Class13

Max Strul

11-9-2022

Table of contents

Outline for setting up RNAseq data

- 1.) Read input files,
  - a.) countdata
  - b.) coldata
- 2.) Check and fix
  - a.) remove zero count genes across all
- 3.)DESEQ
  - a.) plot of Log2fc vs -log(p-value)
  - b.)write csv of results
- 4.) Annotation
- 5.) Pathway analysis

Getting our data ready:

```
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, 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, setdiff, sort, table, tapply,
    union, unique, unsplit, which.max, which.min


Attaching package: 'S4Vectors'

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

    expand.grid, I, unname

Loading required package: IRanges

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
```

ReadInputFiles

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

```
             condition
SRR493366 control_sirna
SRR493367 control_sirna
SRR493368 control_sirna
SRR493369      hoxa1_kd
SRR493370      hoxa1_kd
SRR493371      hoxa1_kd
```

```
countdata = read.csv(countFile, row.names=1)
head(countdata)
```

```
                 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
```

```
length(countdata$SRR493366)
```

```
[1] 19808
```

```
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

Whats the point of having colnames if our data frame can initially just have the experimental conditions?

Check and Fix

```
counts <- countData[rowSums(countData)!=0,]
head(counts)
```

	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

```
length(counts[,1])
```

```
[1] 15975
```

number of genes removed

```
19808-15975
```

```
[1] 3833
```

QC wth PCA

the `prcomp()` function in base R can do some "QC". It will check if there are unique groups differentially separated based on the read counts

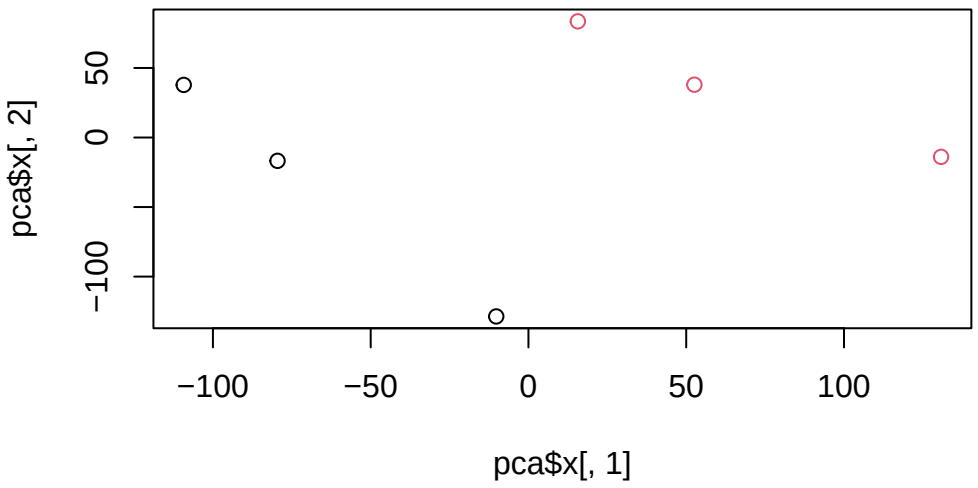
For prcomps we will ensure that we scale our data

```
pca <- prcomp(t(counts),scale=TRUE)
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
```

Here we see the frist two PCS cover about 82 % of variance

```
plot(pca$x[,1],pca$x[,2],col=factor(colData$condition))
```



Why does PCA give QC? Cant a PCA be used to find variance unrelated to the different groups regardless of any other data? its job is to find some way to manipulate the data to find differences in the groups that allow us to separate it out? # DESeq

```
  dds = DESeqDataSetFromMatrix(countData=counts,
                               colData = colData,
                               design=~condition)
```

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

```
  dds = DESeq(dds)
```

```
estimating size factors
```

```
estimating dispersions
```

```
gene-wise dispersion estimates
```

```
mean-dispersion relationship
```

```
final dispersion estimates
```

```
fitting model and testing
```

```
  dds
```

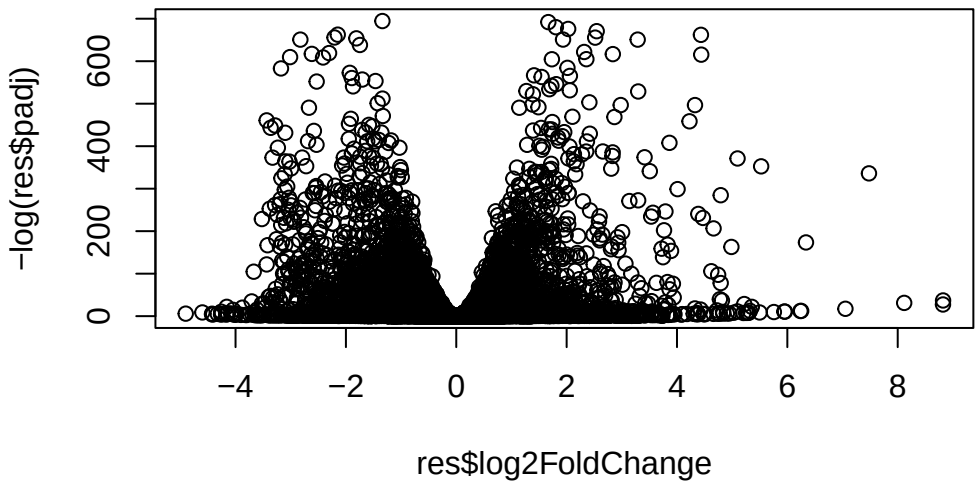
```
class: DESeqDataSet
dim: 15975 6
metadata(1): version
assays(4): counts mu H cooks
rownames(15975): ENSG00000279457 ENSG00000187634 ... ENSG00000276345
  ENSG00000271254
rowData names(22): baseMean baseVar ... deviance maxCooks
colnames(6): SRR493366 SRR493367 ... SRR493370 SRR493371
colData names(2): condition sizeFactor
```

```
res = results(dds)

summary(res)
```

```
out of 15975 with nonzero total read count
adjusted p-value < 0.1
LFC > 0 (up)       : 4349, 27%
LFC < 0 (down)     : 4396, 28%
outliers [1]       : 0, 0%
low counts [2]     : 1237, 7.7%
(mean count < 0)
[1] see 'cooksCutoff' argument of ?results
[2] see 'independentFiltering' argument of ?results
```

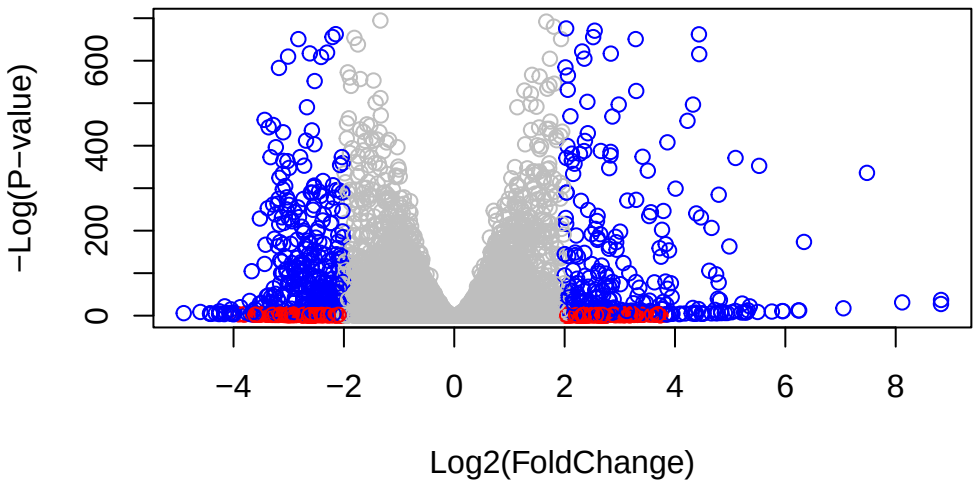
```
plot(res$log2FoldChange,-log(res$padj))
```




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

mycols[abs(res$log2FoldChange)>2] <- "red"
```

```
inds <- (res$padj<0.05 & abs(res$log2FoldChange)>2)
mycols[inds] <- "blue"
plot(res$log2FoldChange, - log(res$padj),col=mycols, xlab="Log2(FoldChange)", ylab="-Log(P
```



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(counts),
                      keytype="ENSEMBL",
                      column="SYMBOL")
```

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

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

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

```
  res$name = mapIds(org.Hs.eg.db,
                    keys=rownames(counts),
                    keytype="ENSEMBL",
                    column="GENENAME")
```

```
'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 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      entrez                  name
                 <numeric> <character> <character>           <character>
ENSG00000279457 6.86555e-01          NA          NA                    NA
```

```
ENSG00000187634 5.15718e-03      SAMD11      148398 sterile alpha motif ..
ENSG00000188976 1.76549e-35       NOC2L       26155 NOC2 like nucleolar ..
ENSG00000187961 1.13413e-07      KLHL17      339451 kelch like family me..
ENSG00000187583 9.19031e-01     PLEKHN1       84069 pleckstrin homology ..
ENSG00000187642 4.03379e-01       PERM1       84808 PPARGC1 and ESRR ind..
```

loading up pathway analysis

Pathway Analysis

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

```
library("gage")
```

```
library("gageData")
data("kegg.sets.hs")
data("sigmet.idx.hs")
kegg.sets.hs=kegg.sets.hs[sigmet.idx.hs]
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
```

gage pathway analysis

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

attributes(keggres)
```

```
$names
[1] "greater" "less"    "stats"
```

```
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
```

```
head(keggres$greater)
```

```
                                    p.geomean stat.mean       p.val
hsa04640 Hematopoietic cell lineage 0.002822776  2.833362 0.002822776
hsa04630 Jak-STAT signaling pathway 0.005202070  2.585673 0.005202070
hsa00140 Steroid hormone biosynthesis 0.007255099  2.526744 0.007255099
hsa04142 Lysosome                   0.010107392  2.338364 0.010107392
hsa04330 Notch signaling pathway    0.018747253  2.111725 0.018747253
hsa04916 Melanogenesis              0.019399766  2.081927 0.019399766
                                        q.val set.size        exp1
hsa04640 Hematopoietic cell lineage 0.3893570       55 0.002822776
hsa04630 Jak-STAT signaling pathway 0.3893570      109 0.005202070
hsa00140 Steroid hormone biosynthesis 0.3893570       31 0.007255099
```

```
hsa04142 Lysosome                     0.4068225      118 0.010107392
hsa04330 Notch signaling pathway      0.4391731       46 0.018747253
hsa04916 Melanogenesis                0.4391731       90 0.019399766
```

Can I create my own gsets? a list of known symbols or entrez iDS? and see what is less or greater?

pathview function

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

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

```
Info: Working in directory /Users/maxstrul/Desktop/BGGN213/Rfiles/Class13/Class13
```

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

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

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

```
Info: Working in directory /Users/maxstrul/Desktop/BGGN213/Rfiles/Class13/Class13
```

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

```
data(go.sets.hs)
data(go.subs.hs)
gobpsets = go.sets.hs[go.subs.hs$BP]
#gobpsets = go.sets.hs#[go.subs.hs$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                       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
```

```r
sig_genes <- res[res$padj <= 0.05 & !is.na(res$padj), "symbol"]
print(paste("Total number of significant genes:", length(sig_genes)))
```

```
[1] "Total number of significant genes: 8147"
```

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

```
#library("clusterProfiler")
#library("tidyverse")
#x <- as.data.frame(res)
#y <- filter(x,log2FoldChange < (-2), padj < 0.05)
```

Analysis from reactome:

```
results_from_reactome <- read.csv("result.csv")
head(results_from_reactome)
```

```
  Pathway.identifier
1      R-HSA-1236977
2       R-HSA-983170
3        R-HSA-69278
4      R-HSA-1640170
5        R-HSA-69618
6       R-HSA-141444
                                                                      Pathway.name
1                                                                Endosomal/Vacuolar pathway
2           Antigen Presentation: Folding, assembly and peptide loading of class I MHC
3                                                                Cell Cycle, Mitotic
4                                                                        Cell Cycle
5                                                                Mitotic Spindle Checkpoint
6 Amplification  of signal from unattached  kinetochores via a MAD2  inhibitory signal
  X.Entities.found X.Entities.total Entities.ratio Entities.pValue Entities.FDR
1               76               82    0.005400777    0.0001670508    0.4211349
2               89              108    0.007113219    0.0018116666    0.8046760
3              409              596    0.039254429    0.0018299063    0.8046760
4              495              734    0.048343542    0.0022879854    0.8046760
5               89              111    0.007310808    0.0037350832    0.8046760
6               77               94    0.006191135    0.0040032064    0.8046760
  X.Reactions.found X.Reactions.total Reactions.ratio Species.identifier
1                 4                 4    0.0002841313               9606
2                15                16    0.0011365251               9606
3               352               352    0.0250035516               9606
4               449               451    0.0320358005               9606
5                 7                 7    0.0004972297               9606
6                 4                 4    0.0002841313               9606
  Species.name
```