RNAseq Mini Project

Jordan Prych (A17080226)

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

```
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
```

```
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
```

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

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

Inspect and Tidy Data

Does the `counts` column data match the `colData` rows?

No, there is an extra column, legnth

```
head(counts)
```

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

```
ENSG00000278566         0
ENSG00000273547         0
ENSG00000187634       258
```

```
head(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
```

```
metadata$id
```

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

```
colnames(counts)
```

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

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

The fix here looks to be removing the first "legnth" column from counts:

```
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) == metadata$id
```

```
[1] TRUE TRUE TRUE TRUE TRUE TRUE
```

Q. How many genes in total?

```
nrow(countData)
```

```
[1] 19808
```

Q. 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).

```
countData = countData[rowSums(countData)>0,]
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

Setup for DESeq

```
library(DESeq2)

dds = DESeqDataSetFromMatrix(countData=countData, 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
```

Running DEseq

```
dds
```

```
class: DESeqDataSet
dim: 15975 6
metadata(1): version
assays(4): counts mu H cooks
rownames(15975): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
  ENSG00000271254
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(3): id condition sizeFactor
```

```
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)     : 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
```

```
res
```

```
log2 fold change (MLE): condition hoxa1 kd vs control sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 15975 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 1.43990e-36
ENSG00000187961  209.6379      0.7297556 0.1318599   5.534326 3.12428e-08
ENSG00000187583   47.2551      0.0405765 0.2718928   0.149237 8.81366e-01
...                   ...            ...       ...        ...          ...
ENSG00000273748  35.30265       0.674387  0.303666   2.220817 2.63633e-02
ENSG00000278817   2.42302      -0.388988  1.130394  -0.344117 7.30758e-01
ENSG00000278384   1.10180       0.332991  1.660261   0.200565 8.41039e-01
ENSG00000276345  73.64496      -0.356181  0.207716  -1.714752 8.63908e-02
ENSG00000271254 181.59590      -0.609667  0.141320  -4.314071 1.60276e-05
                      padj
                 <numeric>
ENSG00000279457 6.86555e-01
ENSG00000187634 5.15718e-03
ENSG00000188976 1.76549e-35
ENSG00000187961 1.13413e-07
ENSG00000187583 9.19031e-01
...                     ...
ENSG00000273748 4.79091e-02
ENSG00000278817 8.09772e-01
ENSG00000278384 8.92654e-01
ENSG00000276345 1.39762e-01
ENSG00000271254 4.53648e-05
```

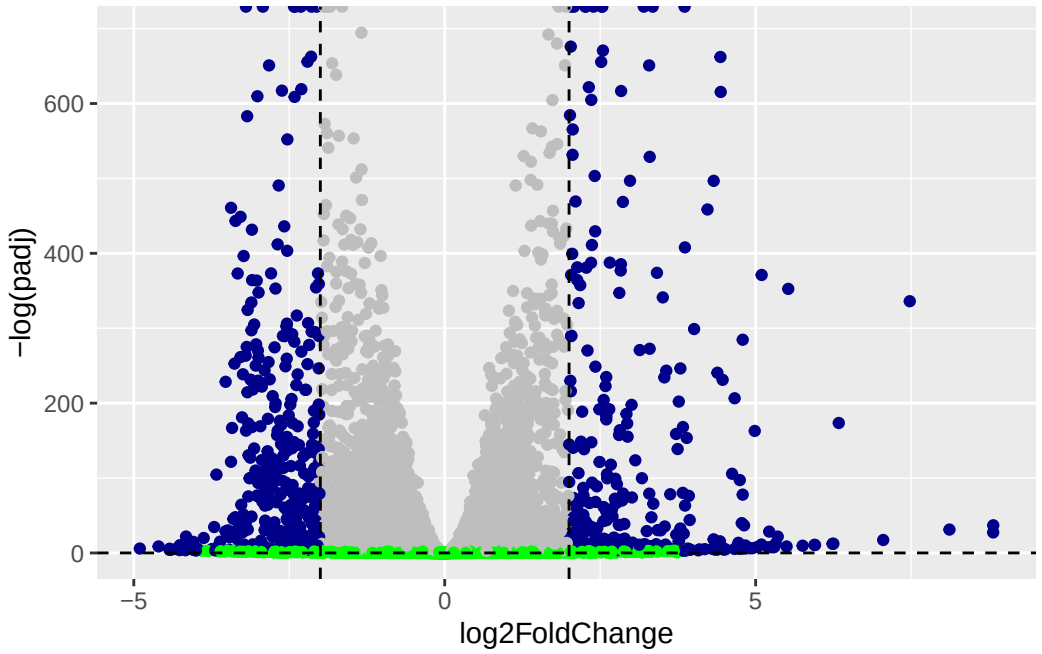
Volcano Plot of Results

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

```
#Color Vectors
mycols <- rep("gray", nrow(res))
mycols[res$log2FoldChange >=2] <- "darkblue"
mycols[res$log2FoldChange <= -2] <- "darkblue"
```

```
mycols[res$padj >0.05] <- "green"

library(ggplot2)

ggplot(res) + aes(log2FoldChange, -log(padj)) + geom_point(col=mycols) + geom_vline(xintercep
```

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



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"
```

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

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

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

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

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

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

```
'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              genename
```

```
                 <numeric> <character> <character>           <character>
ENSG00000279457 6.86555e-01          NA          NA                    NA
ENSG00000187634 5.15718e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76549e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13413e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.19031e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03379e-01       PERM1       84808 PPARGC1 and ESRR ind..
ENSG00000188290 1.30538e-24        HES4       57801 hes family bHLH tran..
ENSG00000187608 2.37452e-02       ISG15        9636 ISG15 ubiquitin like..
ENSG00000188157 4.21963e-16        AGRN      375790                  agrin
ENSG00000237330          NA      RNF223      401934 ring finger protein ..
```

Q. 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.

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)
```

```
library(gageData)

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

head(kegg.sets.hs, 2)
```

```
$`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"
```

We have the Entrez gene Ids and we have the fold change results from DESeq2 analysis

```
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
```

Now, lets run the gage pathway

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

```
                                              p.geomean stat.mean
hsa04110 Cell cycle                         8.995727e-06 -4.378644
hsa03030 DNA replication                    9.424076e-05 -3.951803
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 -3.765330
hsa03013 RNA transport                      1.375901e-03 -3.028500
hsa03440 Homologous recombination           3.066756e-03 -2.852899
hsa04114 Oocyte meiosis                     3.784520e-03 -2.698128
```

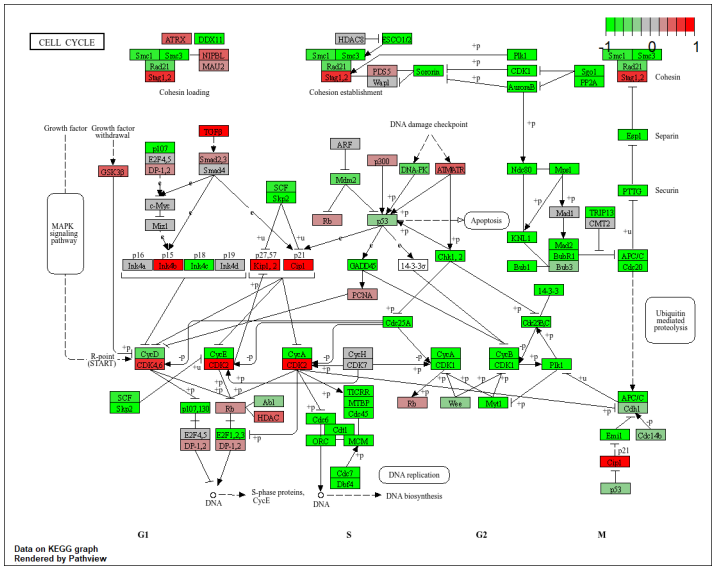
```
                                                      p.val       q.val
hsa04110 Cell cycle                                8.995727e-06 0.001889103
hsa03030 DNA replication                           9.424076e-05 0.009841047
hsa05130 Pathogenic Escherichia coli infection     1.405864e-04 0.009841047
hsa03013 RNA transport                             1.375901e-03 0.072234819
hsa03440 Homologous recombination                  3.066756e-03 0.128803765
hsa04114 Oocyte meiosis                            3.784520e-03 0.132458191
                                                   set.size         exp1
hsa04110 Cell cycle                                     121 8.995727e-06
hsa03030 DNA replication                                 36 9.424076e-05
hsa05130 Pathogenic Escherichia coli infection           53 1.405864e-04
hsa03013 RNA transport                                  144 1.375901e-03
hsa03440 Homologous recombination                        28 3.066756e-03
hsa04114 Oocyte meiosis                                 102 3.784520e-03
```

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

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

```
Info: Working in directory C:/Users/joelp/BIMM143/Class14
```

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



```
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 C:/Users/joelp/BIMM143/Class14
```

```
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] "hsa04060" "hsa05323" "hsa05146" "hsa05332" "hsa04640"
```

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

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

```
Info: Working in directory C:/Users/joelp/BIMM143/Class14
```

```
Info: Writing image file hsa04060.pathview.png
```

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

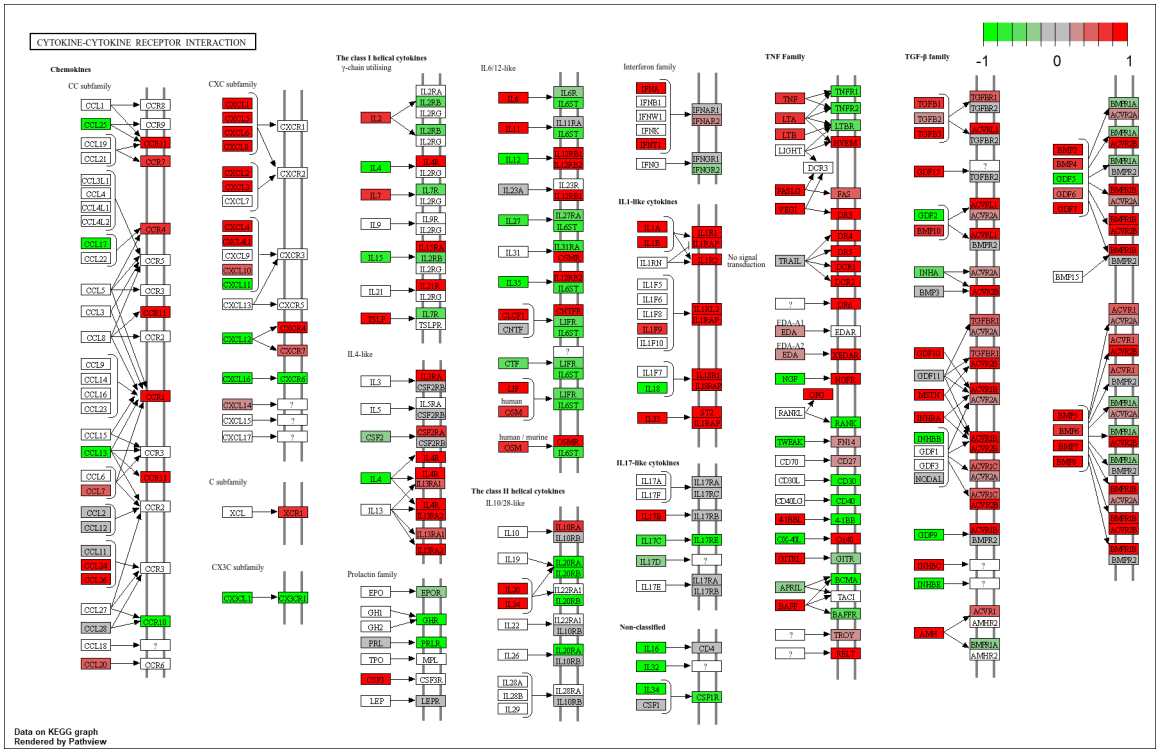
```
Info: Working in directory C:/Users/joelp/BIMM143/Class14
```

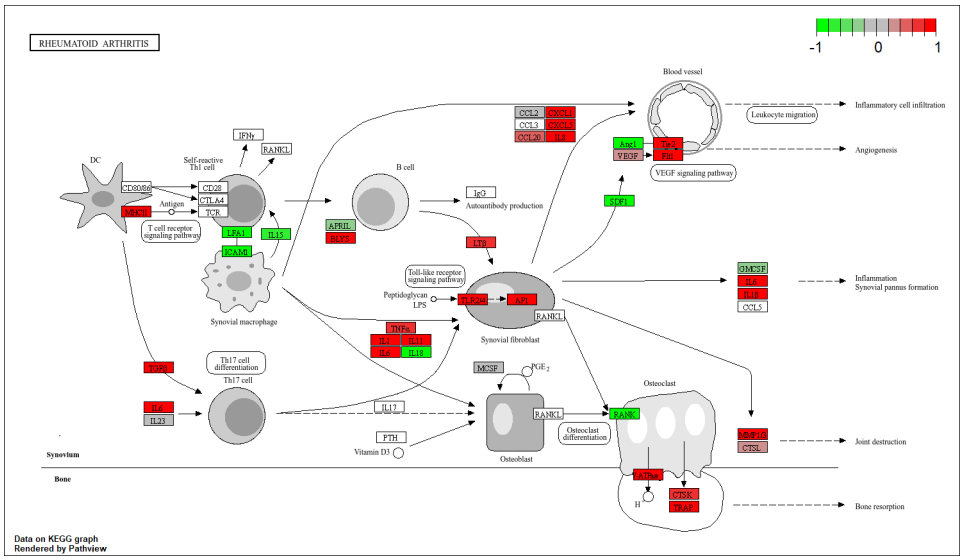
```
Info: Writing image file hsa05323.pathview.png
```

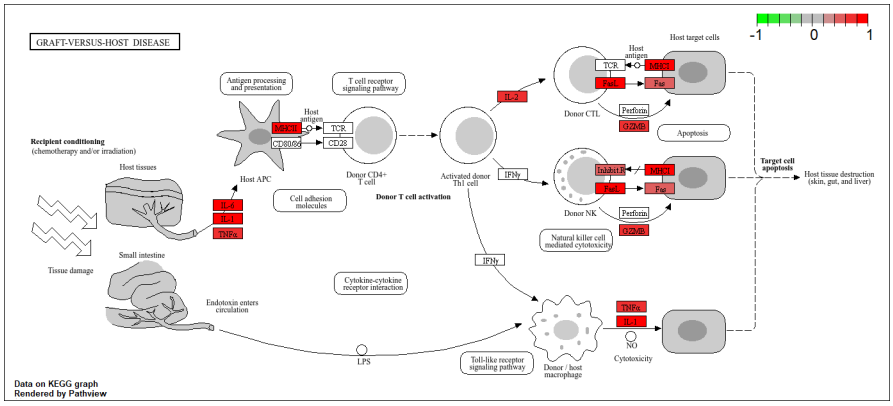
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa05146.pathview.png

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

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa05332.pathview.png

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

Info: Working in directory C:/Users/joelp/BIMM143/Class14

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

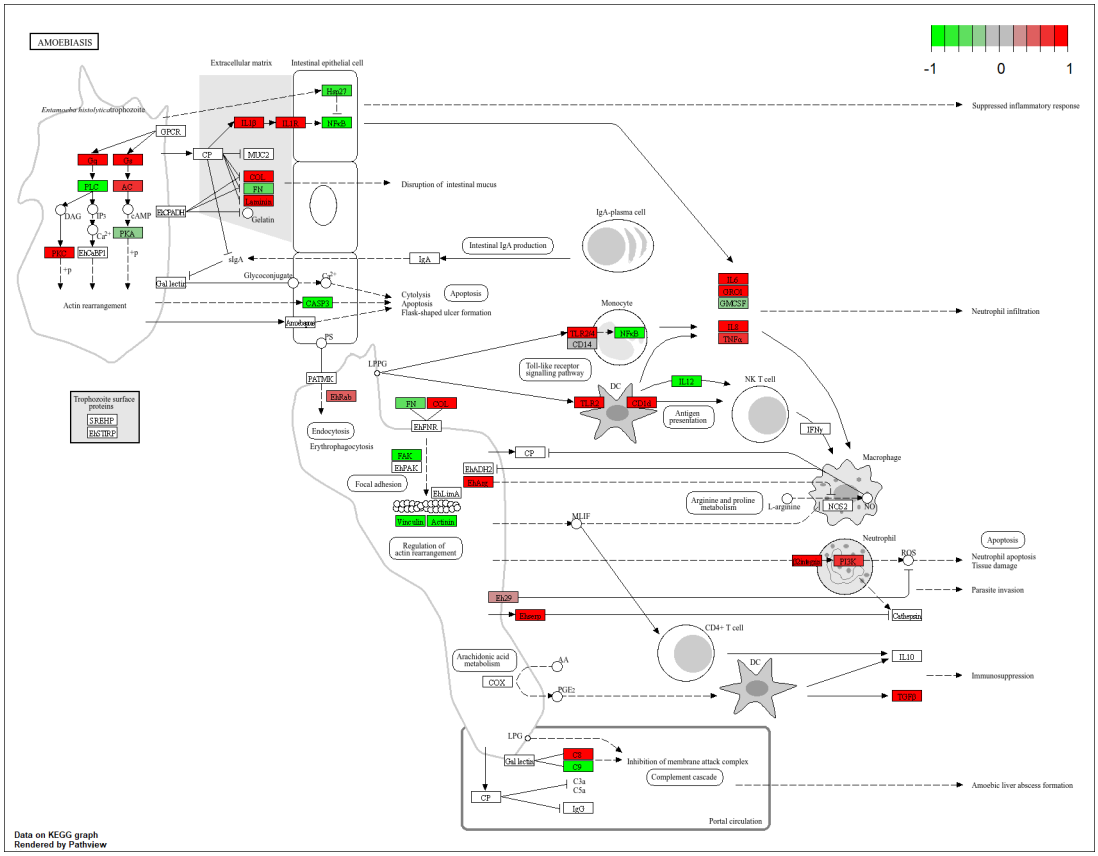
```
pathview(gene.data=foldchanges, pathway.id="hsa04060")
```

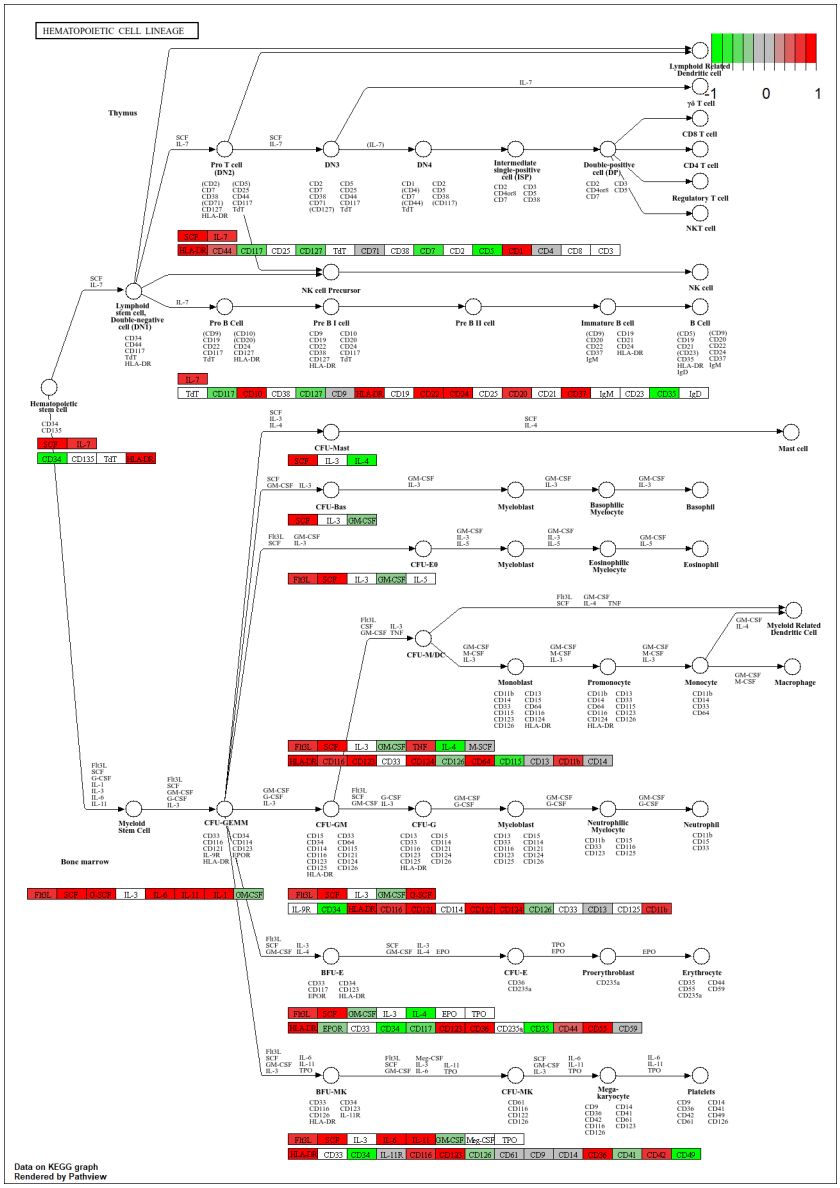
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa04060.pathview.png
```


CYTOKINE-CYTOKINE RECEPTOR INTERACTION
-1
0
1
Chemokines
CC subfamily
CXC subfamily
C subfamily
CX3C subfamily
The class I helical cytokines
γ-chain utilising
IL4-like
Prolactin family
IL6/12-like
The class II helical cytokines
IL10/28-like
Interferon family
IL1-like cytokines
No signal transduction
IL17-like cytokines
Non-classified
TNF Family
TGF-β family
Data on KEGG graph
Rendered by Pathview


RHEUMATOID ARTHRITIS
-1
0
1
Blood vessel
Leukocyte migration
Inflammatory cell infiltration
Angiogenesis
VEGF signaling pathway
DC
Self-reactive Th1 cell
IFNγ
RANKL
CD80/86
CD28
Antigen
CTLA4
TCR
MHC II
T cell receptor signaling pathway
LFA1
ICAM1
IL15
Synovial macrophage
B cell
IgG
Autoantibody production
APRIL
BLYS
LTB
Toll-like receptor signaling pathway
Peptidoglycan
LPS
TLR2/4
APL
RANKL
Synovial fibroblast
TNFα
IL1
IL11
IL6
IL18
STAT1
GM-CSF
IL6
IL18
CCL5
Inflammation
Synovial pannus formation
TOLL
Th17 cell differentiation
Th17 cell
IL6
IL23
IL17
MCSF
PGE2
RANKL
RANK
Osteoclast
Osteoclast differentiation
Osteoblast
PTH
Vitamin D3
Synovium
Bone
MMP13
CTSL
Joint destruction
LAPase
CTSK
TRAP
H
Bone resorption
Data on KEGG graph
Rendered by Pathview


GRAFT-VERSUS-HOST DISEASE
-1
0
1
Host target cells
Antigen processing and presentation
T cell receptor signaling pathway
Host antigen
MHCII
TCR
CD80/86
CD28
Donor CD4+ T cell
Activated donor Th1 cell
IL-2
IFNγ
Donor CTL
TCR
FasL
Fas
Perforin
GZMB
Apoptosis
Donor NK
Inhibit R
MHC I
Natural killer cell mediated cytotoxicity
Target cell apoptosis
Host tissue destruction (skin, gut, and liver)
Recipient conditioning
(chemotherapy and/or irradiation)
Host tissues
Host APC
IL-1
IL-6
TNFα
Cell adhesion molecules
Donor T cell activation
Tissue damage
Small intestine
Endotoxin enters circulation
LPS
Cytokine-cytokine receptor interaction
Toll-like receptor signaling pathway
Donor / host macrophage
TNFα
IL-1
NO
Cytotoxicity
Data on KEGG graph
Rendered by Pathview


AMOEBIASIS
-1
0
1
Extracellular matrix
Intestinal epithelial cell
Entamoeba histolytica trophozoite
Hsp27
IL10
IL1R
NFκB
Suppressed inflammatory response
GPCR
CP
MUC2
COL
FN
Laminin
Gelatin
Disruption of intestinal mucus
Gαi
Gβ
PLC
AC
DAG
IP3
cAMP
EhCPADH
PI3K
EhCaBP1
PIP2
+p
PIPA
sIgA
IgA
IgA-plasma cell
Intestinal IgA production
Oal lectin
Glycoconjugate
Ca2+
Cytolysis
Apoptosis
Flask-shaped ulcer formation
CASP3
Actin rearrangement
Amoebapore
PS
Monocyte
TLR2/4
CD14
MyD88
IL6
CPCa
GMCSF
IL8
TNFα
Neutrophil infiltration
LPPG
FATMK
EhRab
Trophozoite surface proteins
SREHP
EhTIRP
Toll-like receptor signalling pathway
DC
TLR2
CD14
IL12
NK T cell
Antigen presentation
IFNγ
Endocytosis
Erythrophagocytosis
EhPNR
FN
COL
FAK
EhPAK
Focal adhesion
EhADH2
EhADH1
EhLimA
Fimbrin
Actinin
Regulation of actin rearrangement
CP
Macrophage
Arginine and proline metabolism
L-arginine
NOS2
NO
MIF
Neutrophil
Cathepsin
PI3K
ROS
Apoptosis
Neutrophil apoptosis
Tissue damage
Parasite invasion
Eh29
Filamin
Cathepsin
CD4+ T cell
DC
Arachidonic acid metabolism
AA
COX
PGE2
IL10
Immunosuppression
TGFB1
LPG
Oal lectin
C5
C5
C3a
C5a
CP
IgO
Inhibition of membrane attack complex
Complement cascade
Amoebic liver abscess formation
Portal circulation
Data on KEGG graph
Rendered by Pathview



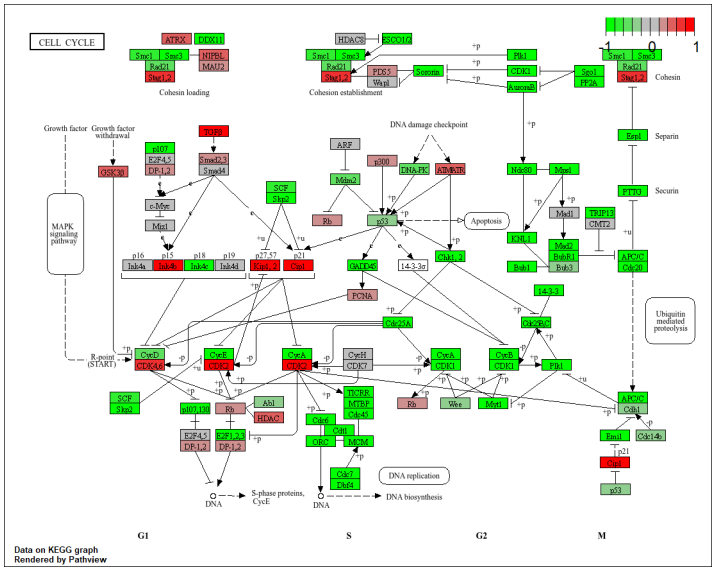
>Q. Can you do the same procedure as above to plot the pathview figures for the top 5 down-reguled pathways?

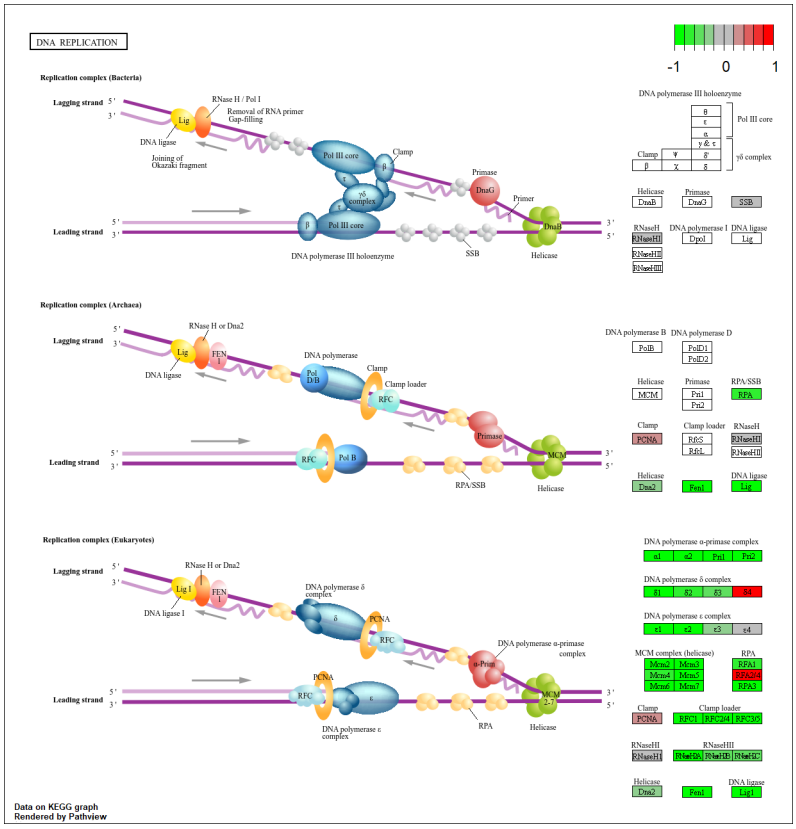
```
keggrespathways <- rownames(keggres$less)[1:5]

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

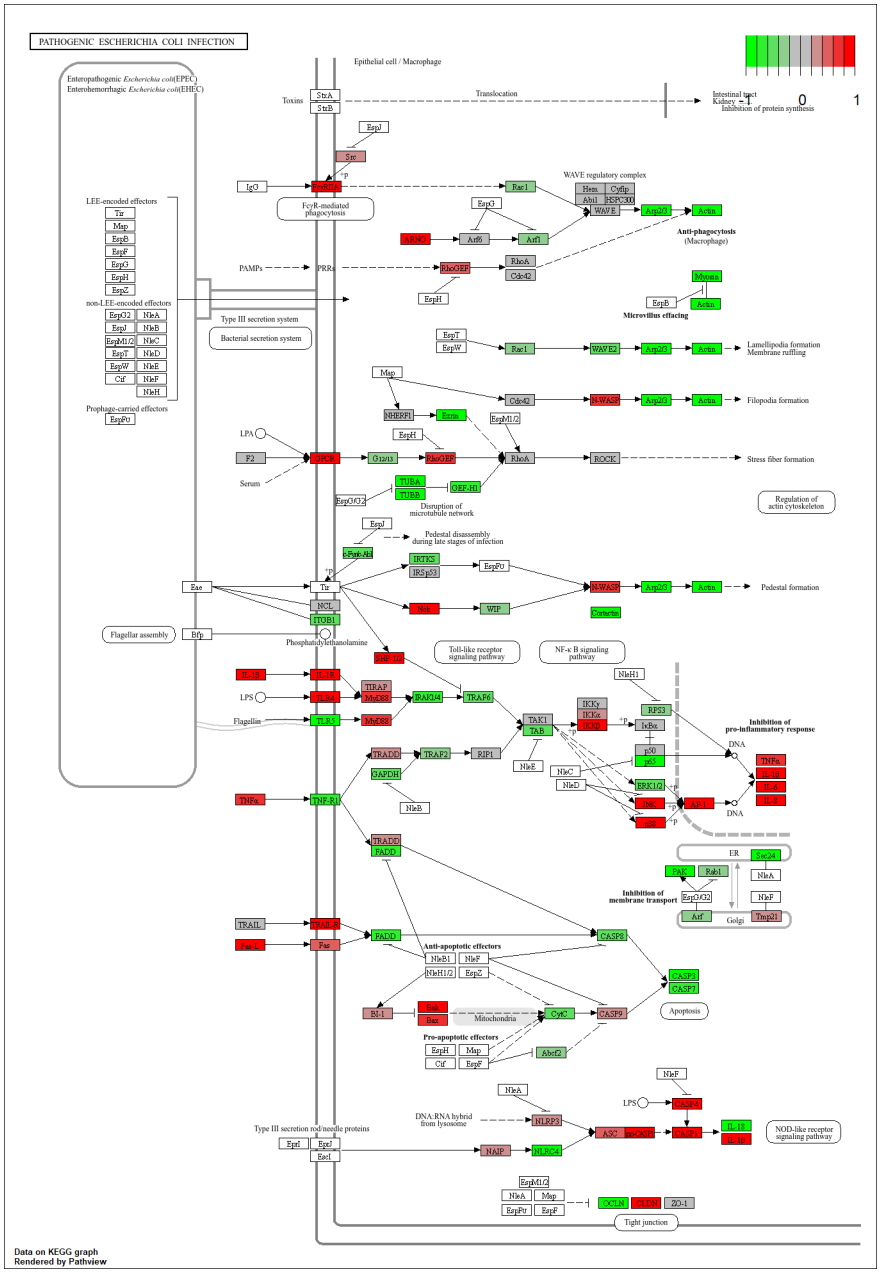
```
[1] "hsa04110" "hsa03030" "hsa05130" "hsa03013" "hsa03440"
```

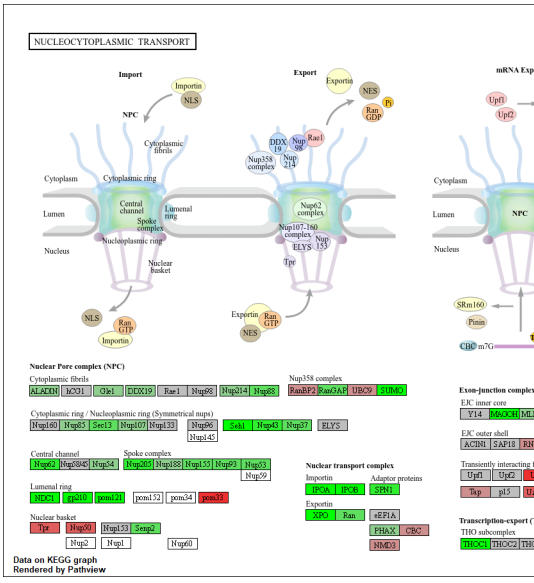
```
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

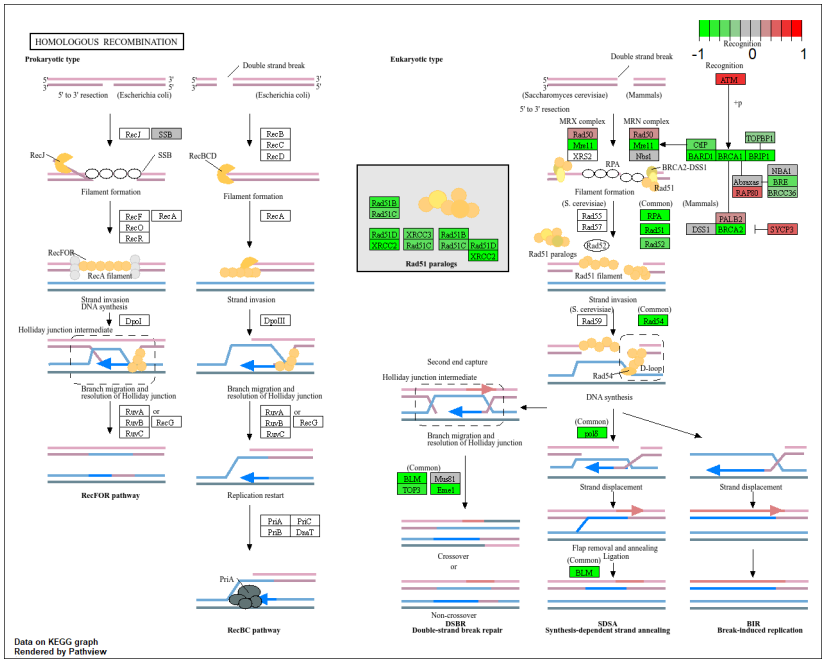
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa04110.pathview.png

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

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa05130.pathview.png

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

Info: Working in directory C:/Users/joelp/BIMM143/Class14

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory C:/Users/joelp/BIMM143/Class14

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


CELL CYCLE
-1
0
1
Cohesin loading
Cohesion establishment
Cohesin
Separin
Securin
Growth factor
Growth factor withdrawal
MAPK signaling pathway
DNA damage checkpoint
Apoptosis
Ubiquitin mediated proteolysis
R-point (START)
DNA replication
DNA biosynthesis
S-phase proteins, CycE
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
DNA ligase
Joining of Okazaki fragment
Pol III core
Clamp
γδ complex
Primase
DnaG
Primer
DnaB
SSB
Helicase
DNA polymerase III holoenzyme
DNA polymerase III holoenzyme
Pol III core
γδ complex
Clamp
Helicase
Primase
SSB
RNaseH
DNA polymerase I
DNA ligase
Replication complex (Archaea)
RNase H or Dna2
DNA polymerase
Clamp
Clamp loader
RFC
Pol B
Primase
MCM
RPA/SSB
Helicase
DNA polymerase B
DNA polymerase D
Replication complex (Eukaryotes)
RNase H or Dna2
DNA ligase I
DNA polymerase δ complex
PCNA
DNA polymerase α-primase complex
DNA polymerase ε complex
RPA
Helicase
DNA polymerase α-primase complex
DNA polymerase δ complex
DNA polymerase ε complex
MCM complex (helicase)
Clamp
Clamp loader
RNaseHII
RNaseHIII
Data on KEGG graph
Rendered by Pathview


PATHOGENIC ESCHERICHIA COLI INFECTION
Epithelial cell / Macrophage
Enteropathogenic Escherichia coli(EPEC)
Enterohemorrhagic Escherichia coli(EHEC)
Toxins
Translocation
Intestinal tract
Kidney
Inhibition of protein synthesis
0
1
LEE-encoded effectors
non-LEE-encoded effectors
Prophage-carried effectors
Type III secretion system
Bacterial secretion system
PAMPs
PRRs
FcγR-mediated phagocytosis
WAVE regulatory complex
Anti-phagocytosis
(Macrophage)
Microvillus effacing
Lamellipodia formation
Membrane ruffling
Filopodia formation
Stress fiber formation
Regulation of actin cytoskeleton
Disruption of microtubule network
Pedestal disassembly during late stages of infection
Pedestal formation
Flagellar assembly
Phosphatidylethanolamine
Toll-like receptor signaling pathway
NF-κB signaling pathway
Inhibition of pro-inflammatory response
Inhibition of membrane transport
Anti-apoptotic effectors
Pro-apoptotic effectors
Mitochondria
Apoptosis
Type III secretion rod/needle proteins
DNA/RNA hybrid from lysosome
NOD-like receptor signaling pathway
Tight junction
Data on KEGG graph
Rendered by Pathview


NUCLEOCYTOPLASMIC TRANSPORT
Import
Export
NPC
Cytoplasm
Cytoplasmic ring
Lumen
Nucleus
Central channel
Spoke complex
Nucleoplasmic ring
Lumenal ring
Nuclear basket
Cytoplasmic fibrils
Nup358 complex
Nup62 complex
Nup107-160 complex
Importin
Exportin
Ran GTP
Nuclear Pore complex (NPC)
Cytoplasmic ring / Nucleoplasmic ring (Symmetrical nups)
Central channel
Spoke complex
Lumenal ring
Nuclear basket
Nuclear transport complex
Importin
Adaptor proteins
Exportin
Transcription-export (
Data on KEGG graph
Rendered by Pathview



Gene Ontology Analysis

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

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

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

lapply(gobpres, head)
```

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

```
GO:0007156 homophilic cell adhesion       0.1951953      113 8.519724e-05
GO:0002009 morphogenesis of an epithelium 0.1951953      339 1.396681e-04
GO:0048729 tissue morphogenesis           0.1951953      424 1.432451e-04
GO:0007610 behavior                       0.1967577      426 1.925222e-04
GO:0060562 epithelial tube morphogenesis  0.3565320      257 5.932837e-04
GO:0035295 tube development               0.3565320      391 5.953254e-04

$less
                                          p.geomean stat.mean        p.val
GO:0048285 organelle fission             1.536227e-15 -8.063910 1.536227e-15
GO:0000280 nuclear division              4.286961e-15 -7.939217 4.286961e-15
GO:0007067 mitosis                       4.286961e-15 -7.939217 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.169934e-14 -7.797496 1.169934e-14
GO:0007059 chromosome segregation        2.028624e-11 -6.878340 2.028624e-11
GO:0000236 mitotic prometaphase          1.729553e-10 -6.695966 1.729553e-10
                                                q.val set.size         exp1
GO:0048285 organelle fission             5.841698e-12      376 1.536227e-15
GO:0000280 nuclear division              5.841698e-12      352 4.286961e-15
GO:0007067 mitosis                       5.841698e-12      352 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.195672e-11      362 1.169934e-14
GO:0007059 chromosome segregation        1.658603e-08      142 2.028624e-11
GO:0000236 mitotic prometaphase          1.178402e-07       84 1.729553e-10

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

```
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj), "symbol"]
print(paste("Total number of significant genes:", length(sig_genes)))
```

```
[1] "Total number of significant genes: 8147"
```

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

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?

The cell cycle, Miotic has the most significant Entities p-value with the value 1.7E-4. Yes, the top result for the KEGG results was the cell cycle, although it has a different p-value of 8.99E-6. KEGG looks at things in the context of complex biological pathways, whereas GO provides a standardizes way to describe gene function, leading to differences in results between the two methods.