Class 14: RNA-seq analysis mini-project

Joseph Elmaghraby (A16788229) 02/20/25

Table of contents

Background

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

##Inspect and Tidy Data

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

```
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
```

```
         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
```

```
colData$id
```

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

```
colnames(counts)
```

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

The fix here looks to be removing the first "length" 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 contData and colData

```
colnames(countData)==colData$id
```

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

Q1. How many gnes 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
```

```
dds<- DESeqDataSetFromMatrix(countData = new.counts,
                             colData= colData,
                             design= ~condition)
```

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

##Run DESeq

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

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

```
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
```

##Volcano plot results

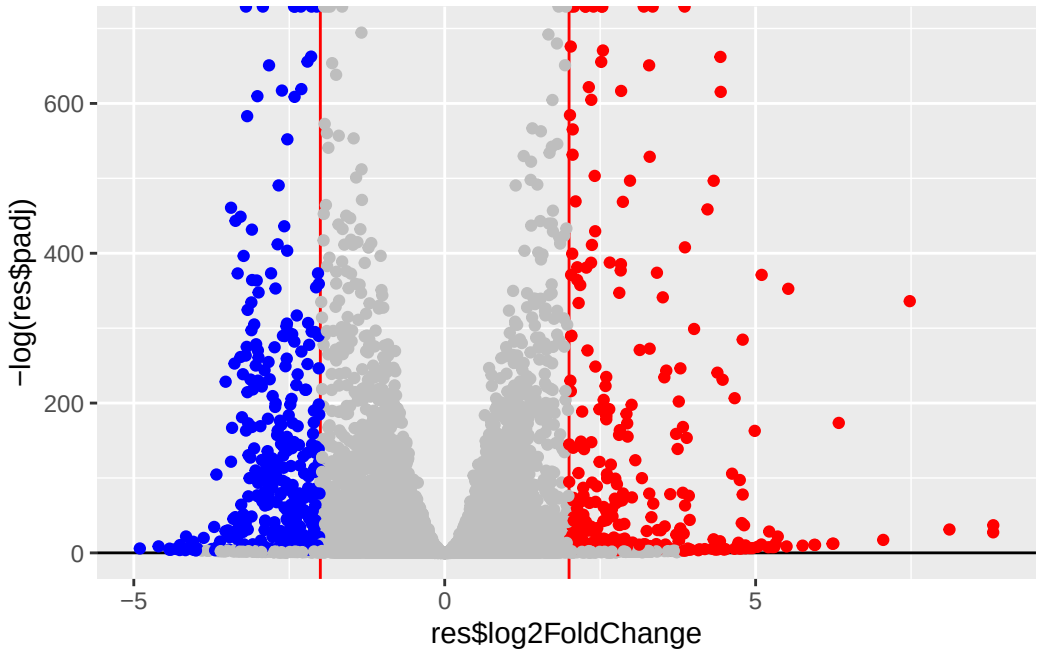
```
library(ggplot2)
```

```
mycols<-rep("gray",nrow(res))
mycols[res$log2FoldChange >= 2] <- "red"
mycols[res$log2FoldChange <= -2]<- "blue"
mycols[res$padj > 0.05]<- "gray"
```

```
library(ggplot2)

ggplot(res)+
  aes(x=res$log2FoldChange, y=-log(res$padj))+
  geom_vline(xintercept = c(-2,2), col="red") +
  geom_hline(yintercept=0.05)+
  geom_point(col=mycols)
```

```
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"
```

Add gene symbol and entrez

```
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
```

```
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 8 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      symbol      entrez
                 <numeric> <character> <character>
ENSG00000279457 6.86555e-01          NA          NA
ENSG00000187634 5.15718e-03      SAMD11      148398
ENSG00000188976 1.76549e-35       NOC2L       26155
ENSG00000187961 1.13413e-07      KLHL17      339451
ENSG00000187583 9.19031e-01     PLEKHN1       84069
ENSG00000187642 4.03379e-01       PERM1       84808
```

##Pathway Analysis

```
library(gage)
```

```
library(gageData)
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).
################################################################################
```

Input vector for `gage()`

```
foldchanges= res$log2FoldChange
names(foldchanges) = res$entrez
```

Load up the KEGG

```
data(kegg.sets.hs)
```

Run Pathway analysis

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

```
head(keggres$less, 3)
```

```
                                              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
                                                    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
                                                  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
```

Cell Cycle figure

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

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

```
Info: Working in directory /Users/josephelmaghraby/Desktop/BIMM 143 /Rstudio/Class 14
```

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

```
pathview(foldchanges, pathway.id = "hsa03030")
```

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

```
Info: Working in directory /Users/josephelmaghraby/Desktop/BIMM 143 /Rstudio/Class 14
```

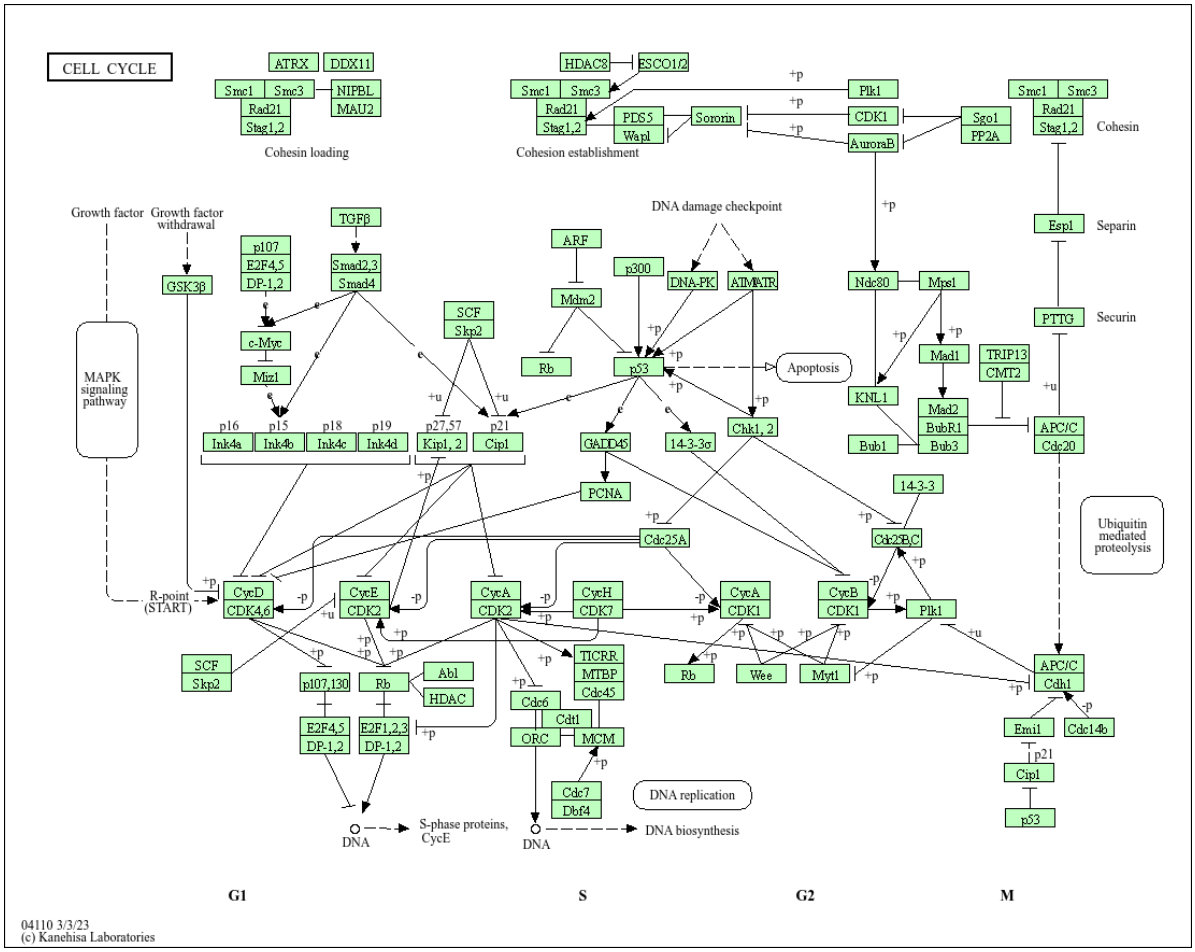
```
Info: Writing image file hsa03030.pathview.png
```

```
pathview(foldchanges, pathway.id = "hsa05130")
```

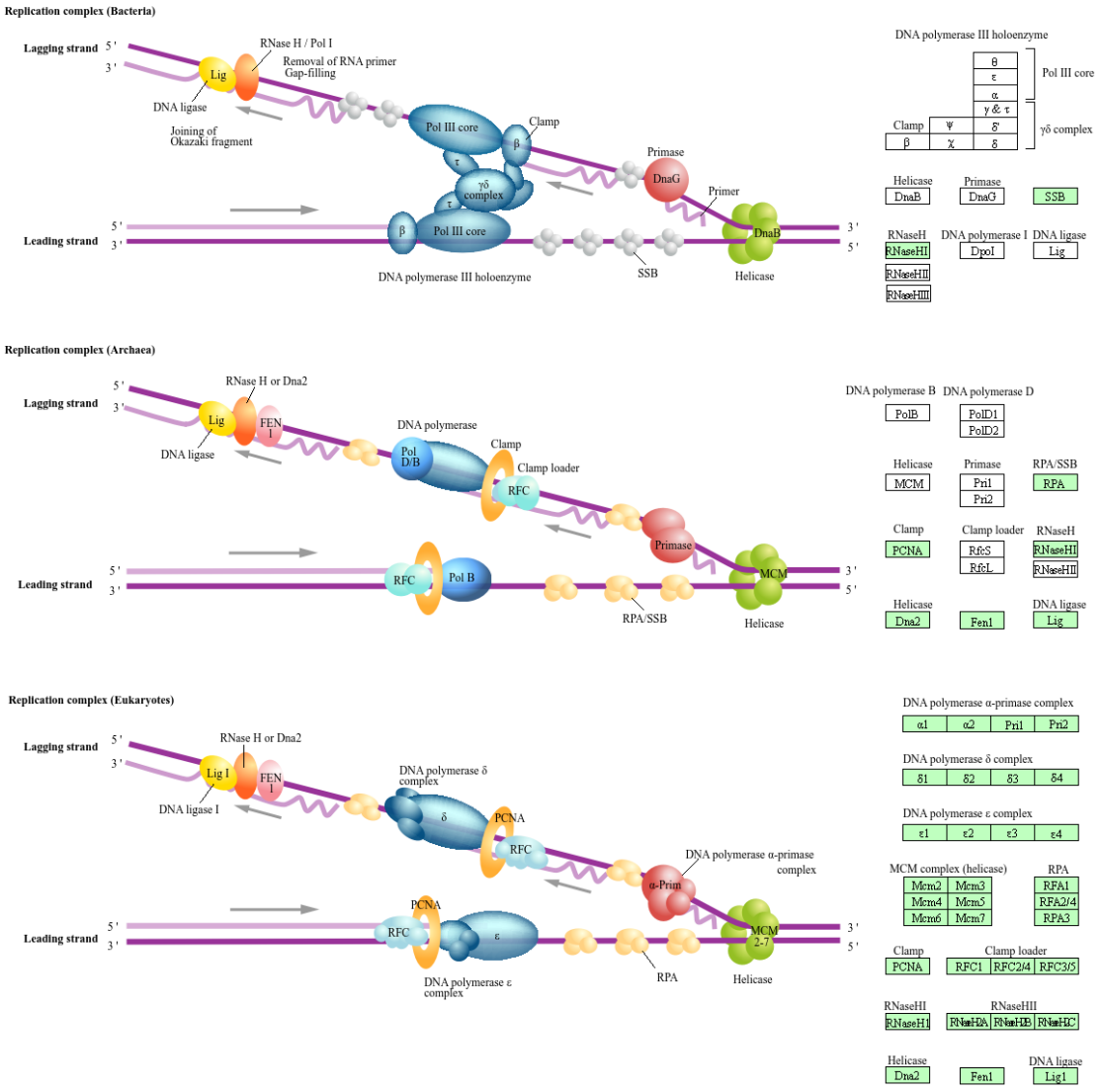
```
'select()' returned 1:1 mapping between keys and columns
```

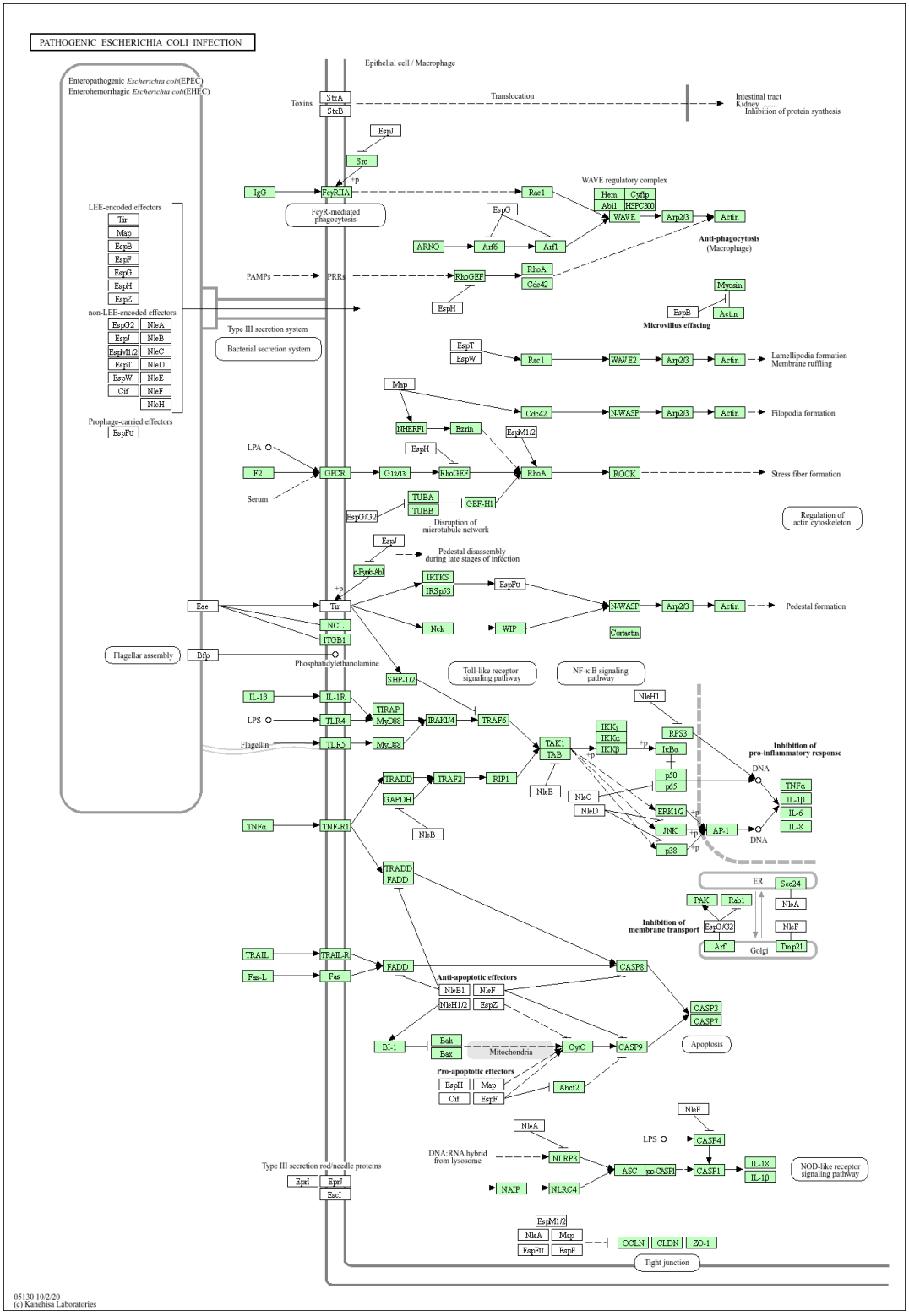
```
Info: Working in directory /Users/josephelmaghraby/Desktop/BIMM 143 /Rstudio/Class 14
```

```
Info: Writing image file hsa05130.pathview.png
```


CELL CYCLE
ATRX
DDX11
Smc1
Smc3
NIPBL
MAU2
Rad21
Stag1,2
Cohesin loading
HDAC8
ESCO1/2
PDS5
Wapl
Sororin
Plk1
CDK1
AuroraB
Sgo1
PP2A
Cohesion establishment
Espl1
Separin
PTTG
Securin
Cohesin
DNA damage checkpoint
Growth factor
Growth factor withdrawal
GSK3β
TGFβ
Smad2,3
Smad4
p107
E2F4,5
DP-1,2
c-Myc
Miz1
MAPK signaling pathway
ARF
Mdm2
p300
DNA-PK
ATM
ATR
SCF
Skp2
Rb
p53
Apoptosis
Ndc80
Mps1
Mad1
TRIP13
CMT2
KNL1
Mad2
BubR1
Bub1
Bub3
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
DNA replication
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 biosynthesis
G1
S
G2
M
04110 3/3/23
(c) Kanehisa Laboratories

DNA REPLICATION




PATHOGENIC ESCHERICHIA COLI INFECTION
Epithelial cell / Macrophage
Enteropathogenic Escherichia coli(EPEC)
Enterohemorrhagic Escherichia coli(EHEC)
Toxins
StxA
StxB
Translocation
Intestinal tract
Kidney
Inhibition of protein synthesis
EspJ
Src
IgG
FcγRIIA
FcγR-mediated phagocytosis
Rac1
WAVE regulatory complex
Hem
Cyfip
Abi1
HSPC300
WAVE
Arp2/3
Actin
Anti-phagocytosis
(Macrophage)
EspG
ARNO
Arf6
Arf1
LEE-encoded effectors
Tir
Map
EspB
EspF
EspG
EspH
EspZ
non-LEE-encoded effectors
EspG2
NleA
EspJ
NleB
EspM1/2
NleC
EspT
NleD
EspW
NleE
Cif
NleF
NleH
Prophage-carried effectors
EspFU
PAMPs
PRRs
RhoGEF
RhoA
Cdc42
EspH
Myosin
EspB
Actin
Microvillus effacing
Type III secretion system
Bacterial secretion system
EspT
EspW
Rac1
WAVE2
Arp2/3
Actin
Lamellipodia formation
Membrane ruffling
Map
Cdc42
N-WASP
Arp2/3
Actin
Filopodia formation
NHERF1
Ezrin
EspM1/2
LPA
F2
Serum
GPCR
G12/13
RhoGEF
RhoA
ROCK
Stress fiber formation
TUBA
TUBB
GEF-H1
EspO/O2
Disruption of microtubule network
Regulation of actin cytoskeleton
EspJ
Pedestal disassembly during late stages of infection
c-Pupt-Abl
IRTKS
IRSp53
EspFU
Eae
Tir
N-WASP
Arp2/3
Actin
Pedestal formation
NCL
ITGB1
Nck
WIP
Cortactin
Flagellar assembly
BfpB
Phosphatidylethanolamine
SHP-1/2
Toll-like receptor signaling pathway
NF-κ B signaling pathway
NleH1
IL-1β
IL-1R
LPS
TLR4
TIRAP
MyD88
IRAK1/4
TRAF6
TAK1
TAB
IKKγ
IKKα
IKKβ
RPS3
IκBα
p50
p65
Inhibition of pro-inflammatory response
DNA
TNFα
IL-1β
IL-6
IL-8
Flagellin
TLR5
MyD88
TRADD
TRAF2
RIP1
GAPDH
NleE
NleC
NleD
ERK1/2
JNK
p38
AP-1
DNA
TNFα
TNF-R1
NleB
TRADD
FADD
ER
Sec24
PAK
Rab1
NleA
Inhibition of membrane transport
EspG/G2
NleF
Arf
Golgi
Tmp21
TRAIL
TRAIL-R
Fas-L
Fas
FADD
CASP8
Anti-apoptotic effectors
NleB1
NleF
NleH1/2
EspZ
CASP3
CASP7
Bi-1
Bak
Bax
Mitochondria
CytC
CASP9
Apoptosis
Pro-apoptotic effectors
EspH
Map
Cif
EspF
Abcf2
NleF
NleA
LPS
CASP4
DNA:RNA hybrid from lysosome
NLRP3
ASC
pro-CASP1
CASP1
IL-18
IL-1β
NOD-like receptor signaling pathway
Type III secretion rod/needle proteins
EprI
EprJ
EscI
NAIP
NLRC4
EspM1/2
NleA
Map
EspFU
EspF
OCLN
CLDN
ZO-1
Tight junction
05130 10/2/20
(c) Kanehisa Laboratories

##Gene Ontology Analysis

Run pathway analysis with 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)

head(gobpres$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
```