Class 14: RNASeq Mini-Project

AUTHOR
Christopher Leone | A16731724

Loading Files and Packages

Here we will perform a complete RNASeq analysis from counts to pathways and biological interpretation.

First, we have to load our packages and read the data into our project:

```
# Loading packages:
library(DESeq2)
library(AnnotationDbi)
library(org.Hs.eg.db)
library(pathview)
library(gage)
library(gageData)
```

```
# Import metadata:
colData <- read.csv("GSE37704_metadata.csv", row.names=1)
head(colData)
```

```
             condition
SRR493366 control_sirna
SRR493367 control_sirna
SRR493368 control_sirna
SRR493369      hoxa1_kd
SRR493370      hoxa1_kd
SRR493371      hoxa1_kd
```

```
# Import countdata:
countData <- read.csv("GSE37704_featurecounts.csv", row.names=1)
head(countData)
```

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

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

Tidying the Data

Let's check the correspondance of colData rows and countData columns.

```
rownames(colData) == colnames(countData)
```

```
Warning in rownames(colData) == colnames(countData): longer object length is
not a multiple of shorter object length
```

```
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
colnames(countData)
```

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

We see that we have an unwanted "length" column in `countData`, so let's remove it:

```
counts <- countData[,2:7]

# Double check that they match:
all(rownames(colData) == colnames(counts))
```

```
[1] TRUE
```

Remove Zero Count Genes

We will have rows in `counts` for genes that are insignificant to us due to having no expression. So, let's clear those genes:

```
# If the rowSums() of a gene is 0, let's remove that gene:
to.keep <- rowSums(counts) != 0
cleanCounts <- counts[to.keep,]
```

> How many genes do we have left after cleaning?

```
nrow(cleanCounts)
```

```
[1] 15975
```

There are 15975 genes remaining after clearing the empty entries.

Setup for DESeq Analysis

```
# We have to use the very long function here
dds <- DESeqDataSetFromMatrix(countData=cleanCounts,
                              colData=colData,
                              design=~condition)
```

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

Running DESeq Analysis

```
# Rewrite the variable with the results:
dds <- DESeq(dds)
```

```
estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

Extract the Results

We just need to declare a new `res` results variable:

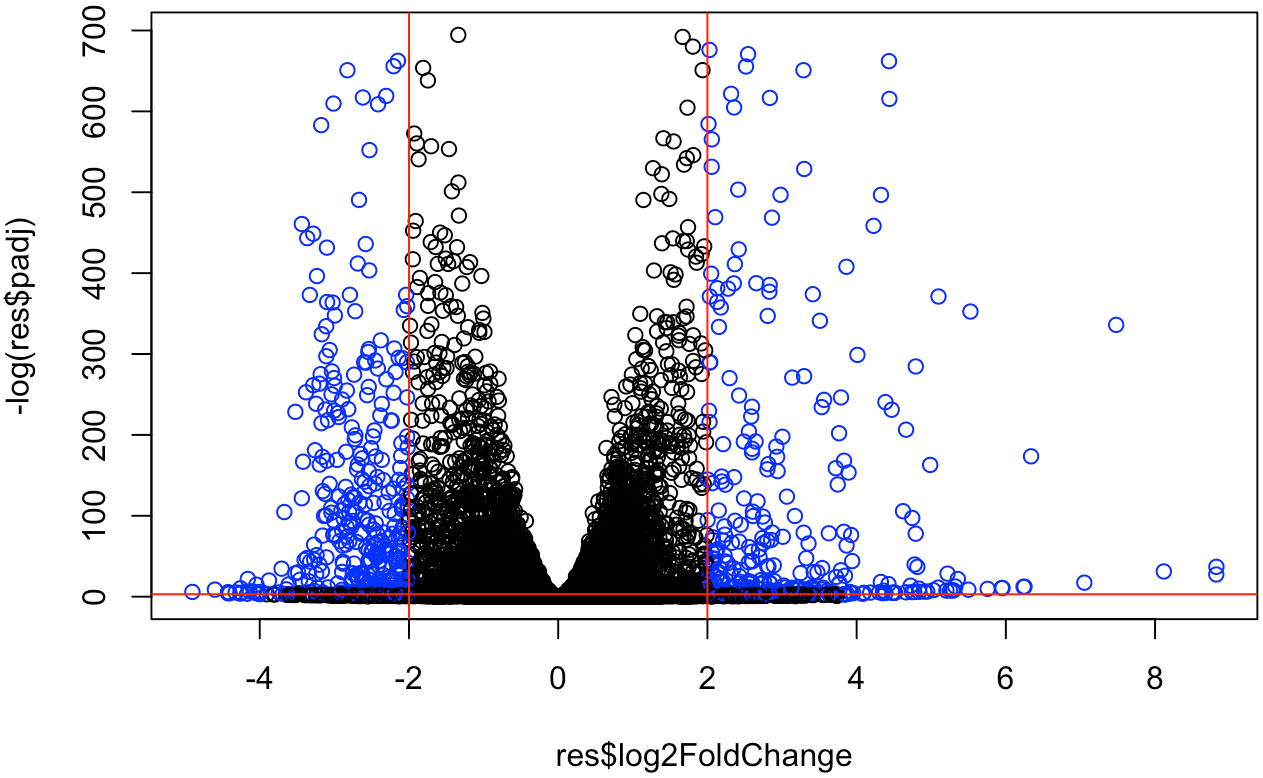
```
res <- results(dds)
head(res)
```

```
log2 fold change (MLE): condition hoxa1 kd vs control sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 6 rows and 6 columns
                 baseMean log2FoldChange     lfcSE       stat      pvalue
                <numeric>      <numeric> <numeric>  <numeric>   <numeric>
ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43989e-36
ENSG00000187961  209.6379      0.7297556 0.1318599   5.534326 3.12428e-08
```

```
ENSG00000187583   47.2551      0.0405765 0.2718928   0.149237 8.81366e-01
ENSG00000187642   11.9798      0.5428105 0.5215599   1.040744 2.97994e-01
                       padj
                    <numeric>
ENSG00000279457 6.86555e-01
ENSG00000187634 5.15718e-03
ENSG00000188976 1.76549e-35
ENSG00000187961 1.13413e-07
ENSG00000187583 9.19031e-01
ENSG00000187642 4.03379e-01
```

Let's remake a volcano plot with our new data:

```
# To color the points:
mycols = rep("black", nrow(res))
mycols[res$log2FoldChange <= -2] <- "blue"
mycols[res$log2FoldChange >= 2] <- "blue"
mycols[res$padj >= 0.05] <- "black"

# The plot:
plot(res$log2FoldChange, -log(res$padj), col=mycols)

# Finally, the boundaries of significant stats:
abline(v=2, col="red")
abline(v=-2, col="red")
abline(h=-log(0.05), col="red")
```



Add Gene Annotations

We have to re-annotate the gene data so we have gene symbols, gene names, and ENTREZ IDs.

```
# Translation to gene symbols:
res$symbol <- mapIds(x=org.Hs.eg.db,
                     keys=rownames(res),
                     keytype="ENSEMBL",
                     column="SYMBOL")
```

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

```
# Translation to gene names:
res$genename <- mapIds(x=org.Hs.eg.db,
                     keys=rownames(res),
                     keytype="ENSEMBL",
                     column="GENENAME")
```

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

```
# Translation to gene ENTREZ IDs:
res$entrezid <- mapIds(x=org.Hs.eg.db,
                       keys=rownames(res),
                       keytype="ENSEMBL",
                       column="ENTREZID")
```

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

Saving Results to a CSV File:

```
write.csv(res, file="class14data.csv")
```

Result Visualization

```
data(kegg.sets.hs)

# Examine the first 2 pathways in this kegg set for humans
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"
```

```
foldchanges = res$log2FoldChange
names(foldchanges) = res$entrezid
head(foldchanges)
```

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

```
# And finally, let's run gage!
keggres = gage(foldchanges, gsets=kegg.sets.hs)
```

```
attributes(keggres)
```

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

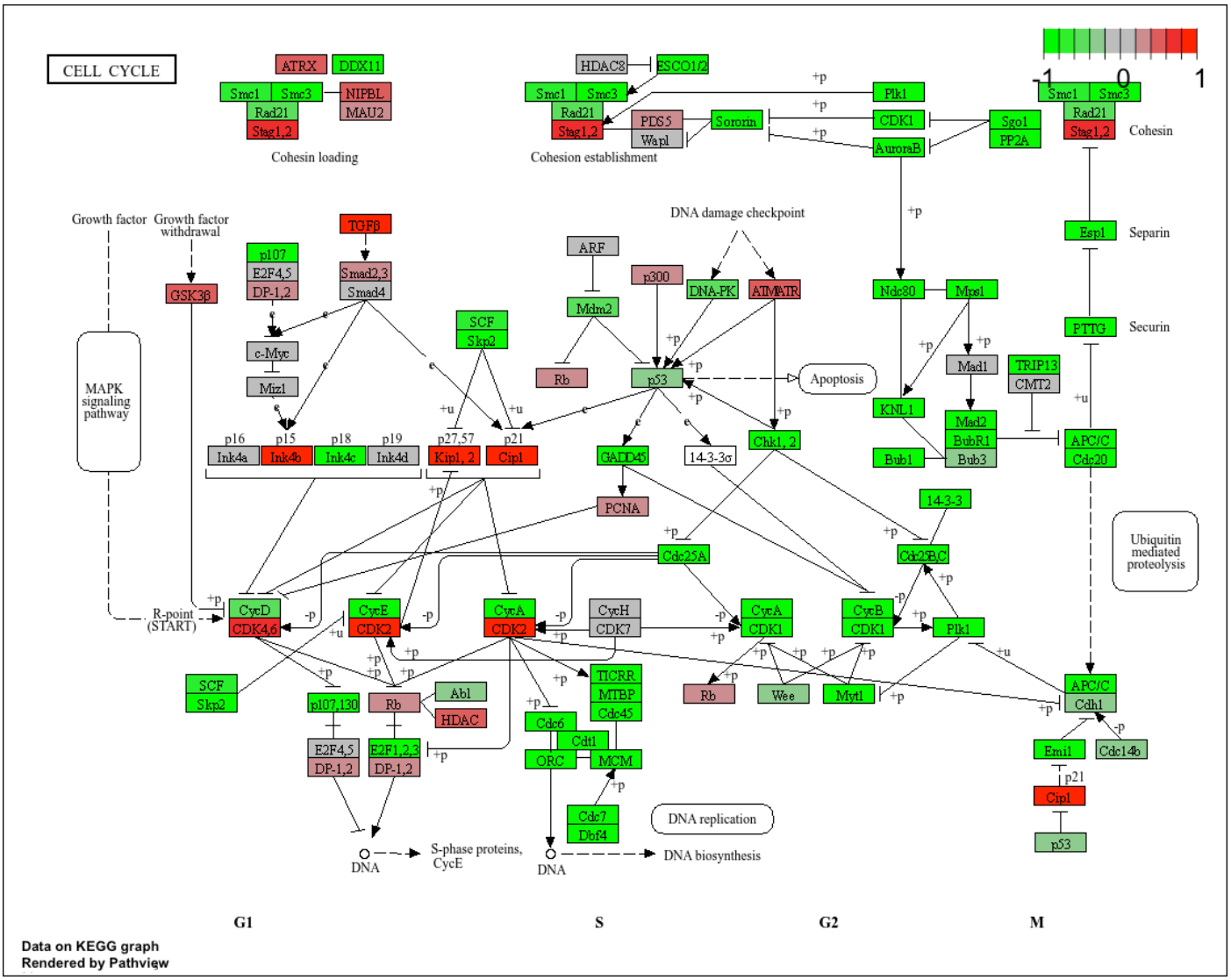
```
# What pathways overlap with what we have annotated?
head(keggres$less)
```

```
                                           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                     1.246882e-03 -3.059466
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                     1.246882e-03 0.065461279
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 1.246882e-03
hsa03440 Homologous recombination                28 3.066756e-03
hsa04114 Oocyte meiosis                         102 3.784520e-03
```

Pathway Analysis

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

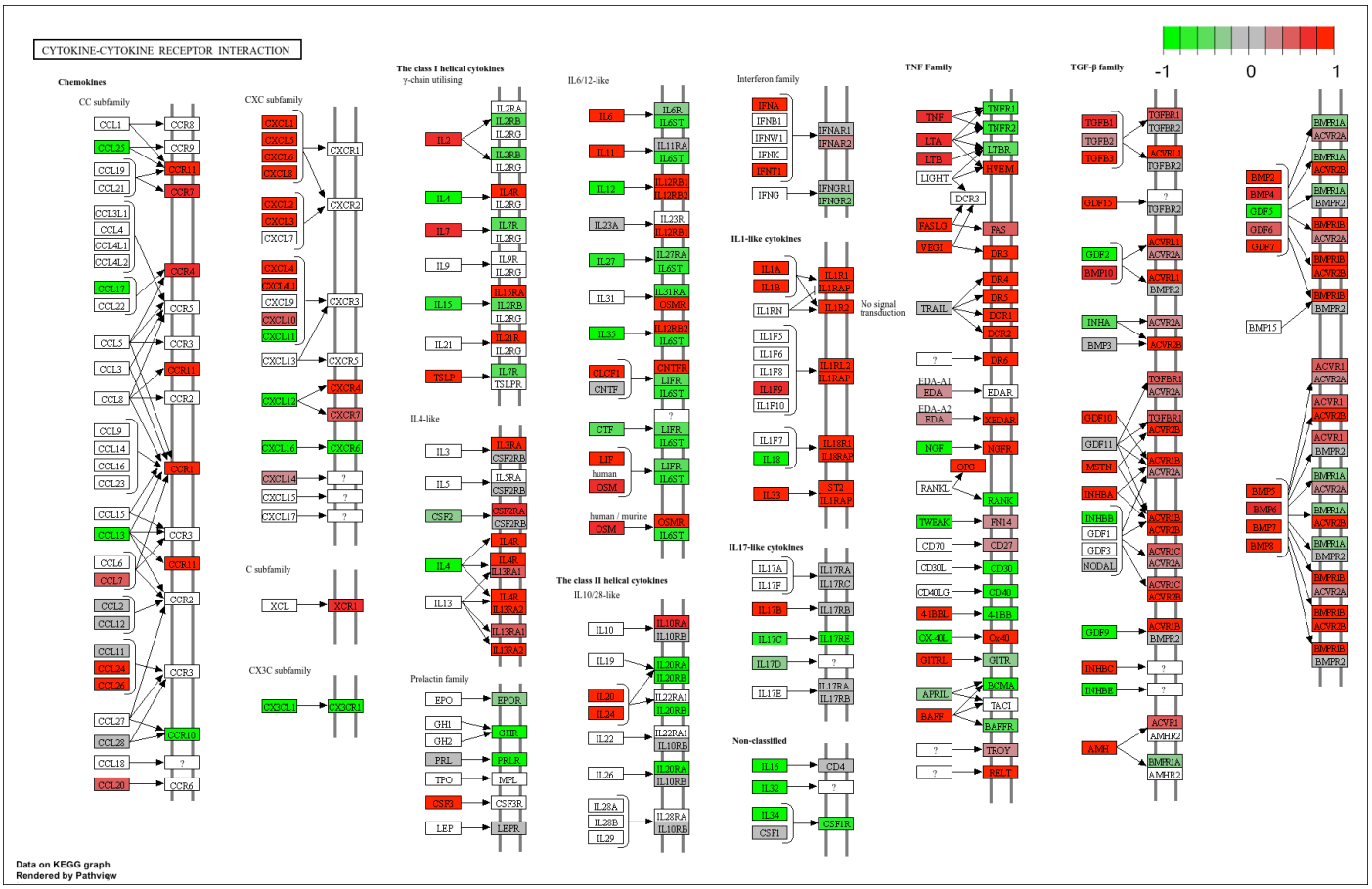
Once we have done that, we can take a look at our first pathway:

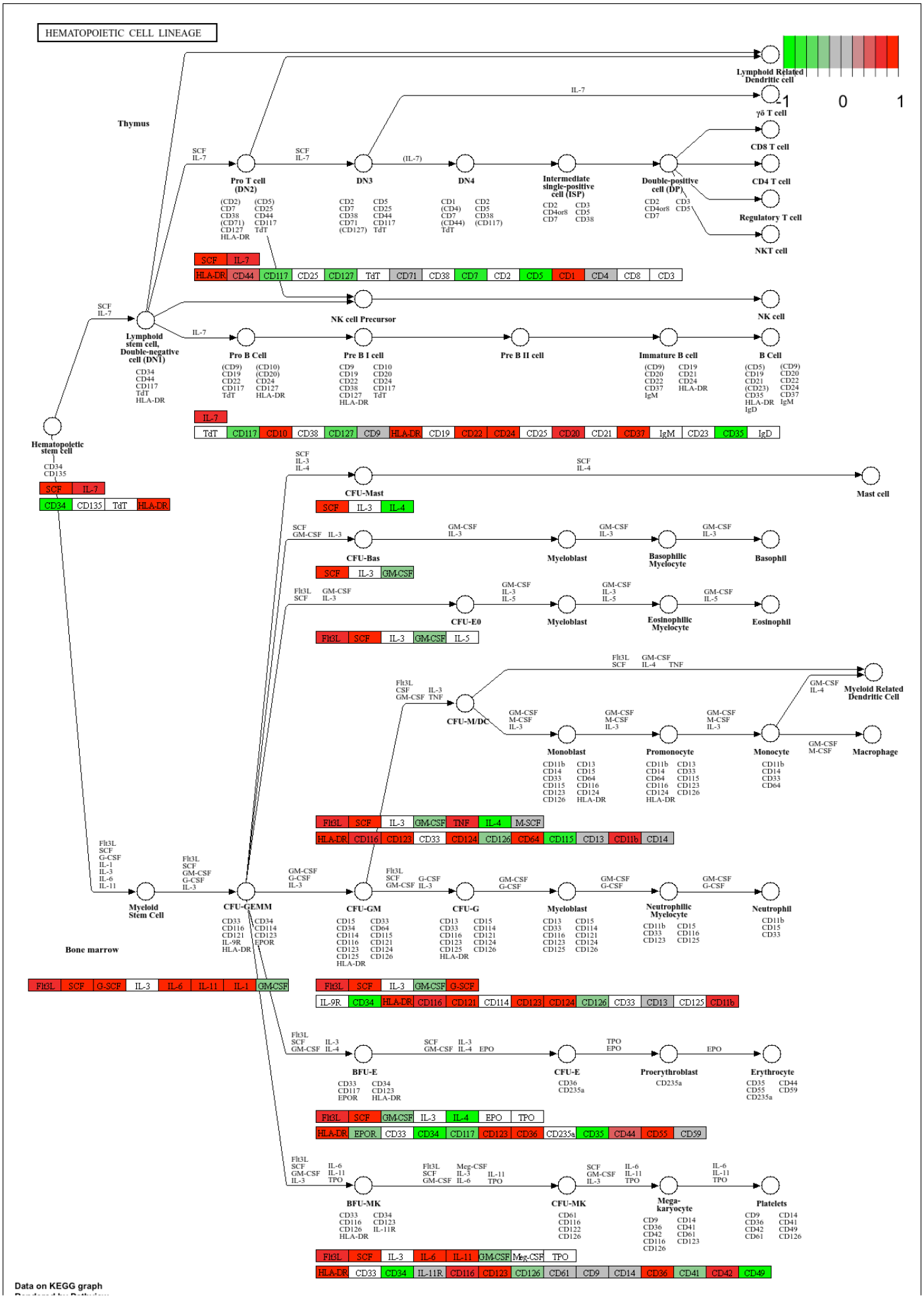


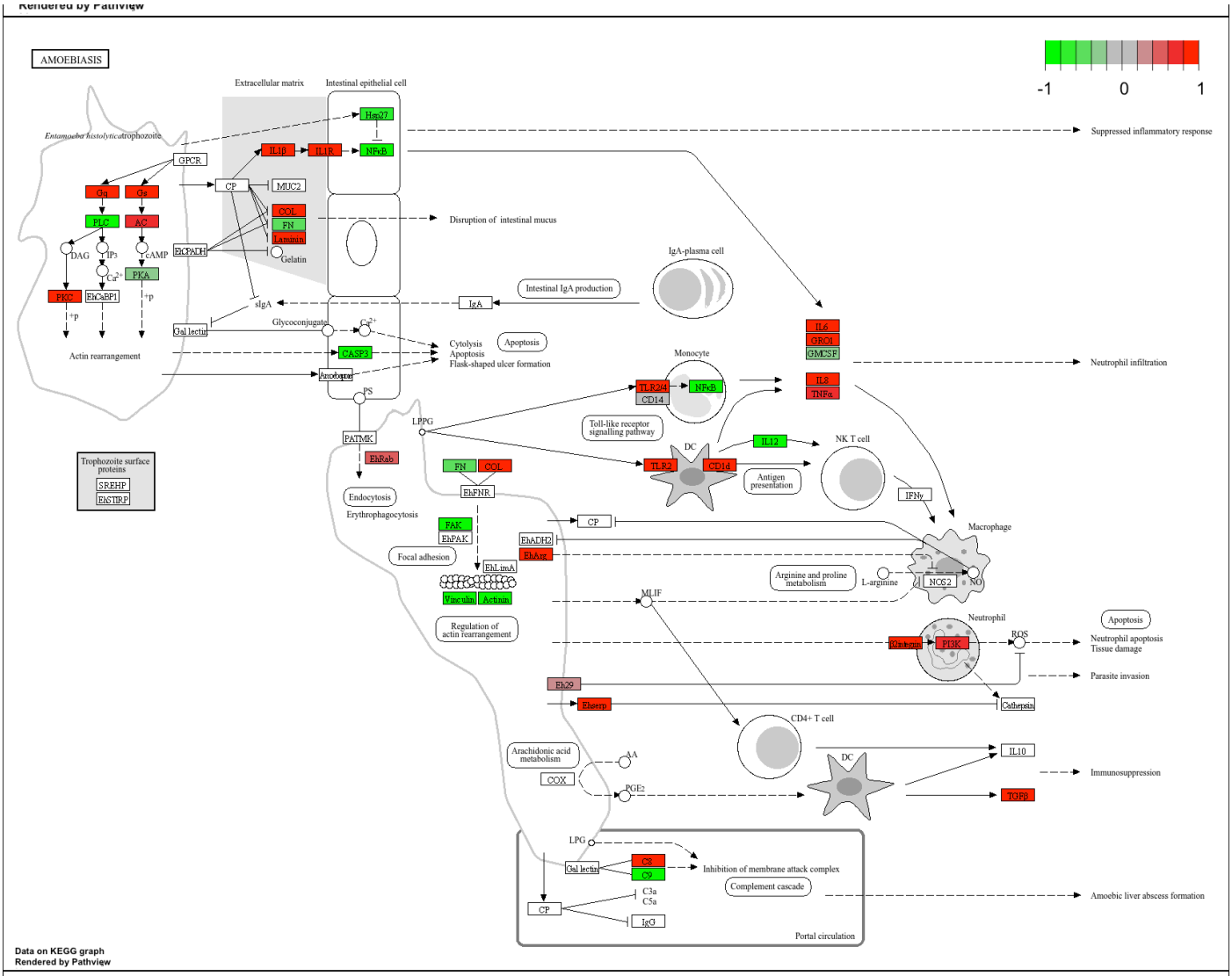
But, I want to see the remaining pathways as well, so we can do this with the following code:

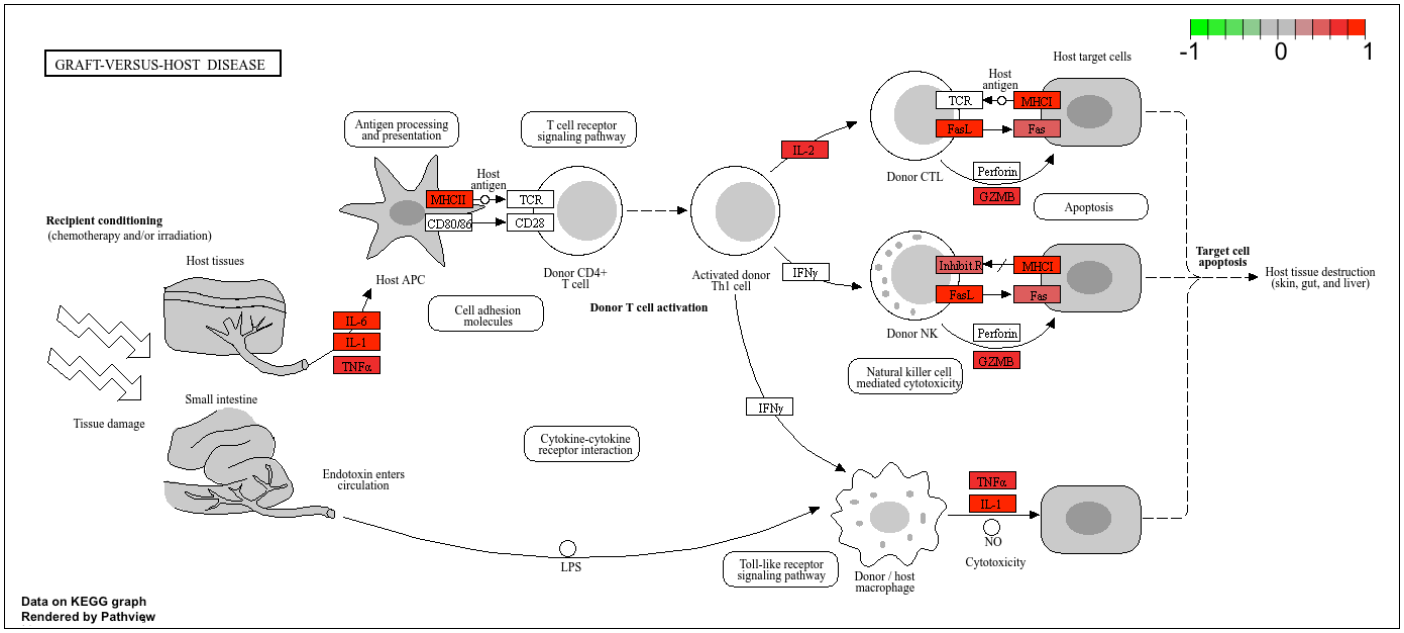
```
# Focus on top 5 upregulated pathways here for demo purposes only
keggrespathways <- rownames(keggres$greater)[1:5]

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

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

And here are the generated KEGG pathway images:


CYTOKINE-CYTOKINE RECEPTOR INTERACTION
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
IL17-like cytokines
Non-classified
TNF Family
TGF-β family
-1
0
1
No signal transduction
Data on KEGG graph
Rendered by Pathview


HEMATOPOIETIC CELL LINEAGE
-1 0 1
Hematopoietic stem cell
CD34 CD135
SCF IL-7
CD34 CD135 TdT HLA-DR
Lymphoid stem cell, Double-negative cell (DN1)
CD34 CD44 CD117 TdT HLA-DR
Thymus
Pro T cell (DN2)
(CD2) (CD5) CD7 CD25 CD38 CD44 (CD71) CD117 CD127 TdT HLA-DR
DN3
CD2 CD5 CD7 CD25 CD38 CD44 CD71 CD117 (CD127) TdT
DN4
CD1 CD2 (CD4) CD5 CD7 CD38 (CD44) (CD117) TdT
Intermediate single-positive cell (ISP)
CD2 CD3 CD4or8 CD5 CD7 CD38
Double-positive cell (DP)
CD2 CD3 CD4or8 CD5 CD7
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
SCF IL-7
HLA-DR CD44 CD117 CD25 CD127 TdT CD71 CD38 CD7 CD2 CD5 CD1 CD4 CD8 CD3
NK cell Precursor
NK cell
Pro B Cell
(CD9) (CD10) CD19 (CD20) CD22 CD24 CD117 CD127 TdT HLA-DR
Pre B I cell
CD9 CD10 CD19 CD20 CD22 CD24 CD38 CD117 CD127 TdT HLA-DR
Pre B II cell
Immature B cell
(CD9) CD19 CD20 CD21 CD22 CD24 CD37 HLA-DR IgM
B Cell
(CD5) (CD9) CD19 CD20 CD21 CD22 (CD23) CD24 CD35 CD37 HLA-DR IgM IgD
IL-7
TdT CD117 CD10 CD38 CD127 CD9 HLA-DR CD19 CD22 CD24 CD25 CD20 CD21 CD37 IgM CD23 CD35 IgD
Lymphoid Related Dendritic cell
γδ T cell
SCF IL-4
Mast cell
CFU-Mast
SCF IL-3 IL-4
SCF GM-CSF IL-3
CFU-Bas
SCF IL-3 GM-CSF
Myeloblast
Basophilic Myelocyte
Basophil
Flt3L SCF GM-CSF IL-3
CFU-E0
Flt3L SCF IL-3 GM-CSF IL-5
Eosinophilic Myelocyte
Eosinophil
Flt3L SCF GM-CSF IL-4 TNF
Myeloid Related Dendritic Cell
Flt3L CSF GM-CSF IL-3 TNF
CFU-M/DC
Monoblast
CD11b CD13 CD14 CD15 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
Promonocyte
CD11b CD13 CD14 CD33 CD64 CD115 CD116 CD123 CD124 CD126 HLA-DR
Monocyte
CD11b CD14 CD33 CD64
Macrophage
Flt3L SCF IL-3 GM-CSF TNF IL-4 M-SCF
HLA-DR CD116 CD123 CD33 CD124 CD126 CD64 CD115 CD13 CD11b CD14
Flt3L SCF G-CSF IL-1 IL-3 IL-6 IL-11
Myeloid Stem Cell
Flt3L SCF GM-CSF G-CSF IL-3
CFU-GEMM
CD33 CD34 CD116 CD114 CD121 CD123 IL-9R EPOR HLA-DR
GM-CSF G-CSF IL-3
CFU-GM
CD15 CD33 CD34 CD64 CD114 CD115 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
Flt3L SCF GM-CSF G-CSF IL-3
CFU-G
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126 HLA-DR
GM-CSF G-CSF
Myeloblast
CD13 CD15 CD33 CD114 CD116 CD121 CD123 CD124 CD125 CD126
Neutrophilic Myelocyte
CD11b CD15 CD33 CD116 CD123 CD125
Neutrophil
CD11b CD15 CD33
Bone marrow
Flt3L SCF G-CSF IL-3 IL-6 IL-11 IL-1 GM-CSF
Flt3L SCF IL-3 GM-CSF G-CSF
IL-9R CD34 HLA-DR CD116 CD121 CD114 CD123 CD124 CD126 CD33 CD13 CD125 CD11b
Flt3L SCF GM-CSF IL-3 IL-4
BFU-E
CD33 CD34 CD117 CD123 EPOR HLA-DR
SCF GM-CSF IL-3 IL-4 EPO
CFU-E
CD36 CD235a
TPO EPO
Proerythroblast
CD235a
EPO
Erythrocyte
CD35 CD44 CD55 CD59 CD235a
Flt3L SCF GM-CSF IL-3 IL-4 EPO TPO
HLA-DR EPOR CD33 CD34 CD117 CD123 CD36 CD235a CD35 CD44 CD55 CD59
Flt3L SCF GM-CSF IL-3 IL-6 IL-11 TPO
BFU-MK
CD33 CD34 CD116 CD123 CD126 IL-11R HLA-DR
Flt3L SCF GM-CSF IL-3 Meg-CSF IL-6 IL-11 TPO
CFU-MK
CD61 CD116 CD122 CD126
SCF GM-CSF IL-3 IL-6 IL-11 TPO
Mega-karyocyte
CD9 CD14 CD36 CD41 CD42 CD61 CD116 CD123 CD126
IL-6 IL-11 TPO
Platelets
CD9 CD14 CD36 CD41 CD42 CD49 CD61 CD126
Flt3L SCF IL-3 IL-6 IL-11 GM-CSF Meg-CSF TPO
HLA-DR CD33 CD34 IL-11R CD116 CD123 CD126 CD61 CD9 CD14 CD36 CD41 CD42 CD49
Data on KEGG graph


Rendered by Pathview
AMOEBIASIS
-1 0 1
Extracellular matrix
Intestinal epithelial cell
Entamoeba histolytica trophozoite
GPCR
Gq
Gs
PLC
AC
DAG
IP3
cAMP
Ca2+
PKA
PKC
EhCaBP1
+p
Actin rearrangement
CP
MUC2
COL
FN
Laminin
EhCPADH
Gelatin
Disruption of intestinal mucus
Hsp27
IL1β
IL1R
NFκB
Suppressed inflammatory response
IgA-plasma cell
Intestinal IgA production
IgA
sIgA
Gal lectin
Glycoconjugate
Ca2+
Cytolysis
Apoptosis
Flask-shaped ulcer formation
Apoptosis
CASP3
Amoebapore
PS
PATMK
EhRab
Trophozoite surface proteins
SREHP
EhSTIRP
LPPG
Toll-like receptor signalling pathway
Monocyte
TLR2/4
CD14
NFκB
IL6
GRO1
GMCSF
IL8
TNFα
Neutrophil infiltration
DC
TLR2
CD1d
IL12
Antigen presentation
NK T cell
IFNγ
Endocytosis
Erythrophagocytosis
FN
COL
EhFNR
FAK
EhPAK
Focal adhesion
EhADH2
EhArg
EhLimA
Vinculin
Actinin
Regulation of actin rearrangement
CP
Macrophage
Arginine and proline metabolism
L-arginine
NOS2
NO
MIF
Neutrophil
Apoptosis
Calreticulin
PI3K
ROS
Neutrophil apoptosis
Tissue damage
Parasite invasion
Eh29
Ehcp5
Cathepsin
CD4+ T cell
DC
IL10
Immunosuppression
TGFβ
Arachidonic acid metabolism
AA
COX
PGE2
LPG
Gal lectin
C8
C9
Inhibition of membrane attack complex
Complement cascade
C3a
C5a
CP
IgG
Amoebic liver abscess formation
Portal circulation
Data on KEGG graph
Rendered by Pathview
RHEUMATOID ARTHRITIS
-1 0 1
Blood vessel
Inflammatory cell infiltration
Leukocyte migration
Angiogenesis
Ang1
Tie2
VEGF
Flt1
VEGF signaling pathway
CCL2
CXCL1
CCL3
CXCL5
CCL20
IL8
DC
CD80/86
MHCII
Antigen
T cell receptor signaling pathway
Self-reactive Th1 cell
IFNγ
RANKL
CD28
CTLA4
TCR
LFA1
IL15
ICAM1
Synovial macrophage
B cell
APRIL
BLYS
IgG
Autoantibody production
SDF1
LTB
Toll-like receptor signaling pathway
Peptidoglycan
LPS
TLR2/4
AP1
RANKL
Synovial fibroblast
GMCSF
IL6
IL1β
CCL5
Inflammation
Synovial pannus formation
TNFα
IL1
IL11
IL6
IL18
TGFβ
Th17 cell differentiation
Th17 cell
IL6
IL23
IL17
PTH
Vitamin D3
MCSF
PGE2
RANKL
Osteoblast
Osteoclast differentiation
Osteoclast
RANK
MMP1/3
CTSL
Joint destruction
Synovium
Bone
LAT
H
CTSK
TRAP
Bone resorption
Data on KEGG graph
Rendered by Pathview



Gene Ontology Subsets

```
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
```

Reactome Analysis

We need to make a little file of our significant genes so that we can upload to the online reactome webpage:

```
# Here are our significant genes:
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"
```

```
# And let's write the file:
write.table(sig_genes, file="significant_genes.txt", row.names=FALSE,
            col.names=FALSE, quote=FALSE)
```

And here is a diagram of one of the pathways we found:

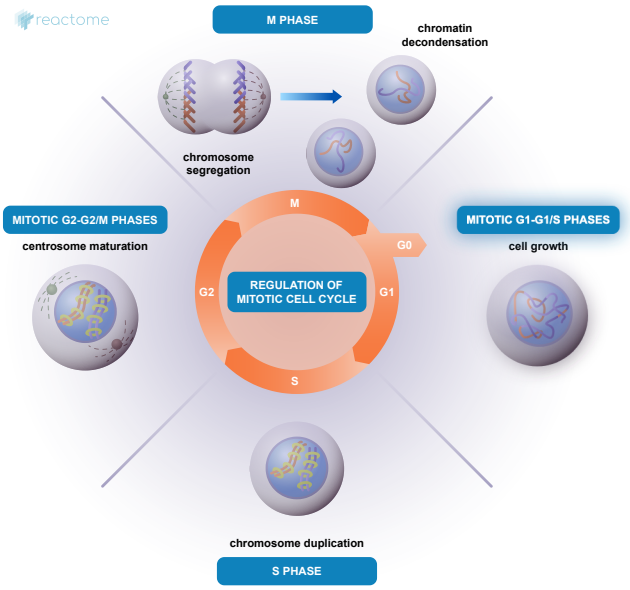


Figure 1: The Mitotic Cell Cycle.