Differential Gene Expression

Workshop on RNA-Seq

NBIS | 23-Nov-2020

NBIS, SciLifeLab

Preparation

- Create the DESeq2 object

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

```
## class: DESeqDataSet
## dim: 14578 12
## metadata(1): version
## assays(1): counts
## rownames(14578): ENSG00000000003 ENSG00000000419 ... ENSG00000266865
##   ENSG00000266876
## rowData names(0):
## colnames(12): Sample_1 Sample_2 ... Sample_11 Sample_12
## colData names(5): Sample_ID Sample_Name Time Replicate Cell
```

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

Size factors

- Normalisation factors are computed

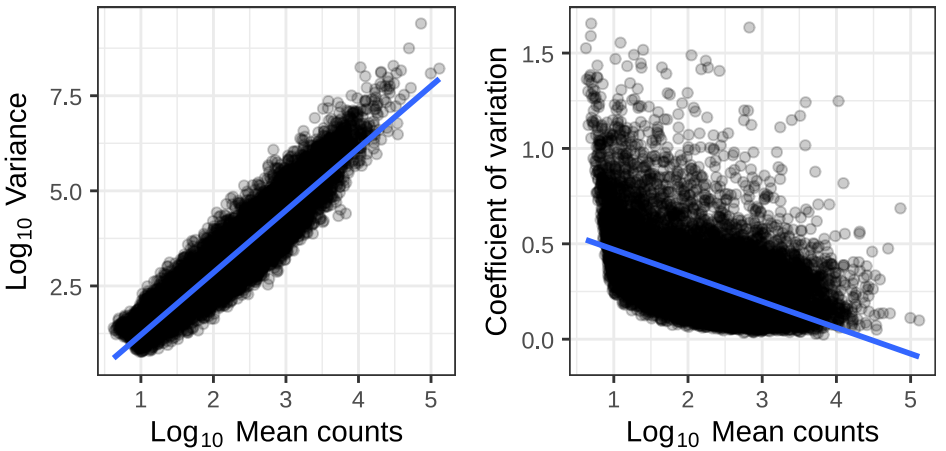
```
d <- DESeq2::estimateSizeFactors(d,type="ratio")
sizeFactors(d)
```

```
##  Sample_1  Sample_2  Sample_3  Sample_4  Sample_5  Sample_6  Sample_7  Sample_8
## 0.9003753 0.8437393 0.5106445 1.1276451 1.0941383 0.8133849 0.7553903 1.1744008
##  Sample_9 Sample_10 Sample_11 Sample_12
## 1.0189325 1.3642797 1.2325485 1.8555904
```

Dispersion

- We need to measure the variability of gene counts

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

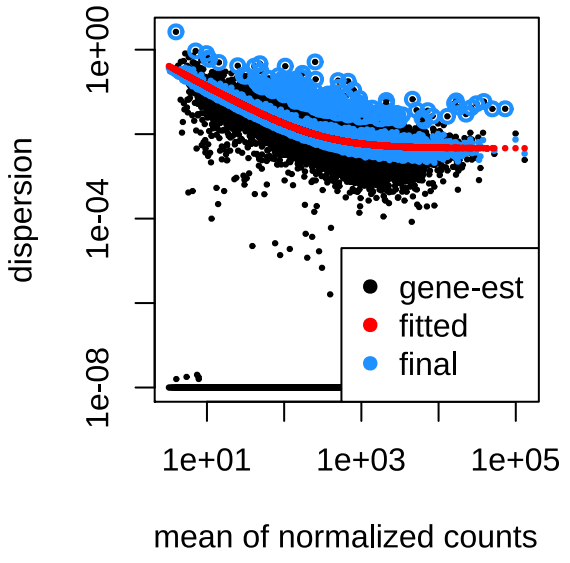


- Dispersion is a measure of variability in gene expression for a given mean

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"      "Time_t2_vs_t0"  "Time_t24_vs_t0" "Time_t6_vs_t0"
```

- 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="Time_t2_vs_t0",alpha=0.05)
summary(res1)
```

```
##
## out of 14578 with nonzero total read count
## adjusted p-value < 0.05
## LFC > 0 (up)       : 413, 2.8%
## LFC < 0 (down)     : 696, 4.8%
## outliers [1]       : 0, 0%
## low counts [2]     : 2261, 16%
## (mean count < 26)
## [1] see 'cooksCutoff' argument of ?results
## [2] see 'independentFiltering' argument of ?results
```

Testing

```
head(res1)
```

```
## log2 fold change (MLE): Time t2 vs t0
## Wald test p-value: Time t2 vs t0
## DataFrame with 6 rows and 6 columns
##                  baseMean log2FoldChange     lfcSE      stat    pvalue
##                 <numeric>      <numeric> <numeric> <numeric> <numeric>
## ENSG00000000003  490.0172      0.2206198 0.1127611  1.956524 0.0504034
## ENSG00000000419  817.7807      0.0592720 0.1014813  0.584068 0.5591746
## ENSG00000000457   82.0788      0.2077486 0.2204049  0.942577 0.3458972
## ENSG00000000460  356.0716     -0.1291864 0.1151392 -1.122002 0.2618616
## ENSG00000001036  919.6068      0.0288827 0.0851501  0.339198 0.7344609
## ENSG00000001084  529.5940      0.2119648 0.0929811  2.279655 0.0226281
##                      padj
##                 <numeric>
## ENSG00000000003  0.263505
## ENSG00000000419  0.830262
## ENSG00000000457  0.689946
## ENSG00000000460  0.612625
## ENSG00000001036  0.909639
## ENSG00000001084  0.159263
```

- 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 : 23-Nov-2020 at 09:29:15