

# Class 14: RNASeq Mini Project

Ashley PID:A17891957

## Table of contents

Background . . . . .	1
Data Import . . . . .	1
Remove zero count genes . . . . .	3
DESeq analysis . . . . .	3
Data Visualization . . . . .	5
Add Annotation . . . . .	6
Pathway Analysis . . . . .	6
KEGG Pathway . . . . .	6
Go terms . . . . .	10
Reactome . . . . .	12
Save Our Results . . . . .	12

## Background

Here we work through a complete RNASeq analysis project. The input data comes from 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")
```

```
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 between metadata and counts columns

```
ncol(counts)
```

```
[1] 7
```

```
nrow(metadata)
```

```
[1] 6
```

Looks like we need to get rid of the first “length” column of our ‘counts’ object.

```

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 count genes

There are lots of genes with zero counts. We can remove these from futher analysis.

```

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 package

```
library(DESeq2)
```

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)
head(res)
```

log2 fold change (MLE): condition hoxa1 kd vs control sirna

Wald test p-value: condition hoxa1 kd vs control sirna

DataFrame with 6 rows and 6 columns

	baseMean	log2FoldChange	lfcSE	stat	pvalue
	<numeric>	<numeric>	<numeric>	<numeric>	<numeric>
ENSG00000279457	29.9136	0.1792571	0.3248215	0.551863	5.81042e-01
ENSG00000187634	183.2296	0.4264571	0.1402658	3.040350	2.36304e-03
ENSG00000188976	1651.1881	-0.6927205	0.0548465	-12.630156	1.43993e-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
                    <numeric>
ENSG00000279457  6.86555e-01
ENSG00000187634  5.15718e-03
ENSG00000188976  1.76553e-35
ENSG00000187961  1.13413e-07
ENSG00000187583  9.19031e-01
ENSG00000187642  4.03379e-01

```

## Data Visualization

Volcano plot

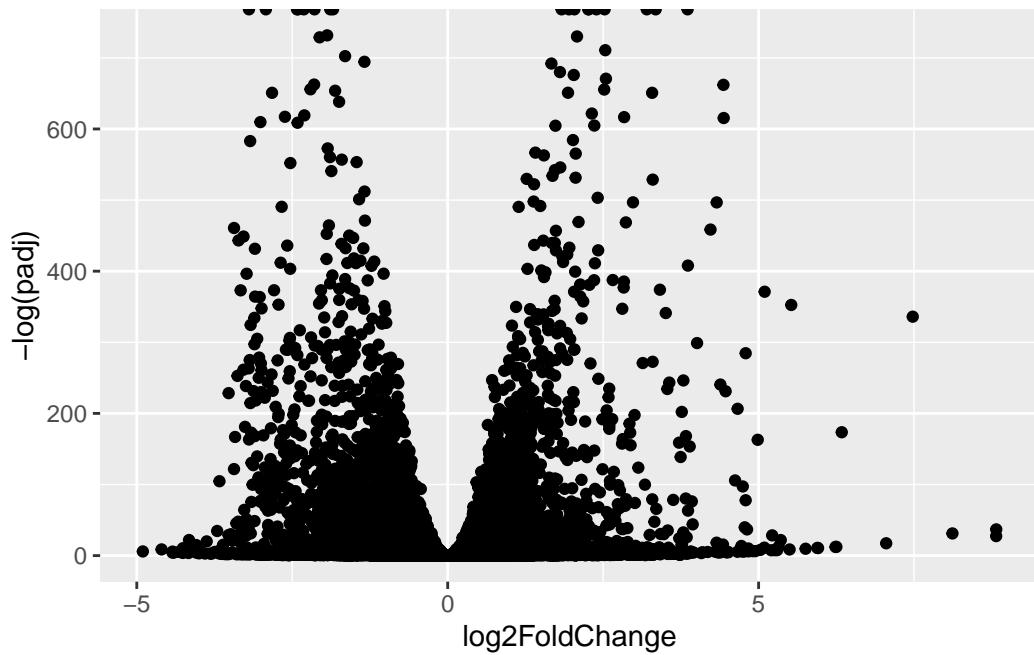
```

library(ggplot2)

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 and P-value and color our subset of genes that make these threshold cut-offs in the plot

## Add Annotation

Add gene symbols and entrez IDs

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

```
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

### KEGG Pathway

Run gage analysis with KEGG

```
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>).

```
#####
```

We need a named vector of fold-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
```

```
library(gage)
library(gageData)

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

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"

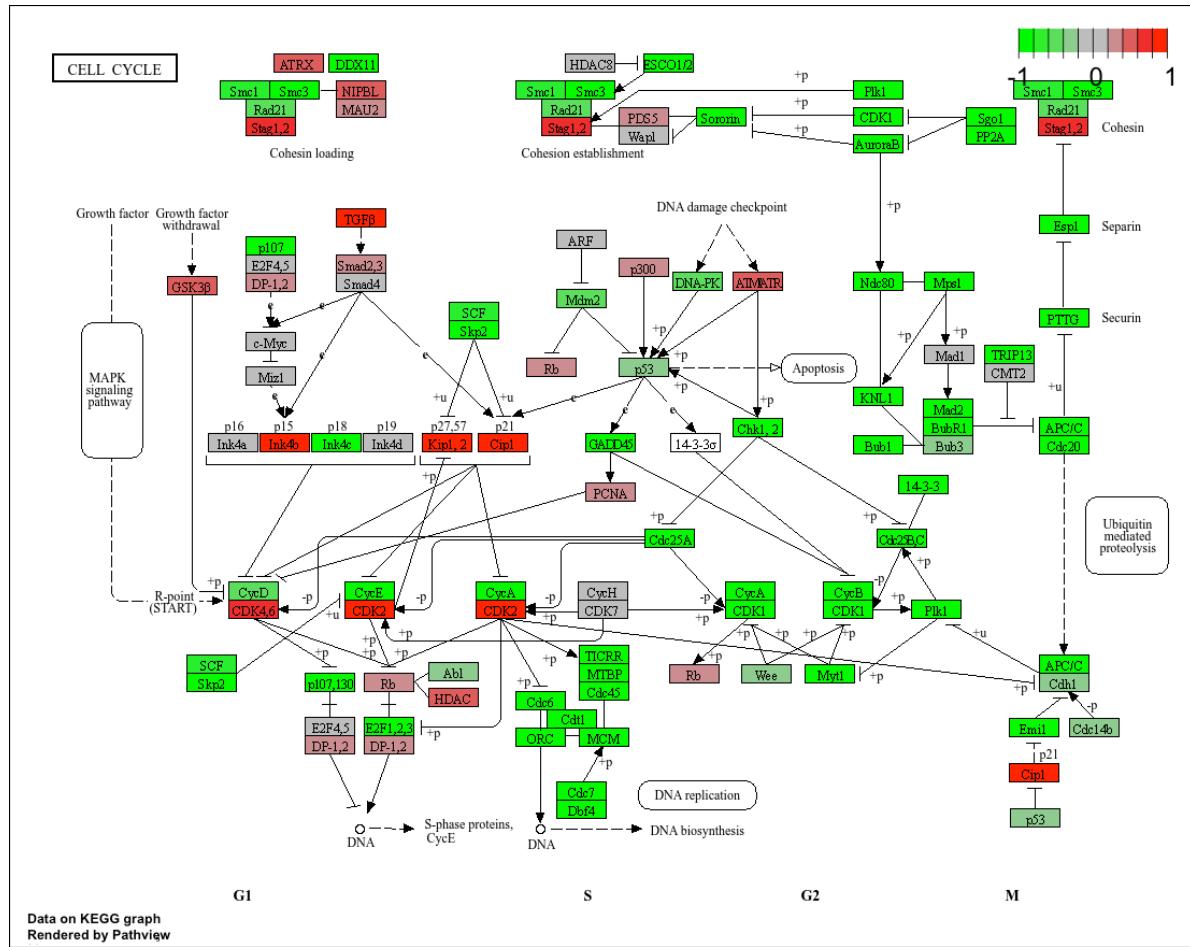
```

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

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

```
Info: Working in directory /Users/ashleymartinez/bimm143/Class 14
```

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

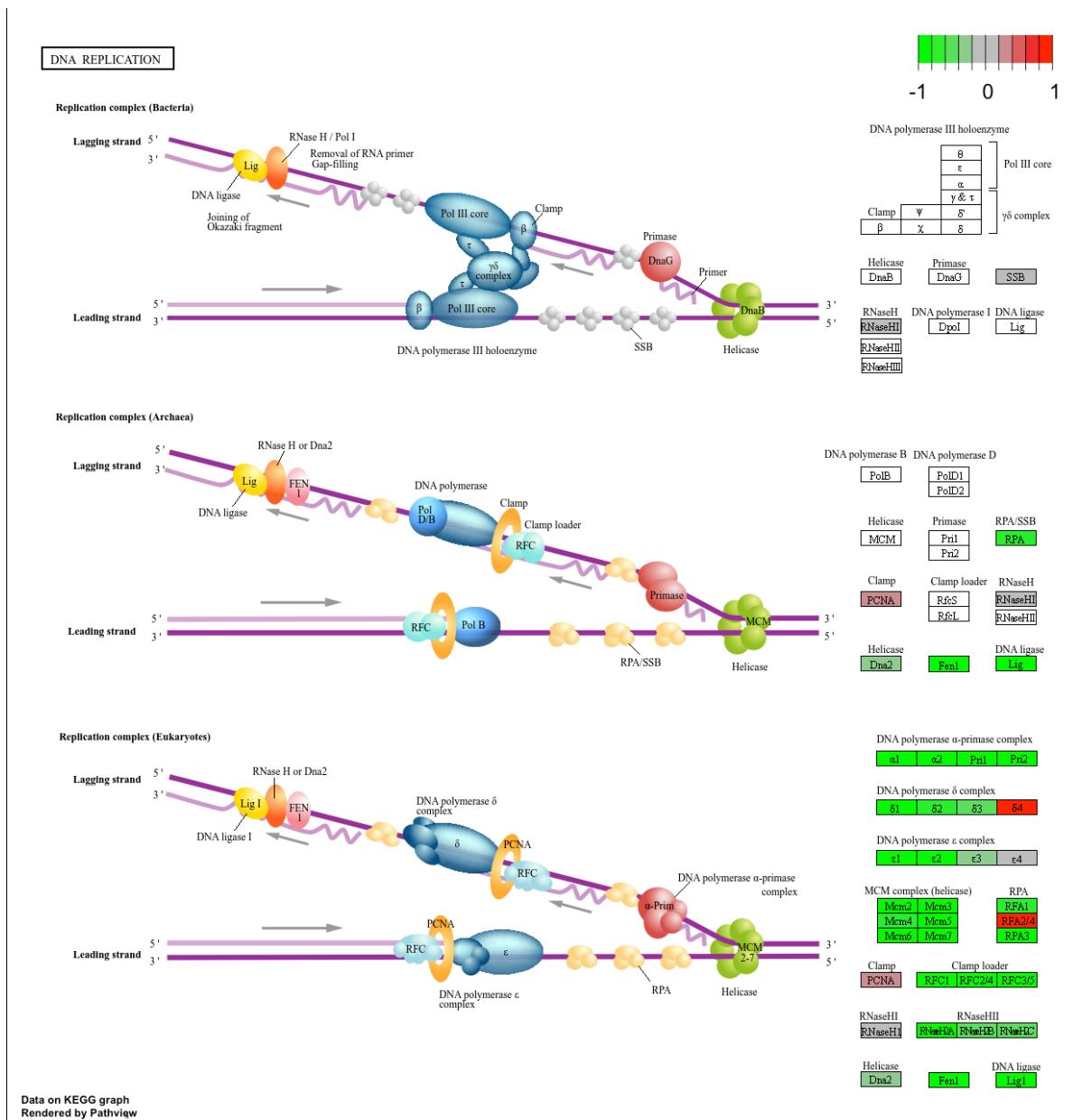


```
pathview(pathway.id="hsa03030", gene.data=foldchanges)
```

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

```
Info: Working in directory /Users/ashleymartinez/bimmm143/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, same.dir=TRUE)

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
```

## Reactome

Lots of folks like the reactome web interface. You can also 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 file:

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

## Save Our Results

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

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?

A: Autophagy has the most significant. It is slightly different from the KEGG results most likely because they use different gene information and mappings.