Lab 14: RNA-seq Analysis Mini Project

Trevor Hoang (A16371830)

Table of contents

Background

The data for this 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?

```
coldata
```

```
         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
```

```
coldata$id
```

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

```
colnames(counts)
```

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

Need to remove first column (length) 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 inttoal

```
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
```

Set up for DESeq

```
#|message: false
library(DESeq2)
```

```
Loading required package: S4Vectors

Loading required package: stats4

Loading required package: BiocGenerics


Attaching package: 'BiocGenerics'
```

```
The following objects are masked from 'package:stats':

    IQR, mad, sd, var, xtabs

The following objects are masked from 'package:base':

    anyDuplicated, aperm, append, as.data.frame, basename, cbind,
    colnames, dirname, do.call, duplicated, eval, evalq, Filter, Find,
    get, grep, grepl, intersect, is.unsorted, lapply, Map, mapply,
    match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
    Position, rank, rbind, Reduce, rownames, sapply, saveRDS, setdiff,
    table, tapply, union, unique, unsplit, which.max, which.min


Attaching package: 'S4Vectors'

The following object is masked from 'package:utils':

    findMatches

The following objects are masked from 'package:base':

    expand.grid, I, unname

Loading required package: IRanges


Attaching package: 'IRanges'

The following object is masked from 'package:grDevices':

    windows

Loading required package: GenomicRanges

Loading required package: GenomeInfoDb

Loading required package: SummarizedExperiment
Loading required package: MatrixGenerics

Loading required package: matrixStats


Attaching package: 'MatrixGenerics'


The following objects are masked from 'package:matrixStats':

    colAlls, colAnyNAs, colAnys, colAvgsPerRowSet, colCollapse,
    colCounts, colCummaxs, colCummins, colCumprods, colCumsums,
    colDiffs, colIQRDiffs, colIQRs, colLogSumExps, colMadDiffs,
    colMads, colMaxs, colMeans2, colMedians, colMins, colOrderStats,
    colProds, colQuantiles, colRanges, colRanks, colSdDiffs, colSds,
    colSums2, colTabulates, colVarDiffs, colVars, colWeightedMads,
    colWeightedMeans, colWeightedMedians, colWeightedSds,
    colWeightedVars, rowAlls, rowAnyNAs, rowAnys, rowAvgsPerColSet,
    rowCollapse, rowCounts, rowCummaxs, rowCummins, rowCumprods,
    rowCumsums, rowDiffs, rowIQRDiffs, rowIQRs, rowLogSumExps,
    rowMadDiffs, rowMads, rowMaxs, rowMeans2, rowMedians, rowMins,
    rowOrderStats, rowProds, rowQuantiles, rowRanges, rowRanks,
    rowSdDiffs, rowSds, rowSums2, rowTabulates, rowVarDiffs, rowVars,
    rowWeightedMads, rowWeightedMeans, rowWeightedMedians,
    rowWeightedSds, rowWeightedVars

Loading required package: Biobase

Welcome to Bioconductor

    Vignettes contain introductory material; view with
    'browseVignettes()'. To cite Bioconductor, see
    'citation("Biobase")', and for packages 'citation("pkgname")'.



Attaching package: 'Biobase'


The following object is masked from 'package:MatrixGenerics':

    rowMedians
```

```
The following objects are masked from 'package:matrixStats':

    anyMissing, rowMedians
```

Set up input object for DESeq

```
dds = DESeqDataSetFromMatrix(countData = new.counts,
                             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, contrast=c("condition", "hoxa1_kd", "control_sirna"))
```

```
head(res)
```

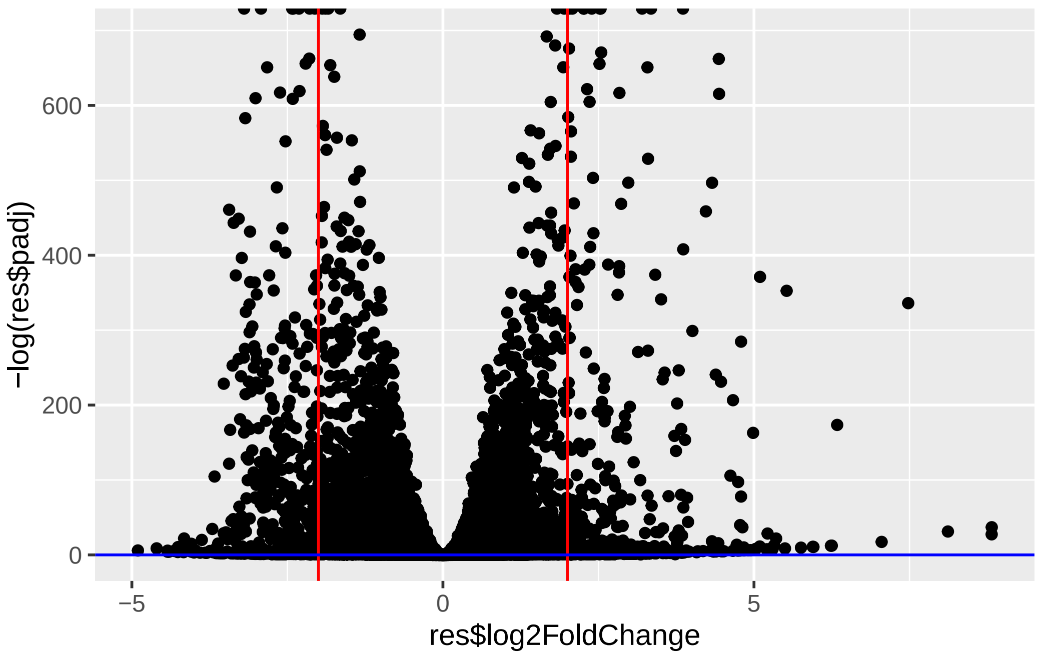
```
log2 fold change (MLE): condition hoxa1_kd vs control_sirna
Wald test p-value: condition hoxa1 kd vs control sirna
DataFrame with 6 rows and 6 columns
                 baseMean log2FoldChange     lfcSE       stat      pvalue
                <numeric>      <numeric> <numeric>  <numeric>   <numeric>
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
```

Plot results

```r
library(ggplot2)

ggplot(res) +
  aes(res$log2FoldChange, -log(res$padj)) +
  geom_point() +
  geom_vline(xintercept = c(2,-2), col ="red") +
geom_hline(yintercept = 0.01, col="blue")
```

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

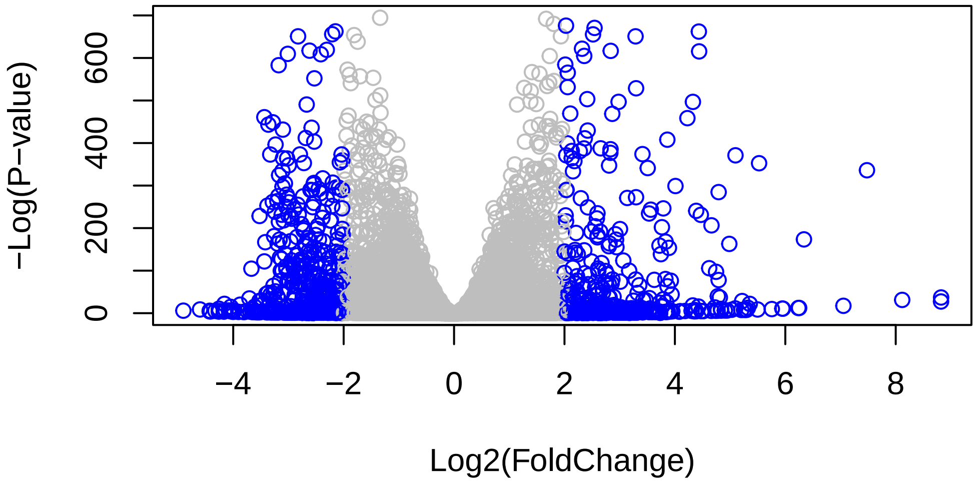


```r
# Make a color vector for all genes
mycols <- rep("gray", nrow(res) )

# Color red the genes with absolute fold change above 2
mycols[ abs(res$log2FoldChange) > 2 ] <- "red"

# Color blue those with adjusted p-value less than 0.01
#  and absolute fold change more than 2
inds <- (-log(res$padj)) & (abs(res$log2FoldChange) > 2 )
mycols[ inds ] <- "blue"

plot( res$log2FoldChange, -log(res$padj), col=mycols, xlab="Log2(FoldChange)", ylab="-Log(P-v
```



Gene annotation

```
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(org.Hs.eg.db,
                    keys=rownames(res),
                    keytype="ENSEMBL",
                    column="ENTREZID")
```

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

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

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

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

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

Pathway Analysis

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

Input vector for `gage()`

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

```
       <NA>      SAMD11       NOC2L      KLHL17     PLEKHN1       PERM1 
 0.17925708  0.42645712 -0.69272046  0.72975561  0.04057653  0.54281049 
```

Load up the KEGG geneset

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

Run pathway analysis

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

Cell cycle figure

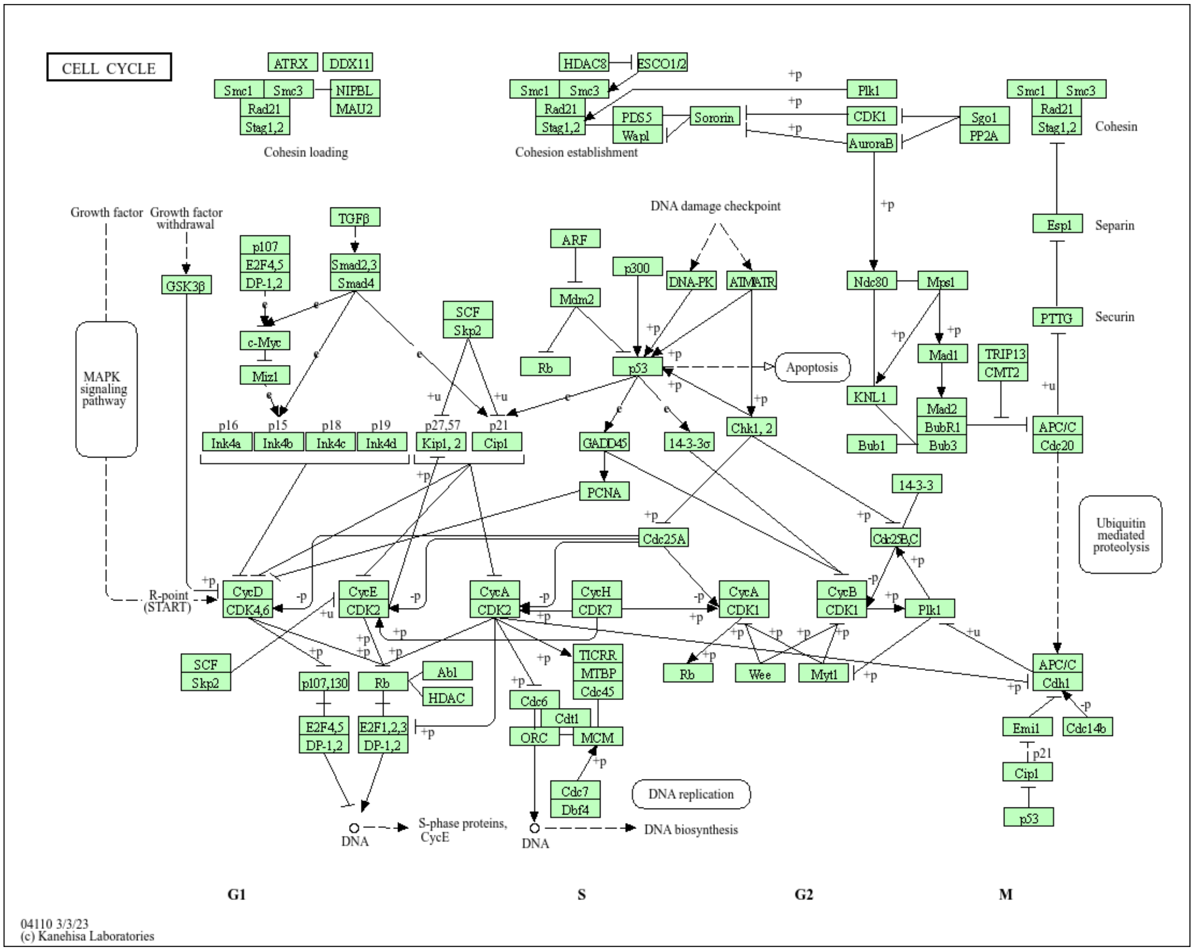
```
pathview(gene.data=foldchanges, pathway.id="hsa04110")
```

```
Warning: None of the genes or compounds mapped to the pathway!
Argument gene.idtype or cpd.idtype may be wrong.

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

Info: Working in directory C:/Users/treer/OneDrive/Documents/BIMM 143/Lab 14

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



Gene Ontology analysis

```
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 q.val
GO:0000002 mitochondrial genome maintenance           NA       NaN    NA    NA
GO:0000003 reproduction                               NA       NaN    NA    NA
GO:0000012 single strand break repair                 NA       NaN    NA    NA
GO:0000018 regulation of DNA recombination            NA       NaN    NA    NA
GO:0000019 regulation of mitotic recombination        NA       NaN    NA    NA
GO:0000022 mitotic spindle elongation                 NA       NaN    NA    NA
                                               set.size exp1
GO:0000002 mitochondrial genome maintenance           0   NA
GO:0000003 reproduction                               0   NA
GO:0000012 single strand break repair                 0   NA
GO:0000018 regulation of DNA recombination            0   NA
GO:0000019 regulation of mitotic recombination        0   NA
GO:0000022 mitotic spindle elongation                 0   NA

$less
                                               p.geomean stat.mean p.val q.val
GO:0000002 mitochondrial genome maintenance           NA       NaN    NA    NA
GO:0000003 reproduction                               NA       NaN    NA    NA
GO:0000012 single strand break repair                 NA       NaN    NA    NA
GO:0000018 regulation of DNA recombination            NA       NaN    NA    NA
GO:0000019 regulation of mitotic recombination        NA       NaN    NA    NA
GO:0000022 mitotic spindle elongation                 NA       NaN    NA    NA
                                               set.size exp1
GO:0000002 mitochondrial genome maintenance           0   NA
GO:0000003 reproduction                               0   NA
GO:0000012 single strand break repair                 0   NA
GO:0000018 regulation of DNA recombination            0   NA
GO:0000019 regulation of mitotic recombination        0   NA
GO:0000022 mitotic spindle elongation                 0   NA

$stats
                                               stat.mean exp1
GO:0000002 mitochondrial genome maintenance          NaN   NA
GO:0000003 reproduction                              NaN   NA
GO:0000012 single strand break repair                NaN   NA
GO:0000018 regulation of DNA recombination           NaN   NA
GO:0000019 regulation of mitotic recombination       NaN   NA
GO:0000022 mitotic spindle elongation                NaN   NA
```

```r
head(gobpres$less)
```

```
                                              p.geomean stat.mean p.val q.val
GO:0000002 mitochondrial genome maintenance          NA       NaN    NA    NA
GO:0000003 reproduction                              NA       NaN    NA    NA
GO:0000012 single strand break repair                NA       NaN    NA    NA
GO:0000018 regulation of DNA recombination           NA       NaN    NA    NA
GO:0000019 regulation of mitotic recombination       NA       NaN    NA    NA
GO:0000022 mitotic spindle elongation                NA       NaN    NA    NA
                                              set.size exp1
GO:0000002 mitochondrial genome maintenance          0   NA
GO:0000003 reproduction                              0   NA
GO:0000012 single strand break repair                0   NA
GO:0000018 regulation of DNA recombination           0   NA
GO:0000019 regulation of mitotic recombination       0   NA
GO:0000022 mitotic spindle elongation                0   NA
```