Pathway Analysis

Laura Siemens

2023-12-08

```
library("DESeq2")
```

```
## Loading required package: S4Vectors
```

```
## Warning: package 'S4Vectors' was built under R version 4.3.2
```

```
## 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
```

```
library("dplyr")
```

```
## Warning: package 'dplyr' was built under R version 4.3.2

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:Biobase':
##
##     combine

## The following object is masked from 'package:matrixStats':
##
##     count

## The following objects are masked from 'package:GenomicRanges':
##
##     intersect, setdiff, union

## The following object is masked from 'package:GenomeInfoDb':
##
##     intersect

## The following objects are masked from 'package:IRanges':
##
##     collapse, desc, intersect, setdiff, slice, union

## The following objects are masked from 'package:S4Vectors':
##
##     first, intersect, rename, setdiff, setequal, union

## The following objects are masked from 'package:BiocGenerics':
##
##     combine, intersect, setdiff, union

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
expData <- read.csv("C:\\BMEG 310\\rnaseq_data_shared.csv")
clinical_clusters <- read.csv(file = "C:\\BMEG 310\\clinical_with_groups.csv")

#preparing coldata
colData <- data.frame(group = clinical_clusters[,c("subgroup")],
                              row.names=clinical_clusters$PATIENT_ID)
colData$group <- as.factor((colData$group))

#Preparing countdata
set.seed(1)

countData <- expData
genes<- countData$X
countData <- countData[,c(-1)]
rownames(countData) <- genes
patients <- colnames(countData)
countData <- as.data.frame(sapply(countData, as.numeric))
colnames(countData) <- substr(patients, 1, 12)
rownames(countData) <- genes

#removes duplicates of patients
countData <- countData[,!duplicated(colnames(countData)) ]
countData <- countData[rowSums(countData)>1,]#removes low expression counts
countData <- countData[,order(colnames(countData))]

#Removing outliers
outliers <- c('TCGA.CC.A3M9','TCGA.CC.A7II','TCGA.FV.A4ZP',
              'TCGA.DD.AACK', 'TCGA.CC.A7IJ', 'TCGA.CC.A1HT','TCGA.G3.A7M9',
              'TCGA.G3.A7M6')
countData<-countData[,!(colnames(countData) %in% outliers)]

colRow <- rownames(colData)
colRow <- gsub('-', '.', colRow)
rownames(colData) <-colRow
colRow <- colRow[!colRow %in%outliers]
colData <- colData[!rownames(colData)%in% outliers,]
colData <- data.frame(group = colData,
                              row.names=colRow)
```

```
#Differential gene expression
dds = DESeqDataSetFromMatrix(countData=countData,
                             colData=colData,
                             design=~group)
```

```
## converting counts to integer mode
```

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

```
## estimating size factors

## estimating dispersions
```

```
## gene-wise dispersion estimates

## mean-dispersion relationship

## final dispersion estimates

## fitting model and testing

## -- replacing outliers and refitting for 5457 genes
## -- DESeq argument 'minReplicatesForReplace' = 7
## -- original counts are preserved in counts(dds)

## estimating dispersions

## fitting model and testing
```

```
res <- results(dds)

#Filtering based on thresholds found in volcano plot
res.table <- table(res$padj < 0.01 & abs(res$log2FoldChange) > 1.5)
rownames(res) <- gsub("\\..*","",rownames(res))
resSig <- subset(res, padj < 0.01 & abs(res$log2FoldChange) > 1.5)
```

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

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

##
## Attaching package: 'AnnotationDbi'

## The following object is masked from 'package:dplyr':
##
##     select
```

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

```
##
```

```
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")
```

```
#mapping filtered genes
resSig$symbol = mapIds(org.Hs.eg.db,
                    keys=row.names(resSig),
                    column="SYMBOL",
                    keytype="ENSEMBL",
                    multiVals="first")
```

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

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

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

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

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

```
#Pathway analysis with filtered genes
#Focus on signaling and metabolic pathways only
data(kegg.sets.hs)
data(sigmet.idx.hs)
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]

foldchangesSig = resSig$log2FoldChange
names(foldchangesSig) = resSig$entrez

# Get the results
keggresSig = gage(foldchangesSig, gsets=kegg.sets.hs)

## Focus on top 5 upregulated pathways
keggrespathwaysSigUp <- rownames(keggresSig$greater)[1:5]
keggresDownpathwaysSigDown <- rownames(keggresSig$less)[1:5]

# Extract the 8 character long IDs part of each string
keggresidsSigUp = substr(keggrespathwaysSigUp, start=1, stop=8)
keggresidsSigDown = substr(keggresDownpathwaysSigDown, start=1, stop=8)

#checking if pathways are significant
head(keggresSig$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      2   NA
## hsa00230 Purine metabolism                    0   NA
## hsa04514 Cell adhesion molecules (CAMs)       1   NA
## hsa04010 MAPK signaling pathway               2   NA
## hsa04012 ErbB signaling pathway               1   NA
```

```
head(keggresSig$greater)
```

```
##                                        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      2   NA
## hsa00230 Purine metabolism                    0   NA
## hsa04514 Cell adhesion molecules (CAMs)       1   NA
## hsa04010 MAPK signaling pathway               2   NA
## hsa04012 ErbB signaling pathway               1   NA
```

```
#No differences in pathways were found so this part is commented out
#pathview(gene.data=foldchangesSig, pathway.id=keggresidsSigUp, species="hsa")
#pathview(gene.data=foldchangesSig, pathway.id = keggresidsSigDown,
#species = "hsa")
```

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

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

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

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

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

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

```
#Pathway analysis with unfiltered genes
# Focus on signaling and metabolic pathways only
data(kegg.sets.hs)
data(sigmet.idx.hs)
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]

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

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

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

```
## Focus on top 5 upregulated
keggrespathwaysUp <- rownames(keggres$greater)[1:5]
## Focus on top 5 downregulated
keggrespathwaysDown <- rownames(keggres$less)[1:5]

# Extract the 8 character long IDs part of each string
keggresidsUp = substr(keggrespathwaysUp, start=1, stop=8)
keggresidsDown = substr(keggrespathwaysDown, start=1, stop=8)

#checking to see if top upregulated and down regulated are significant
head(keggres$less)
```

```
##                                             p.geomean stat.mean
## hsa04610 Complement and coagulation cascades 3.118303e-09 -6.226849
## hsa00071 Fatty acid metabolism               2.950824e-05 -4.294474
## hsa00982 Drug metabolism - cytochrome P450   1.345140e-04 -3.747753
## hsa04976 Bile secretion                      1.460894e-04 -3.718870
## hsa00830 Retinol metabolism                  6.409657e-04 -3.313128
## hsa00350 Tyrosine metabolism                 1.027774e-03 -3.191363
##                                                    p.val        q.val set.size
## hsa04610 Complement and coagulation cascades 3.118303e-09 5.114018e-07       69
## hsa00071 Fatty acid metabolism               2.950824e-05 2.419675e-03       43
## hsa00982 Drug metabolism - cytochrome P450   1.345140e-04 5.989665e-03       70
## hsa04976 Bile secretion                      1.460894e-04 5.989665e-03       71
## hsa00830 Retinol metabolism                  6.409657e-04 2.102367e-02       62
## hsa00350 Tyrosine metabolism                 1.027774e-03 2.809250e-02       41
##                                                     exp1
## hsa04610 Complement and coagulation cascades 3.118303e-09
```

```
## hsa00071 Fatty acid metabolism                2.950824e-05
## hsa00982 Drug metabolism - cytochrome P450   1.345140e-04
## hsa04976 Bile secretion                       1.460894e-04
## hsa00830 Retinol metabolism                   6.409657e-04
## hsa00350 Tyrosine metabolism                  1.027774e-03
```

```r
head(keggres$greater)
```

```
##                                     p.geomean stat.mean        p.val
## hsa04110 Cell cycle               5.010868e-06  4.511918 5.010868e-06
## hsa03030 DNA replication          1.553537e-04  3.861782 1.553537e-04
## hsa03440 Homologous recombination 2.377534e-03  2.966464 2.377534e-03
## hsa03430 Mismatch repair          1.201095e-02  2.367758 1.201095e-02
## hsa04142 Lysosome                 1.204639e-02  2.271530 1.204639e-02
## hsa03013 RNA transport            1.223501e-02  2.266934 1.223501e-02
##                                           q.val set.size         exp1
## hsa04110 Cell cycle               0.0008217823      124 5.010868e-06
## hsa03030 DNA replication          0.0127390044       36 1.553537e-04
## hsa03440 Homologous recombination 0.1299718424       28 2.377534e-03
## hsa03430 Mismatch repair          0.3092192913       23 1.201095e-02
## hsa04142 Lysosome                 0.3092192913      121 1.204639e-02
## hsa03013 RNA transport            0.3092192913      149 1.223501e-02
```

```r
#Significant pathways were found without filtering of data
#commented out for sake of knitting
#pathview(gene.data=foldchanges, pathway.id=keggresidsUp, species="hsa")
#pathview(gene.data=foldchanges, pathway.id = keggresidsDown, species = "hsa")
```