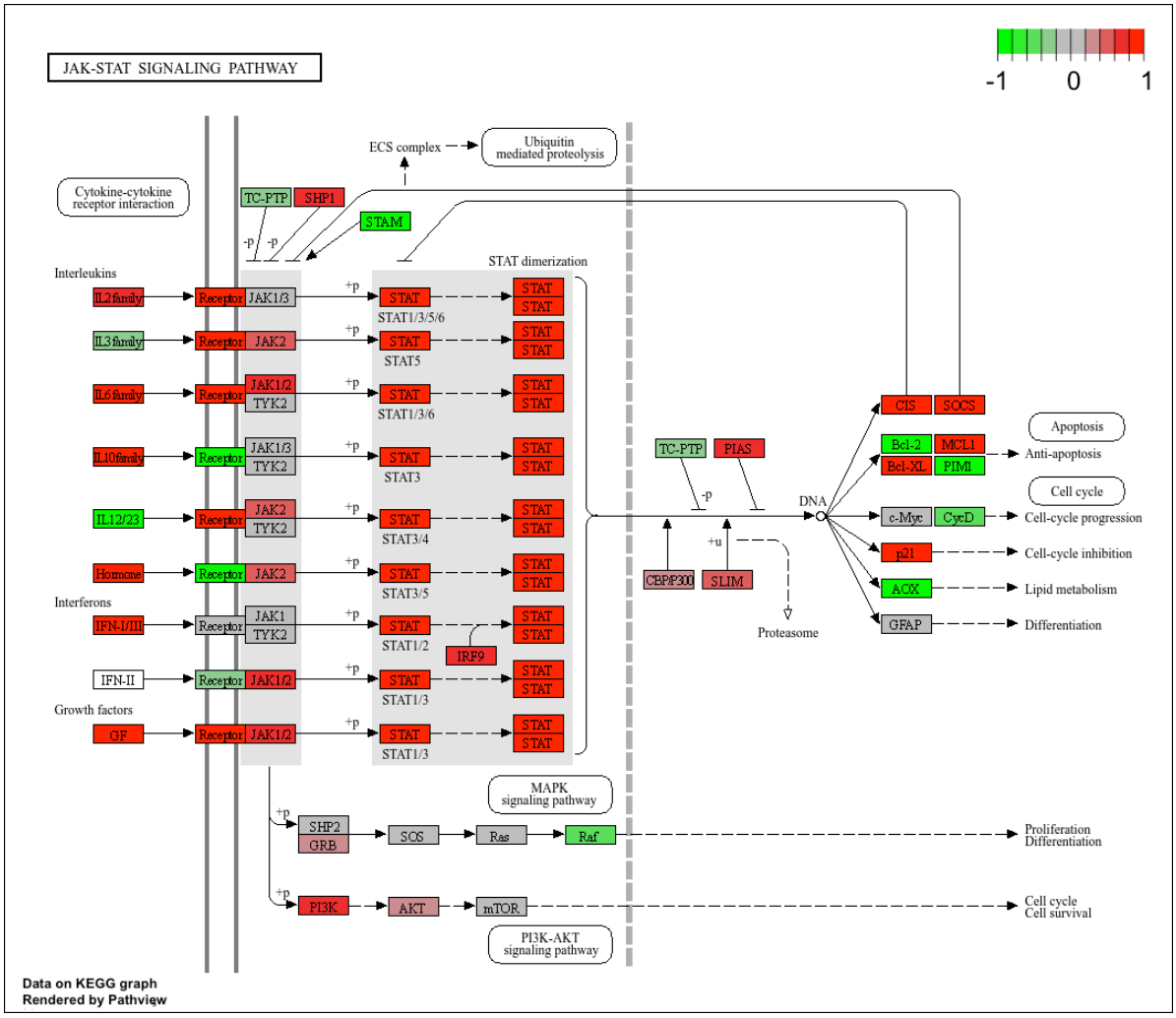
```
'select()' returned 1:1 mapping between keys and columns
```

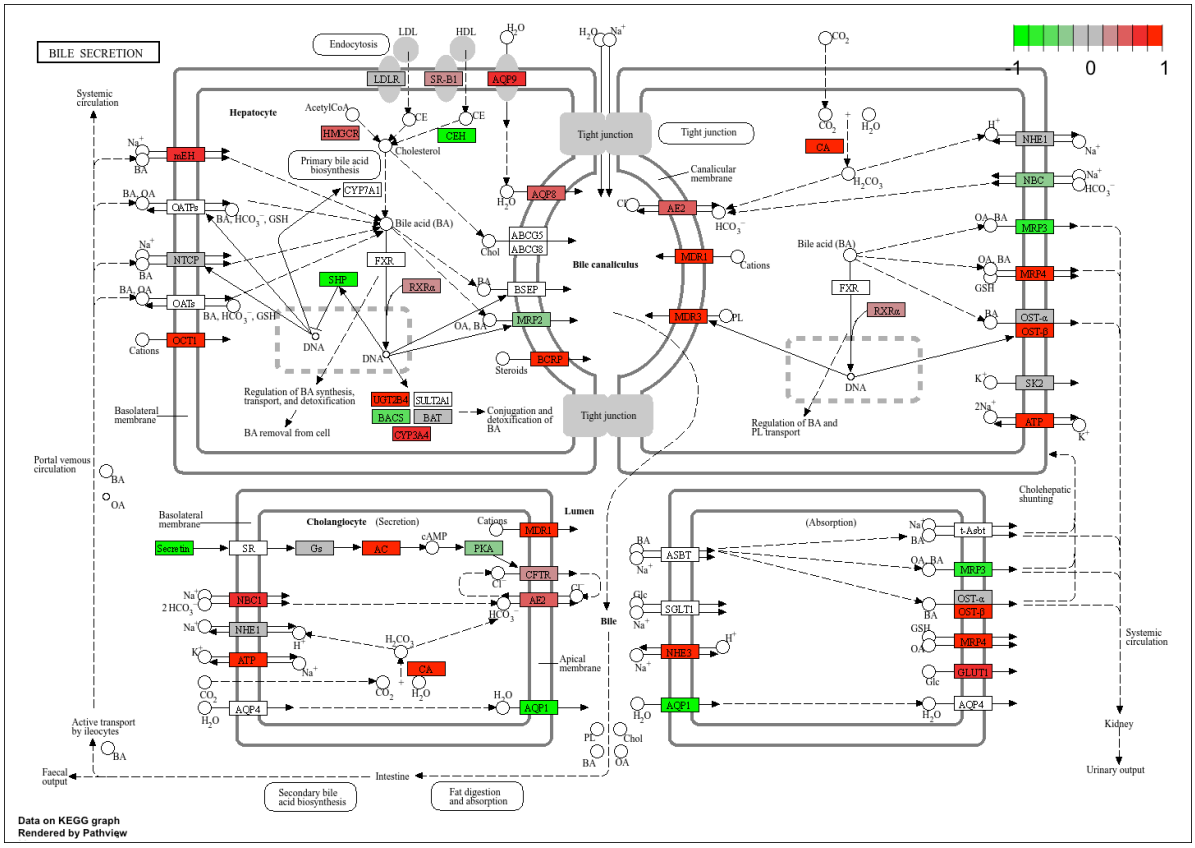
```
Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

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


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


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


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


JAK-STAT SIGNALING PATHWAY
-1
0
1
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
TC-PTP
SHP1
STAM
-p
-p
STAT dimerization
Interleukins
IL2 family
IL3 family
IL6 family
IL10 family
IL12/23
Hormone
Interferons
IFN-I/III
IFN-II
Growth factors
GF
Receptor
JAK1/3
JAK2
JAK1/2
TYK2
JAK1
+p
STAT
STAT1/3/5/6
STAT5
STAT1/3/6
STAT3
STAT3/4
STAT3/5
STAT1/2
STAT1/3
IRF9
TC-PTP
PIAS
-p
+u
CBP/P300
SLIM
Proteasome
DNA
CIS
SOCS
Bcl-2
MCL1
Bcl-XL
PIM1
c-Myc
CycD
p21
AOX
GFAP
Apoptosis
Anti-apoptosis
Cell cycle
Cell-cycle progression
Cell-cycle inhibition
Lipid metabolism
Differentiation
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



> Q8. Can you do the same procedure as above to plot the pathview figures for the top 5 down-reguled pathways?

```
keggrespathways.down <- rownames(keggres$less)[1:5]
keggresids = substr(keggrespathways.down, start=1, stop=8)
keggresids
```

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

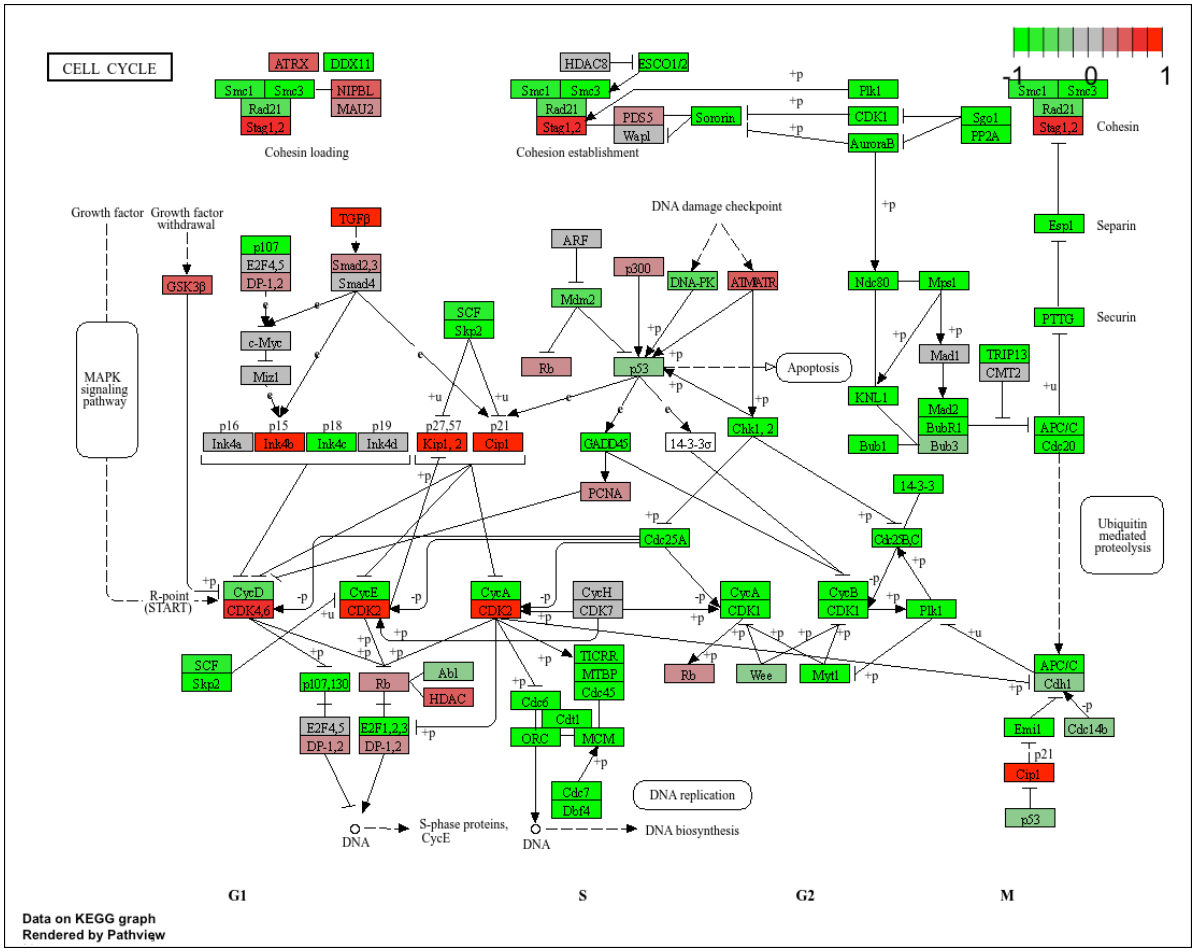
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

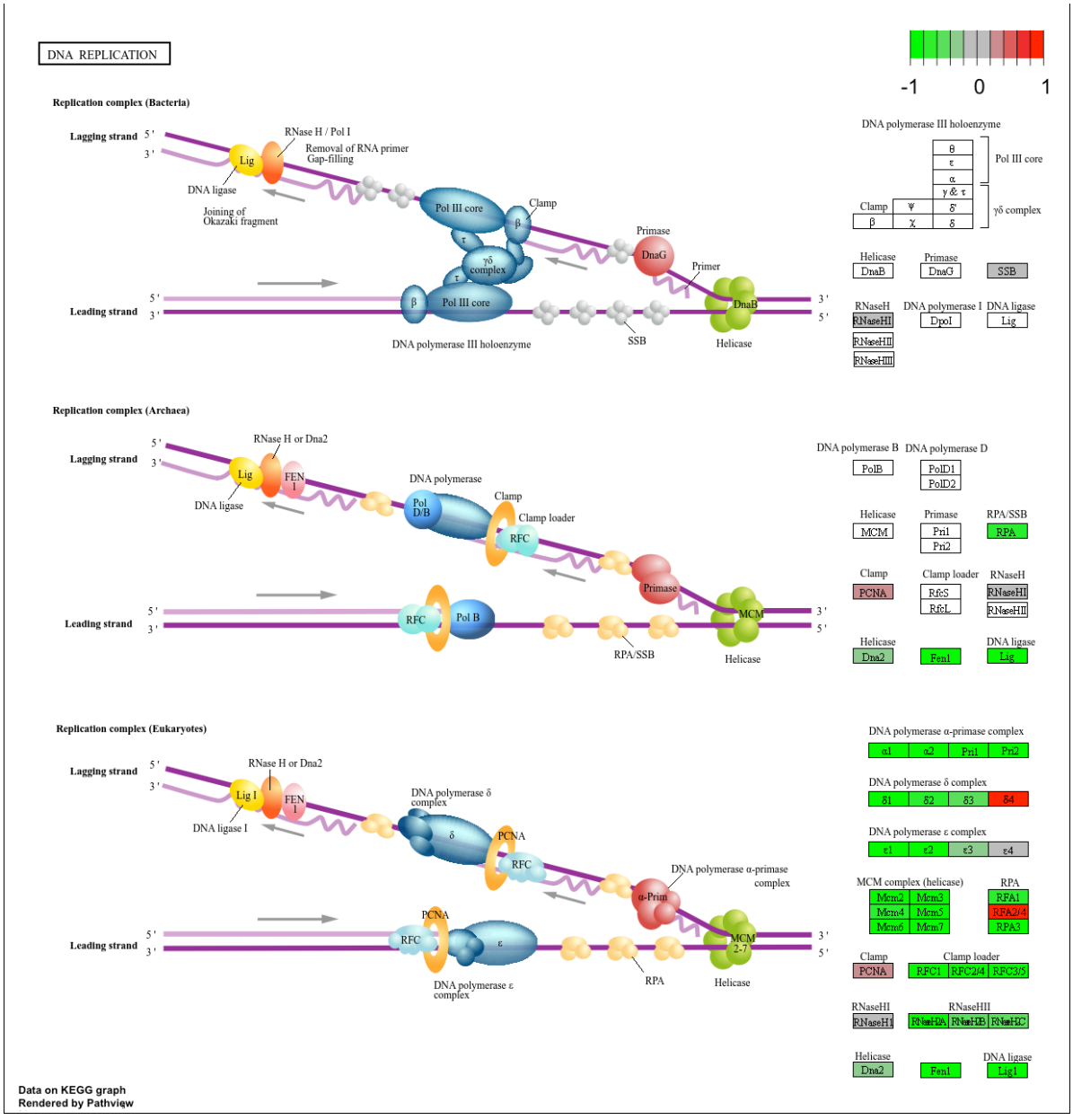
```
'select()' returned 1:1 mapping between keys and columns
```

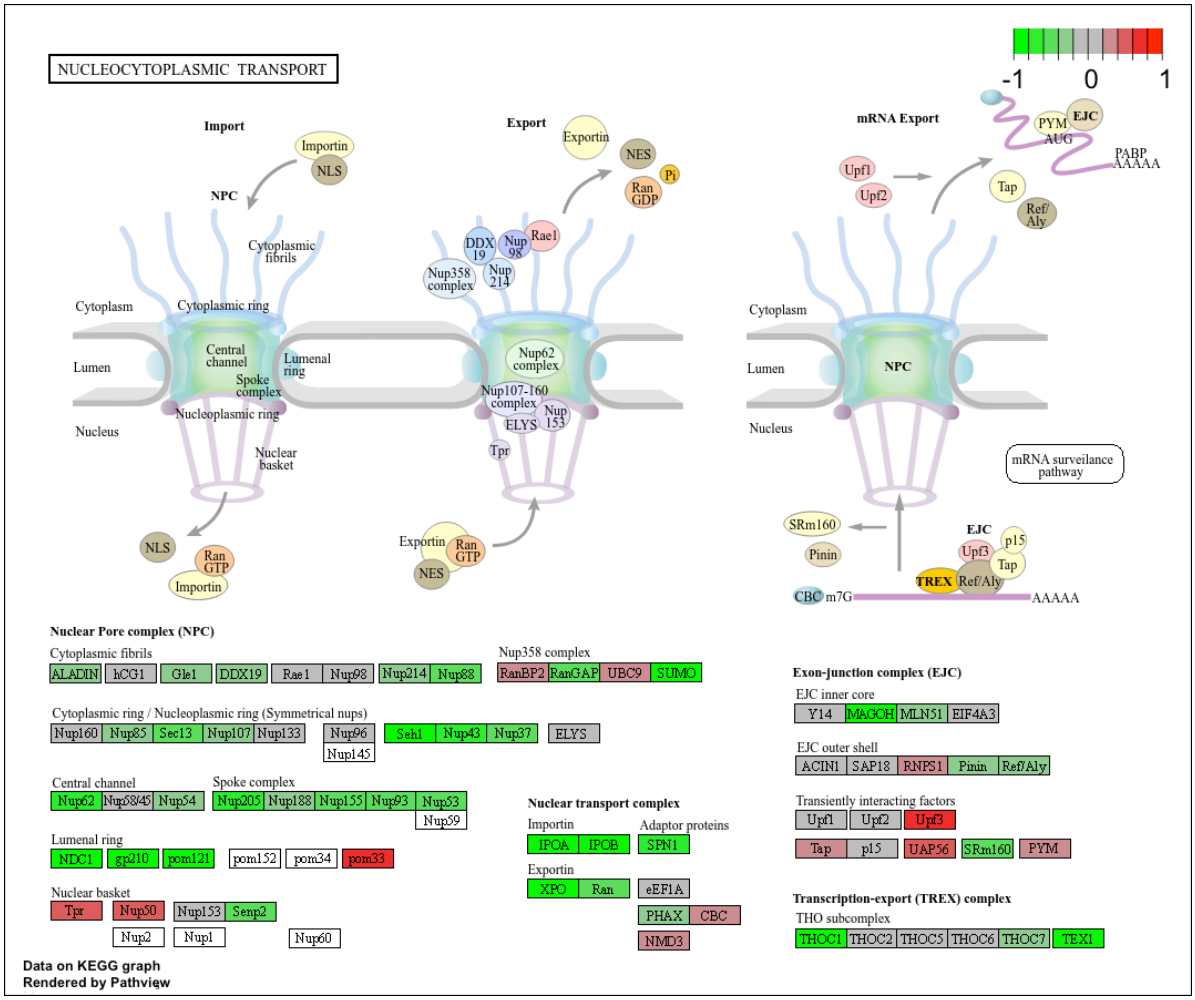
```
Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20
```

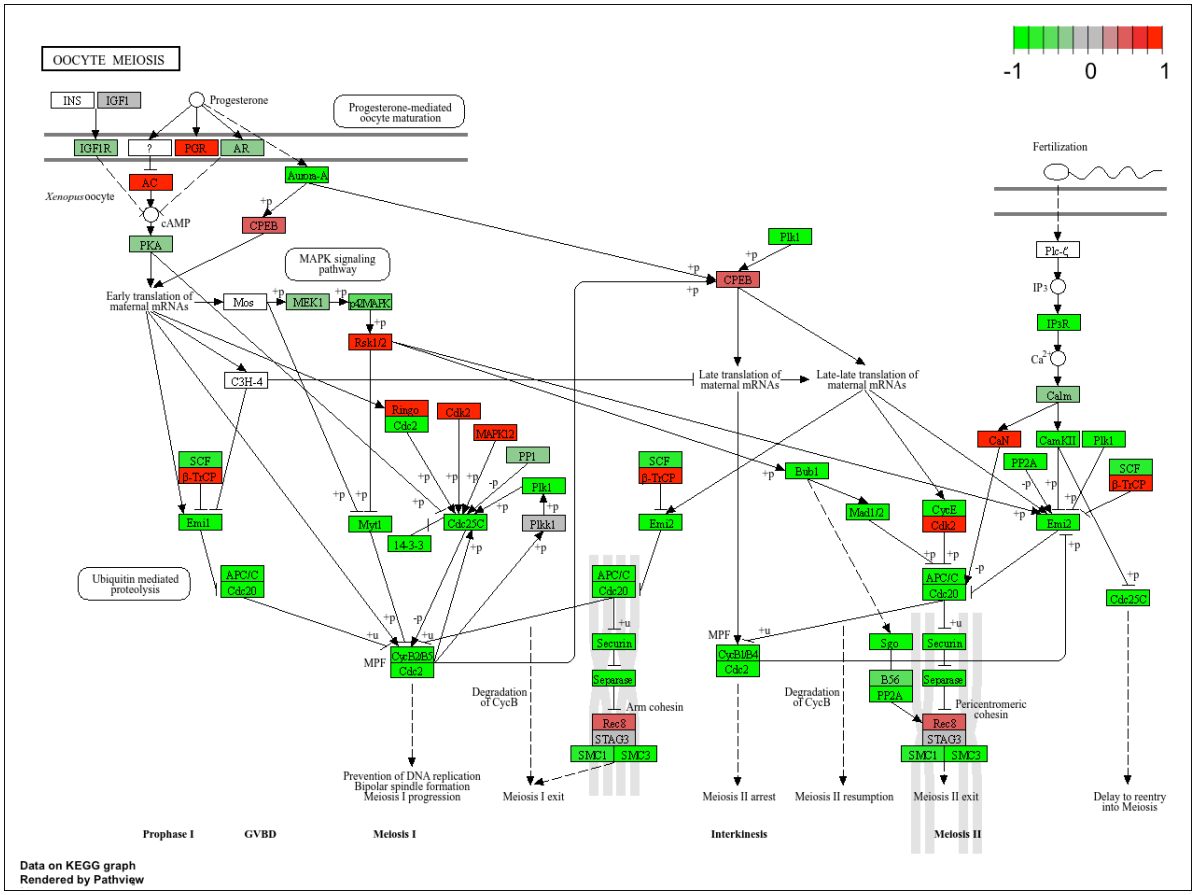
```
Info: Writing image file hsa04110.pathview.png
```

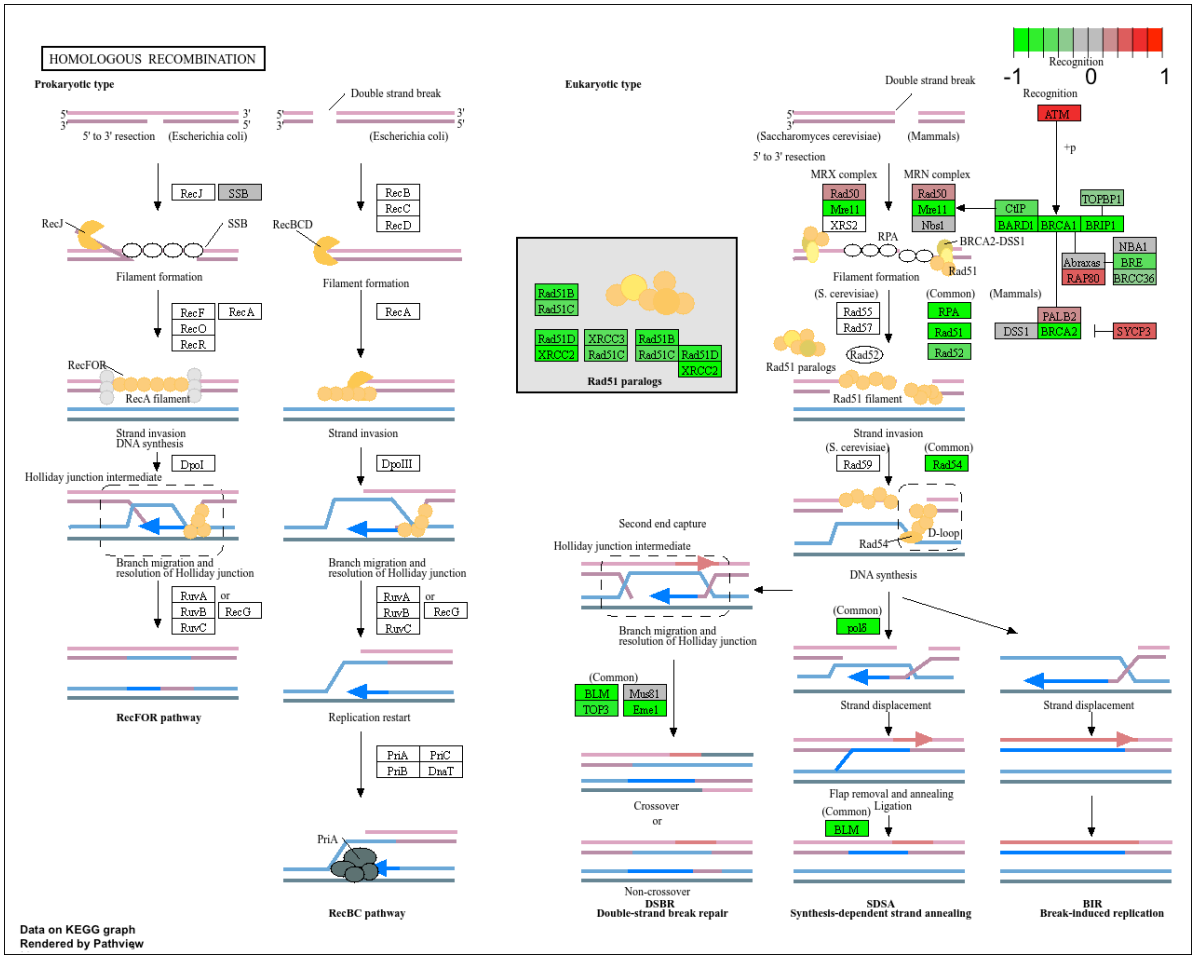
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

Info: Writing image file hsa04114.pathview.png

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

Info: Working in directory /Users/jessica/Documents/BIMM143/Lab 14 - Thu 2.20

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


CELL CYCLE
-1
0
1
ATRX
DDX11
Smc1
Smc3
NIPBL
MAU2
Rad21
Stag1,2
Cohesin loading
HDAC8
ESCO1/2
Smc1
Smc3
Rad21
Stag1,2
PDS5
Wapl
Sororin
Cohesion establishment
Plk1
CDK1
AuroraB
Sgo1
PP2A
Smc1
Smc3
Rad21
Stag1,2
Cohesin
Espl1
Separin
PTTG
Securin
Growth factor
Growth factor withdrawal
TGFβ
p107
E2F4,5
DP-1,2
Smad2,3
Smad4
GSK3β
c-Myc
Miz1
MAPK signaling pathway
SCF
Skp2
ARF
Mdm2
Rb
p53
p300
DNA-PK
ATM/ATR
DNA damage checkpoint
Apoptosis
Ndc80
Mps1
KNL1
Mad1
Mad2
BubR1
Bub1
Bub3
TRIP13
CMT2
APC/C
Cdc20
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Kip1, 2
p21
Cip1
GADD45
14-3-3σ
Chk1, 2
PCNA
14-3-3
Cdc25A
Cdc25B,C
Ubiquitin mediated proteolysis
R-point
(START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plk1
SCF
Skp2
p107,130
Rb
Abl
HDAC
TICRR
MTBP
Cdc45
Cdc6
Cdt1
ORC
MCM
Cdc7
Dbf4
Rb
Wee
Myt1
APC/C
Cdh1
Emi1
Cdc14b
p21
Cip1
p53
E2F4,5
DP-1,2
E2F1,2,3
DP-1,2
DNA
S-phase proteins, CycE
DNA
DNA replication
DNA biosynthesis
G1
S
G2
M
Data on KEGG graph
Rendered by Pathview


DNA REPLICATION
-1
0
1
Replication complex (Bacteria)
Lagging strand
Leading strand
RNase H / Pol I
Removal of RNA primer
Gap-filling
Lig
DNA ligase
Joining of
Okazaki fragment
Pol III core
Clamp
β
τ
γδ
complex
Primase
DnaG
Primer
DnaB
SSB
Helicase
DNA polymerase III holoenzyme
DNA polymerase III holoenzyme
θ
ε
α
γ & τ
Pol III core
Clamp
ψ
δ'
γδ complex
β
χ
δ
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
RNaseHI
RNaseHII
RNaseHIII
DNA polymerase I
DpoI
DNA ligase
Lig
Replication complex (Archaea)
RNase H or Dna2
Lig
FEN
1
DNA ligase
DNA polymerase
Pol
D/B
Clamp
Clamp loader
RFC
Primase
MCM
RPA/SSB
Helicase
DNA polymerase B
PolB
DNA polymerase D
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
Rfc-S
Rfc-L
RNaseH
RNaseHI
RNaseHII
Helicase
Dna2
Fen1
DNA ligase
Lig
Replication complex (Eukaryotes)
RNase H or Dna2
Lig I
FEN
1
DNA ligase I
DNA polymerase δ
complex
δ
PCNA
RFC
DNA polymerase α-primase
complex
α-Prim
MCM
2-7
RPA
Helicase
DNA polymerase ε
complex
ε
DNA polymerase α-primase complex
α1
α2
Pri1
Pri2
DNA polymerase δ complex
δ1
δ2
δ3
δ4
DNA polymerase ε complex
ε1
ε2
ε3
ε4
MCM complex (helicase)
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
RPA
RFA1
RFA2/4
RFA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHI
RNaseH1
RNaseHII
RNaseHIII
Helicase
Dna2
Fen1
DNA ligase
Lig1
Data on KEGG graph
Rendered by Pathview


NUCLEOCYTOPLASMIC TRANSPORT
-1
0
1
Import
Importin
NLS
NPC
Cytoplasmic fibrils
Cytoplasm
Cytoplasmic ring
Central channel
Lumenal ring
Lumen
Spoke complex
Nucleoplasmic ring
Nucleus
Nuclear basket
NLS
Ran GTP
Importin
Export
Exportin
NES
Pi
Ran GDP
DDX 19
Nup 98
Rae1
Nup358 complex
Nup 214
Nup62 complex
Nup107-160 complex
ELYS
Nup 153
Tpr
Exportin
Ran GTP
NES
mRNA Export
EJC
PYM
AUG
PABP
AAAAA
Upf1
Upf2
Tap
Ref/ Aly
Cytoplasm
Lumen
NPC
Nucleus
mRNA surveilance pathway
SRm160
Pinin
EJC
p15
Upf3
Tap
TREX
Ref/Aly
CBC
m7G
AAAAA
Nuclear Pore complex (NPC)
Cytoplasmic fibrils
ALADIN
hCG1
Gle1
DDX19
Rae1
Nup98
Nup214
Nup88
Nup358 complex
RanBP2
RanGAP
UBC9
SUMO
Cytoplasmic ring / Nucleoplasmic ring (Symmetrical nups)
Nup160
Nup85
Sec13
Nup107
Nup133
Nup96
Nup145
Seh1
Nup43
Nup37
ELYS
Central channel
Nup62
Nup58/45
Nup54
Spoke complex
Nup205
Nup188
Nup155
Nup93
Nup53
Nup59
Lumenal ring
NDC1
gp210
pom121
pom152
pom34
pom33
Nuclear basket
Tpr
Nup50
Nup153
Senp2
Nup2
Nup1
Nup60
Nuclear transport complex
Importin
IPOA
IPOB
Adaptor proteins
SPN1
Exportin
XPO
Ran
eEF1A
PHAX
CBC
NMD3
Exon-junction complex (EJC)
EJC inner core
Y14
MAGOH
MLN51
EIF4A3
EJC outer shell
ACIN1
SAP18
RNPS1
Pinin
Ref/Aly
Transiently interacting factors
Upf1
Upf2
Upf3
Tap
p15
UAP56
SRm160
PYM
Transcription-export (TREX) complex
THO subcomplex
THOC1
THOC2
THOC5
THOC6
THOC7
TEX1
Data on KEGG graph
Rendered by Pathview


OOCYTE MEIOSIS
-1
0
1
INS
IGF1
Progesterone
Progesterone-mediated oocyte maturation
IGF1R
?
PGR
AR
AC
Xenopus oocyte
cAMP
PKA
Aurora-A
CPEB
MAPK signaling pathway
Early translation of maternal mRNAs
Mos
MEK1
p-MAPK
Rsk1/2
C3H-4
Ringo
Cdk2
Cdc2
MAPK12
PP1
Plk1
Plkk1
SCF
β-TrCP
Emi1
Myt1
14-3-3
Cdc25C
Ubiquitin mediated proteolysis
APC/C
Cdc20
MPF
CycB2/B5
Cdc2
Degradation of CycB
Securin
Separase
Arm cohesin
Rec8
STAG3
SMC1
SMC3
Prevention of DNA replication
Bipolar spindle formation
Meiosis I progression
Meiosis I exit
Late translation of maternal mRNAs
Late-late translation of maternal mRNAs
Emi2
Btrb1
Mad1/2
CycE
Cdk2
CycB/B4
Cdc2
Sgo
B56
PP2A
Pericentromeric cohesin
Meiosis II arrest
Meiosis II resumption
Meiosis II exit
Fertilization
Plc-ζ
IP3
IP3R
Ca2+
Calm
CaN
CamKII
Delay to reentry into Meiosis
Prophase I
GVBD
Meiosis I
Interkinesis
Meiosis II
Data on KEGG graph
Rendered by Pathview



Gene Ontology (GO)

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

```
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        1.734864e-05  4.210777 1.734864e-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        0.07584825      137 1.734864e-05
GO:0048729 tissue morphogenesis            0.08347021      483 5.407952e-05
GO:0002009 morphogenesis of an epithelium  0.08347021      382 5.727599e-05
GO:0030855 epithelial cell differentiation 0.16449701      299 2.053700e-04
GO:0060562 epithelial tube morphogenesis   0.16449701      289 2.927804e-04
GO:0048598 embryonic morphogenesis         0.16449701      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               2.618901e-12      386 6.626774e-16
GO:0000280 nuclear division                2.618901e-12      362 1.797050e-15
GO:0007067 mitosis                         2.618901e-12      362 1.797050e-15
GO:0000087 M phase of mitotic cell cycle   5.199689e-12      373 4.757263e-15
GO:0007059 chromosome segregation          9.459800e-09      146 1.081862e-11
GO:0051301 cell division                   6.352901e-08      479 8.718528e-11

$stats
                                           stat.mean     exp1
GO:0007156 homophilic cell adhesion         4.210777 4.210777
GO:0048729 tissue morphogenesis             3.888470 3.888470
GO:0002009 morphogenesis of an epithelium   3.878706 3.878706
GO:0030855 epithelial cell differentiation  3.554776 3.554776
GO:0060562 epithelial tube morphogenesis    3.458463 3.458463
```

```
GO:0048598 embryonic morphogenesis          3.446527 3.446527
```

Reactome Analysis

```
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)
```

> Q9 and 10. What pathway has the most significant "Entities p-value"? Do the most significant pathways listed match your previous KEGG results? What factors could cause differences between the two methods?

The most significant pathway is Cell Cycle, Mitotic with an "Entities p-value" of 1.69e-4. Yes, the most significant pathways listed using the website do match the previous KEGG results since the top result for the KEGG results was also the cell cycle, but there is a different p-value of 7.08e-6. The difference in the two methods is that gene ontology is a more standardized compared to KEGG which provides a deeper analysis of gene function and interaction. In other words, KEGG considers how genes interact within complex biological pathways instead of only considering gene function at a basic level like GO.