Class 14: RNA Seq Mini Project

Genomics involves a lot of data, and using the traditional significance level of 0.05 would mean that around a 1000 genes would be deemed significant by chance. This is why we used an adjusted p-value.

A volcano plot shows the fold change over the -log of the p-value.

Run a complete analysis of RNA-seq

##Data import

```
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
```

```r
metaFile <- "GSE37704_metadata.csv"
countFile <- "GSE37704_featurecounts.csv"
coldata<-read.csv(metaFile, row.names=1)
```

```
countdata<-read.csv(countFile,row.names=1)
```

Q1. Removing the first column of countdata

```
countData<- 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

```
head(coldata)
```

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

Q2. 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).

```
zero.vals <- which(countData[,]==0, arr.ind=TRUE)

to.rm <- unique(zero.vals[,1])
countData <- countData[-to.rm,]
head(countData)
```

	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

```
head(coldata)
```

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

Running 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: 13282 6
metadata(1): version
assays(4): counts mu H cooks
rownames(13282): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
  ENSG00000271254
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

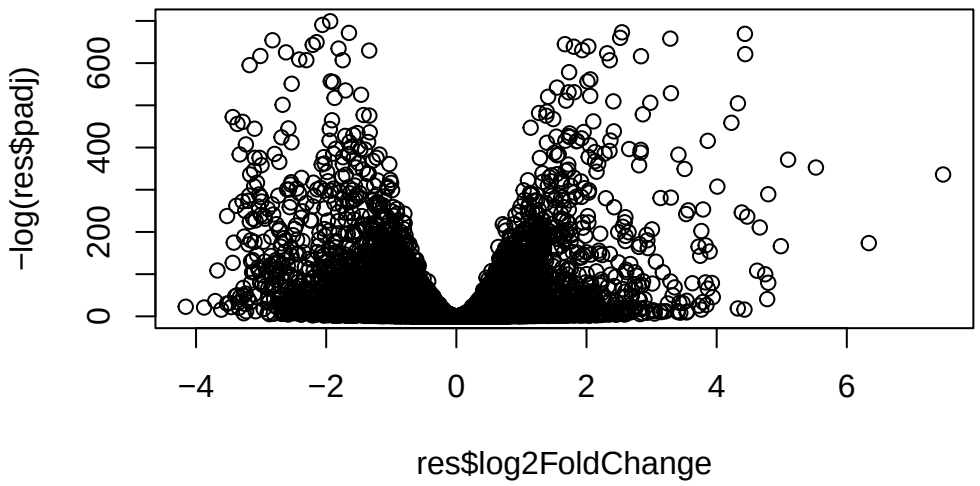
Q. Call the summary() function on your results to get a sense of how many genes are up or down-regulated at the default 0.1 p-value cutoff.

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

```
out of 13282 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4333, 33%
LFC < 0 (down)     : 4400, 33%
outliers [1]       : 0, 0%
low counts [2]     : 0, 0%
(mean count < 1)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

##Volcano Plot

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



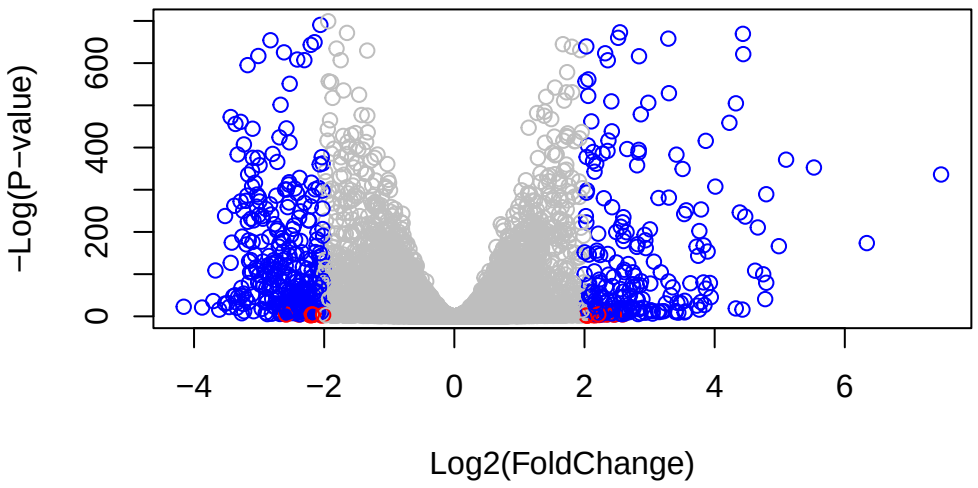
Q2. Q. Improve this plot by completing the below code, which adds color and axis labels

```
# 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(
```



Q3. Q. Use the mapIDs() function multiple times to add SYMBOL, ENTREZID and GENENAME annotation to our results by completing the code below.

```
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, 1)
```

```
log2 fold change (MLE): condition hoxa1_kd vs control_sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 1 row and 9 columns
                 baseMean log2FoldChange     lfcSE      stat    pvalue
                <numeric>      <numeric> <numeric> <numeric> <numeric>
ENSG00000279457   29.9136       0.180304  0.312157  0.577607  0.563529
                     padj      symbol      entrez        name
                <numeric> <character> <character> <character>
ENSG00000279457  0.647026          NA          NA          NA
```

Q4. 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")
```

Section 2: Pathway Analysis

KEGG pathways

```
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]

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

```
     1266     54855      1465      2034      2150      6659
-2.422683  3.201858 -2.313713 -1.887999  3.344480  2.392257
```

```
# 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"
```

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

foldchanges = res$log2FoldChange
```

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

```
     1266     54855      1465      2034      2150      6659
-2.422683  3.201858 -2.313713 -1.887999  3.344480  2.392257
```

```
  attributes(keggres)
```

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

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

```
                                      p.geomean stat.mean        p.val
hsa04110 Cell cycle                3.548176e-06 -4.604234 3.548176e-06
hsa03030 DNA replication           3.992330e-05 -4.191094 3.992330e-05
hsa04114 Oocyte meiosis            2.332810e-04 -3.564509 2.332810e-04
hsa03440 Homologous recombination  2.248158e-03 -2.967340 2.248158e-03
hsa03013 RNA transport             4.162613e-03 -2.662235 4.162613e-03
hsa00670 One carbon pool by folate 8.202725e-03 -2.535331 8.202725e-03
                                           q.val set.size         exp1
hsa04110 Cell cycle                0.0005535155      118 3.548176e-06
hsa03030 DNA replication           0.0031140177       36 3.992330e-05
hsa04114 Oocyte meiosis            0.0121306145       95 2.332810e-04
hsa03440 Homologous recombination  0.0876781678       28 2.248158e-03
hsa03013 RNA transport             0.1298735381      140 4.162613e-03
hsa00670 One carbon pool by folate 0.2115248982       17 8.202725e-03
```

```
  # Look at the first few down (less) pathways
  head(keggres$greater,5)
```

```
                                                    p.geomean stat.mean
hsa04142 Lysosome                                  0.0002611924  3.526257
hsa04640 Hematopoietic cell lineage                0.0033457600  2.794759
hsa04974 Protein digestion and absorption          0.0094536531  2.397179
hsa00603 Glycosphingolipid biosynthesis - globo series 0.0157738311  2.301868
hsa04380 Osteoclast differentiation                0.0200800436  2.067310
```

```
                                                           p.val      q.val
hsa04142 Lysosome                                    0.0002611924 0.04074602
hsa04640 Hematopoietic cell lineage                  0.0033457600 0.26096928
hsa04974 Protein digestion and absorption            0.0094536531 0.49158996
hsa00603 Glycosphingolipid biosynthesis - globo series 0.0157738311 0.51128105
hsa04380 Osteoclast differentiation                  0.0200800436 0.51128105
                                                     set.size         exp1
hsa04142 Lysosome                                         108 0.0002611924
hsa04640 Hematopoietic cell lineage                        40 0.0033457600
hsa04974 Protein digestion and absorption                  42 0.0094536531
hsa00603 Glycosphingolipid biosynthesis - globo series     12 0.0157738311
hsa04380 Osteoclast differentiation                        92 0.0200800436
```

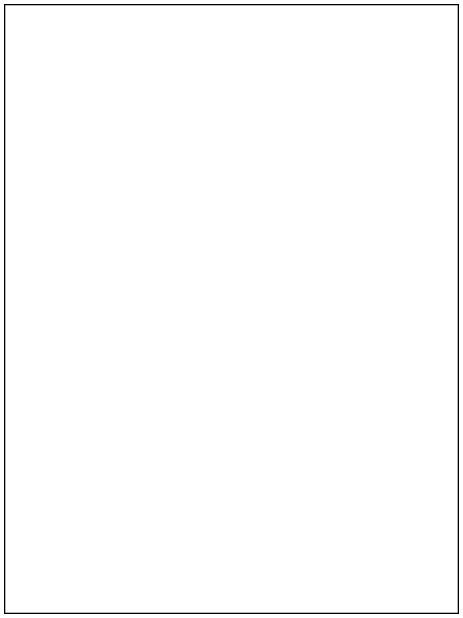
```
pathview(gene.data=foldchanges, pathway.id="hsa04110")
```

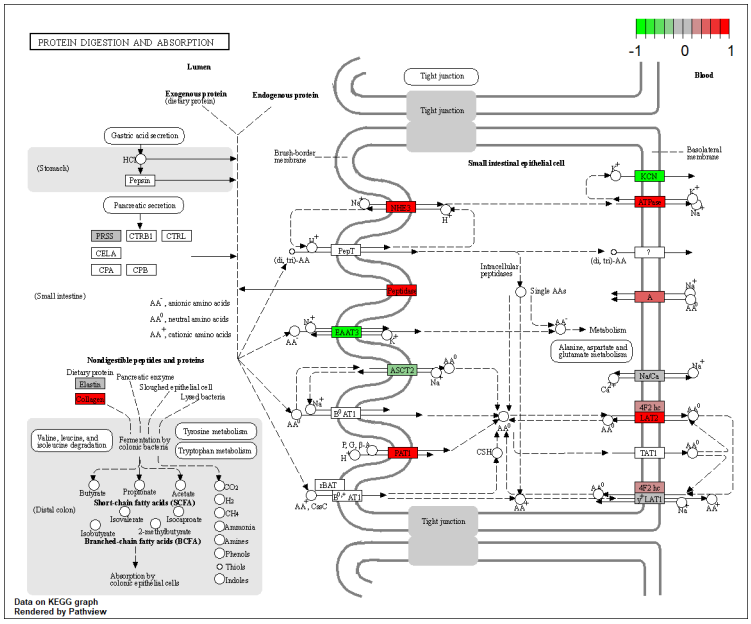
```
'select()' returned 1:1 mapping between keys and columns
```

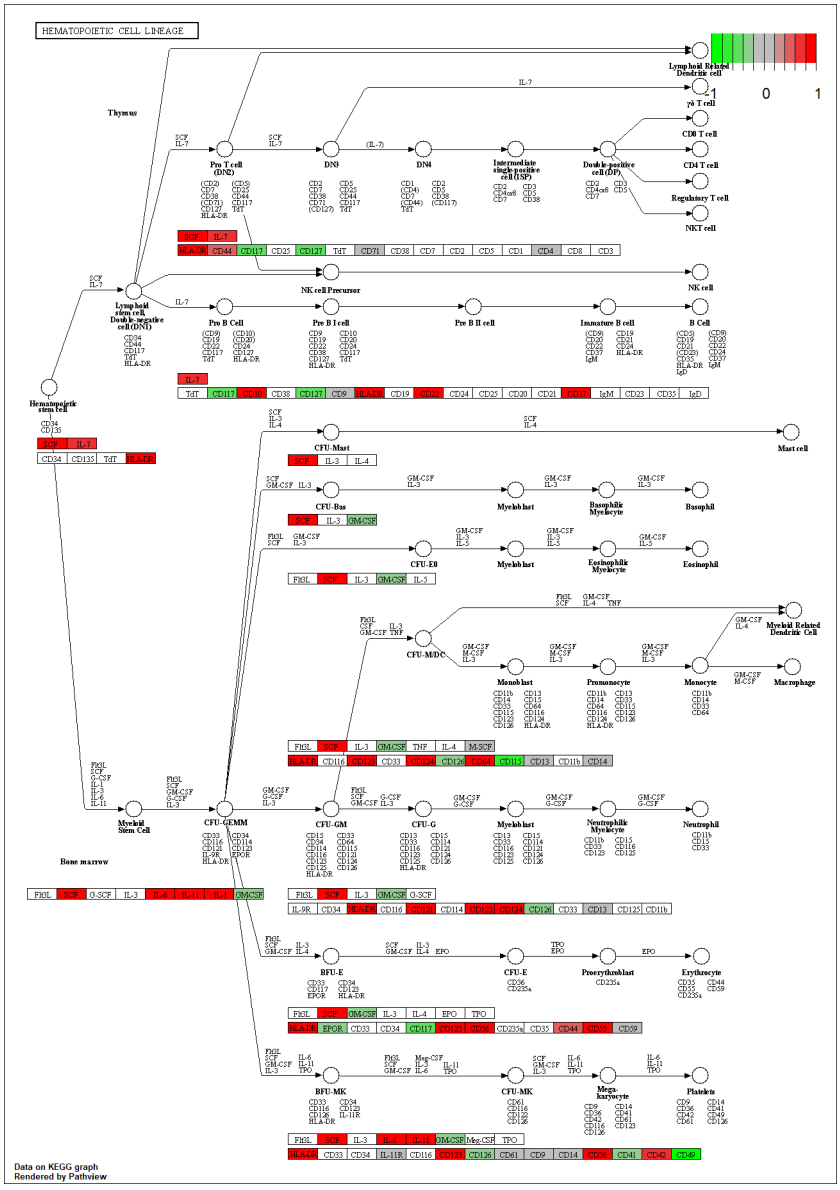
```
Info: Working in directory C:/Users/neha2/Desktop/Winter 2024/BIMM 143/Class14
```

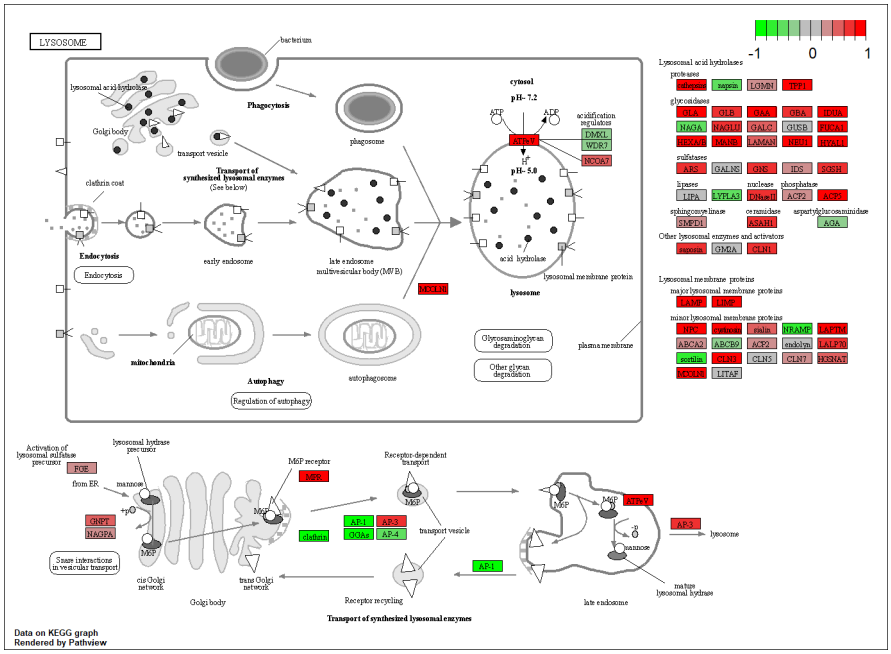
```
Info: Writing image file hsa04110.pathview.png
```

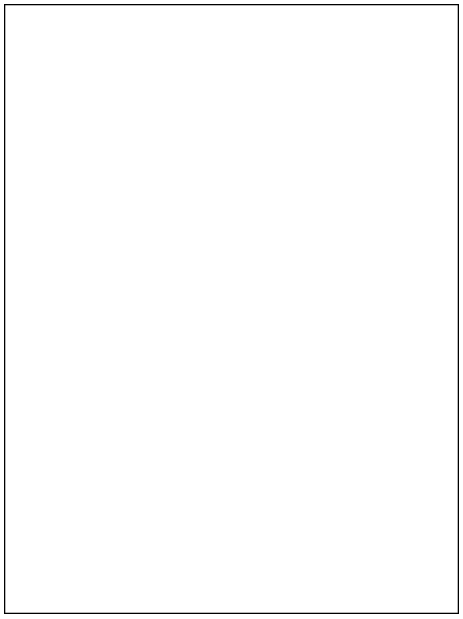
```
#length(foldchanges)
```




PROTEIN DIGESTION AND ABSORPTION
-1
0
1
Lumen
Blood
Exogenous protein
(dietary protein)
Endogenous protein
Tight junction
Gastric acid secretion
(Stomach)
HCl
Pepsin
Brush-border membrane
Small intestinal epithelial cell
Basolateral membrane
KCNJ1
ATPase
Pancreatic secretion
PRSS
CTRB1
CTRL
CELA
CPA
CPB
NHE3
PepT1
(di, tri)-AA
?
Intracellular peptidases
Single AAs
(Small intestine)
AA-, anionic amino acids
AA0, neutral amino acids
A+, cationic amino acids
Cytoplasm
A
LAAT1
Metabolism
Alanine, aspartate and glutamate metabolism
Nondigestible peptides and proteins
Dietary protein
Pancreatic enzyme
Sloughed epithelial cell
Elastin
Lysed bacteria
Collagen
ASCT2
NaPCs
B0AT1
4F2 hc
LAT2
Valine, leucine, and isoleucine degradation
Fermentation by colonic bacteria
Tyrosine metabolism
Tryptophan metabolism
PAT1
TAT1
CSH
rBAT
b0,+AT
4F2 hc
y+LAT1
Butyrate
Propionate
Acetate
Short-chain fatty acids (SCFA)
(Distal colon)
Isovalerate
Isocaproate
Isobutyrate
2-methylbutyrate
Branched-chain fatty acids (BCFA)
Absorption by colonic epithelial cells
CO2
H2
CH4
Ammonia
Amines
Phenols
Thiols
Indoles
Data on KEGG graph
Rendered by Pathview


HEMATOPOIETIC CELL LINEAGE
Lymphoid Related Dendritic cell
-1
0
1
Thymus
IL-7
γδ T cell
CD8 T cell
CD4 T cell
Regulatory T cell
NKT cell
Pro T cell (DN2)
DN3
DN4
Intermediate single-positive cell (ISP)
Double-positive cell (DP)
NK cell Precursor
NK cell
Lymphoid stem cell, Double-negative cell (DN1)
Pre B Cell
Pre B I cell
Pre B II cell
Immature B cell
B Cell
Hematopoietic stem cell
CFU-Mast
Mast cell
CFU-Bas
Myelocyte
Basophil
CFU-E0
Eosinophilic Myelocyte
Eosinophil
CFU-MDC
Monoblast
Promonocyte
Monocyte
Myeloid Related Dendritic Cell
Macrophage
Myeloid Stem Cell
CFU-GEMM
CFU-GM
CFU-G
Neutrophilic Myelocyte
Neutrophil
Bone marrow
BFU-E
CFU-E
Proerythroblast
Erythrocyte
BFU-MK
CFU-MK
Mega-karyocyte
Platelets
Data on KEGG graph
Rendered by Pathview





```
## 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] "hsa04142" "hsa04640" "hsa04974" "hsa00603" "hsa04380"
```

```
length(keggresids)
```

```
[1] 5
```

```
#pathview(gene.data=foldchanges, pathway.id=keggresids)
```

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

Yes. In the code : `keggrespathways <- rownames(keggres$greater)[1:5]` we would change the `greater` to `less` to look at downregulated pathways.

Section 3: Gene Ontology

```
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
GO:0007156 homophilic cell adhesion                7.523307e-05  3.873939
GO:0016339 calcium-dependent cell-cell adhesion    8.556504e-04  3.340855
GO:0010817 regulation of hormone levels            1.058523e-03  3.091986
GO:0048729 tissue morphogenesis                    1.389102e-03  3.002504
GO:0008285 negative regulation of cell proliferation 1.443571e-03  2.989717
GO:0051047 positive regulation of secretion        1.877703e-03  2.927781
                                                          p.val     q.val
GO:0007156 homophilic cell adhesion                7.523307e-05 0.2796413
```

```
GO:0016339 calcium-dependent cell-cell adhesion      8.556504e-04 0.5718590
GO:0010817 regulation of hormone levels              1.058523e-03 0.5718590
GO:0048729 tissue morphogenesis                      1.389102e-03 0.5718590
GO:0008285 negative regulation of cell proliferation 1.443571e-03 0.5718590
GO:0051047 positive regulation of secretion          1.877703e-03 0.5718590
                                                     set.size         exp1
GO:0007156 homophilic cell adhesion                        90 7.523307e-05
GO:0016339 calcium-dependent cell-cell adhesion            24 8.556504e-04
GO:0010817 regulation of hormone levels                   225 1.058523e-03
GO:0048729 tissue morphogenesis                           347 1.389102e-03
GO:0008285 negative regulation of cell proliferation      386 1.443571e-03
GO:0051047 positive regulation of secretion               130 1.877703e-03

$less
                                          p.geomean stat.mean        p.val
GO:0000279 M phase                     6.451975e-18 -8.738701 6.451975e-18
GO:0048285 organelle fission           1.832907e-16 -8.369971 1.832907e-16
GO:0000280 nuclear division            2.627088e-16 -8.340038 2.627088e-16
GO:0007067 mitosis                     2.627088e-16 -8.340038 2.627088e-16
GO:0000087 M phase of mitotic cell cycle 9.244549e-16 -8.166584 9.244549e-16
GO:0007059 chromosome segregation      2.502912e-12 -7.264756 2.502912e-12
                                              q.val set.size         exp1
GO:0000279 M phase                     2.398199e-14      467 6.451975e-18
GO:0048285 organelle fission           2.441221e-13      360 1.832907e-16
GO:0000280 nuclear division            2.441221e-13      338 2.627088e-16
GO:0007067 mitosis                     2.441221e-13      338 2.627088e-16
GO:0000087 M phase of mitotic cell cycle 6.872398e-13      348 9.244549e-16
GO:0007059 chromosome segregation      1.550554e-09      135 2.502912e-12

$stats
                                                    stat.mean     exp1
GO:0007156 homophilic cell adhesion                  3.873939 3.873939
GO:0016339 calcium-dependent cell-cell adhesion      3.340855 3.340855
GO:0010817 regulation of hormone levels              3.091986 3.091986
GO:0048729 tissue morphogenesis                      3.002504 3.002504
GO:0008285 negative regulation of cell proliferation 2.989717 2.989717
GO:0051047 positive regulation of secretion          2.927781 2.927781
```

```r
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: 8186"
```

```
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"? Do the most significant pathways listed match your previous KEGG results? What factors could cause differences between the two methods?

The cell cycle has the most significant "Entities p-value". These are not the same as in the KEGG results. This would be because the databases that the data are drawn from are different with different data.