Differential gene expression analysis
DESeq2 and edgeR

Reference Transcriptome

De novo Transcriptome

HTSeq

Counts

DESeq2

Differentially Expressed Genes

EdgeR

MA plots

Volcano plots

R Statistical Analysis Package
Differential gene expression analysis
The R Project

• What is R???
  ◦ https://www.r-project.org/
  ◦ https://cran.r-project.org/web/packages/

• How to run R
  ◦ Unix command line
  ◦ Rstudio
    – plots are a little more convenient
  ◦ R GUI
    – slightly less convenient for plots and help

• On PC, run as administrator
  ◦ makes installing packages easier
  ◦ installation of base package takes a few minutes with fast internet
  ◦ update all/some/none? [a/s/n]: answer a
Rstudio on Scholar

Add to .bash_profile

```bash
if [[ -n $RSTUDIO_MULTI_SESSION ]]; then
    export RSTUDIO_SERVER=1
    module purge
    module load gcc/5.2.0
fi
```

- log out of scholar and log back in
Differential gene expression analysis

Bioconductor

• load Bioconductor first, then use biocLite to load other packages
  ○ DESeq2
Differential gene expression analysis

DESeq2

```r
> source("http://bioconductor.org/biocLite.R")
> biocLite("DESeq2")
> library("DESeq2")

# set a working directory - this is on my desktop PC
> getwd()
> setwd('../Desktop/hpc')
> list.files()
[1] "SRR5295840.count" "SRR5295841.count" "SRR5295842.count" "SRR5295843.count" "SRR5295844.count" "SRR5295845.count"

# make a list of the count files produced by htseq, my files from htseq are called SRR<something>.count
# list a variable by typing its name
> files <- grep("SRR.*count",list.files(),value=TRUE)
> files
[1] "SRR5295840.count" "SRR5295841.count" "SRR5295842.count" "SRR5295843.count" "SRR5295844.count" "SRR5295845.count"

# alternatively, create a list of file names by hand
> files <- c("SRR5295840.count","SRR5295841.count","SRR5295842.count","SRR5295843.count","SRR5295844.count","SRR5295845.count")

# make a list of the sample names by extracting the substring before .count
> samples <- sub("(*).count","\1",files)
> samples
[1] "SRR5295840" "SRR5295841" "SRR5295842" "SRR5295843" "SRR5295844" "SRR5295845"

# create a dataframe, called matrix, with the sample names, files, and conditions
> matrix <- data.frame(sampleName=samples, fileName=files, condition=c("co10","co10","co10","xrn3","xrn3","xrn3"))
> matrix
          sampleName     fileName condition
1    SRR5295840    SRR5295840.count     co10
2    SRR5295841    SRR5295841.count     co10
3    SRR5295842    SRR5295842.count     co10
4    SRR5295843    SRR5295843.count   xrn3
5    SRR5295844    SRR5295844.count   xrn3
6    SRR5295845    SRR5295845.count   xrn3

# read data into DESeq2, this is a special method designed to work with HTSeq files
> ds<-DESeqDataSetFromHTSeqCount(sampleTable=matrix, design= ~ condition)
```
**Differential gene expression analysis**

**DESeq2**

```r
> ds
class: DESeqDataSet
dim: 5537 6
metadata(1): version
assays(1): counts
rownames(5537): XLOC_000001 XLOC_000002 ... XLOC_005536 XLOC_005537
rowData names(0):
colnames(6): SRR5295840 SRR5295841 ... SRR5295844 SRR5295845
colData names(1): condition

# look at the first 10 rows
> counts(ds[1:10,])

<table>
<thead>
<tr>
<th>SRR5295840</th>
<th>SRR5295841</th>
<th>SRR5295842</th>
<th>SRR5295843</th>
<th>SRR5295844</th>
<th>SRR5295845</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_000001</td>
<td>159</td>
<td>168</td>
<td>162</td>
<td>193</td>
<td>193</td>
</tr>
<tr>
<td>XLOC_000002</td>
<td>139</td>
<td>135</td>
<td>121</td>
<td>144</td>
<td>136</td>
</tr>
<tr>
<td>XLOC_000003</td>
<td>88</td>
<td>86</td>
<td>100</td>
<td>74</td>
<td>94</td>
</tr>
<tr>
<td>XLOC_000004</td>
<td>280</td>
<td>367</td>
<td>359</td>
<td>353</td>
<td>527</td>
</tr>
<tr>
<td>XLOC_000005</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XLOC_000006</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XLOC_000007</td>
<td>1265</td>
<td>1557</td>
<td>1650</td>
<td>1130</td>
<td>1882</td>
</tr>
<tr>
<td>XLOC_000008</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XLOC_000009</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>XLOC_000010</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

# total counts in each sample
> colSums(counts(ds[,1:6]))

<table>
<thead>
<tr>
<th>SRR5295840</th>
<th>SRR5295841</th>
<th>SRR5295842</th>
<th>SRR5295843</th>
<th>SRR5295844</th>
<th>SRR5295845</th>
</tr>
</thead>
<tbody>
<tr>
<td>1167699</td>
<td>1396648</td>
<td>1415193</td>
<td>1097648</td>
<td>1749632</td>
<td>1756563</td>
</tr>
</tbody>
</table>

> plot(counts(ds[,1]),counts(ds[,2]),pch=20)
> plot(log10(counts(ds[,1])),log10(counts(ds[,2])),pch=20)
```

• some rows(genes) have no counts!
**Differential gene expression analysis**

**DESeq2**

- very nice, variation in samples much less than between samples
- all plots are straight and narrow
- no side peaks

> pairs(log(counts(ds[,1:6])), ch=20, cex=0.5)
Differential gene expression analysis

DESeq2

• scatterplot with transparency

```r
> plot(log(counts(ds[,1])), log(counts(ds[,4])), pch=20, col=rgb(1,0,0,0.05))
```
Differential gene expression analysis
DESeq2
• histogram of raw counts, sample 1

```r
> hist(log(counts(ds[, 1])), breaks=seq(-5, 15, by=0.25))
```
Differential gene expression analysis

**DESeq2**

```r
# DESeq2 is run in a single command
> ds <- DESeq(ds)
estimating size factors
estimating dispersions
gene-wise dispersion estimates
mean-dispersion relationship
final dispersion estimates
fitting model and testing

# results are extracted using the results function
> diff <- results(ds, contrast=c("condition", "col0", "xrn3"))
> diff
log2 fold change (MLE): condition col0 vs xrn3
Wald test p-value: condition col0 vs xrn3
DataFrame with 5537 rows and 6 columns
baseMean log2FoldChange lfcSE stat pvalue padj
<numeric>      <numeric> <numeric>  <numeric>  <numeric> <numeric>
XLOC_000001 183.6774317 -0.00520159 0.2037746 -0.0255262 0.97963525 0.9913485
XLOC_000002 139.5829475     0.19412803 0.2381469  0.8151608 0.41498025 0.6374767
XLOC_000003  89.1867684     0.43603209 0.2339736  1.8635951 0.06237859 0.1889913
XLOC_000004 396.1177489 -0.16924631 0.1254198 -1.3494381 0.17719629 0.3873928
XLOC_000005   0.4142939 -1.94867374 3.7428816  0.5206346 0.60262133        NA
...                 ...            ...       ...        ...        ...       ...
...                 ...            ...       ...        ...        ...       ...
XLOC_005533  425.037261 0.06634836 0.1316014 -0.5041617 0.6141478 0.7908928
XLOC_005534  294.727806 0.22979472 0.1504705  1.5271750 0.1267175 0.3090301
XLOC_005535 141.943955 -0.21796147 0.2141253 -1.0179156 0.3087181 0.5465699
XLOC_005536  3.006644 1.23035506 1.1834939  1.0395956 0.2985278 0.5359263
XLOC_005537  0.000000          NA        NA        NA        NA        NA
```
# show in order of adjusted P Value (FDR)
> diff[order(diff$padj),]

log2 fold change (MLE): condition col0 vs xrn3
Wald test p-value: condition col0 vs xrn3
 DataFrame with 5537 rows and 6 columns

<table>
<thead>
<tr>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_003949</td>
<td>1616.8668</td>
<td>1.467015</td>
<td>0.08966453</td>
<td>16.36115</td>
<td>3.622011e-60</td>
</tr>
<tr>
<td>XLOC_002300</td>
<td>2437.8786</td>
<td>2.079183</td>
<td>0.13740850</td>
<td>15.13140</td>
<td>1.005348e-51</td>
</tr>
<tr>
<td>XLOC_003951</td>
<td>744.0812</td>
<td>1.685829</td>
<td>0.11217557</td>
<td>15.02848</td>
<td>4.778314e-51</td>
</tr>
<tr>
<td>XLOC_004746</td>
<td>1835.1037</td>
<td>2.040852</td>
<td>0.13814181</td>
<td>14.77360</td>
<td>2.168122e-49</td>
</tr>
<tr>
<td>XLOC_005175</td>
<td>151.3101</td>
<td>-4.756780</td>
<td>0.34378165</td>
<td>-13.83663</td>
<td>1.532172e-43</td>
</tr>
</tbody>
</table>

# show in order of fold change
> diff[order(-log2FoldChange),]

log2 fold change (MLE): condition col0 vs xrn3
Wald test p-value: condition col0 vs xrn3
 DataFrame with 5537 rows and 6 columns

<table>
<thead>
<tr>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>XLOC_004641</td>
<td>36.61127</td>
<td>-8.468820</td>
<td>1.231776</td>
<td>-6.875294</td>
<td>6.186217e-12</td>
</tr>
<tr>
<td>XLOC_002996</td>
<td>18.95773</td>
<td>-7.538905</td>
<td>1.281068</td>
<td>-5.884860</td>
<td>3.983903e-09</td>
</tr>
<tr>
<td>XLOC_001569</td>
<td>18.34071</td>
<td>-7.488098</td>
<td>1.257228</td>
<td>-5.956037</td>
<td>2.584281e-09</td>
</tr>
<tr>
<td>XLOC_005165</td>
<td>16.97992</td>
<td>-7.355236</td>
<td>1.278347</td>
<td>-5.753706</td>
<td>8.730770e-09</td>
</tr>
<tr>
<td>XLOC_002717</td>
<td>30.32408</td>
<td>-7.232830</td>
<td>1.223487</td>
<td>-5.911654</td>
<td>3.386886e-09</td>
</tr>
</tbody>
</table>

# histogram of mean expression levels
hist(log(diff[,"baseMean"],breaks=seq(-5,15,by=0.25),main="mean expression")
Differential gene expression analysis

**DESeq2**

- count distribution should be approximately log-normal, we see a big tail on the left (low expression) side
- counts in this region are poorly measured
- DESeq incorporates a test for outliers (Cook's cutoff) but only works for larger number of samples
- We need a manual cutoff by eye, i say 6 counts

```R
> hist(log10(diff[, "baseMean"]), breaks=seq(-2, 6, by=0.1), main="mean expression")
```
**DESeq2**

- The idea of independent filtering is to filter out those tests from the procedure that have no, or little chance of showing significant evidence, without even looking at their test statistic.
- Typically, this results in increased detection power at the same experiment-wide type I error (false discovery rate).
- A good choice for a filtering criterion is one that
  - is statistically independent from the test statistic under the null hypothesis,
  - is correlated with the test statistic under the alternative, and
  - does not notably change the dependence structure – if any – between the test statistics of nulls and alternatives.
- Filtering by total counts generally meets these criteria
**DESeq2**

- **manual cutoff**

```r
> dds = ds[rowSums(counts(ds[,1:6]))>40,]
> hist(log(counts(dds[,1])), breaks=seq(-5,15,by=0.25))
> dds
class: DESeqDataSet
dim: 2862 6
metadata(1): version
assays(3): counts mu cooks
rownames(2862): XLOC_000001 XLOC_000002 ... XLOC_005534 XLOC_005535
rowData names(21): baseMean baseVar ... deviance maxCooks
colnames(6): SRR5295840 SRR5295841 ... SRR5295844 SRR5295845
colData names(2): condition sizeFactor
```

- **rerun DESeq on reduced set**
  - better able to see differences
  - fewer spurious results
DESeq2

- PCA plot with variance stabilizing transformation

```r
# Simple PCA to check whether the data make sense, i.e., do the replicates seem more like each other than the different samples
> vsd <- varianceStabilizingTransformation(dds)
> plotPCA(vsd, intgroup = c("condition"))
```
Differential gene expression analysis

DESeq2

• dispersion estimates

```r
# plot estimated dispersion
plotDispEsts(dds)
```
**Differential gene expression analysis**

**DESeq2**

# MA plot is a traditional way to look at DGE results

```r
> plotMA(diff, ylim=c(-9, 9))
> abline(h=2, col="blue")
> abline(h=-2, col="blue")
```

# volcano plots are another traditional method

```r
> with(diff, plot(log2FoldChange, -log10(padj), pch=20))
> with(subset(diff, padj<0.01), points(log2FoldChange, -log10(padj), pch=20, col="red"))
> with(subset(diff, abs(log2FoldChange)>1.0 & padj<0.01), points(log2FoldChange, -log10(padj), pch=20, col="green"))
> abline(h=2, col="blue")
> abline(v=c(1, -1), col="blue")
```
• Sequentially build up plots
**DESeq2**

- Write out differentially expressed genes
- Actually you should write them all
- With cutoffs from the volcano plot, 203 DEGs

\[ \text{padj} < 0.01 \ \text{abs(log2FoldChange)} > 1.0 \]

```r
# select significant genes with large fold change
> select=diff[,"padj"]<0.01 & abs(diff[,"log2FoldChange"])>1,]
> select

log2 fold change (MLE): condition col0 vs xrn3
Wald test p-value: condition col0 vs xrn3

DataFrame with 203 rows and 6 columns

<table>
<thead>
<tr>
<th></th>
<th>baseMean</th>
<th>log2FoldChange</th>
<th>lfcSE</th>
<th>stat</th>
<th>pvalue</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>col0</td>
<td>87.42532</td>
<td>1.008224</td>
<td>0.247339</td>
<td>4.076285</td>
<td>4.576083e-05</td>
<td>3.956722e-04</td>
</tr>
<tr>
<td>col0</td>
<td>373.19005</td>
<td>1.707859</td>
<td>0.163566</td>
<td>10.441377</td>
<td>1.604657e-25</td>
<td>1.837011e-23</td>
</tr>
<tr>
<td>col0</td>
<td>599.59472</td>
<td>1.73120</td>
<td>0.145523</td>
<td>8.061368</td>
<td>7.544543e-16</td>
<td>3.659743e-14</td>
</tr>
<tr>
<td>col0</td>
<td>499.76517</td>
<td>1.210535</td>
<td>0.156936</td>
<td>10.441377</td>
<td>1.604657e-25</td>
<td>1.837011e-23</td>
</tr>
<tr>
<td>col0</td>
<td>15.96232</td>
<td>-1.847492</td>
<td>0.568850</td>
<td>-3.247767</td>
<td>1.163143e-03</td>
<td>6.807598e-03</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>col0</td>
<td>13.93252</td>
<td>1.882495</td>
<td>0.600901</td>
<td>3.132788</td>
<td>1.731547e-03</td>
<td>9.439403e-03</td>
</tr>
<tr>
<td>col0</td>
<td>124.16093</td>
<td>-1.166974</td>
<td>0.207447</td>
<td>-5.625409</td>
<td>1.850691e-08</td>
<td>3.152784e-07</td>
</tr>
</tbody>
</table>

# write results to file

> write.table( select, file="c:/users/mgribsko/Desktop/hpc/deseq.selected", row.names=T, col.names=T, sep=\"\t\"")