

Differential Gene Expression

Workshop on RNA-Seq

Roy Francis | 25-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: 17529 6
## metadata(1): version
## assays(1): counts
## rownames(17529): ENSMUSG000000051951 ENSMUSG000000025902 ...
## ENSMUSG000000063897 ENSMUSG000000095742
## 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

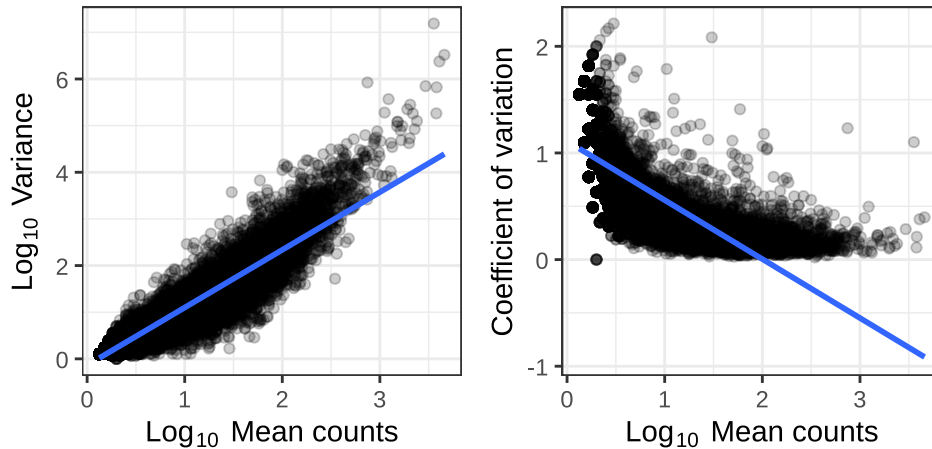
```
d <- DESeq2::estimateSizeFactors(d,type="ratio")
sizeFactors(d)
```

```
## DSSd00_1 DSSd00_2 DSSd00_3 DSSd07_1 DSSd07_2 DSSd07_3
## 1.0153287 0.9597123 0.9984645 1.0358322 1.0787888 0.9988912
```

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)
```

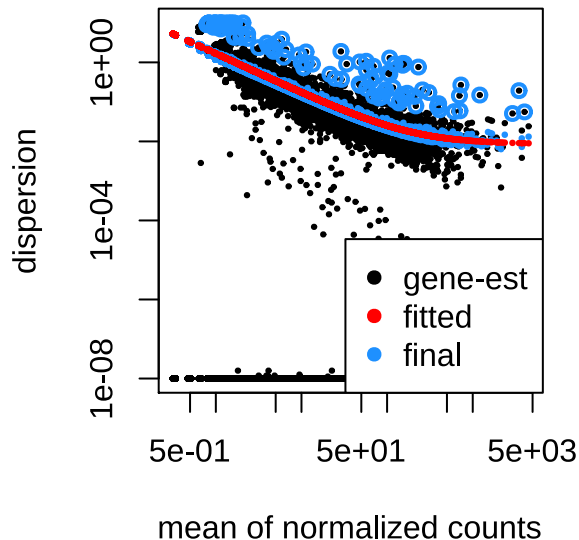


- Dispersion is a measure of variability in gene expression for a given mean

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 17529 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]    : 9176, 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>
## ENSMUSG000000051951  1.304695      0.658757  1.658587  0.397180  0.691235
## ENSMUSG000000025902  1.492510     -1.080732  1.521456 -0.710327  0.477501
## ENSMUSG000000102269  0.327753      1.813628  3.064237  0.591869  0.553938
## ENSMUSG000000098104 18.839038      0.209551  0.425672  0.492284  0.622519
## ENSMUSG000000103922  2.649333     -1.200658  1.135654 -1.057240  0.290402
## ENSMUSG000000033845 23.324913      0.657772  0.401765  1.637204  0.101588
##           padj
##           <numeric>
## ENSMUSG000000051951      NA
## ENSMUSG000000025902      NA
## ENSMUSG000000102269      NA
## ENSMUSG000000098104  0.930371
## ENSMUSG000000103922      NA
## ENSMUSG000000033845  0.529152
```

- 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: 📅 25-Nov-2020 at 🕒 08:49:38

2020 • SciLifeLab • NBIS