Class 14: Pathway Analysis from RNA-Seq Results

Clarissa Savko (PID: A69028482)

Table of contents

```r
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, setdiff, sort,
    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

Warning: package 'GenomeInfoDb' was built under R version 4.3.2

Loading required package: SummarizedExperiment
Loading required package: MatrixGenerics

Loading required package: matrixStats

Warning: package 'matrixStats' was built under R version 4.3.2

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
```

Data Import

```
metaFile <- "GSE37704_metadata.csv"
countFile <- "GSE37704_featurecounts.csv"

# Import metadata and take a peak
colData = read.csv(metaFile, row.names=1)
head(colData)
```

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

Data Tidying

```
countData = read.csv(countFile, 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
```

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

```
countData <- as.matrix(countData[,-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
```

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).

How many genes do we have to start with?

```
nrow(countData)
```

```
[1] 19808
```

1) Find the rowSums() this will be zero for any genes with no count data
2) Find the zero sum genes
3) Remove them

```
to.rm.inds <- rowSums(countData) == 0
countData <- countData[!to.rm.inds,]
nrow(countData)
```

```
[1] 15975
```

DESeq setup and analysis

```
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
```

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

```
estimating size factors

estimating dispersions

gene-wise dispersion estimates

mean-dispersion relationship

final dispersion estimates

fitting model and testing
```

Save Results

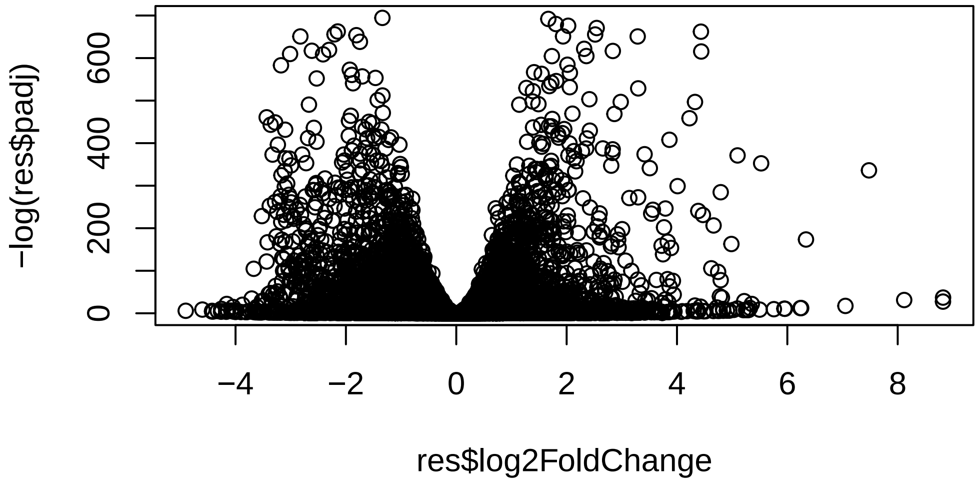
Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

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

Visualize

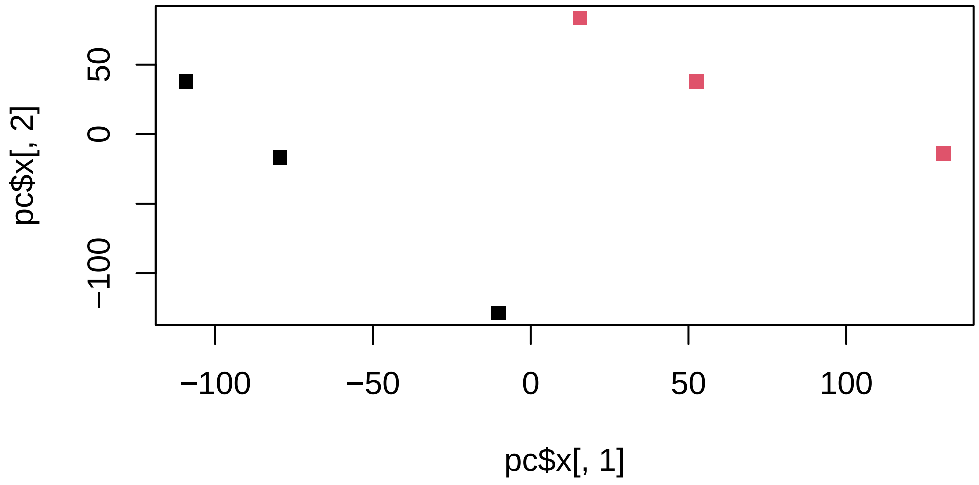
```
plot( res$log2FoldChange, -log(res$padj) )
```



```
pc <- prcomp(t(countData), scale= T)
summary(pc)
```

```
Importance of components:
                           PC1     PC2      PC3      PC4      PC5       PC6
Standard deviation     87.7211 73.3196 32.89604 31.15094 29.18417 7.373e-13
Proportion of Variance  0.4817  0.3365  0.06774  0.06074  0.05332 0.000e+00
Cumulative Proportion   0.4817  0.8182  0.88594  0.94668  1.00000 1.000e+00
```

```
plot(pc$x[,1], pc$x[,2], col=as.factor(colData$condition), pch=15)
```



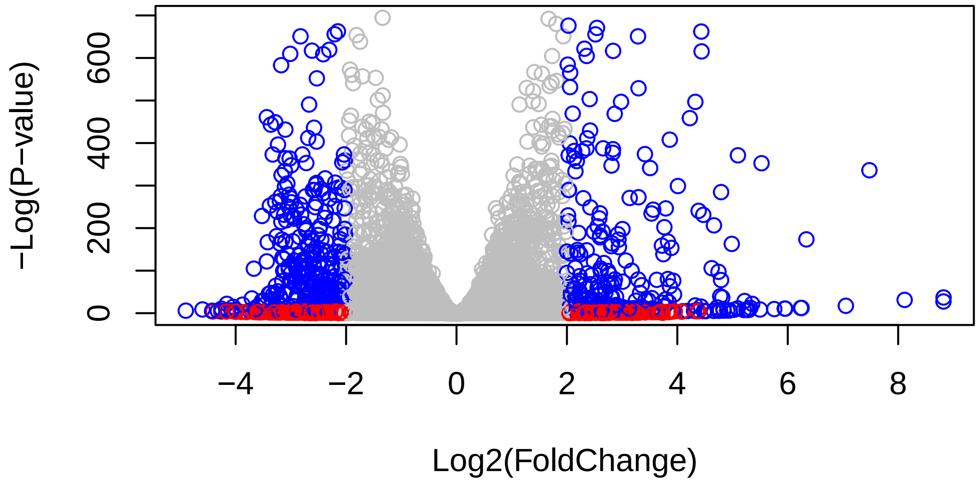
Q. Improve this plot by completing the below code, which adds color and axis labels

```
# Make a color vector for all genes
mycols <- rep("gray", nrow(res) )

# Color red the genes with absolute fold change above 2
mycols[ abs(res$log2FoldChange) > 2 ] <- "red"

# Color blue those with adjusted p-value less than 0.01
#  and absolute fold change more than 2
inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"

plot( res$log2FoldChange, -log(res$padj), col= mycols, xlab="Log2(FoldChange)", ylab="-Log
```



Annotation Data

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

```
library("AnnotationDbi")
```

```
Warning: package 'AnnotationDbi' was built under R version 4.3.2
```

```
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"
```

```r
res$symbol = mapIds(org.Hs.eg.db,
                    keys= row.names(res),
                    keytype="ENSEMBL",
                    column= "SYMBOL",
                    multiVals="first")
```

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

```r
res$entrez = mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype="ENSEMBL",
                    column="ENTREZID",
                    multiVals="first")
```

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

```r
res$name =   mapIds(org.Hs.eg.db,
                    keys=row.names(res),
                    keytype= "ENSEMBL",
                    column= "GENENAME",
                    multiVals="first")
```

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

```r
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                   name
                  <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.

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

Geneset enrichment/pathway analysis

```r
library(gage)
library(gageData)
library(pathview)
```

The `gage()` function wants a "vector of importance" in our case here it will be fold-change values with associated entrez gene names.

```r
foldchange <- res$log2FoldChange
names(foldchange) = res$entrez
```

```
data(kegg.sets.hs)
keggres = gage(foldchange, 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
```

hsa04110 cell cycle

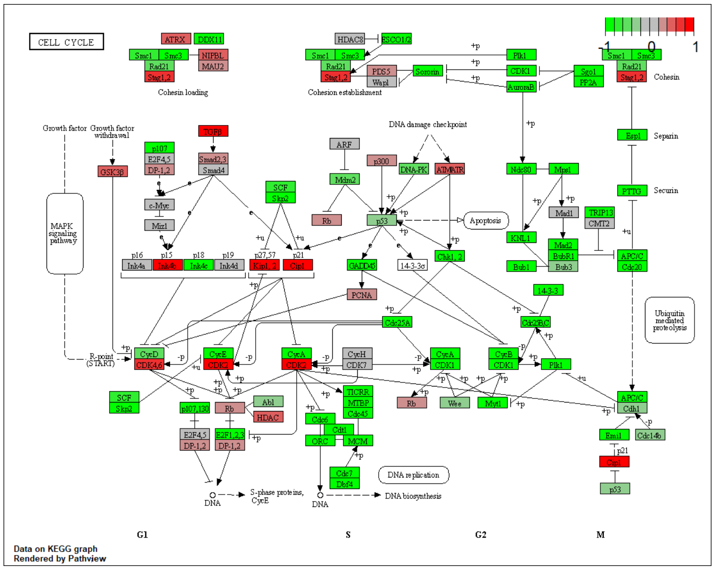
```
pathview(gene.data= foldchange, pathway.id = "hsa04110")
```

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

```
Info: Working in directory C:/Users/clari/Desktop/BGGN213 Bioinformatics/R files BGGN213/Cla
```

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

Have a look at my figure (**?@fig-cellcycle**)



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

```
pathview(gene.data= foldchange, pathway.id = "hsa03030")
```

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

Info: Working in directory C:/Users/clari/Desktop/BGGN213 Bioinformatics/R files BGGN213/Cla

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

```
pathview(gene.data= foldchange, pathway.id = "hsa05130")
```

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

Info: Working in directory C:/Users/clari/Desktop/BGGN213 Bioinformatics/R files BGGN213/Cla

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

```
pathview(gene.data= foldchange, pathway.id = "hsa03013")
```

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

```
Info: Working in directory C:/Users/clari/Desktop/BGGN213 Bioinformatics/R files BGGN213/Cla
```

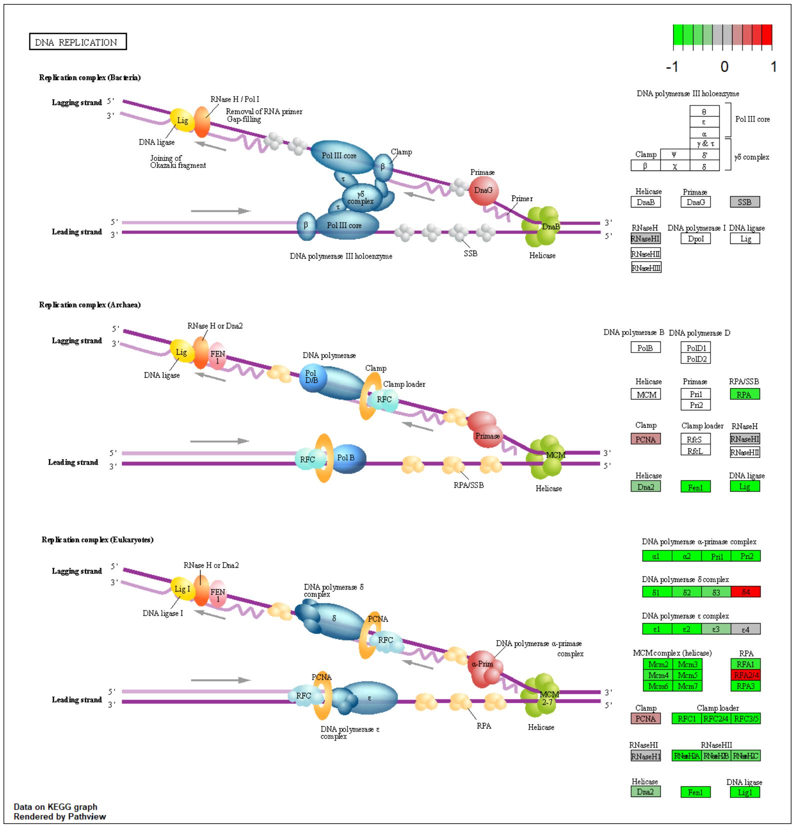
```
Info: Writing image file hsa03013.pathview.png
```

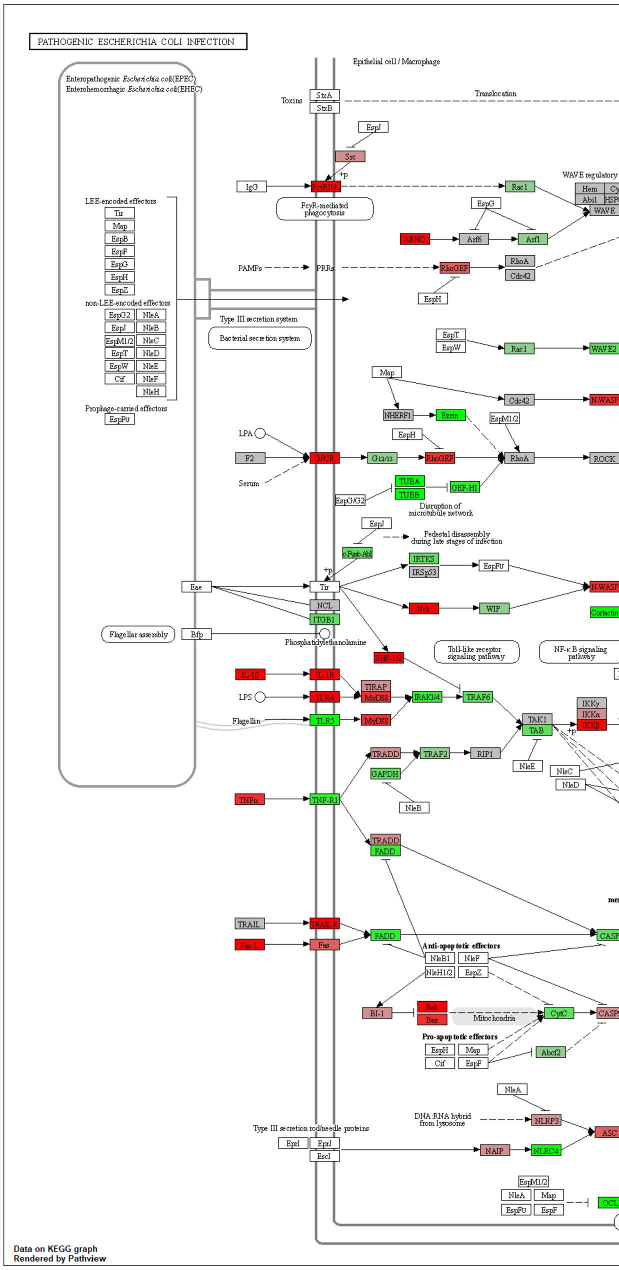
```r
pathview(gene.data= foldchange, pathway.id = "hsa03440")
```

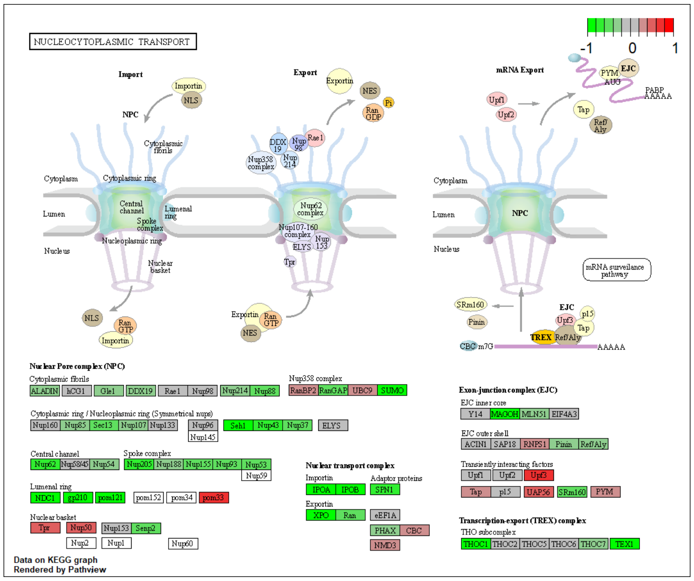
```
'select()' returned 1:1 mapping between keys and columns
```

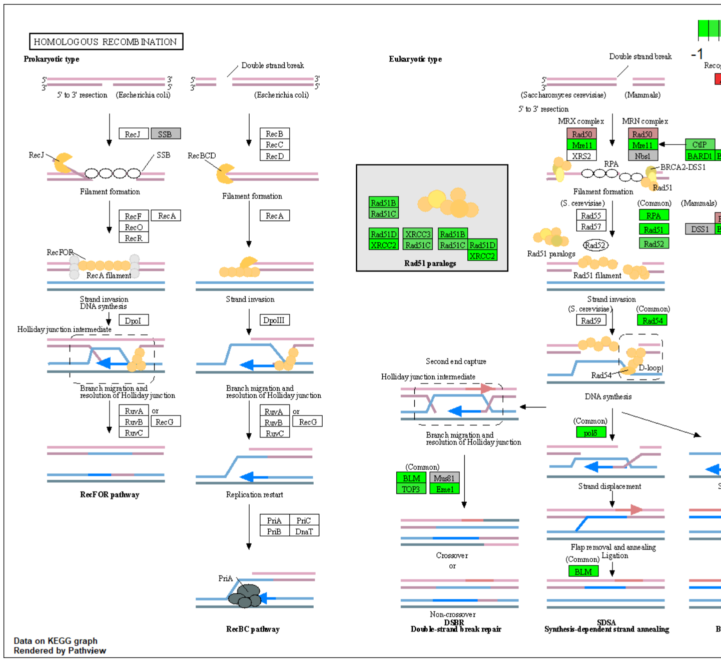
```
Info: Working in directory C:/Users/clari/Desktop/BGGN213 Bioinformatics/R files BGGN213/Cla
```

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


DNA REPLICATION
-1
0
1
Replication complex (Bacteria)
Lagging strand
Leading strand
RNase H / Pol I
Removal of RNA primer
Gap-filling
Lig
DNA ligase
Joining of Okazaki fragment
Pol III core
Clamp
β
τ
γδ complex
Primase
DnaG
Primer
DnaB
Helicase
SSB
DNA polymerase III holoenzyme
θ
ε
α
γ & τ
δ'
δ
ψ
χ
Pol III core
γδ complex
Helicase
DnaB
Primase
DnaG
SSB
RNaseH
RNaseHI
RNaseHII
RNaseHIII
DNA polymerase I
DpoI
DNA ligase
Lig
Replication complex (Archaea)
RNase H or Dna2
FEN 1
DNA ligase
DNA polymerase
Pol D/B
Clamp loader
RFC
Primase
MCM
RPA/SSB
Pol B
DNA polymerase B
PolB
DNA polymerase D
PolD1
PolD2
Helicase
MCM
Primase
Pri1
Pri2
RPA/SSB
RPA
Clamp
PCNA
Clamp loader
RfcS
RfcL
RNaseH
RNaseHI
RNaseHII
Helicase
Dna2
Fen1
DNA ligase
Lig
Replication complex (Eukaryotes)
RNase H or Dna2
Lig I
FEN 1
DNA ligase I
DNA polymerase δ complex
δ
PCNA
RFC
α-Prim
DNA polymerase α-primase complex
MCM 2-7
ε
DNA polymerase ε complex
RPA
Helicase
DNA polymerase α-primase complex
α1
α2
Pri1
Pri2
DNA polymerase δ complex
δ1
δ2
δ3
δ4
DNA polymerase ε complex
ε1
ε2
ε3
ε4
MCM complex (helicase)
Mcm2
Mcm3
Mcm4
Mcm5
Mcm6
Mcm7
RPA
RFA1
RFA2/4
RPA3
Clamp
PCNA
Clamp loader
RFC1
RFC2/4
RFC3/5
RNaseHI
RNaseHII
RNaseH2A
RNaseH2B
RNaseH2C
Helicase
Dna2
Fen1
DNA ligase
Lig1
Data on KEGG graph
Rendered by Pathview


PATHOGENIC ESCHERICHIA COLI INFECTION
Epithelial cell / Macrophage
Enteropathogenic Escherichia coli (EPEC)
Enterohemorrhagic Escherichia coli (EHEC)
Toxins
StxA
StxB
Translocation
EspJ
Src
+p
IgG
FcγRIIA
FcγR-mediated phagocytosis
Rac1
WAVE regulatory
Hem
Abi1
WAVE
EspG
ARNO
Arf6
Arf1
LEE-encoded effectors
Tir
Map
EspB
EspF
EspG
EspH
EspZ
non-LEE-encoded effectors
EspG2
EspJ
EspM1/2
EspT
EspW
Cif
NleA
NleB
NleC
NleD
NleE
NleF
NleH
Prophage-carried effectors
EspFu
PAMPs
PRRs
RhoGEF
RhoA
Cdc42
EspH
Type III secretion system
Bacterial secretion system
EspT
EspW
Rac1
WAVE2
Map
Cdc42
N-WASP
NHERF1
Ezrin
EspM1/2
EspH
LPA
F2
GPCR
G12/13
RhoGEF
RhoA
ROCK
Serum
EspG/G2
TUBA
TUBB
GEF-H1
Disruption of microtubule network
EspJ
c-Fyn/c-Abl
Pedestal disassembly during late stages of infection
IRTKS
IRSp53
EspFu
Eae
+p
Tir
NCL
ITGB1
Nck
WIP
N-WASP
Cortactin
Flagellar assembly
Bfp
Phosphatidylethanolamine
SHP-1/2
Toll-like receptor signaling pathway
NF-κB signaling pathway
IL-1β
IL-1R
TIRAP
MyD88
IRAK1/4
TRAF6
LPS
TLR4
Flagellin
TLR5
MyD88
TAK1
TAB
IKKγ
IKKα
IKKβ
+p
TRADD
TRAF2
RIP1
NleE
NleC
NleD
GAPDH
NleB
TNFα
TNF-R1
TRADD
FADD
TRAIL
TRAIL-R
Fas-L
Fas
FADD
CASP
Anti-apoptotic effectors
NleB1
NleF
NleH1/2
EspZ
Bl-1
Bak
Bax
Mitochondria
CytC
CASP
Pro-apoptotic effectors
EspH
Map
Cif
EspF
Abcf2
NleA
DNA:RNA hybrid from lysosome
NLRP3
ASC
Type III secretion rod/needle proteins
EprI
EprJ
EscI
NAIP
NLRC4
EspM1/2
NleA
Map
EspFu
EspF
OCL
Data on KEGG graph
Rendered by Pathview


NUCLEOCYTOPLASMIC TRANSPORT
-1
0
1
Import
Export
mRNA Export
NPC
Cytoplasm
Lumen
Nucleus
Importin
Exportin
Nuclear Pore complex (NPC)
Nuclear transport complex
Exon-junction complex (EJC)
Transcription-export (TREX) complex
Data on KEGG graph
Rendered by Pathview



Gene Ontology

```
data(go.sets.hs)
data(go.subs.hs)
gobpsets = go.sets.hs[go.subs.hs$BP]
gobpres = gage(foldchange, 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       0.1952430      113 8.519724e-05
GO:0002009 morphogenesis of an epithelium 0.1952430      339 1.396681e-04
GO:0048729 tissue morphogenesis           0.1952430      424 1.432451e-04
```

```
GO:0007610 behavior                       0.1968058      426 1.925222e-04
GO:0060562 epithelial tube morphogenesis  0.3566193      257 5.932837e-04
GO:0035295 tube development               0.3566193      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             5.843127e-12      376 1.536227e-15
GO:0000280 nuclear division              5.843127e-12      352 4.286961e-15
GO:0007067 mitosis                       5.843127e-12      352 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.195965e-11      362 1.169934e-14
GO:0007059 chromosome segregation        1.659009e-08      142 2.028624e-11
GO:0000236 mitotic prometaphase          1.178690e-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

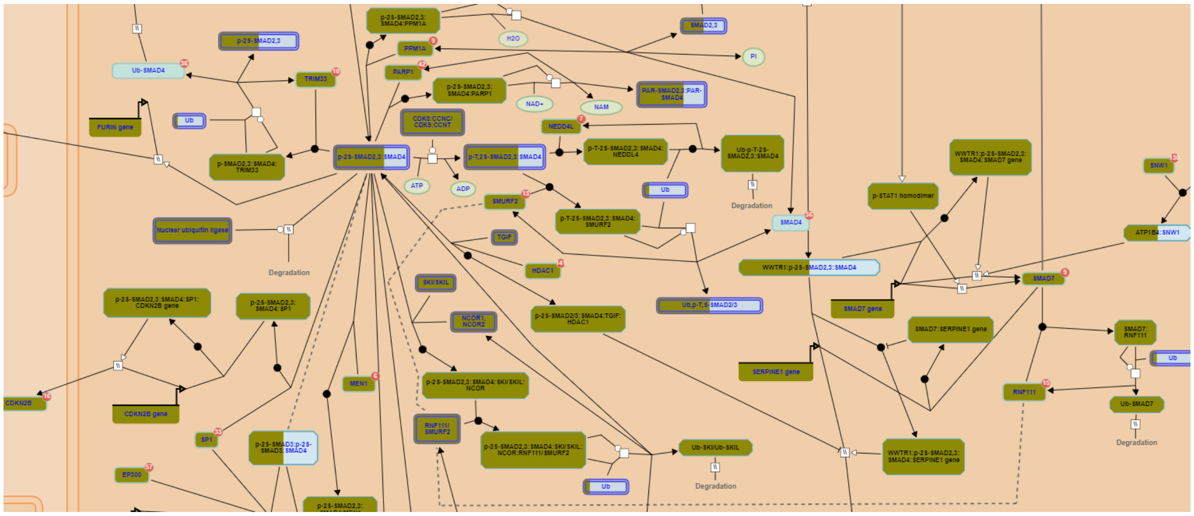
We will use the online version of Reactome. It wants a list of your genes. We will write this out from R here:

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

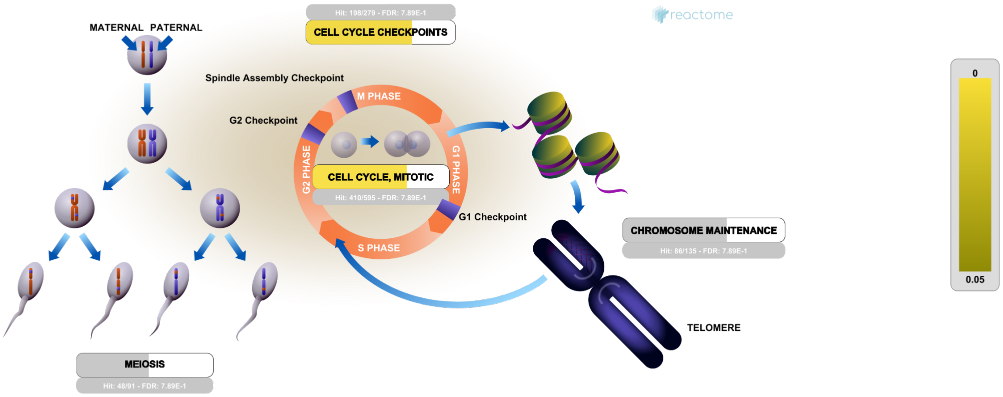
```
[1] "Total number of significant genes: 8147"
```

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

Have a look at my figure (**?@fig-SMAD**)



Have a look at my figure (**?@fig-cellcycle**)



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?

Mitotic cell cycle. This matches the most downregulated pathway in the KEGG results. They are using different databases as references so they will show different results.