Class 12: RNA-seq mini project

Sam Altshuler (PID: A59010373)

2/25/2022

1. Input our counts and metadata files

- Check the format and fix if necessary

2. Run differential expression analysis

- Setup that object required for deseq()
- Run deseq()

3. Add annotation

- Gene names and entrezIDs

4. Volcano plot
5. Pathway analysis
6. Save Results!

```
library(DESeq2)
library(ggplot2)
library(gage)
library(gageData)
library(pathview)
library(AnnotationDbi)
library(org.Hs.eg.db)
```

Input counts and metadata

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

Check that the metadata matches the column names of the counts data

```
all(colnames(countData) == row.names(colData))
```

```
## [1] TRUE
```

Get rid of the zeroes

```
# add across each row and if it's not zero (greater than zero), keep it
ct_data <- countData[rowSums(countData) > 0,]
head(ct_data)
```

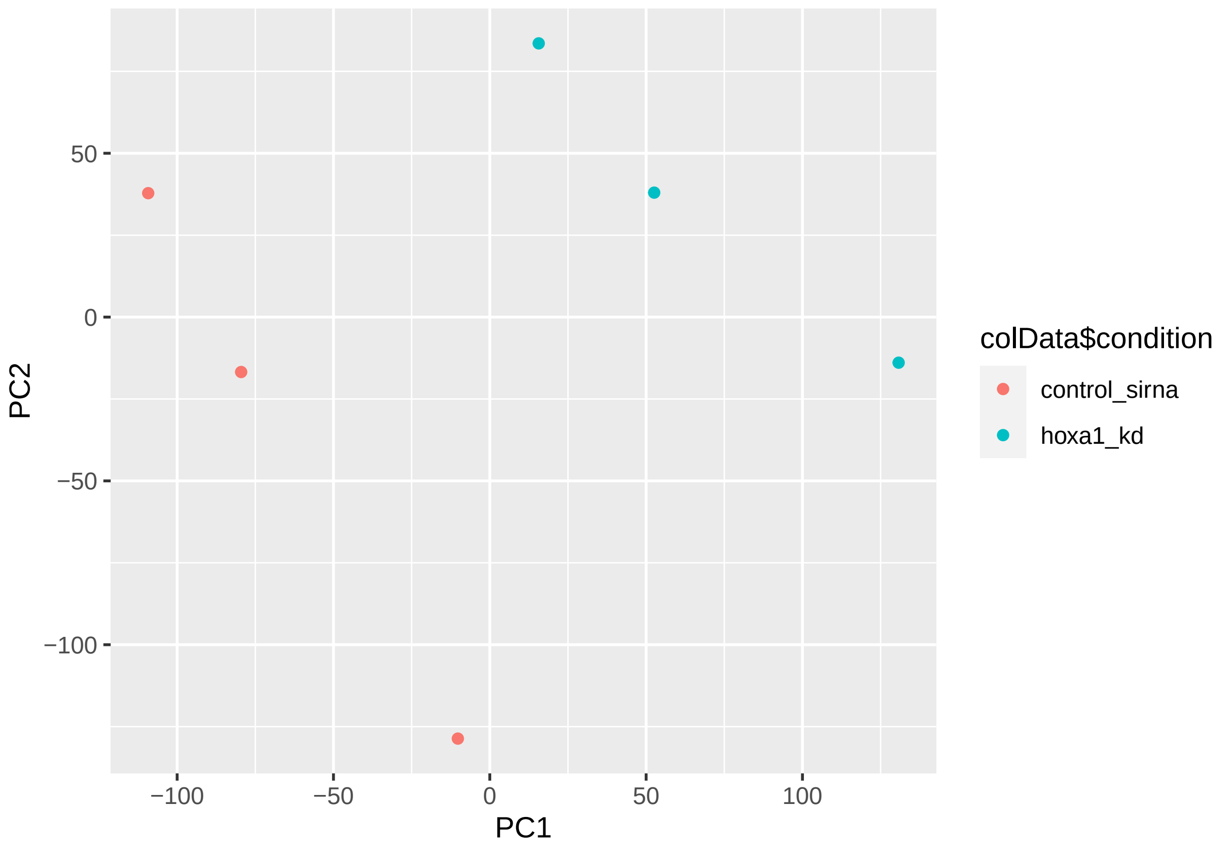
```
##                 SRR493366 SRR493367 SRR493368 SRR493369 SRR493370 SRR493371
## ENSG00000279457        23        28        29        29        28        46
## ENSG00000187634       124       123       205       207       212       258
## ENSG00000188976      1637      1831      2383      1226 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

Let's do a PCA as a QC. This should show us a difference between the control and the experimental condition

```
# remember to transpose the data so the conditions are the rows rather than the columns
pca <- prcomp(t(ct_data), scale = T)
summary(pca)
```

```
## Importance of components:
##                            PC1     PC2      PC3      PC4      PC5       PC6
## Standard deviation     87.7211 73.3196 32.89604 31.15094 29.18417 6.648e-13
## Proportion of Variance  0.4817  0.3365  0.06774  0.06074  0.05332 0.000e+00
## Cumulative Proportion   0.4817  0.8182  0.88594  0.94668  1.00000 1.000e+00
```

```
# pca$x is where the data is stored
ggplot(as.data.frame(pca$x), aes(x= PC1, y = PC2, color = colData$condition))+
  geom_point()
```



The control and knockdown condition are clearly two separate clusters! QC successful.

Time for DESeq analysis

Like lots of BioConductor functions, it wants our data in a specific organized way.

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

dds <- DESeq(dds)
```

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 0.3248216   0.551863 5.81042e-01
## ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 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
##                   <numeric>
## ENSG00000279457 6.86555e-01
## ENSG00000187634 5.15718e-03
## ENSG00000188976 1.76549e-35
## ENSG00000187961 1.13413e-07
## ENSG00000187583 9.19031e-01
## ENSG00000187642 4.03379e-01
```

Add the annotations

Again we will use the AnnotationDbi package to add gene SYMBOLs and entrezIDs.

```
#Store the correctly mapped IDs as a column in the results data frame
res$symbol <- mapIds(org.Hs.eg.db,
                     key = row.names(res), # what are the values you are trying to map
                     keytype = "ENSEMBL", # what is the format of the values
                     column = "SYMBOL", # what are we mapping two
                     multiVals = "first") # if there are multiple values in the symbol, choose the firs
```

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

```
res$gene_name <- mapIds(org.Hs.eg.db,
                     key = row.names(res),
                     keytype = "ENSEMBL",
                     column = "GENENAME",
                     multiVals = "first")
```

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

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

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

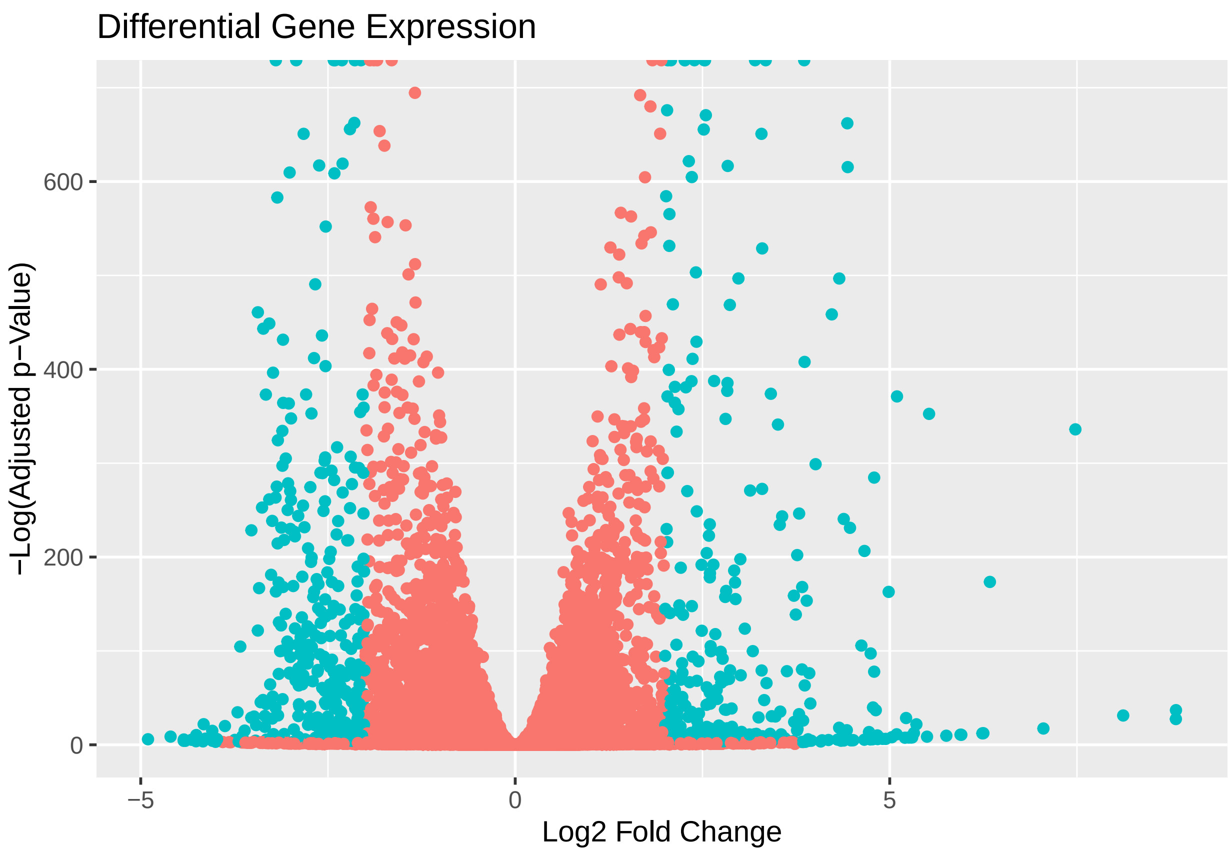
```
#Check the results data frame to confirm that the columns were added correctly
head(res)
```

```
## 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 9 columns
##                   baseMean log2FoldChange     lfcSE       stat      pvalue
##                  <numeric>      <numeric> <numeric>  <numeric>   <numeric>
## ENSG00000279457   29.9136      0.1792571 0.3248216   0.551863 5.81042e-01
## ENSG00000187634  183.2296      0.4264571 0.1402658   3.040350 2.36304e-03
## ENSG00000188976 1651.1881     -0.6927205 0.0548465 -12.630158 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              gene_name      entrez
##                   <numeric> <character>            <character> <character>
## ENSG00000279457 6.86555e-01      WASH9P WAS protein family h..   102723897
## ENSG00000187634 5.15718e-03      SAMD11 sterile alpha motif ..      148398
## ENSG00000188976 1.76549e-35       NOC2L NOC2 like nucleolar ..       26155
## ENSG00000187961 1.13413e-07      KLHL17 kelch like family me..      339451
## ENSG00000187583 9.19031e-01     PLEKHN1 pleckstrin homology ..       84069
## ENSG00000187642 4.03379e-01       PERM1 PPARGC1 and ESRR ind..       84808
```

Volcano plot

```r
# Use Size column to dictate coloring (size = significance)
res$size <- abs(res$log2FoldChange) >2 & res$padj < 0.05
ggplot(as.data.frame(res))+
  aes(x = log2FoldChange, y = -log(padj), color = size)+
  geom_point()+
  xlab("Log2 Fold Change") +
  ylab("-Log(Adjusted p-Value)")+
  labs(title = "Differential Gene Expression")+
  theme(legend.position = "none")
```

```
## Warning: Removed 1237 rows containing missing values (geom_point).
```



Pathway analysis

Use gage() again to start this pathway analysis! Using Kegg and GO genesets here

```
foldchange <- res$log2FoldChange
names(foldchange) <- res$entrez
```

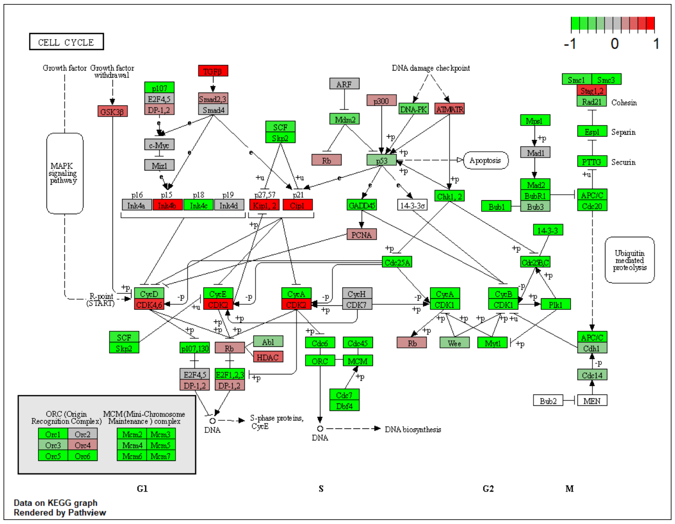
```
data(kegg.sets.hs)
data(sigmet.idx.hs)

# Focus on signaling and metabolic pathways only
kegg.sets.hs <- kegg.sets.hs[sigmet.idx.hs]

# Get the results
keggres <-  gage(foldchange, gsets=kegg.sets.hs)
```

Let's look at the down regulated data.

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



Gene Ontology, Reactome

To use GO we just pass in the GO genesets to the gage function in place of 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(foldchange, 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                       2.195494e-04  3.530241 2.195494e-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                       0.2243795      427 2.195494e-04
## GO:0060562 epithelial tube morphogenesis  0.3711390      257 5.932837e-04
## GO:0035295 tube development               0.3711390      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                        3.530241 3.530241
## GO:0060562 epithelial tube morphogenesis   3.261376 3.261376
## GO:0035295 tube development                3.253665 3.253665
```

Save results

```
write.csv(res, file = "022522_deseq_results.csv")
```