# class14: RNA-Seq analysis mini-project

Thomas Bailey

## Table of contents



### Background

The data for for hands-on session comes from GEO entry: GSE37704, which is associated with the following publication:

> Trapnell C, Hendrickson DG, Sauvageau M, Goff L et al. "Differential analysis of gene regulation at transcript resolution with RNA-seq". Nat Biotechnol 2013 Jan;31(1):46-53. PMID: 23222703

The authors report on differential analysis of lung fibroblasts in response to loss of the developmental transcription factor HOXA1. Their results and others indicate that HOXA1 is required for lung fibroblast and HeLa cell cycle progression. In particular their analysis show that "loss of HOXA1 results in significant expression level changes in thousands of individual transcripts, along with isoform switching events in key regulators of the cell cycle". For our session we have used their Sailfish gene-level estimated counts and hence are restricted to protein-coding genes only.

## Data Import

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

## Inspect and tidy data

Does the `counts` columns match the `colData` rows?

```
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
```

```
colData$id
```

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

```
colnames(counts)
```

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

The fix here looks to be removing the first "length" column from counts:

```
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 |

check for matching countData and colData

```
colnames(countData) == colData$id
```

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

Q1. How many genes in total

```
nrow(countData)
```

```
[1] 19808
```

Q2. Filter to remove zero count genes(rows where there are zero counts in all columns). How many genes are left?

```
to.keep.inds <- rowSums(countData) >0
```

```
new.counts <- countData[to.keep.inds,]
```

```
nrow(new.counts)
```

```
[1] 15975
```

## Setup for DESeq

```
library(DESeq2)
```

Setup input object for DESeq

```
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
```

## Run DESeq

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

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

```
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>
ENSG00000186092    0.0000             NA        NA        NA         NA
ENSG00000279928    0.0000             NA        NA        NA         NA
ENSG00000279457   29.9136       0.179257  0.324822  0.551863 0.58104205
ENSG00000278566    0.0000             NA        NA        NA         NA
ENSG00000273547    0.0000             NA        NA        NA         NA
ENSG00000187634  183.2296       0.426457  0.140266  3.040350 0.00236304
                     padj
                <numeric>
ENSG00000186092         NA
ENSG00000279928         NA
ENSG00000279457 0.68707978
ENSG00000278566         NA
ENSG00000273547         NA
ENSG00000187634 0.00516278
```

### Volcano plot of results

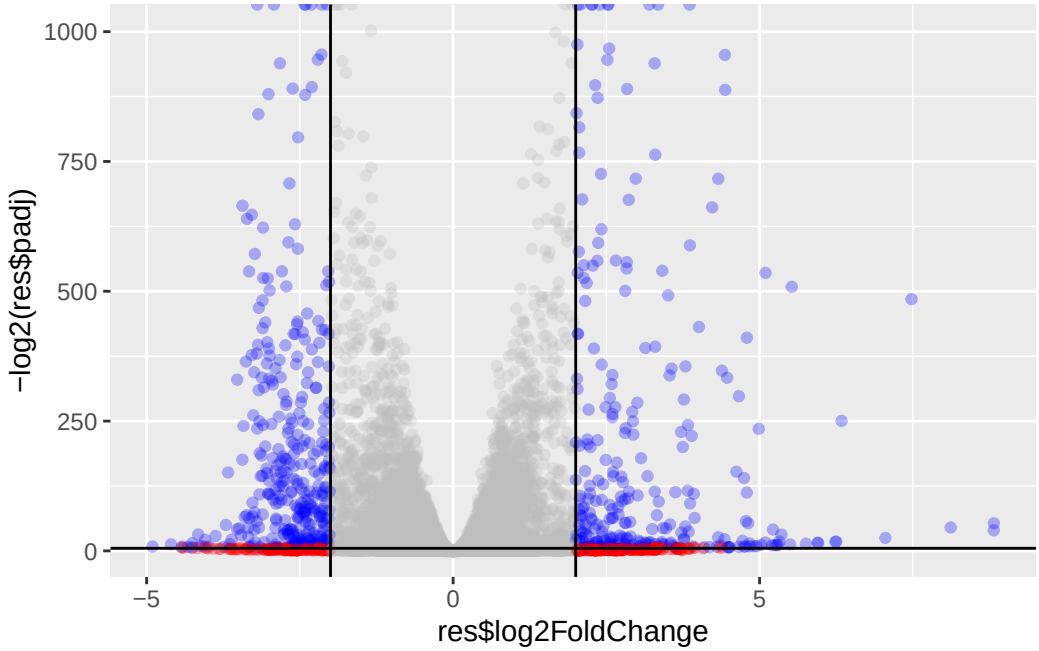
```
library(ggplot2)
```

```
mycols <- rep("gray", nrow(res) )


mycols <- rep("gray", nrow(res))
 mycols[ abs(res$log2FoldChange) > 2 ] <- "red"
 inds <- (res$padj < 0.01) & (abs(res$log2FoldChange) > 2 )
 mycols[ inds ] <- "blue"

ggplot(res) + aes( x = res$log2FoldChange, y = -log2(res$padj)) + geom_point(alpha = 0.3, col
```

```
Warning: Removed 5054 rows containing missing values or values outside the scale range
(`geom_point()`).
```




## Gene annotation

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

```
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).
##############################################################################
```

Add gene SYMBOL and ENTREZID

```
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
```

```
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 8 columns
                 baseMean log2FoldChange     lfcSE      stat     pvalue
                <numeric>      <numeric> <numeric> <numeric>  <numeric>
ENSG00000186092    0.0000             NA        NA        NA         NA
ENSG00000279928    0.0000             NA        NA        NA         NA
ENSG00000279457   29.9136       0.179257  0.324822  0.551863 0.58104205
ENSG00000278566    0.0000             NA        NA        NA         NA
ENSG00000273547    0.0000             NA        NA        NA         NA
ENSG00000187634  183.2296       0.426457  0.140266  3.040350 0.00236304
                     padj      symbol      entrez
                <numeric> <character> <character>
ENSG00000186092         NA       OR4F5       79501
ENSG00000279928         NA          NA          NA
ENSG00000279457 0.68707978          NA          NA
ENSG00000278566         NA          NA          NA
ENSG00000273547         NA          NA          NA
ENSG00000187634 0.00516278      SAMD11      148398
```

## Pathway analysis

```
library(gage)
```

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

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

Load up the KEGG genesets

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

Run pathway analysis

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

```
head(keggres$less, 7)
```

```
                                           p.geomean stat.mean
hsa04110 Cell cycle                        7.077982e-06 -4.432593
hsa03030 DNA replication                   9.424076e-05 -3.951803
hsa05130 Pathogenic Escherichia coli infection 1.076420e-04 -3.835716
hsa03013 RNA transport                     1.048017e-03 -3.112129
hsa04114 Oocyte meiosis                    2.563806e-03 -2.827297
hsa03440 Homologous recombination          3.066756e-03 -2.852899
hsa00010 Glycolysis / Gluconeogenesis      4.360092e-03 -2.663825
                                                  p.val       q.val
hsa04110 Cell cycle                        7.077982e-06 0.001507610
hsa03030 DNA replication                   9.424076e-05 0.007642585
hsa05130 Pathogenic Escherichia coli infection 1.076420e-04 0.007642585
hsa03013 RNA transport                     1.048017e-03 0.055806908
hsa04114 Oocyte meiosis                    2.563806e-03 0.108869849
hsa03440 Homologous recombination          3.066756e-03 0.108869849
hsa00010 Glycolysis / Gluconeogenesis      4.360092e-03 0.132671377
                                           set.size         exp1
hsa04110 Cell cycle                             124 7.077982e-06
```

```
hsa03030 DNA replication                             36 9.424076e-05
hsa05130 Pathogenic Escherichia coli infection       55 1.076420e-04
hsa03013 RNA transport                              149 1.048017e-03
hsa04114 Oocyte meiosis                             112 2.563806e-03
hsa03440 Homologous recombination                    28 3.066756e-03
hsa00010 Glycolysis / Gluconeogenesis                65 4.360092e-03
```

```
head(keggres$greater, 7)
```

```
                                                  p.geomean stat.mean
hsa04740 Olfactory transduction                 6.512007e-08  5.345472
hsa04060 Cytokine-cytokine receptor interaction 8.703597e-08  5.313429
hsa05323 Rheumatoid arthritis                   4.392802e-05  4.030693
hsa05332 Graft-versus-host disease              1.685049e-04  3.771387
hsa04640 Hematopoietic cell lineage             2.654205e-04  3.542990
hsa05320 Autoimmune thyroid disease             3.092317e-04  3.540808
hsa00140 Steroid hormone biosynthesis           6.106061e-04  3.334857
                                                       p.val        q.val
hsa04740 Olfactory transduction                 6.512007e-08 9.269331e-06
hsa04060 Cytokine-cytokine receptor interaction 8.703597e-08 9.269331e-06
hsa05323 Rheumatoid arthritis                   4.392802e-05 3.118889e-03
hsa05332 Graft-versus-host disease              1.685049e-04 8.972885e-03
hsa04640 Hematopoietic cell lineage             2.654205e-04 1.097773e-02
hsa05320 Autoimmune thyroid disease             3.092317e-04 1.097773e-02
hsa00140 Steroid hormone biosynthesis           6.106061e-04 1.857987e-02
                                                set.size         exp1
hsa04740 Olfactory transduction                      354 6.512007e-08
hsa04060 Cytokine-cytokine receptor interaction      263 8.703597e-08
hsa05323 Rheumatoid arthritis                         87 4.392802e-05
hsa05332 Graft-versus-host disease                    36 1.685049e-04
hsa04640 Hematopoietic cell lineage                   86 2.654205e-04
hsa05320 Autoimmune thyroid disease                   49 3.092317e-04
hsa00140 Steroid hormone biosynthesis                 54 6.106061e-04
```

Cell Cycle figure

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

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

```
Info: Working in directory C:/Users/thoma/OneDrive/Desktop/BIMM143 winter/Class14
```

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

Cytokine-cytokine receptor interaction

```
pathview(foldchanges, pathway.id = "hsa00140")
```

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

Info: Working in directory C:/Users/thoma/OneDrive/Desktop/BIMM143 winter/Class14

Info: Writing image file hsa00140.pathview.png
```

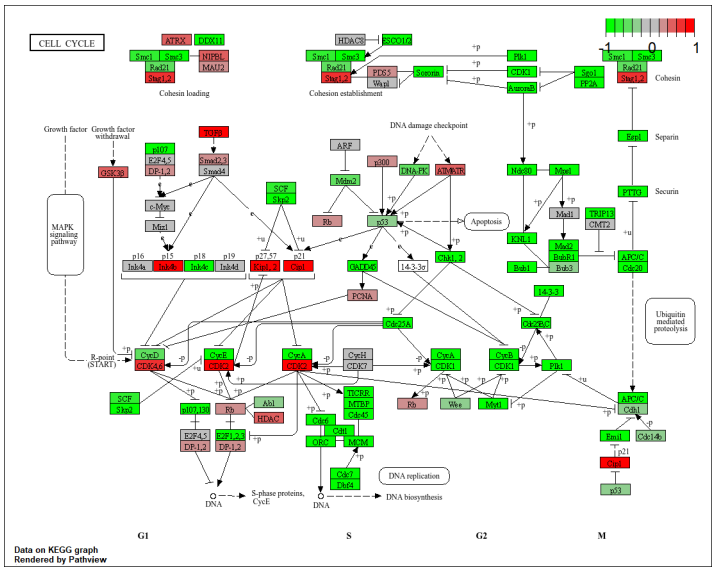
Steroid hormone biosynthesis

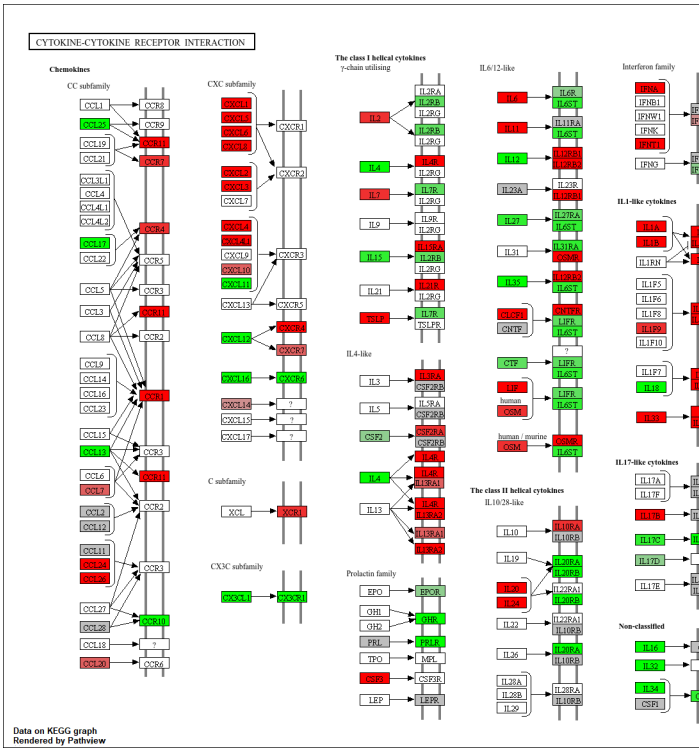
```
pathview(foldchanges, pathway.id = "hsa04060")
```

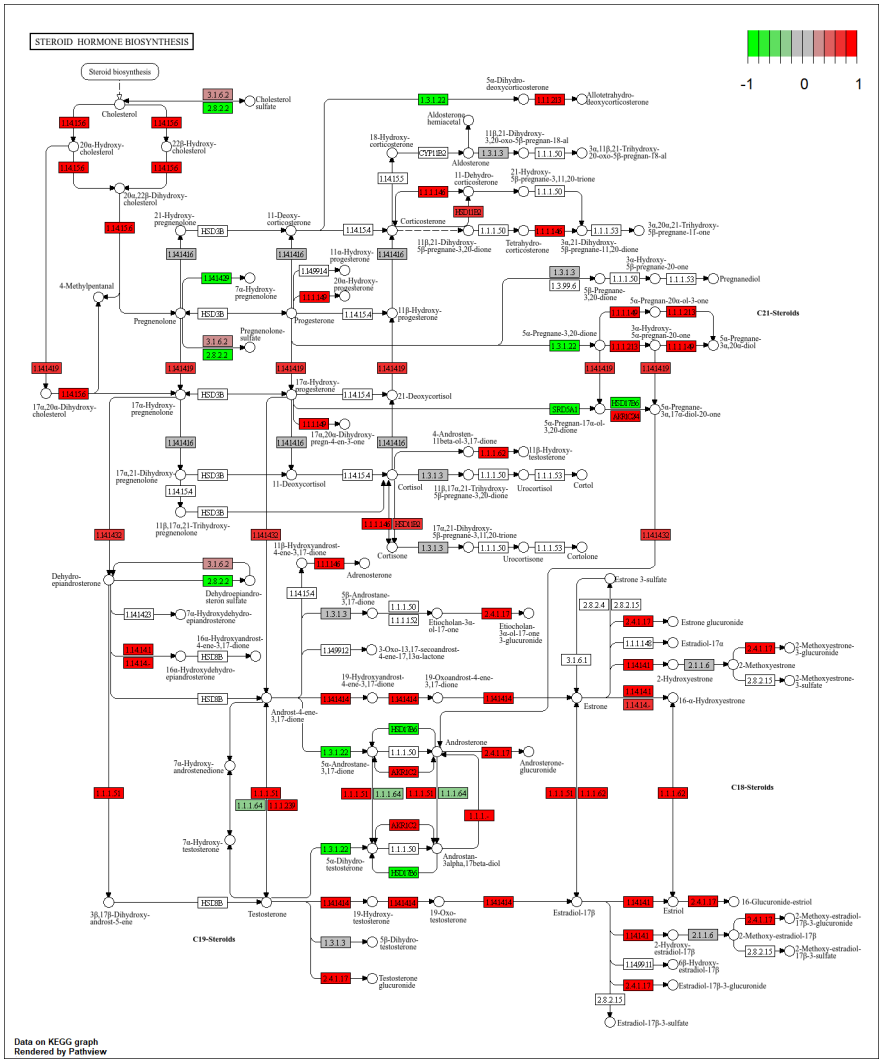
```
'select()' returned 1:1 mapping between keys and columns

Info: Working in directory C:/Users/thoma/OneDrive/Desktop/BIMM143 winter/Class14

Info: Writing image file hsa04060.pathview.png
```






add 3 pictures of cell cycles

## Gene Ontology analysis

Run pathway analysis with GO

```
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)

head(gobpres$less)
```

```
                                          p.geomean stat.mean        p.val
GO:0048285 organelle fission              6.386337e-16 -8.175381 6.386337e-16
GO:0000280 nuclear division               1.726380e-15 -8.056666 1.726380e-15
GO:0007067 mitosis                        1.726380e-15 -8.056666 1.726380e-15
GO:0000087 M phase of mitotic cell cycle  4.593581e-15 -7.919909 4.593581e-15
GO:0007059 chromosome segregation         9.576332e-12 -6.994852 9.576332e-12
GO:0051301 cell division                  8.718528e-11 -6.455491 8.718528e-11
                                                 q.val set.size          exp1
GO:0048285 organelle fission              2.515911e-12      386 6.386337e-16
GO:0000280 nuclear division               2.515911e-12      362 1.726380e-15
GO:0007067 mitosis                        2.515911e-12      362 1.726380e-15
GO:0000087 M phase of mitotic cell cycle  5.020784e-12      373 4.593581e-15
GO:0007059 chromosome segregation         8.373545e-09      146 9.576332e-12
GO:0051301 cell division                  6.352901e-08      479 8.718528e-11
```