Differential Gene Expression

Workshop on RNA-Seq

Julie Lorent | 15-Mar-2024
NBIS, SciLifeLab

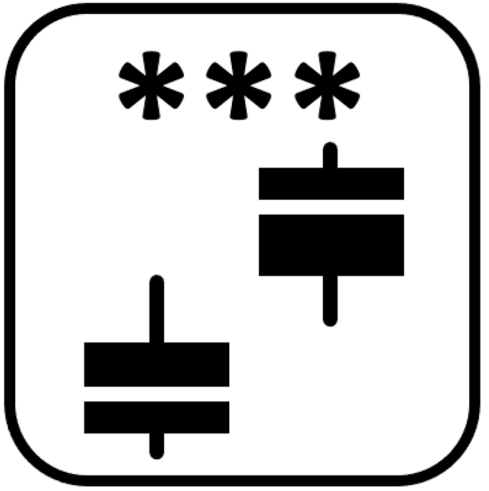
What I'll talk about in this lecture

- Compare gene expression between 2 groups of samples
    - while accounting for differences in sequencing depth
    - by testing for differences in means with appropriate estimation of count data variability
- Calculate log fold changes
- Step by step description of DESeq2 analysis

What we will discuss in the next lecture

- More information on count data distributions
- Batch effects
- Advanced designs

Couldn’t we just use a Student’s t test for each gene?



- May have few replicates -> "borrow" information from other genes
- Multiple testing issues -> Correction
- Distribution is not normal -> negative binomial methods

Preparation

- DEseq2 (and edgeR) take as input raw counts and metadata.
- Create the DESeq2 object

```
library(DESeq2)
mr$Group <- factor(mr$Group)
d <- DESeqDataSetFromMatrix(countData=cf,colData=mr,design=~Group)
d
```

```
## class: DESeqDataSet
## dim: 10573 6
## metadata(1): version
## assays(1): counts
## rownames(10573): ENSMUSG00000098104 ENSMUSG00000033845 ...
##   ENSMUSG00000063897 ENSMUSG00000095742
## rowData names(0):
## colnames(6): DSSd00_1 DSSd00_2 ... DSSd07_2 DSSd07_3
## colData names(7): SampleName SampleID ... Group Replicate
```

- Categorical variables must be factors
- Building GLM models: `~var`, `~covar+var`

Size factors

- Objective of the differential gene expression: compare concentration of cDNA fragments from each gene between conditions/samples.
- Data we have: read counts which depend on these concentration, but also on sequencing depth
- Total count can be influenced by a few highly variable genes
- For this reason, DESeq2 uses size factors (median-of-ratios) instead of total count as normalization factors to account for differences in sequencing depth
- Normalisation factors are computed as follows:

```
d <- DESeq2::estimateSizeFactors(d,type="ratio")
sizeFactors(d)
```

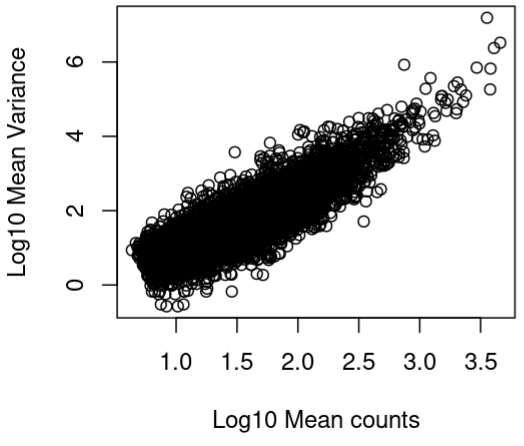
```
##  DSSd00_1  DSSd00_2  DSSd00_3  DSSd07_1  DSSd07_2  DSSd07_3
## 1.0136617 0.9570561 0.9965245 1.0354178 1.0780855 1.0017753
```

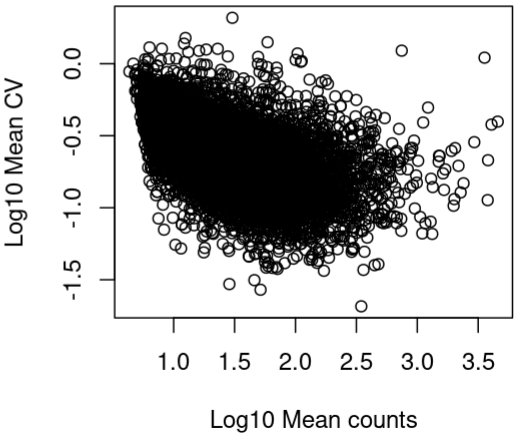
Negative binomial distribution

- RNAseq data is not normally distributed neither as raw counts nor using simple transformations
- DESeq2 and edgeR instead assume negative binomial distributions.
- Given this assumption, to test for differential expression, one need to get a good estimate of the dispersion (variability given the mean).

Dispersion

- Dispersion is a measure of spread or variability in the data.
- Variance is a classical measure of dispersion which is usually not used for negative binomial distributions because of its relationship to the mean
- The DESeq2 dispersion approximates the coefficient of variation for genes with moderate to high count values and is corrected for genes with low count values

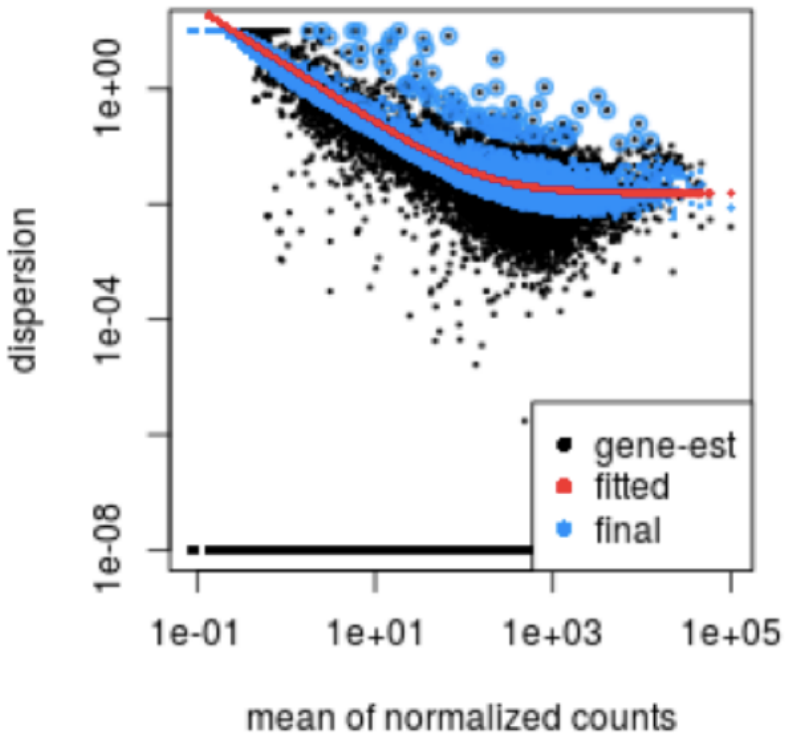


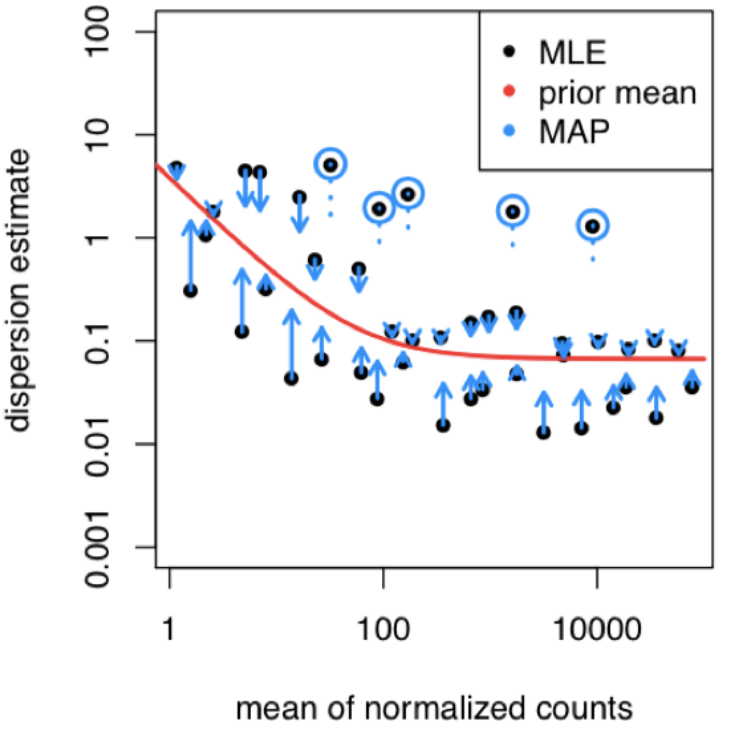


Dispersion

- RNAseq experiments typically have few replicates
- To improve the dispersion estimation in this case, we "borrow" information from other genes with similar mean values

```
d <- DESeq2::estimateDispersions(d)
```





Testing

- Log2 fold changes changes are computed after GLM fitting `FC = counts group B / counts group A`

```
dg <- nbinomWaldTest(d)
resultsNames(dg)
```

```
## [1] "Intercept"           "Group_day07_vs_day00"
```

- Use `results()` to customise/return results
  - Set coefficients using `contrast` or `name`
  - Filtering results by fold change using `lfcThreshold`
  - `cooksCutoff` removes outliers
  - `independentFiltering` removes low count genes
  - `pAdjustMethod` sets method for multiple testing correction
  - `alpha` set the significance threshold

Testing

```
res <- results(dg,name="Group_day07_vs_day00",alpha=0.05)
summary(res)
```

```
##
## out of 10573 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 193, 1.8%
## LFC < 0 (down)     : 238, 2.3%
## outliers [1]       : 1, 0.0095%
## low counts [2]     : 4920, 47%
## (mean count < 21)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

- Alternative way to specify contrast

```
results(dg,contrast=c("Group","day07","day00"),alpha=0.05)
```

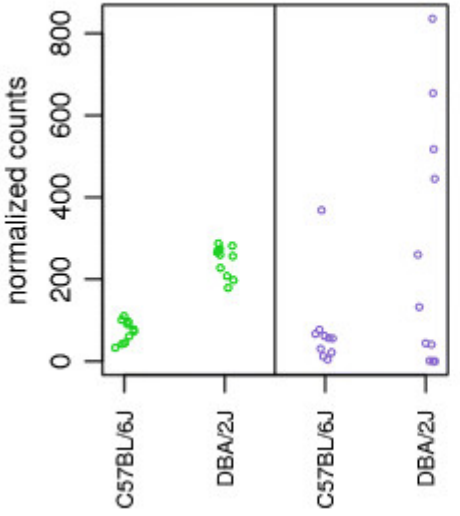
Testing

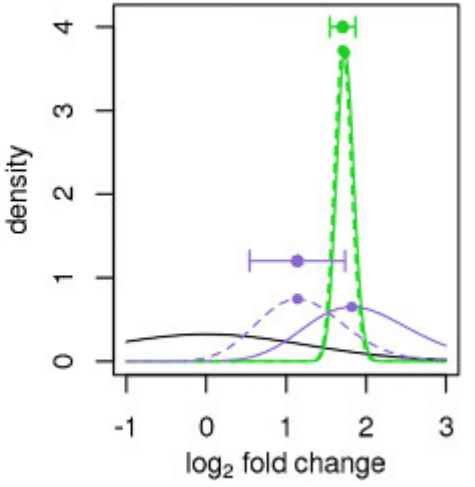
```
head(res)
```

```
## log2 fold change (MLE): Group day07 vs day00
## Wald test p-value: Group day07 vs day00
## DataFrame with 6 rows and 6 columns
##                     baseMean log2FoldChange     lfcSE      stat    pvalue
##                    <numeric>      <numeric> <numeric> <numeric> <numeric>
## ENSMUSG00000098104   18.8505       0.205656  0.401543  0.512164 0.6085362
## ENSMUSG00000033845   23.3333       0.653565  0.379627  1.721596 0.0851426
## ENSMUSG00000025903   37.1016       0.672348  0.298923  2.249232 0.0244977
## ENSMUSG00000033793   33.3673       0.144833  0.305139  0.474646 0.6350394
## ENSMUSG00000025907   22.3875       0.821006  0.376414  2.181125 0.0291742
## ENSMUSG00000051285   21.1485       0.452451  0.378725  1.194669 0.2322163
##                         padj
##                    <numeric>
## ENSMUSG00000098104        NA
## ENSMUSG00000033845  0.377432
## ENSMUSG00000025903  0.177491
## ENSMUSG00000033793  0.886264
## ENSMUSG00000025907  0.201741
## ENSMUSG00000051285        NA
```

Testing

- Use `lfcShrink()` to correct fold changes for genes with high dispersion or low counts
- Does not change number of DE genes





Acknowledgements

- RNA-seq analysis Bioconductor vignette
- DGE Workshop by HBC training

Thank you. Questions?

Graphics from freepik.com
Created: 15-Mar-2024 • SciLifeLab • NBIS