Class 14: Pathway Analysis from RNA-Seq

Hailey Wheeler (A13312713)

Table of contents

I want to get rid of the first "length" column in counts 3
Remove zeros 5
Run DESeq2 5
Improve Plot below 7
Gene Ontology 17

```
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

Loading required package: GenomicRanges

Loading required package: GenomeInfoDb

Loading required package: SummarizedExperiment

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

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
```

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

```
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
```

I want to get rid of the first "length" column in counts

```
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
```

```
colnames(countData)
```

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

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
```

```
colnames(countData)
```

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

```
dim(countData)
```

```
[1] 19808     6
```

Remove zeros

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

```
[1] 15975
```

Run DESeq2

```
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: 15975 6
metadata(1): version
assays(4): counts mu H cooks
rownames(15975): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
  ENSG00000271254
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

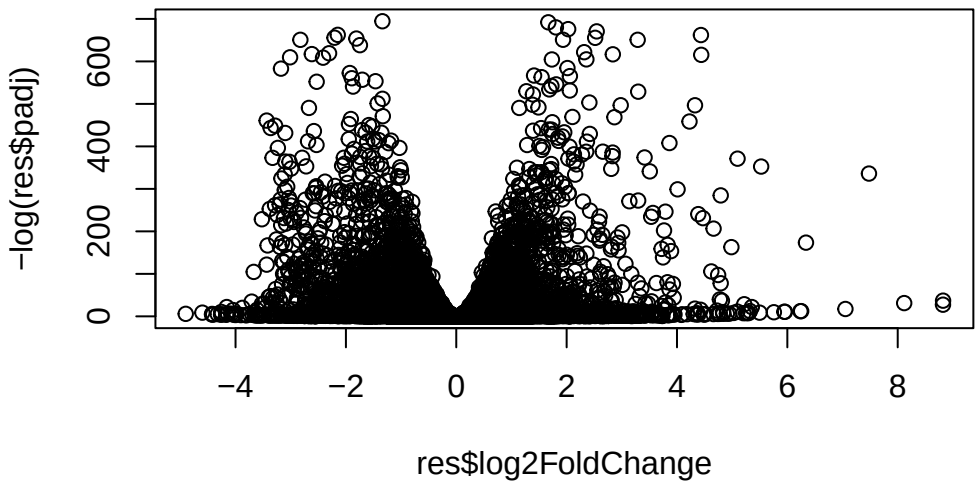
```
res = results(dds)
```

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)     : 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
```

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



Improve Plot below

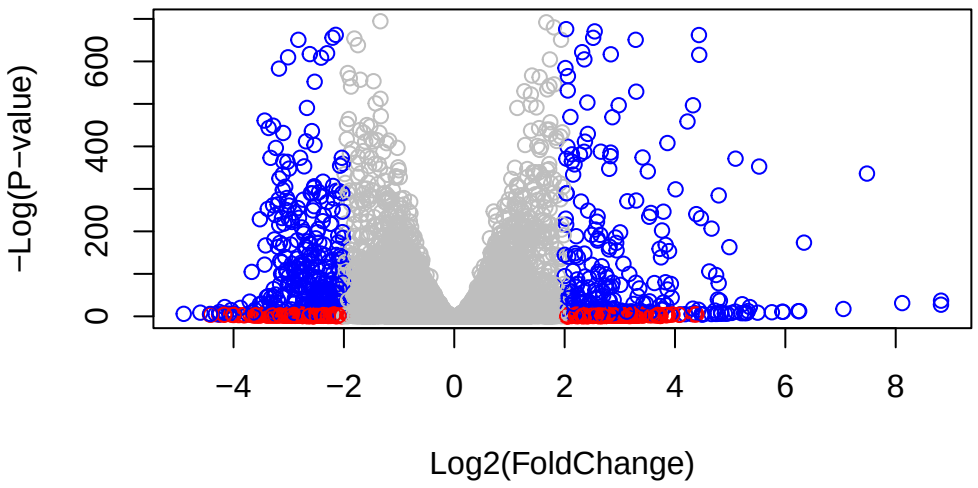
Q. Improve this plot by completing the below code, which adds color and axis labels

```
# 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
```




```
library("AnnotationDbi")
library("org.Hs.eg.db")
```

```
columns(org.Hs.eg.db)
```

```
 [1] "ACCNUM"       "ALIAS"        "ENSEMBL"      "ENSEMBLPROT"  "ENSEMBLTRANS"
 [6] "ENTREZID"     "ENZYME"       "EVIDENCE"     "EVIDENCEALL"  "GENENAME"
[11] "GENETYPE"     "GO"           "GOALL"        "IPI"          "MAP"
[16] "OMIM"         "ONTOLOGY"     "ONTOLOGYALL"  "PATH"         "PFAM"
[21] "PMID"         "PROSITE"      "REFSEQ"       "SYMBOL"       "UCSCKG"
[26] "UNIPROT"
```

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

```
##res$symbol = mapIds(org.Hs.eg.db,
                     # keys=row.names(counts),
                     # keytype="ENSEMBL",
```

```
                    # column="SYMBOL",
                    # multiVals="first")

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

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

# head(res, 10)
```

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")
```

```
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"
```

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

```
[1] -2.422719  3.201955 -2.313738 -2.059631 -1.888019 -1.649792
```

```
# Get the results
keggres = gage(foldchanges, gsets=kegg.sets.hs)
attributes(keggres)
```

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

```
head(keggres$less)
```

```
                                      p.geomean stat.mean p.val q.val
hsa00232 Caffeine metabolism                 NA       NaN    NA    NA
hsa00983 Drug metabolism - other enzymes     NA       NaN    NA    NA
hsa00230 Purine metabolism                   NA       NaN    NA    NA
hsa04514 Cell adhesion molecules (CAMs)      NA       NaN    NA    NA
hsa04010 MAPK signaling pathway              NA       NaN    NA    NA
hsa04012 ErbB signaling pathway              NA       NaN    NA    NA
                                      set.size exp1
hsa00232 Caffeine metabolism                 0   NA
hsa00983 Drug metabolism - other enzymes     0   NA
hsa00230 Purine metabolism                   0   NA
hsa04514 Cell adhesion molecules (CAMs)      0   NA
hsa04010 MAPK signaling pathway              0   NA
hsa04012 ErbB signaling pathway              0   NA
```

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

```
Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

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

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

```
Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

'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 /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

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

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

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

```
[1] "hsa00232" "hsa00983" "hsa00230" "hsa04514" "hsa04010"
```

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

```
Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa00232.pathview.png

Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa00983.pathview.png

Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa00230.pathview.png

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)
```

```
Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)


Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa04514.pathview.png

Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14
```

```
Info: Writing image file hsa04010.pathview.png
```

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

```
downkeggrespathways <- rownames(keggres$less)[1:5]
downkeggresids = substr(downkeggrespathways, start=1, stop=8)
pathview(gene.data=foldchanges, pathway.id=downkeggresids, species="hsa")
```

```
Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa00232.pathview.png

Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa00983.pathview.png

Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa00230.pathview.png
```

```
Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)

Warning in cbind(blk.ind, j): number of rows of result is not a multiple of
vector length (arg 2)
```

```
Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.
```

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

```
Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14
```

```
Info: Writing image file hsa04514.pathview.png

Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory /Users/haileywheeler/Documents/Bioinformatics/R Studio/Class 14

Info: Writing image file hsa04010.pathview.png
```

Gene Ontology

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

```
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 q.val
GO:0000002 mitochondrial genome maintenance        NA       NaN    NA    NA
GO:0000003 reproduction                            NA       NaN    NA    NA
GO:0000012 single strand break repair              NA       NaN    NA    NA
GO:0000018 regulation of DNA recombination         NA       NaN    NA    NA
GO:0000019 regulation of mitotic recombination     NA       NaN    NA    NA
GO:0000022 mitotic spindle elongation              NA       NaN    NA    NA
                                           set.size exp1
GO:0000002 mitochondrial genome maintenance        0   NA
GO:0000003 reproduction                            0   NA
GO:0000012 single strand break repair              0   NA
GO:0000018 regulation of DNA recombination         0   NA
GO:0000019 regulation of mitotic recombination     0   NA
GO:0000022 mitotic spindle elongation              0   NA

$less
```

```
                                          p.geomean stat.mean p.val q.val
GO:0000002 mitochondrial genome maintenance        NA       NaN    NA    NA
GO:0000003 reproduction                            NA       NaN    NA    NA
GO:0000012 single strand break repair              NA       NaN    NA    NA
GO:0000018 regulation of DNA recombination         NA       NaN    NA    NA
GO:0000019 regulation of mitotic recombination     NA       NaN    NA    NA
GO:0000022 mitotic spindle elongation              NA       NaN    NA    NA
                                          set.size exp1
GO:0000002 mitochondrial genome maintenance        0   NA
GO:0000003 reproduction                            0   NA
GO:0000012 single strand break repair              0   NA
GO:0000018 regulation of DNA recombination         0   NA
GO:0000019 regulation of mitotic recombination     0   NA
GO:0000022 mitotic spindle elongation              0   NA

$stats
                                          stat.mean exp1
GO:0000002 mitochondrial genome maintenance       NaN   NA
GO:0000003 reproduction                           NaN   NA
GO:0000012 single strand break repair             NaN   NA
GO:0000018 regulation of DNA recombination        NaN   NA
GO:0000019 regulation of mitotic recombination    NaN   NA
GO:0000022 mitotic spindle elongation             NaN   NA
```

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

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

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 "Entities p-value" is 'Cell Cycle'. This is identical to the KEGG results. The main factor that can cause differences between the two methods is that Reactome has a larger profile and increased specificity.