Class 14 Pathway Analysis

Ani A16647613

Differnetial Expression Analysis

First we are going to download Biocmanager and DESeq2 packages in our console then add them to our library

```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

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
```

Now to load our metadata and count files:

```r
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
```

```r
# importing countdata
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
```

```r
# to remove the nonessential "length" column
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
```

```
# Now we will implement a filter to exclude rows with gene counts of 0 across all the samp
to.keep.inds <- rowSums(countData) > 0
nonzerocounts <- countData[to.keep.inds,]
```

```
head(nonzerocounts)
```

```
                SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
ENSG00000279457        23        28        29        29        28        46
ENSG00000187634       124       123       205       207       212       258
ENSG00000188976      1637      1831      2383      1226      1326      1504
ENSG00000187961       120       153       180       236       255       357
ENSG00000187583        24        48        65        44        48        64
ENSG00000187642         4         9        16        14        16        16
```

Now that we have cleared out the unnecessary data with zero values and the first column, we can begin to run DESeq2

Running DESeq2

We already have DESeq2 added to our library. To set up the necessary `DESEq()` function to run the pipeline, we have to use `DESeqDataSet`

```
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: 19808 6
metadata(1): version
assays(4): counts mu H cooks
rownames(19808): ENSG00000186092 ENSG00000279928 ... ENSG00000277475
  ENSG00000268674
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

We will be specifically looking into the knockdown HoxA1 (`hoxa1_kd`) and siRNA (`control_sirna`) from the colData file and use `resultNames(dds)` to observe the results:

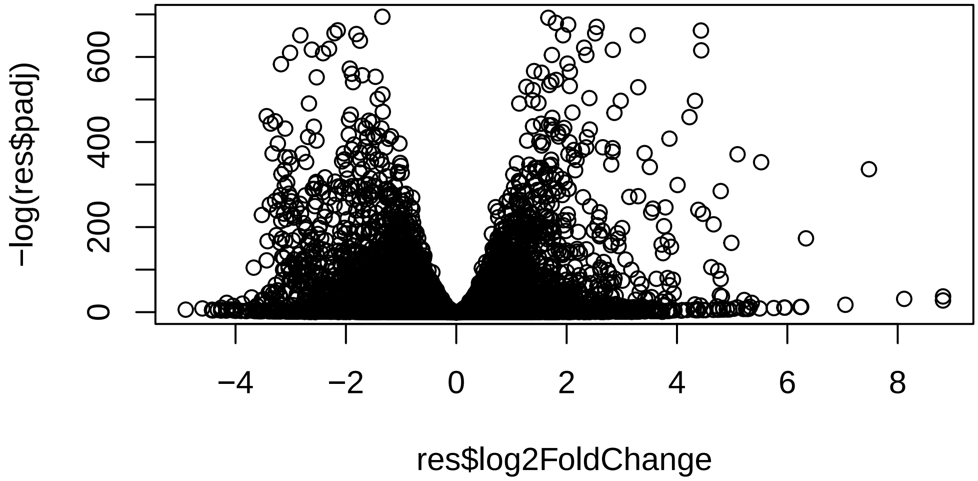
```
res = results(dds,
            contrast=c("condition",
                       "hoxa1_kd",
                       "control_sirna"))

summary(res)
```

```

out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4349, 27%
LFC < 0 (down)     : 4393, 27%
```

```
outliers [1]       : 0, 0%
low counts [2]     : 1221, 7.6%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

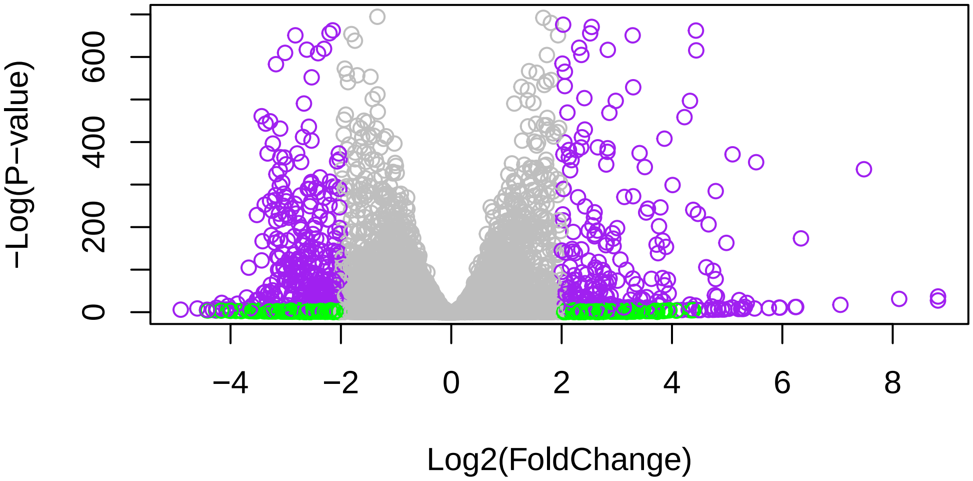
With this data, we can create a **volcano plot** for visualization purposes:

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



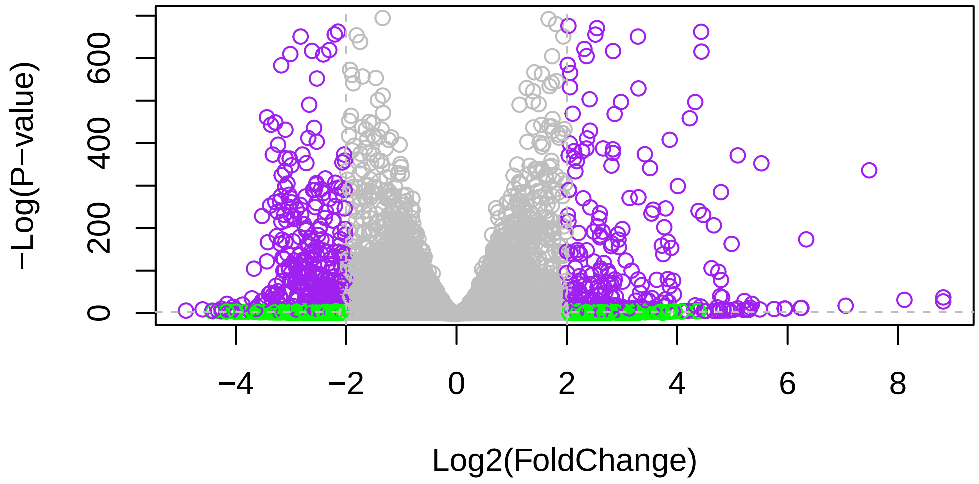
For better visualization of the significant datapoints, we will add colors and axis labels (I wanted to change the colors to make it appear different from the example, but still show distinct separation)

```
mycols <- rep("gray", nrow(res) )
mycols[ abs(res$log2FoldChange) > 2 ] <- "green"
inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "purple"

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



For my own purposes for understanding, I would like to add cutoff marks for better visualization of the volcano plot significant points shown above:

```
plot( res$log2FoldChange, -log(res$padj), col=mycols, xlab="Log2(FoldChange)", ylab="-Log(
abline(v=c(-2,2), col="gray", lty=2)
abline(h=-log(0.1), col="gray", lty=2)
```



To annotate the Entrz Gene IDs and reorder pvalues to safe as a new file:

```
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"
```

```
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
```

```
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
```

```
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
```

```
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>
ENSG00000186092    0.0000             NA        NA         NA          NA
ENSG00000279928    0.0000             NA        NA         NA          NA
ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
ENSG00000278566    0.0000             NA        NA         NA          NA
ENSG00000273547    0.0000             NA        NA         NA          NA
ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 1.43990e-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 0.5215598   1.040744 2.97994e-01
```

```
                         padj      symbol      entrez                   name
                    <numeric> <character> <character>            <character>
ENSG00000186092            NA       OR4F5       79501 olfactory receptor f..
ENSG00000279928            NA          NA          NA                     NA
ENSG00000279457 6.87080e-01          NA          NA                     NA
ENSG00000278566            NA          NA          NA                     NA
ENSG00000273547            NA          NA          NA                     NA
ENSG00000187634 5.16278e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76741e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13536e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.18988e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03817e-01       PERM1       84808 PPARGC1 and ESRR ind..
```

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

** Pathway Analysis

After downloading the KEGG packages, we can begin to work to establish a pathview of a biological pathway.

```
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)
```

```r
library(gageData)

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

Focusing on metabolic pathways:

```r
kegg.sets.hs <- kegg.sets.hs[sigmet.idx.hs]
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"  
```

We should be able to map out the IDS of the Entrez genes to name the fold change vectors.

```
foldchanges <- res$log2FoldChange
names(foldchanges) <- res$entrez
head(foldchanges)
```

```
     1266     54855      1465     51232      2034      2317 
-2.422719  3.201955 -2.313738 -2.059631 -1.888019 -1.649792 
```

Now to run the gage pathway analysis Here is how we can see the results:

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

```
attributes(keggres)
```

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

Now we are gonna specifically look at the first few down pathways signified by `less`

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

```
                                         p.geomean stat.mean        p.val
hsa04110 Cell cycle                   7.077982e-06 -4.432593 7.077982e-06
hsa03030 DNA replication              9.424076e-05 -3.951803 9.424076e-05
hsa03013 RNA transport                1.160132e-03 -3.080629 1.160132e-03
hsa04114 Oocyte meiosis               2.563806e-03 -2.827297 2.563806e-03
hsa03440 Homologous recombination     3.066756e-03 -2.852899 3.066756e-03
hsa00010 Glycolysis / Gluconeogenesis 4.360092e-03 -2.663825 4.360092e-03
                                            q.val set.size         exp1
hsa04110 Cell cycle                   0.001160789      124 7.077982e-06
hsa03030 DNA replication              0.007727742       36 9.424076e-05
hsa03013 RNA transport                0.063420543      149 1.160132e-03
hsa04114 Oocyte meiosis               0.100589607      112 2.563806e-03
hsa03440 Homologous recombination     0.100589607       28 3.066756e-03
hsa00010 Glycolysis / Gluconeogenesis 0.119175854       65 4.360092e-03
```

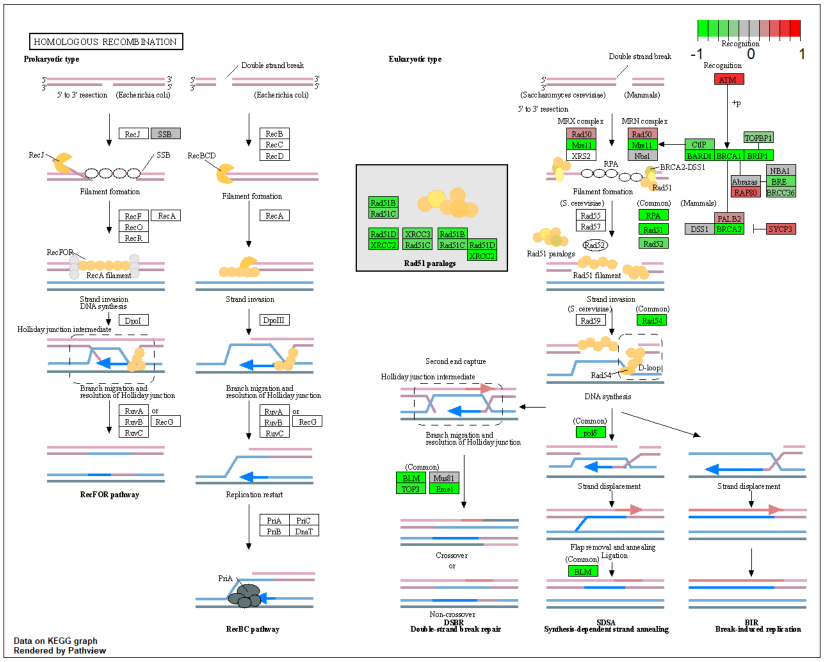
We can use the pathway IDs to view the pathway figure: Let's look at Homologous Recombination for example!

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

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

```
Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys
```

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



Here's a different PDF based on outputs of the same data:

```
pathview(gene.data=foldchanges, pathway.id="hsa03440", kegg.native=FALSE)
```

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

```
Info: Edge type: PPrel without subtype sepecified!
```

```
Warning in .subtypeDisplay(object): Given subtype 'unknown' is not found!
```

```
Info: Edge type: PPrel without subtype sepecified!

Warning in .subtypeDisplay(object): Given subtype 'unknown' is not found!

Info: Edge type: PPrel without subtype sepecified!

Warning in .subtypeDisplay(object): Given subtype 'unknown' is not found!

Info: Edge type: PPrel without subtype sepecified!

Warning in .subtypeDisplay(object): Given subtype 'unknown' is not found!

Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys

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

To focus on the top 5 upregulated pathways without the long IDs of each sample:

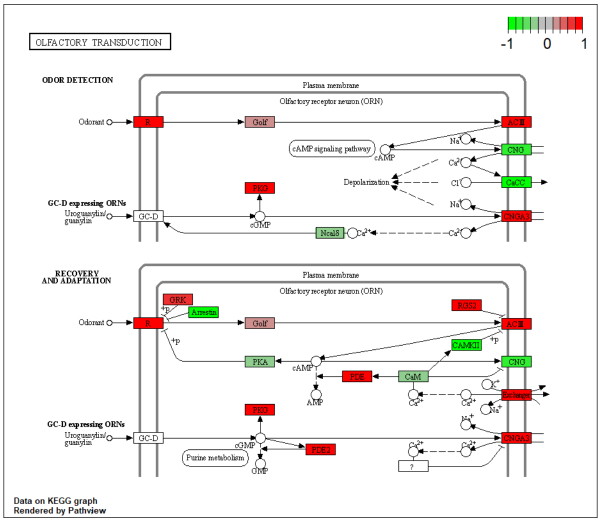
```r
keggrespathways <- rownames(keggres$greater)[1:5]

keggresids <- substr(keggrespathways, start=1, stop=8)

keggresids
```

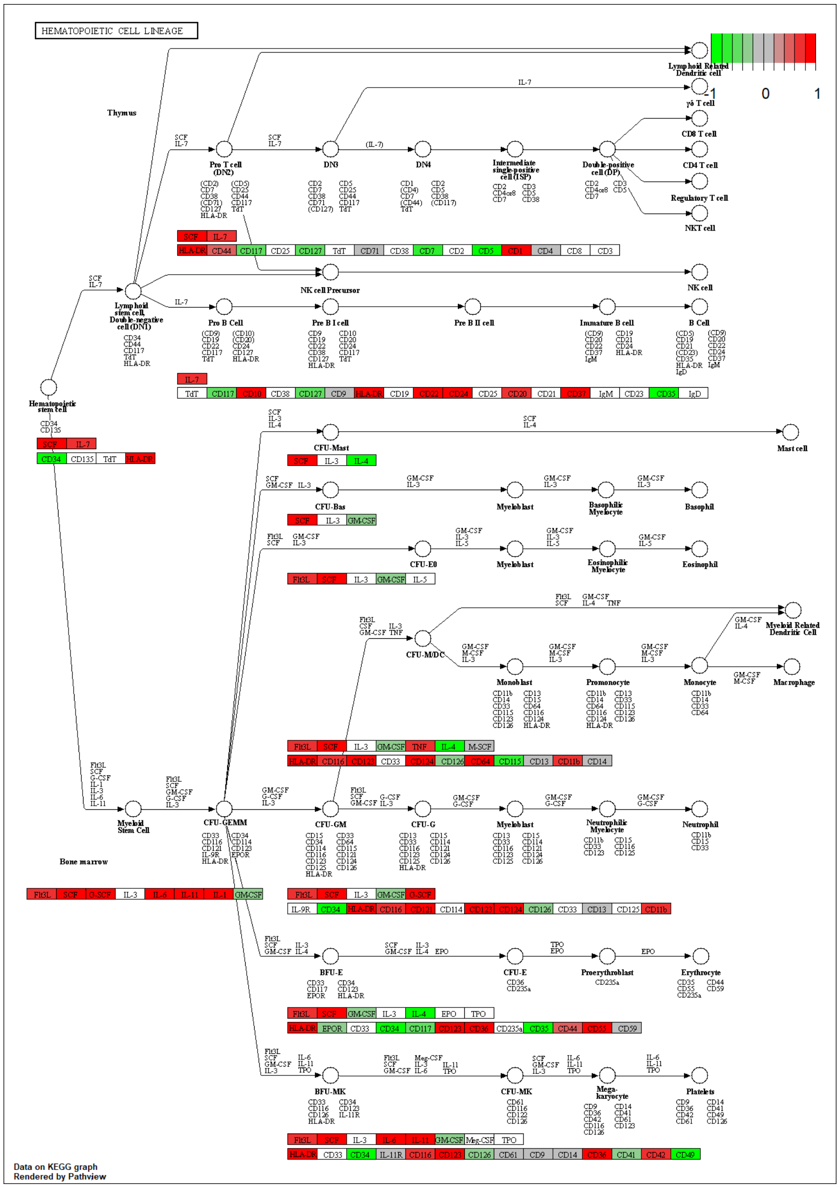
```
[1] "hsa04740" "hsa04640" "hsa00140" "hsa04630" "hsa04976"
```

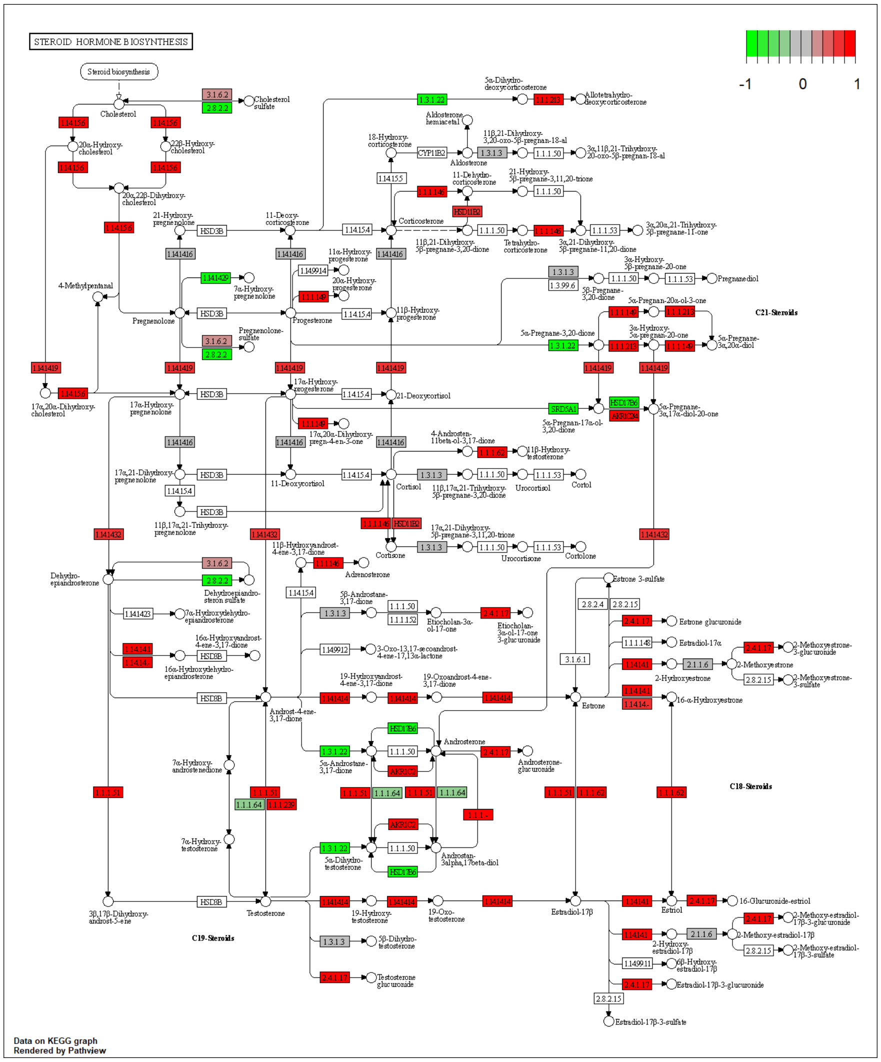
```r
pathview(gene.data=foldchanges, pathway.id=keggresids, species="hsa")
```

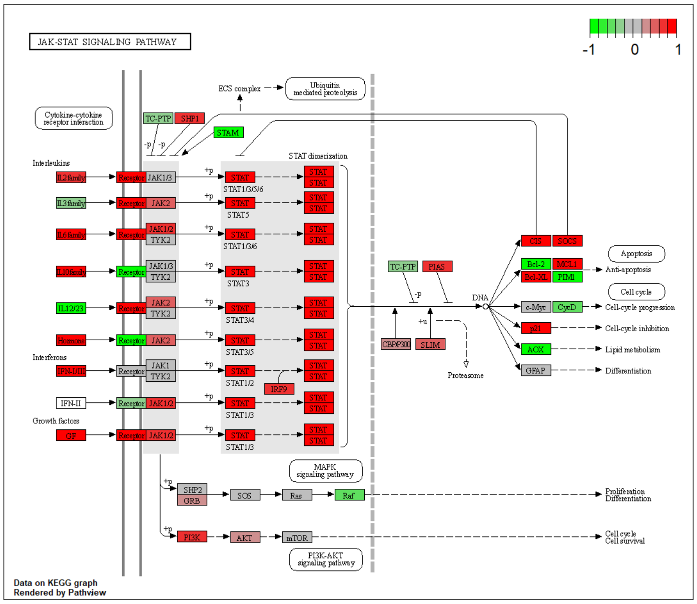
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys

Info: Writing image file hsa04740.pathview.png

Info: some node width is different from others, and hence adjusted!

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

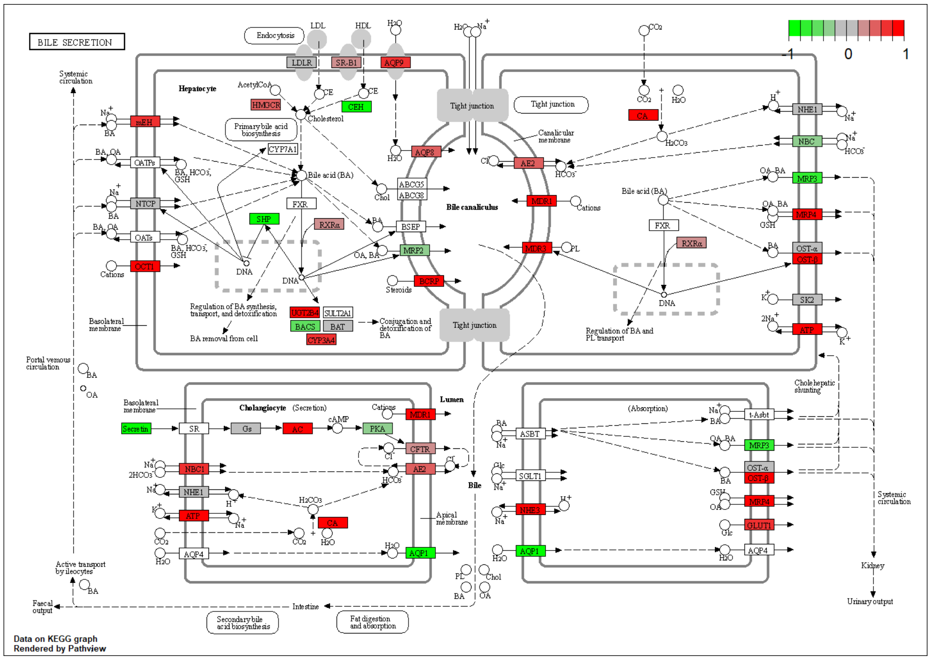
```
Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys

Info: Writing image file hsa04640.pathview.png

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

Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys

Info: Writing image file hsa00140.pathview.png

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

Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys

Info: Writing image file hsa04630.pathview.png

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

Info: Working in directory C:/Users/anban/OneDrive/Desktop/R lab/Class 14 Redo Pathway Analys

Info: Writing image file hsa04976.pathview.png
```




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


STEROID HORMONE BIOSYNTHESIS
-1
0
1
Steroid biosynthesis
Cholesterol
Cholesterol sulfate
20α-Hydroxy-cholesterol
22β-Hydroxy-cholesterol
20α,22β-Dihydroxy-cholesterol
4-Methylpentanal
17α,20α-Dihydroxy-cholesterol
21-Hydroxy-pregnenolone
Pregnenolone
7α-Hydroxy-pregnenolone
Pregnenolone-sulfate
17α-Hydroxy-pregnenolone
17α,21-Dihydroxy-pregnenolone
11β,17α,21-Trihydroxy-pregnenolone
11-Deoxy-corticosterone
11α-Hydroxy-progesterone
20α-Hydroxy-progesterone
Progesterone
17α-Hydroxy-progesterone
17α,20α-Dihydroxy-pregn-4-en-3-one
11-Deoxycortisol
18-Hydroxy-corticosterone
Aldosterone hemiacetal
Aldosterone
Corticosterone
11β-Hydroxy-progesterone
21-Deoxycortisol
Cortisol
Cortisone
5α-Dihydro-deoxycorticosterone
Allotetrahydro-deoxycorticosterone
11β,21-Dihydroxy-3,20-oxo-5β-pregnan-18-al
3α,11β,21-Trihydroxy-20-oxo-5β-pregnan-18-al
11-Dehydro-corticosterone
21-Hydroxy-5β-pregnane-3,11,20-trione
11β,21-Dihydroxy-5β-pregnane-3,20-dione
Tetrahydro-corticosterone
3α,21-Dihydroxy-5β-pregnane-11,20-dione
3α,20α,21-Trihydroxy-5β-pregnane-11-one
3α-Hydroxy-5β-pregnane-20-one
5β-Pregnane-3,20-dione
Pregnanediol
5α-Pregnan-20α-ol-3-one
5α-Pregnane-3,20-dione
3α-Hydroxy-5α-pregnan-20-one
5α-Pregnane-3α,20α-diol
C21-Steroids
5α-Pregnan-17α-ol-3,20-dione
5α-Pregnane-3α,17α-diol-20-one
4-Androsten-11beta-ol-3,17-dione
11β-Hydroxy-testosterone
11β,17α,21-Trihydroxy-5β-pregnane-3,20-dione
Urocortisol
Cortol
17α,21-Dihydroxy-5β-pregnane-3,11,20-trione
Urocortisone
Cortolone
Dehydro-epiandrosterone
Dehydroepiandro-steron sulfate
7α-Hydroxydehydro-epiandrosterone
16α-Hydroxyandrost-4-ene-3,17-dione
16α-Hydroxydehydro-epiandrosterone
11β-Hydroxyandrost-4-ene-3,17-dione
Adrenosterone
5β-Androstane-3,17-dione
Etiocholan-3α-ol-17-one
Etiocholan-3α-ol-17-one 3-glucuronide
3-Oxo-13,17-secoandrost-4-ene-17,13α-lactone
19-Hydroxyandrost-4-ene-3,17-dione
19-Oxoandrost-4-ene-3,17-dione
Androst-4-ene-3,17-dione
7α-Hydroxy-androstenedione
7α-Hydroxy-testosterone
5α-Androstane-3,17-dione
Androsterone
Androsterone-glucuronide
5α-Dihydro-testosterone
Androstan-3alpha,17beta-diol
3β,17β-Dihydroxy-androst-5-ene
Testosterone
19-Hydroxy-testosterone
19-Oxo-testosterone
5β-Dihydro-testosterone
Testosterone glucuronide
C19-Steroids
Estrone 3-sulfate
Estrone glucuronide
Estradiol-17α
2-Methoxyestrone-3-glucuronide
2-Methoxyestrone
2-Hydroxyestrone
2-Methoxyestrone-3-sulfate
Estrone
16-α-Hydroxyestrone
C18-Steroids
Estradiol-17β
Estriol
16-Glucuronide-estriol
2-Methoxy-estradiol-17β-3-glucuronide
2-Methoxy-estradiol-17β
2-Hydroxy-estradiol-17β
2-Methoxy-estradiol-17β-3-sulfate
6β-Hydroxy-estradiol-17β
Estradiol-17β-3-glucuronide
Estradiol-17β-3-sulfate
3.1.6.2
2.8.2.2
1.14.15.6
1.14.14.19
1.14.14.16
1.14.14.29
1.1.1.149
1.14.99.14
1.14.15.4
1.14.15.5
HSD3B
CYP11B2
HSD11B2
1.1.1.146
1.3.1.22
1.3.1.3
1.1.1.50
1.1.1.53
1.1.1.213
1.3.99.6
SRD5A1
HSD17B6
AKR1C24
1.1.1.62
1.14.14.32
1.14.14.23
1.14.14.1
1.14.14.-
1.1.1.152
1.14.99.12
1.14.14.14
2.4.1.17
AKR1C2
1.1.1.51
1.1.1.64
1.1.1.239
1.1.1.-
2.8.2.4
2.8.2.15
3.1.6.1
1.1.1.148
2.1.1.6
1.14.99.11
Data on KEGG graph
Rendered by Pathview





Gene Ontology (GO)

This secion does a similar feature to the one previously but with gene ontology

```r
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        1.624062e-05  4.226117 1.624062e-05
GO:0048729 tissue morphogenesis            5.407952e-05  3.888470 5.407952e-05
GO:0002009 morphogenesis of an epithelium  5.727599e-05  3.878706 5.727599e-05
GO:0030855 epithelial cell differentiation 2.053700e-04  3.554776 2.053700e-04
GO:0060562 epithelial tube morphogenesis   2.927804e-04  3.458463 2.927804e-04
GO:0048598 embryonic morphogenesis         2.959270e-04  3.446527 2.959270e-04
                                                q.val set.size         exp1
GO:0007156 homophilic cell adhesion        0.07102022      138 1.624062e-05
GO:0048729 tissue morphogenesis            0.08348930      483 5.407952e-05
GO:0002009 morphogenesis of an epithelium  0.08348930      382 5.727599e-05
GO:0030855 epithelial cell differentiation 0.16453464      299 2.053700e-04
GO:0060562 epithelial tube morphogenesis   0.16453464      289 2.927804e-04
GO:0048598 embryonic morphogenesis         0.16453464      498 2.959270e-04

$less
                                            p.geomean stat.mean        p.val
GO:0048285 organelle fission             6.626774e-16 -8.170439 6.626774e-16
GO:0000280 nuclear division              1.797050e-15 -8.051200 1.797050e-15
GO:0007067 mitosis                       1.797050e-15 -8.051200 1.797050e-15
GO:0000087 M phase of mitotic cell cycle 4.757263e-15 -7.915080 4.757263e-15
GO:0007059 chromosome segregation        1.081862e-11 -6.974546 1.081862e-11
GO:0051301 cell division                 8.718528e-11 -6.455491 8.718528e-11
                                                q.val set.size         exp1
GO:0048285 organelle fission             2.619500e-12      386 6.626774e-16
GO:0000280 nuclear division              2.619500e-12      362 1.797050e-15
GO:0007067 mitosis                       2.619500e-12      362 1.797050e-15
GO:0000087 M phase of mitotic cell cycle 5.200878e-12      373 4.757263e-15
GO:0007059 chromosome segregation        9.461963e-09      146 1.081862e-11
GO:0051301 cell division                 6.354354e-08      479 8.718528e-11
```

```
$stats
                                              stat.mean     exp1
GO:0007156 homophilic cell adhesion            4.226117 4.226117
GO:0048729 tissue morphogenesis                3.888470 3.888470
GO:0002009 morphogenesis of an epithelium      3.878706 3.878706
GO:0030855 epithelial cell differentiation     3.554776 3.554776
GO:0060562 epithelial tube morphogenesis       3.458463 3.458463
GO:0048598 embryonic morphogenesis             3.446527 3.446527
```

Reactome Analysis

We will be using this section to analyze the database with biological molecules and their contribution to pathways/processes

```
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: 8146"
```

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

> Q. What pathway has the most significant "Entities p-value"?

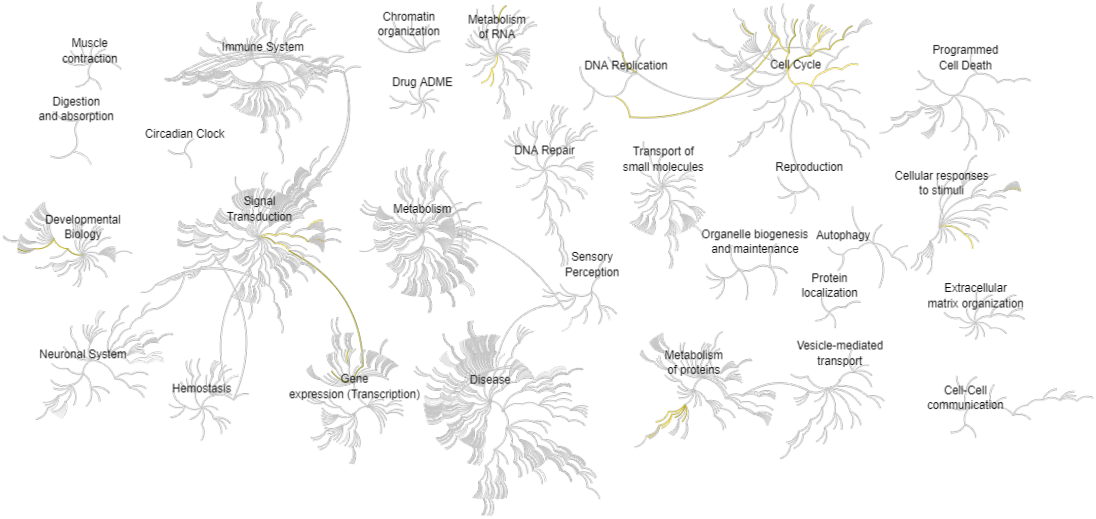
Cell cycle, Mitotic (with a pvalue of 3.88E-4)

> Do the pathways on this list match the previous KEGG results?

Partially but for the most part not all, they have some overlapping like cell cycle and DNA replication but others like homologous recombination and oocyte meiosis are found in the KEGG results but not from Reatome Analysis

> What factors could cause differences between the two methods

A factor I could think of is that we are not specifying in the reactome analysis whether we want to look at down versus up regulated pathways, instead with reactome analysis we are looking at how the pathways and certain molecules function together in general. Once you use the website to analyze the data, you can zoom in to see the corresponding molecules to the processes and their respective processes.


Muscle
contraction
Immune System
Chromatin
organization
Metabolism
of RNA
DNA Replication
Cell Cycle
Programmed
Cell Death
Digestion
and absorption
Drug ADME
Circadian Clock
DNA Repair
Transport of
small molecules
Reproduction
Cellular responses
to stimuli
Signal
Transduction
Metabolism
Developmental
Biology
Organelle biogenesis
and maintenance
Autophagy
Sensory
Perception
Protein
localization
Extracellular
matrix organization
Neuronal System
Hemostasis
Gene
expression (Transcription)
Disease
Metabolism
of proteins
Vesicle-mediated
transport
Cell-Cell
communication