Class 14 RNA Mini-project

Zainab Fatima (PID: A16880407)

2025-02-20

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, stringsAsFactors = F, header =
colData <- read.csv("GSE37704_metadata.csv", stringsAsFactors = F, header = T)
```

Inspect and tidy

> Q. Complete the code below to remove the troublesome first column from count-Data

```
colnames(counts)
```

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

```
#need to remove length column
```

```
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) %in% colData$id
```

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

```
colData[match(colnames(countData),colData$id),]
```

```
         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
```

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).

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

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).

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

head(new.counts)
```

	SRR493366	SRR493367	SRR493368	SRR493369	SRR493370	SRR493371
ENSG00000279457	23	28	29	29	28	46
ENSG00000187634	124	123	205	207	212	258
ENSG00000188976	1637	1831	2383	1226	1326	1504
ENSG00000187961	120	153	180	236	255	357
ENSG00000187583	24	48	65	44	48	64
ENSG00000187642	4	9	16	14	16	16

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

```
[1] 15975
```

Setup for DESeq

```
library(DESeq2)
```

Run DESeq

```
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
```

```
dds
```

```
class: DESeqDataSet
dim: 19808 6
metadata(1): version
assays(4): counts mu H cooks
rownames(19808): ENSG00000186092 ENSG00000279928 ... ENSG00000277475
  ENSG00000268674
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(3): id condition sizeFactor
```

```
res <- results(dds, contrast=c("condition", "hoxa1_kd", "control_sirna"))
```

```
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
```

Q. 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
```

Volcano plot of results

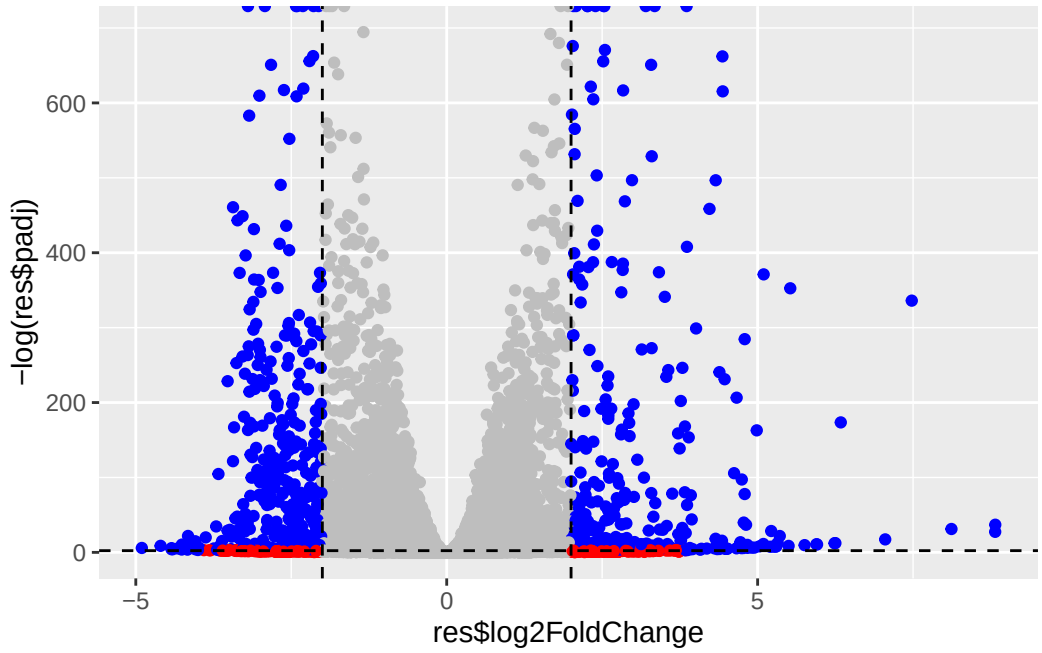
Q. Improve this plot by completing the below code, which adds color and axis labels

```
library(ggplot2)
```

```
mycols <- rep("gray", nrow(res))
#if my log2fold change >2 --> change it to red
mycols[ abs(res$log2FoldChange) > 2 ]  <- "red"
inds <- (res$padj < 0.05) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"
```

```
ggplot(res) +
  aes(x = res$log2FoldChange, y = -log(res$padj)) +
  geom_point(col = mycols) +
   geom_vline(xintercept = 2, linetype = "dashed") +
  geom_vline (xintercept = -2, linetype = "dashed") +
  geom_hline (yintercept = -log(0.1), linetype = "dashed")
```

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



Gene annotations

Q. 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)
```

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

```
'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
```

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

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

```
head(res)
```

```
log2 fold change (MLE): condition hoxa1_kd vs control_sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 6 rows and 9 columns
                 baseMean log2FoldChange     lfcSE      stat     pvalue
                <numeric>      <numeric> <numeric> <numeric>  <numeric>
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      symbol               genename      entrez
                <numeric> <character>            <character> <character>
ENSG00000186092         NA       OR4F5 olfactory receptor f..       79501
ENSG00000279928         NA          NA                     NA          NA
ENSG00000279457 0.68707978          NA                     NA          NA
ENSG00000278566         NA          NA                     NA          NA
ENSG00000273547         NA          NA                     NA          NA
ENSG00000187634 0.00516278      SAMD11 sterile alpha motif ..      148398
```

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.

```
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)
```

```
# 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"
```

The main `gage()` function requires a named vector of fold changes, where the names of the values are the Entrez gene IDs.

```
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, let's run the `gage` pathway analysis.

```
#Get the results

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

```
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
```

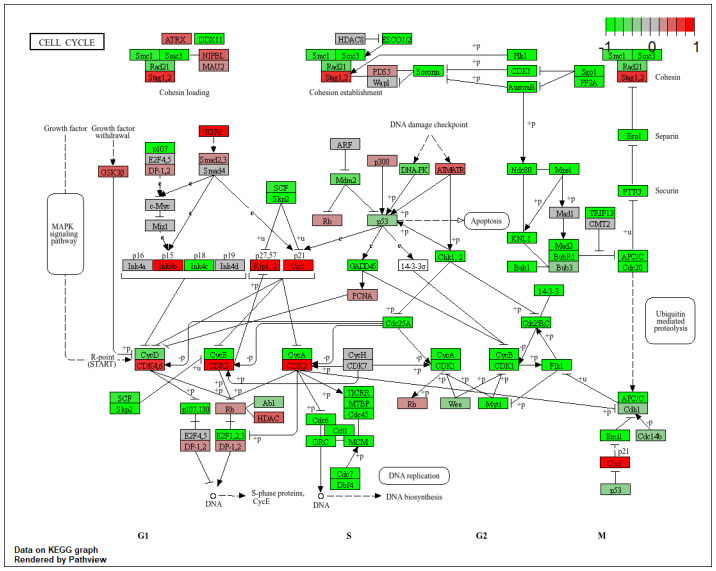
```
hsa03013 RNA transport                  0.063420543      149 1.160132e-03
hsa04114 Oocyte meiosis                 0.100589607      112 2.563806e-03
hsa03440 Homologous recombination       0.100589607       28 3.066756e-03
hsa00010 Glycolysis / Gluconeogenesis 0.119175854       65 4.360092e-03
```

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

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

```
Info: Working in directory C:/BIMM 143/Class 14
```

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



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

```
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:/BIMM 143/Class 14
```

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

Now, let's pull up the top 5 upregulated pathways.

```
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 C:/BIMM 143/Class 14
```

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

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

```
Info: Working in directory C:/BIMM 143/Class 14
```

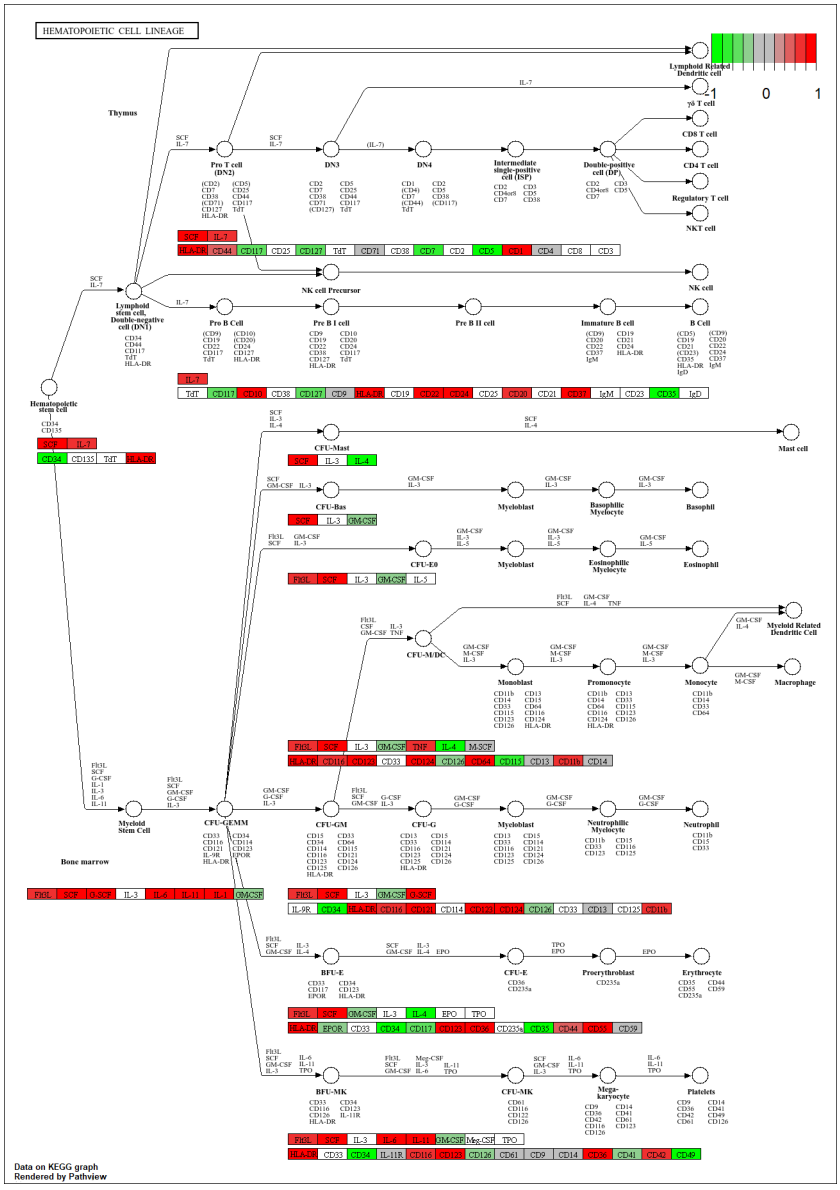
```
Info: Writing image file hsa04640.pathview.png
```

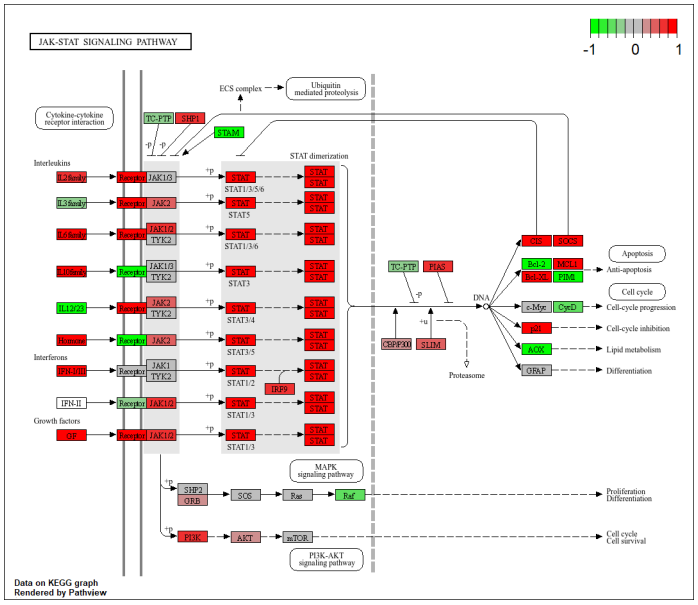
```
'select()' returned 1:1 mapping between keys and columns
```

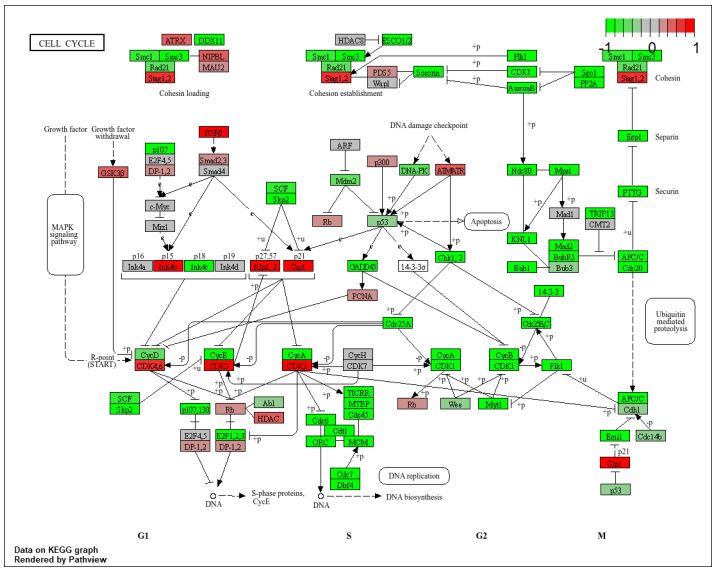
```
Info: Working in directory C:/BIMM 143/Class 14
```

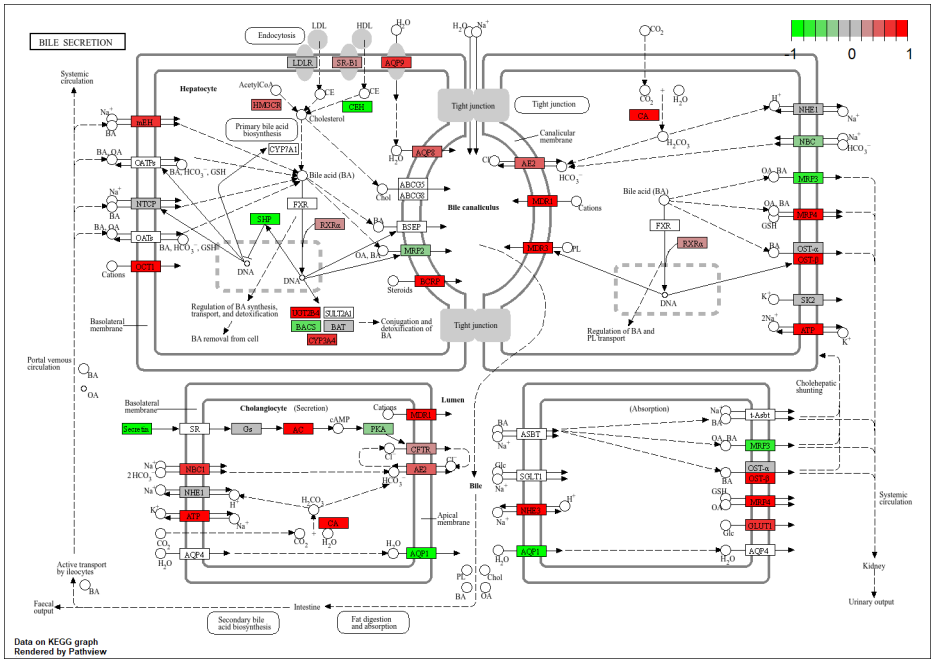
```
Info: Writing image file hsa00140.pathview.png
```

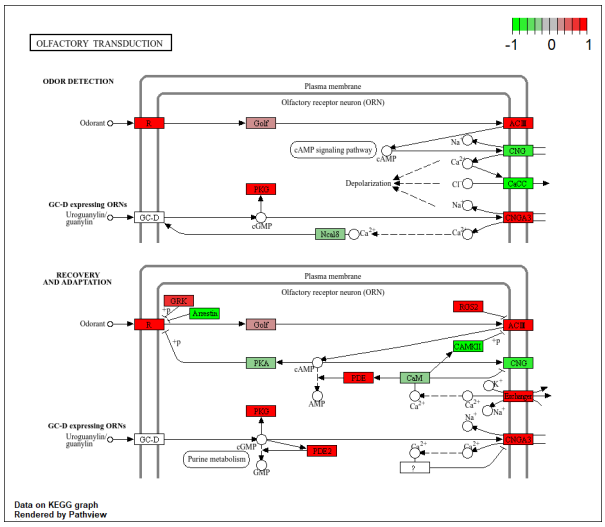
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/BIMM 143/Class 14

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/BIMM 143/Class 14

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


HEMATOPOIETIC CELL LINEAGE
Lymphoid Related Dendritic cell
-1
0
1
Thymus
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF
IL-7
Pro T cell (DN2)
DN3
DN4
Intermediate single-positive cell (ISP)
Double-positive cell (DP)
(CD2) (CD5) CD7 CD25 CD38 CD44 (CD71) CD117 CD127 TdT HLA-DR
CD2 CD5 CD7 CD25 CD38 CD44 CD71 CD117 (CD127) TdT
CD1 (CD4) CD7 (CD44) TdT
CD2 CD5 CD38 (CD117)
CD2 CD3 CD4es8 CD7 CD5 CD38
CD2 CD3 CD4es8 CD7 CD5
SCF
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
CD34 CD44 CD117 TdT HLA-DR
(CD9) CD19 CD22 CD117 TdT
(CD10) (CD20) CD24 CD127 HLA-DR
CD9 CD19 CD22 CD38 CD127 HLA-DR
CD10 CD20 CD24 CD117 TdT
(CD9) CD20 CD22 CD37 IgM
CD19 CD21 CD24 HLA-DR
(CD5) CD19 CD21 (CD21) CD23 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
Hematopoietic stem Cell
CD34 CD135
CD
IL-3
CD34
CD135
TdT
HLA-DR
SCF
IL-3
IL-4
CFU-Mast
Mast cell
SCF
IL-3
IL-4
SCF
GM-CSF
IL-3
CFU-Bas
GM-CSF
IL-3
Myeloblast
Basophilic Myelocyte
Basophil
SCF
IL-3
GMCSF
Flt3L
SCF
GM-CSF
IL-3
CFU-E0
GM-CSF
IL-3
IL-5
Myeloblast
Eosinophilic Myelocyte
Eosinophil
Flt3L
SCF
IL-3
GMCSF
IL-5
Flt3L
SCF
IL-4
TNF
Flt3L
CSF
GM-CSF
IL-3
TNF
CFU-M/DC
GM-CSF
M-CSF
IL-3
Monoblast
Promonocyte
Monocyte
Myeloid Related Dendritic Cell
Macrophage
GM-CSF
IL-4
GM-CSF
M-CSF
CD11b CD13 CD14 CD33 CD115 CD123 HLA-DR
CD11b CD13 CD14 CD64 CD115 CD123 HLA-DR
CD11b CD14 CD33 CD64
Flt3L
SCF
IL-3
GMCSF
TNF
IL-4
M-SCF
HLA-DR
CD116
CD123
CD33
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
GM-CSF
G-CSF
IL-3
CFU-GEMM
Flt3L
SCF
GM-CSF
IL-3
G-CSF
GM-CSF
CFU-G
GM-CSF
G-CSF
Myeloblast
GM-CSF
G-CSF
Neutrophilic Myelocyte
GM-CSF
G-CSF
Neutrophil
CD33 CD34 CD44 CD123 CD90R IL-9R HLA-DR
CD15 CD33 CD14 CD116 CD123 CD125 HLA-DR
CD13 CD15 CD14 CD16 CD123 CD25 HLA-DR
CD13 CD15 CD16 CD21 CD123 CD124 CD126
CD33 CD15 CD16 CD21 CD123 CD124 CD125 CD126
CD11b CD15 CD33 CD116 CD123 CD125
CD11b CD15 CD33
Bone marrow
Flt3L
SCF
G-SCF
IL-3
IL-6
IL-11
IL-1
GMCSF
Flt3L
SCF
IL-3
GMCSF
G-SCF
IL-9R
CD34
HLA-DR
CD116
CD123
CD114
CD123
CD124
CD126
CD33
CD13
CD123
CD11b
Flt3L
SCF
GM-CSF
IL-3
IL-3
IL-4
BFU-E
SCF
GM-CSF
IL-3
IL-4
EPO
CFU-E
TPO
EPO
Proerythroblast
EPO
Erythrocyte
CD33 CD117 EPOR
CD34 CD123 HLA-DR
CD36 CD235a
CD235a
CD35 CD55 CD235a
CD44 CD59
Flt3L
SCF
GMCSF
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
Flt3L
SCF
GM-CSF
IL-3
Megr-CSF
IL-6
IL-11
TPO
CFU-MK
SCF
GM-CSF
IL-3
IL-6
IL-11
TPO
Mega-karyocyte
IL-6
IL-11
TPO
Platelets
CD31 CD34 CD116 CD123 CD126 HL-A-DR
CD34 CD123 IL-11R
CD61 CD116 CD122 CD126
CD9 CD14 CD36 CD41 CD42 CD61 CD116 CD123 CD126
CD9 CD14 CD36 CD41 CD42 CD61 CD126
Flt3L
SCF
IL-3
IL-6
IL-11
GM-CSF
Megr-CSF
TPO
HLA-DR
CD33
CD34
IL-11R
CD116
CD123
CD126
CD61
CD9
CD14
CD36
CD41
CD42
CD49
Data on KEGG graph
Rendered by Pathview


JAK-STAT SIGNALING PATHWAY
-1
0
1
ECS complex
Ubiquitin mediated proteolysis
Cytokine-cytokine receptor interaction
TC-PTP
SHP1
STAM
STAT dimerization
Interleukins
IL2Rαβγ
IL3Rαβγ
IL6Rαβγ
IL10Rαβ
IL12Rβ1
Hormone
Interferons
IFN-I(II)
IFN-II
Growth factors
GH
Receptor
JAK1/3
JAK2
JAK1/2
TYK2
JAK1
STAT1/2/3/4/5/6
STAT1/3/5/6
STAT5
STAT1/3/6
STAT3
STAT3/4
STAT3/5
STAT1/2
STAT1/3
IRF9
SOCS
CIS
Bcl-2
Bcl-XL
Mcl-1
Flt3l
Apoptosis
Anti-apoptosis
Cell cycle
Cell-cycle progression
Cell-cycle inhibition
Lipid metabolism
Differentiation
c-Myc
CyeD
p21
ACO
GFAP
PIAS
DNA
CBP/P300
SLIM
Proteasome
MAPK signaling pathway
SHP2
GRB
SOS
Ras
Raf
Proliferation
Differentiation
PI3K
AKT
mTOR
PI3K-AKT signaling pathway
Cell cycle
Cell survival
Data on KEGG graph
Rendered by Pathview


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
Sororin
Wapl
Cohesin establishment
Plk1
CDK1
Aurora B
Spo1
PP2A
Cohesin
Separin
Securin
PTTG
DNA damage checkpoint
ARF
p300
Mdm2
DNA-PK
ATM/ATR
Rb
p53
Apoptosis
Mdp80
Mps1
Mad1
TRIP13
CMT2
ICBL1
Mad2
BubR1
Bub1
Bub3
APC/C
Cdc20
14-3-3σ
Chk1_2
GADD45
14-3-3
PCNA
Cdc25A
Cdc28/2
Ubiquitin mediated proteolysis
Growth factor
Growth factor withdrawal
TGFβ
p107
E2F4,5
DP1,2
Smad2,3
Smad4
G0/G1
MAPK signaling pathway
c-Myc
Mxi1
SCF
Skp2
p16
Ink4a
p15
Ink4b
p18
Ink4c
p19
Ink4d
p27,57
Cip1,2
p21
Cip1
R-point
(START)
CycD
CDK4,6
CycE
CDK2
CycA
CDK2
CycH
CDK7
CycA
CDK1
CycB
CDK1
Plg1
SCF
Skp2
p107,130
Rb
Abl
HDAC
TCBP
MTBP
Cdc45
Cdc6
Cdt1
Mcm4
Cdc7
Dbf4
E2F4,5
DP1,2
E2F1,2,3
DP1,2
Rb
Wee
Myt1
APC/C
Cdh1
Emi1
Cdc14b
p21
Cip1
p53
DNA
S-phase proteins, CycE
DNA
DNA replication
DNA biosynthesis
G1
S
G2
M
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?

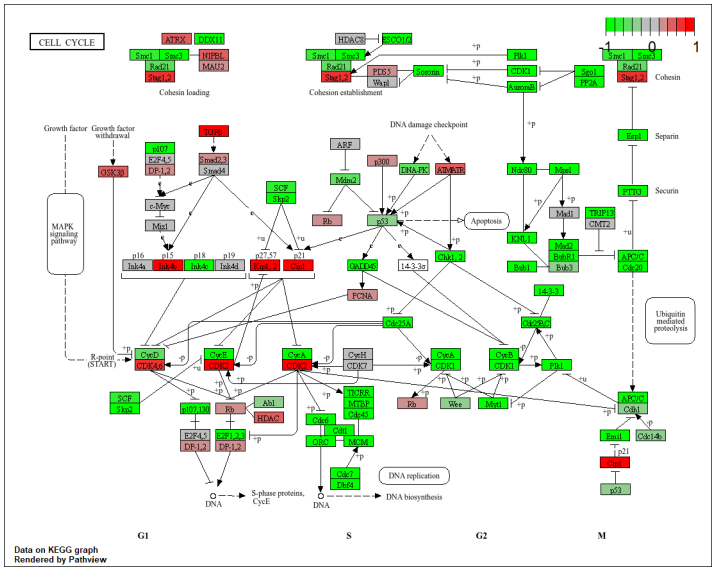
```
keggrespathways2 <- rownames(keggres$less)[1:5]
```

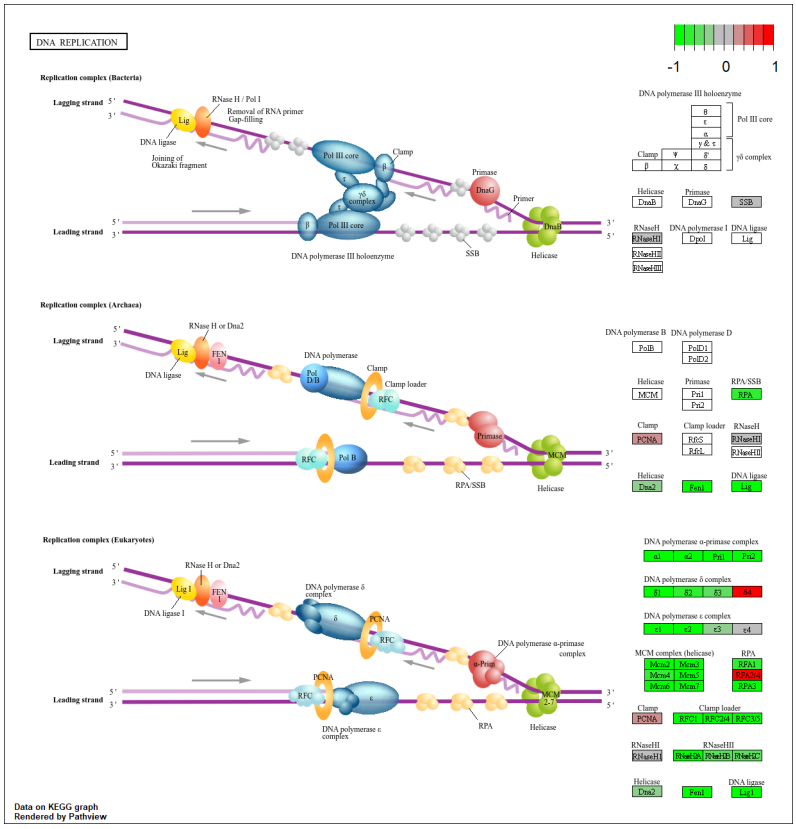
```
keggresids2 = substr(keggrespathways2, start=1, stop=8)
keggresids2
```

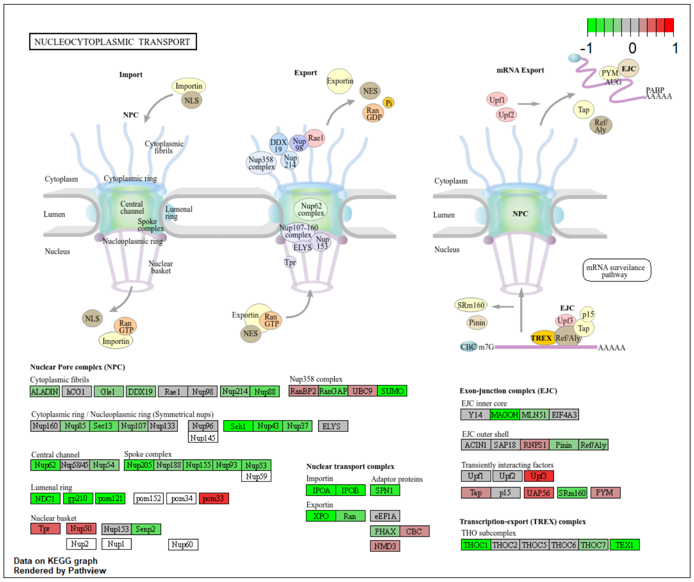
```
[1] "hsa04110" "hsa03030" "hsa03013" "hsa04114" "hsa03440"
```

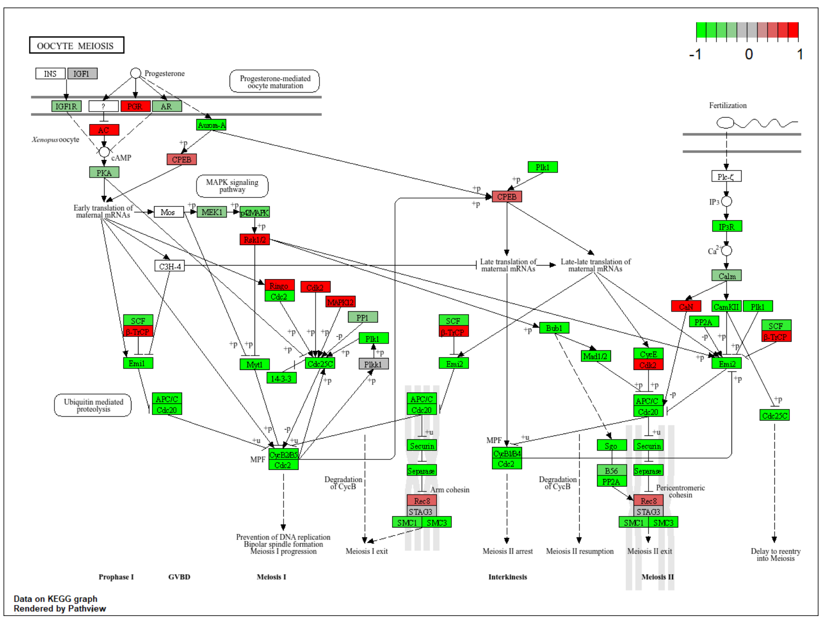
```
pathview(gene.data=foldchanges, pathway.id=keggresids2, species="hsa")
```

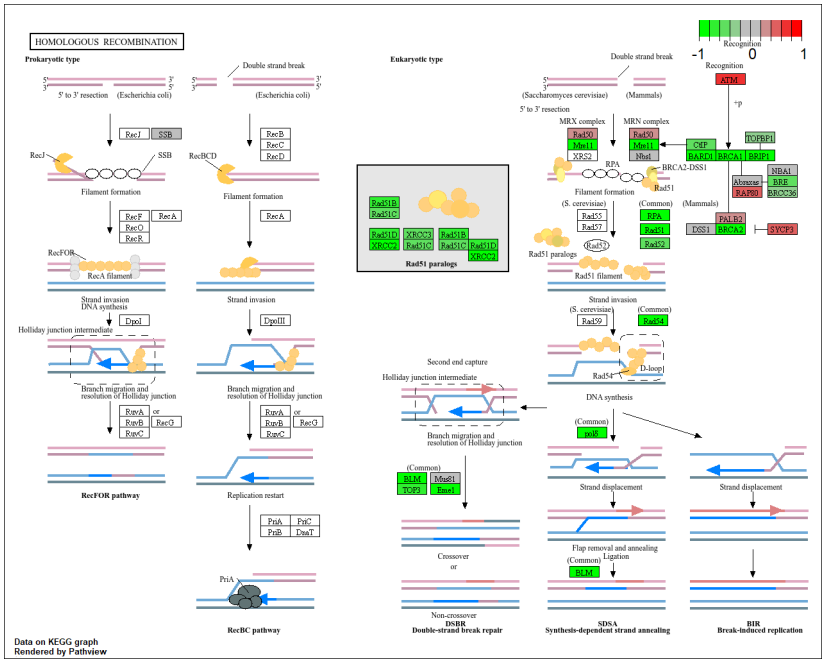
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/BIMM 143/Class 14

Info: Writing image file hsa04110.pathview.png

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

Info: Working in directory C:/BIMM 143/Class 14

Info: Writing image file hsa03030.pathview.png

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

Info: Working in directory C:/BIMM 143/Class 14

Info: Writing image file hsa03013.pathview.png

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

Info: Working in directory C:/BIMM 143/Class 14

Info: Writing image file hsa04114.pathview.png

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

Info: Working in directory C:/BIMM 143/Class 14

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


CELL CYCLE
-1 0 1
Cohesin loading
Cohesion establishment
Cohesin
Separin
Securin
DNA damage checkpoint
Growth factor
Growth factor withdrawal
MAPK signaling pathway
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
Primase
DnaG
Primer
DnaB
SSB
Helicase
DNA polymerase III holoenzyme
DNA polymerase III holoenzyme
Pol III core
γδ complex
Clamp
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
DNA polymerase I
DNA ligase
Lig
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
PolB
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
RfcS
RfcL
RNaseH
Helicase
Dna2
Fen1
DNA ligase
Lig
Replication complex (Eukaryotes)
RNase H or Dna2
DNA ligase I
DNA polymerase δ complex
PCNA
RFC
DNA polymerase α-primase complex
α-Prim
MCM 2-7
RPA
Helicase
DNA polymerase ε complex
DNA polymerase α-primase complex
DNA polymerase δ complex
DNA polymerase ε complex
MCM complex (helicase)
RPA
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHII
RNaseHIII
Helicase
Dna2
Fen1
DNA ligase
Lig1
Data on KEGG graph
Rendered by Pathview


NUCLEOCYTOPLASMIC TRANSPORT
-1 0 1
Import
Export
mRNA Export
Cytoplasm
Lumen
Nucleus
NPC
Nuclear Pore complex (NPC)
Cytoplasmic fibrils
Nup358 complex
Exon-junction complex (EJC)
EJC inner core
EJC outer shell
Transiently interacting factors
Nuclear transport complex
Importin
Exportin
Adaptor proteins
Central channel
Spoke complex
Lumenal ring
Nuclear basket
Transcription-export (TREX) complex
THO subcomplex
mRNA surveilance pathway
Data on KEGG graph
Rendered by Pathview


OOCYTE MEIOSIS
-1 0 1
Progesterone
Progesterone-mediated oocyte maturation
MAPK signaling pathway
Fertilization
Early translation of maternal mRNAs
Late translation of maternal mRNAs
Late-late translation of maternal mRNAs
Ubiquitin mediated proteolysis
Degradation of CycB
Prevention of DNA replication
Bipolar spindle formation
Meiosis I progression
Meiosis I exit
Meiosis II arrest
Meiosis II resumption
Meiosis II exit
Delay to reentry into Meiosis
Arm cohesin
Pericentromeric cohesin
Prophase I
GVBD
Meiosis I
Interkinesis
Meiosis II
Data on KEGG graph
Rendered by Pathview



Gene Ontology (GO)

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

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

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

lapply(gobpres, head)
```

```
$greater
                                           p.geomean stat.mean        p.val
GO:0007156 homophilic cell adhesion        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

```r
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=
```

Then, to perform pathway analysis online go to the Reactome website (https://reactome.org/PathwayBrowser/#

> 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 most significant entities p-value was the mitotic cell cycle. The most significant pathways mostly match the previous KEGG results. The differences can be caused due to variations in database structure, pathway coverage, and gene annotations. KEGG focuses on metabolic and signaling pathways, while Reactome provides more detailed molecular interactions and hierarchical biological processes.

GO online (optional)

> 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 pathway with the most significant p-value is the cell cycle. This matches the previous KEGG results but it also includes the trachea formation pathway. The differences can arise due to differences in databases and gene annotations.