# Differential Gene Expression

## Workshop on RNA-Seq

**Roy Francis** | 26-Nov-2020
NBIS, SciLifeLab

# Preparation

- Create the DESeq2 object

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

```
## class: DESeqDataSet
## dim: 17515 6
## metadata(1): version
## assays(1): counts
## rownames(17515): 0610010F05Rik 0610010K14Rik ... Zyg11a Zzz3
## 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

- Normalisation factors are computed

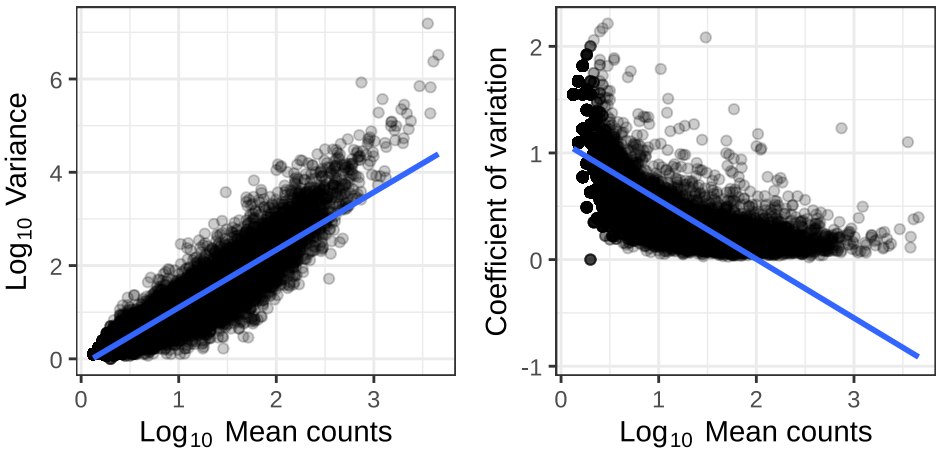
```
d <- DESeq2::estimateSizeFactors(d,type="ratio")
sizeFactors(d)
```

```
##  DSSd00_1  DSSd00_2  DSSd00_3  DSSd07_1  DSSd07_2  DSSd07_3
## 1.0153287 0.9597101 0.9984645 1.0358161 1.0787996 0.9988740
```

# Dispersion

- We need to measure the variability of gene counts

```
dm <- apply(cf,1,mean)
dv <- apply(cf,1,var)
cva <- function(x) sd(x)/mean(x)
dc <- apply(cf,1,cva)
```

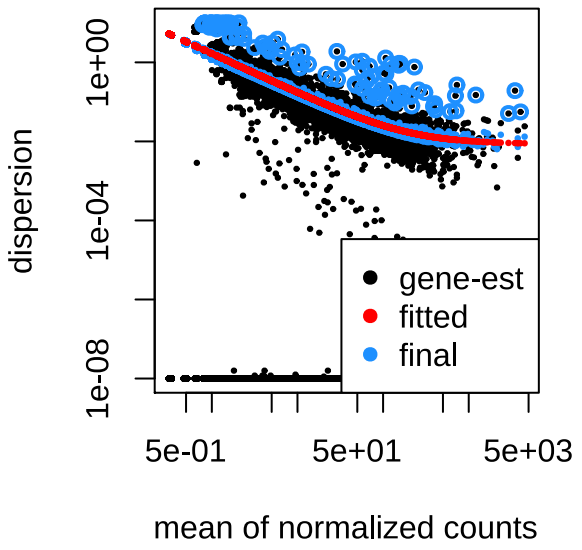



- Dispersion is a measure of variability in gene expression for a given mean

# Dispersion

- Dispersion is unreliable for low mean counts
- Genes with similar mean values must have similar dispersion
- Estimate likely (ML) dispersion for each gene based on counts
- Fit a curve through the gene-wise estimates
- Shrink dispersion towards the curve

```
d <- DESeq2::estimateDispersions(d)
{par(mar=c(4,4,1,1))
plotDispEsts(d)}
```



# Testing

- Log2 fold changes changes are computed after GLM fitting

```
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

```
res1 <- results(dg,name="Group_day07_vs_day00",alpha=0.05)
summary(res1)
```

```
##
## out of 17515 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 194, 1.1%
## LFC < 0 (down)     : 217, 1.2%
## outliers [1]       : 1, 0.0057%
## low counts [2]     : 9169, 52%
## (mean count < 10)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

# Testing

```
head(res1)
```

```
## 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>
## 0610010F05Rik  39.80868       0.324551  0.301846  1.075222 2.82275e-01
## 0610010K14Rik 438.41972      -0.768534  0.171606 -4.478485 7.51746e-06
## 0610033M10Rik 335.66354       0.393472  0.137554  2.860493 4.22982e-03
## 0610040F04Rik  37.06121      -0.922058  0.334462 -2.756844 5.83622e-03
## 0610040J01Rik  10.80725      -0.243347  0.555218 -0.438291 6.61175e-01
## 1010001B22Rik   1.31554       0.661815  1.601288  0.413302 6.79386e-01
##                      padj
##                 <numeric>
## 0610010F05Rik 0.778682265
## 0610010K14Rik 0.000531637
## 0610033M10Rik 0.075101892
## 0610040F04Rik 0.093122837
## 0610040J01Rik 0.940267958
## 1010001B22Rik          NA
```

- Use `lfcShrink()` to correct fold changes for high dispersion genes

# Thank you. Questions?

R version 4.0.3 (2020-10-10)

Platform: x86_64-pc-linux-gnu (64-bit)

OS: Ubuntu 18.04.5 LTS

---

Built on : 26-Nov-2020 at 16:52:22

**2020** • SciLifeLab • NBIS