Class 14: RNASeq mini project

Rachel Galleta(A16859649)

Table of contents

Background

here we work thought a complete RNASeq analysis project .The input data comes form a knock-down experiment of a HOX gene.

Data import

Reading the`counts` and `metadata` CSV files

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

check on data structure

```
head(counts)
```

```
                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
```

```
metadata
```

```
         id     condition
1 SRR493366 control_sirna
2 SRR493367 control_sirna
3 SRR493368 control_sirna
4 SRR493369      hoxa1_kd
5 SRR493370      hoxa1_kd
6 SRR493371      hoxa1_kd
```

some book-keeping is required as there looks to be a mis match btween metadat rows and couts

```
ncol(counts)
```

```
[1] 7
```

```
nrow(metadata)
```

```
[1] 6
```

look like we need to get rid of the first length column of our `counts` objects.

```
cleancounts <- counts[,-1]
```

```
colnames(cleancounts)
```

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

```
metadata$id
```

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

```
all(colnames(cleancounts)==metadata$id)
```

```
[1] TRUE
```

Remove Zero counts genes

there are lots of genes with zero counts. We can Remove these form further anaylis.

```
head(cleancounts)
```

```
                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.keep.inds <-rowSums(cleancounts)> 0
nonzero_counts<-cleancounts[to.keep.inds,]
```

DESeq Analysis

load the packages

```
library(DESeq2)
```

```
Warning: package 'IRanges' was built under R version 4.4.2
```

```
Warning: package 'GenomeInfoDb' was built under R version 4.4.2
```

```
Warning: package 'MatrixGenerics' was built under R version 4.4.2
```

```
Warning: package 'matrixStats' was built under R version 4.4.3
```

Setup DESeq

```
dds<- DESeqDataSetFromMatrix(
                countData= nonzero_counts,
                colData= metadata,
                design= ~condition)
```

```
Warning in DESeqDataSet(se, design = design, ignoreRank): some variables in
design formula are characters, converting to factors
```

Run DESeq

```
dds<-DESeq (dds)
```

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

Get results

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

Data Visualization

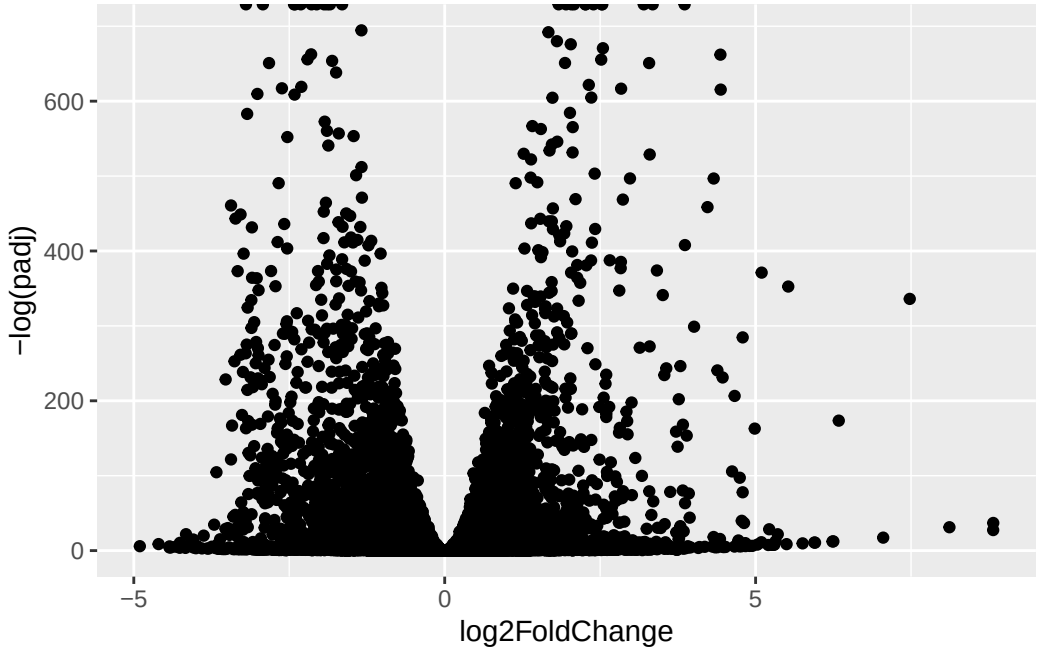
Volcano plot

```
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.4.3
```

```
ggplot(res) +
  aes(log2FoldChange,-log(padj))+
  geom_point()
```

```
Warning: Removed 1237 rows containing missing values or values outside the scale range
(`geom_point()`).
```

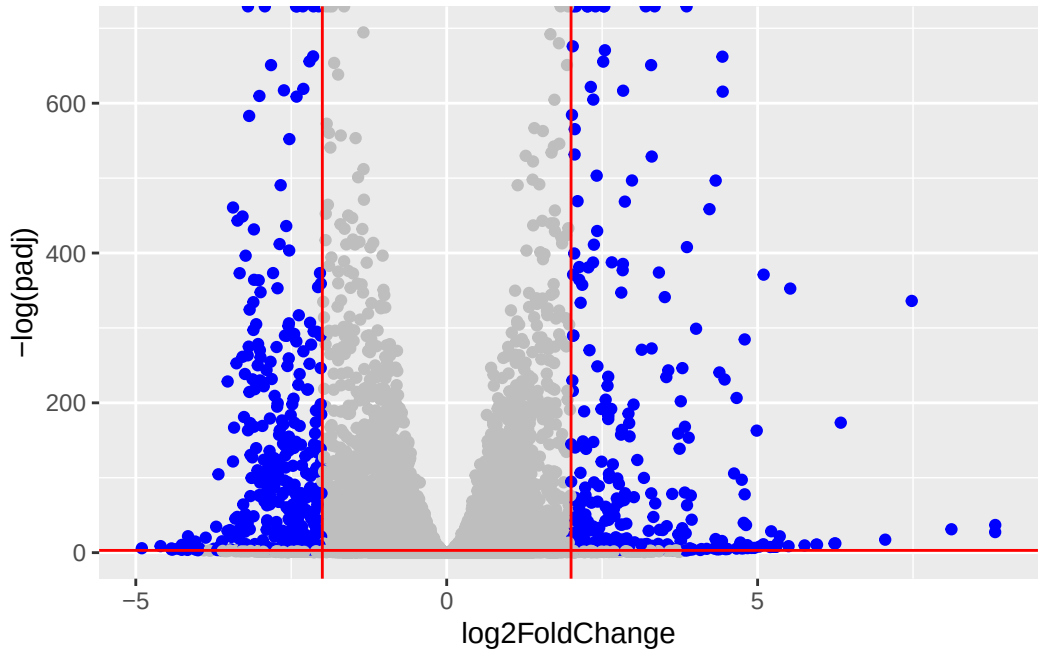


Add threshold lines for fold-change, p-value and color our subset of genes that make threshold cut offs in the plot.

```
mycols<- rep("gray" , nrow(res))
mycols[ abs(res$log2FoldChange) >2 ]<- "blue"
mycols[res$padj>0.05] <-"gray"

ggplot(res) +
  aes(log2FoldChange,-log(padj))+
  geom_point(col=mycols)+
  geom_vline(xintercept= c(-2,2), col="red")+
  geom_hline(yintercept=-log(0.05),col="red")
```

```
Warning: Removed 1237 rows containing missing values or values outside the scale range
(`geom_point()`).
```



Add Annotation

Add gene symbols and entrez ids

```
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(x=org.Hs.eg.db,
       keys = row.names(res),
       keytype= "ENSEMBL",
       column= "SYMBOL")
```

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

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

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

Pathway Analysis

Run gage analysis with KEGG

```
library(gage)
library (gageData)
library(pathview)
```

we named vector of the change values as input for gage

```
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
```

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

```
head(keggres$less,5)
```

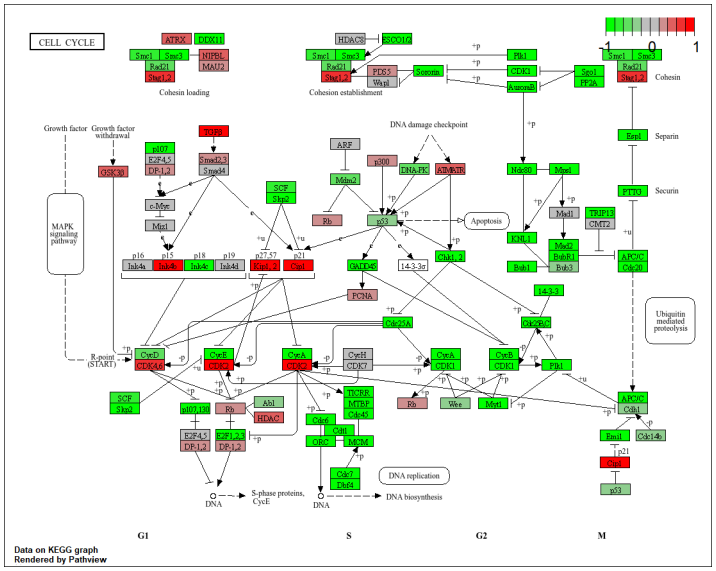
```
                                                  p.geomean stat.mean
hsa04110 Cell cycle                             8.995727e-06 -4.378644
hsa03030 DNA replication                        9.424076e-05 -3.951803
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 -3.765330
hsa03013 RNA transport                          1.246882e-03 -3.059466
hsa03440 Homologous recombination               3.066756e-03 -2.852899
                                                       p.val       q.val
hsa04110 Cell cycle                             8.995727e-06 0.001889103
hsa03030 DNA replication                        9.424076e-05 0.009841047
hsa05130 Pathogenic Escherichia coli infection 1.405864e-04 0.009841047
hsa03013 RNA transport                          1.246882e-03 0.065461279
hsa03440 Homologous recombination               3.066756e-03 0.128803765
                                                set.size         exp1
hsa04110 Cell cycle                                  121 8.995727e-06
hsa03030 DNA replication                              36 9.424076e-05
hsa05130 Pathogenic Escherichia coli infection        53 1.405864e-04
hsa03013 RNA transport                               144 1.246882e-03
hsa03440 Homologous recombination                     28 3.066756e-03
```

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

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

```
Info: Working in directory C:/Users/rache/Documents/BIMM 143/class 14
```

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

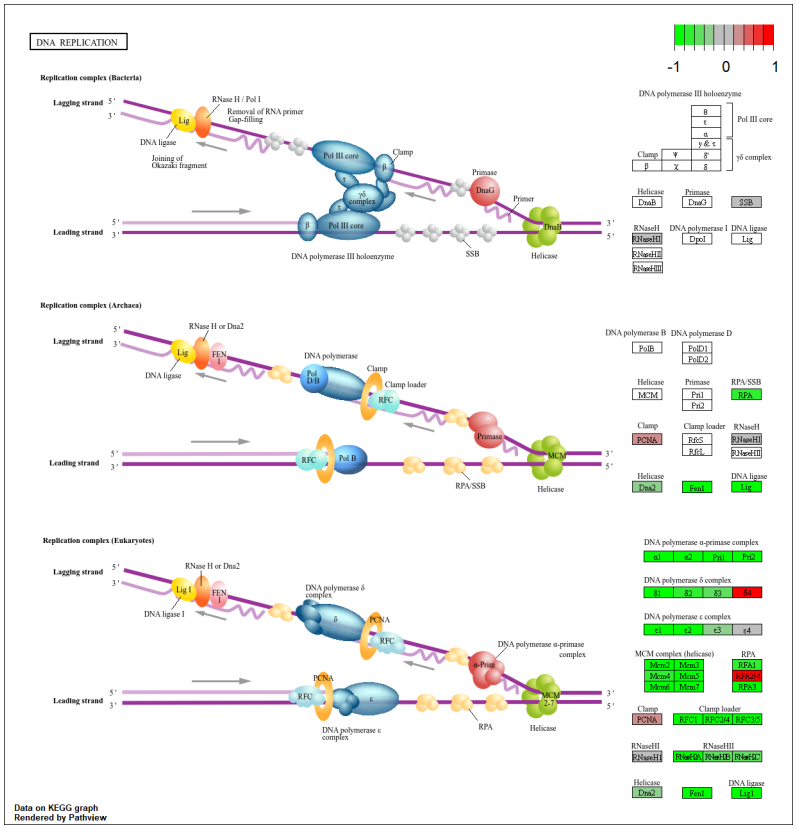



```
pathview(pathway.id = "hsa03030", gene.data=foldchanges)
```

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

```
Info: Working in directory C:/Users/rache/Documents/BIMM 143/class 14
```

```
Info: Writing image file hsa03030.pathview.png
```



Go terms

same analysis but using GO genesets rather than KEGG

```
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)

lapply(gobpres, head)
```

```
$greater
                                            p.geomean stat.mean        p.val
GO:0007156 homophilic cell adhesion      8.519724e-05  3.824205 8.519724e-05
```

```
GO:0002009 morphogenesis of an epithelium 1.396681e-04  3.653886 1.396681e-04
GO:0048729 tissue morphogenesis           1.432451e-04  3.643242 1.432451e-04
GO:0007610 behavior                       1.925222e-04  3.565432 1.925222e-04
GO:0060562 epithelial tube morphogenesis  5.932837e-04  3.261376 5.932837e-04
GO:0035295 tube development               5.953254e-04  3.253665 5.953254e-04
                                              q.val set.size         exp1
GO:0007156 homophilic cell adhesion       0.1951953      113 8.519724e-05
GO:0002009 morphogenesis of an epithelium 0.1951953      339 1.396681e-04
GO:0048729 tissue morphogenesis           0.1951953      424 1.432451e-04
GO:0007610 behavior                       0.1967577      426 1.925222e-04
GO:0060562 epithelial tube morphogenesis  0.3565320      257 5.932837e-04
GO:0035295 tube development               0.3565320      391 5.953254e-04

$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

$stats
                                          stat.mean     exp1
GO:0007156 homophilic cell adhesion        3.824205 3.824205
GO:0002009 morphogenesis of an epithelium  3.653886 3.653886
GO:0048729 tissue morphogenesis            3.643242 3.643242
GO:0007610 behavior                        3.565432 3.565432
GO:0060562 epithelial tube morphogenesis   3.261376 3.261376
GO:0035295 tube development                3.253665 3.253665
```

```
head(gobpres$less,4)
```

```
                                            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
                                                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
```

Reactome

Lots of folks like the reactome web interface. You can run this as an R function but lets look at the website first< https://reactome.org/

The website wants a text file with one gene symbol per line of the genes you want to map to pathways.

```
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj),]$symbol
head(sig_genes)
```

```
ENSG00000187634 ENSG00000188976 ENSG00000187961 ENSG00000188290 ENSG00000187608
       "SAMD11"         "NOC2L"        "KLHL17"          "HES4"         "ISG15"
ENSG00000188157
         "AGRN"
```

```
#res$symbol
```

and write out to a fil:

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

Save our results

```
write.csv(res,file="myresults.csv")
```