Class 14

Tim

Import Data

```
counts <- read.csv("GSE37704_featurecounts.csv", row.names = 1)
MetaData <- read.csv("GSE37704_metadata.csv")
```

counts MetaData

```
colnames(counts)
```

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

```
MetaData$id
```

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

Data Cleanup

we can get rid of the first column in `counts` to make these match

```
countData <- counts[,-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

```
MetaData$id
```

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

```
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
```

to check they all match

```
all(colnames(countData) == MetaData$id)
```

```
[1] TRUE
```

```
to.keep.inds <- rowSums(countData) > 0
cleanCounts <- countData[to.keep.inds,]
head(cleanCounts)
```

```
                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
```

```
#rowSums(countData) == 0
```

setup for DESeup

```
library(DESeq2)
```

```
dds <- DESeqDataSetFromMatrix(countData = cleanCounts,
                              colData = MetaData,
                              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
```

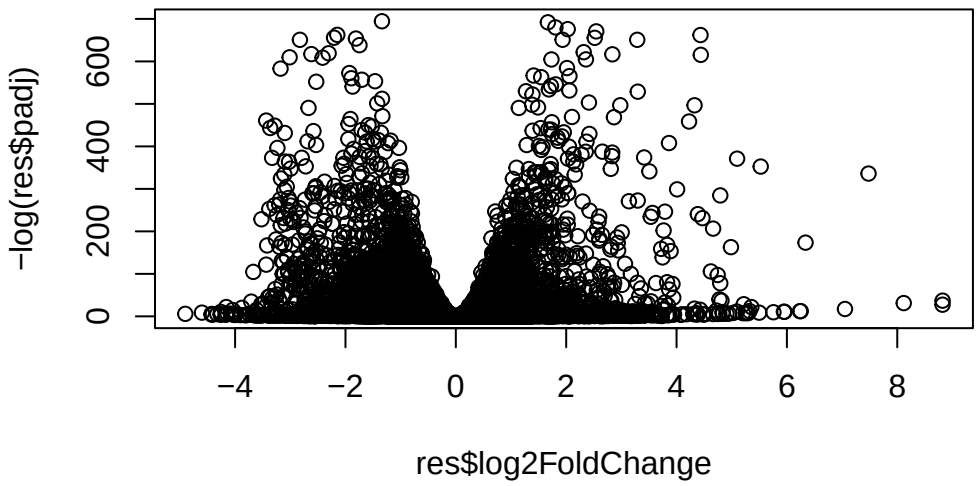
```
fitting model and testing
```

##DESeq

```
res <- results(dds)
```

##Data Vis

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



Section 2 ### Anotation of genes

translate Ensemble IDs in my `res` object to Entrez and gene symbol formats

use the AnnotationDbi package and its `mapIds()` function

```
library(gage)
```

```
library(gageData)
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).
##############################################################################
```

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

```
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"
```

lets map to "SYMBOL", "ENTREZID", "GENENAME" fro our "ENSEMBL" ids ##Pathway Analysis

```
res$genename <- mapIds(org.Hs.eg.db,
                       keys=rownames(res),
                       keytype = "ENSEMBL",
                       column = "GENENAME")
```

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

```
res$symbol <- mapIds(org.Hs.eg.db,
                       keys=rownames(res),
                       keytype = "ENSEMBL",
                       column = "SYMBOL")
```

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

```
res$entrez <- mapIds(org.Hs.eg.db,
                       keys=rownames(res),
                       keytype = "ENSEMBL",
                       column = "ENTREZID")
```

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

Before going any further lets focus in on a subset of "top" hits.

we can use as a starting point log2FC of +2/-2 and a adjusted p-value of 0.05

```
top.inds <- (abs(res$log2FoldChange) > 2 & (res$padj < 0.05))
top.inds[is.na(top.inds)] <- FALSE
```

```
top.genes <- res[top.inds,]
write.csv(top.genes,file = "top_geneset.csv")
```

```
# focus on signaling and metabolic pathways only
kegg.sets.hs = kegg.sets.hs[sigmet.idx.hs]
```

the **gage** function wants a vector of importance as input with gene names and labels - KEGG speaks

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

```
       <NA>      148398       26155      339451       84069       84808
 0.17925708  0.42645712 -0.69272046  0.72975561  0.04057653  0.54281049
```

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

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

```
                                       p.geomean stat.mean        p.val
hsa04110 Cell cycle                 8.995727e-06 -4.378644 8.995727e-06
hsa03030 DNA replication            9.424076e-05 -3.951803 9.424076e-05
hsa03013 RNA transport              1.246882e-03 -3.059466 1.246882e-03
hsa03440 Homologous recombination   3.066756e-03 -2.852899 3.066756e-03
hsa04114 Oocyte meiosis             3.784520e-03 -2.698128 3.784520e-03
hsa00010 Glycolysis / Gluconeogenesis 8.961413e-03 -2.405398 8.961413e-03
                                           q.val set.size         exp1
hsa04110 Cell cycle                 0.001448312      121 8.995727e-06
hsa03030 DNA replication            0.007586381       36 9.424076e-05
hsa03013 RNA transport              0.066915974      144 1.246882e-03
```

```
hsa03440 Homologous recombination     0.121861535       28 3.066756e-03
hsa04114 Oocyte meiosis               0.121861535      102 3.784520e-03
hsa00010 Glycolysis / Gluconeogenesis 0.212222694       53 8.961413e-03
```

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

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

```
Info: Working in directory C:/Users/timha/OneDrive/Desktop/BGGN 213/HW/14/Class14
```

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

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

gobpsets = go.sets.hs[go.subs.hs$BP]
gores <- gage(foldchanges, gsets = gobpsets)
```

```
head(gores$less)
```

```
                                      p.geomean stat.mean        p.val
GO:0048285 organelle fission          1.536227e-15 -8.063910 1.536227e-15
GO:0000280 nuclear division           4.286961e-15 -7.939217 4.286961e-15
GO:0007067 mitosis                    4.286961e-15 -7.939217 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.169934e-14 -7.797496 1.169934e-14
GO:0007059 chromosome segregation     2.028624e-11 -6.878340 2.028624e-11
GO:0000236 mitotic prometaphase       1.729553e-10 -6.695966 1.729553e-10
                                             q.val set.size         exp1
GO:0048285 organelle fission          5.841698e-12      376 1.536227e-15
GO:0000280 nuclear division           5.841698e-12      352 4.286961e-15
GO:0007067 mitosis                    5.841698e-12      352 4.286961e-15
GO:0000087 M phase of mitotic cell cycle 1.195672e-11      362 1.169934e-14
GO:0007059 chromosome segregation     1.658603e-08      142 2.028624e-11
GO:0000236 mitotic prometaphase       1.178402e-07       84 1.729553e-10
```

to run reactome online we need to make a little text

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

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

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