

# Data Preprocessing

---

Workshop on RNA-Seq

Roy Francis | 27-Nov-2020

NBIS, SciLifeLab

# Raw data

- Raw count table

```
##          DSSd00_1 DSSd00_2 DSSd00_3 DSSd07_1 DSSd07_2 DSSd07_3
## ENSMUSG00000102693      0      0      0      0      0      0
## ENSMUSG00000064842      0      0      0      0      0      0
## ENSMUSG00000051951      0      1      2      0      3      2
## ENSMUSG00000102851      0      0      0      0      0      0
## ENSMUSG00000103377      0      0      0      0      0      0
## ENSMUSG00000104017      0      0      0      0      0      0
```

- Metadata

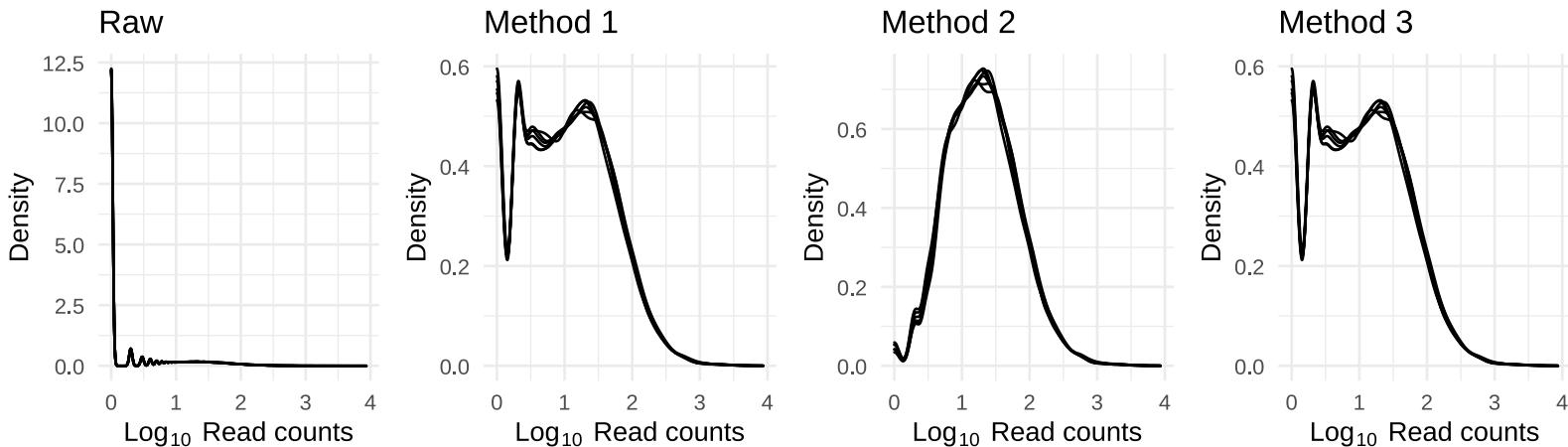
```
##   SampleName   SampleID No Model Day Group Replicate
## DSSd00_1   DSSd00_1 KI_PC1606_01  1  DSS  0 day00        1
## DSSd00_2   DSSd00_2 KI_PC1606_02  2  DSS  0 day00        2
## DSSd00_3   DSSd00_3 KI_PC1606_03  3  DSS  0 day00        3
## DSSd07_1   DSSd07_1 KI_PC1606_13 13  DSS  7 day07        1
## DSSd07_2   DSSd07_2 KI_PC1606_14 14  DSS  7 day07        2
## DSSd07_3   DSSd07_3 KI_PC1606_15 15  DSS  7 day07        3
```

# Preprocessing

- Remove genes and samples with low counts

```
cf1 <- cr[rowSums(cr>0) >= 2, ]  
cf2 <- cr[rowSums(cr>3) >= 2, ]  
cf3 <- cr[rowSums(edgeR::cpm(cr)>1) >= 2, ]
```

- Inspect distribution

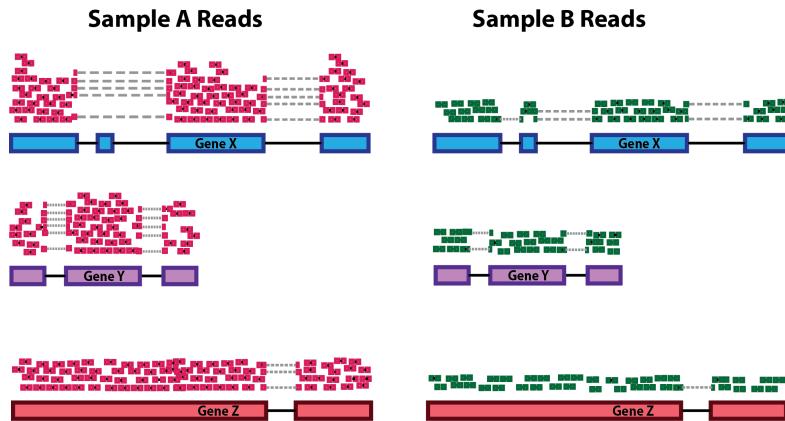


- Inspect the number of rows (genes)

```
## Raw: 55487, Method 1: 17529, Method 2: 12481, Method 3: 17529
```

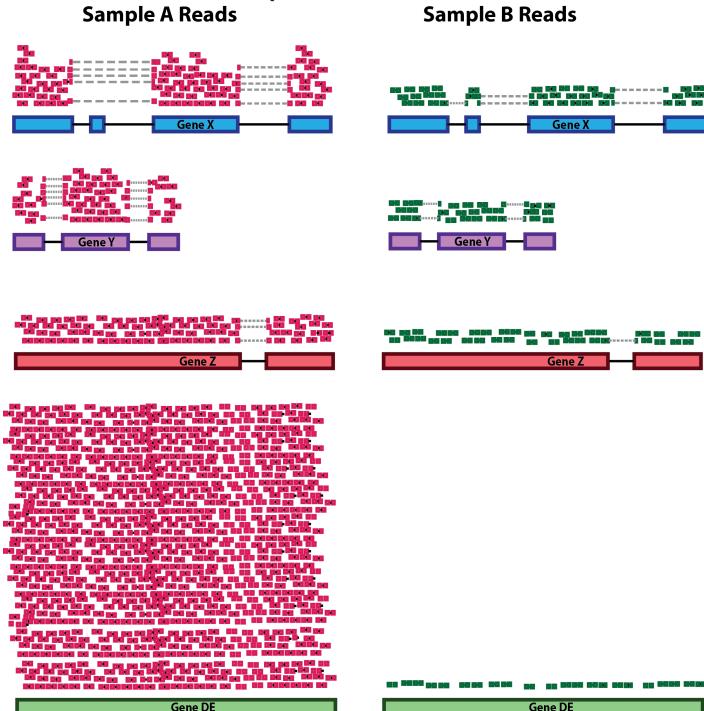
# Normalisation

- Make counts comparable across samples
- Control for sequencing depth



```
##      A B A_tc B_tc
## x 20 6 Inf Inf
## y 25 6 Inf Inf
## z 15 4 Inf Inf
```

- Control for compositional bias

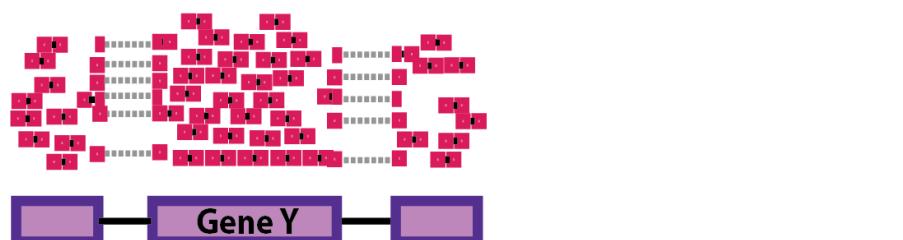
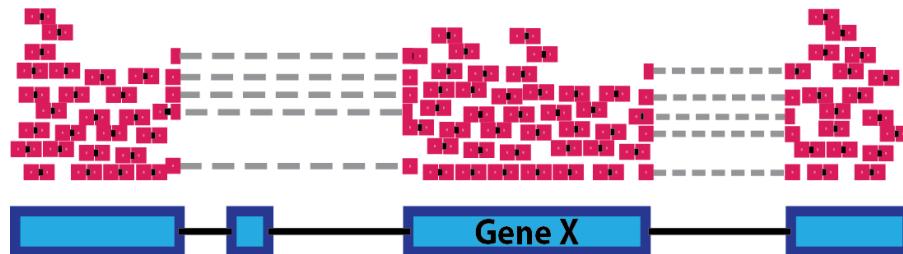


```
##      A B A_tc B_tc
## x 0 20 NaN Inf
## y 25 25 Inf Inf
## z 15 4 Inf Inf
## de 100 2 Inf Inf
```

# Normalisation

- Make counts comparable across features (genes)

## Sample A Reads



```
##   counts gene_length norm_counts
## x      50          10        5
## y      25           5        5
```

- Bring counts to a human-friendly scale

# Normalisation

## Normalisation by library size

- Assumes total expression is the same under different experimental conditions
- Methods include TC, RPKM, FPKM, TPM
- RPKM, FPKM and TPM control for sequencing depth and gene length
- TPM enables better comparison between samples and between experiments

## Normalisation by distribution

- Assumes technical effects are same for DE and non-DE genes
- Assumes number of over and under-expressed genes are roughly same across conditions
- Corrects for compositional bias
- Methods include Q, UQ, M, RLE, TMM, MRN
- `edgeR::calcNormFactors()` implements TMM, TMMwzp, RLE & UQ
- `DESeq2::estimateSizeFactors()` implements median ratio method (RLE)
- Does not correct for gene length
- `geTMM` is gene length corrected TMM

# Normalisation

## Normalisation by testing

- A more robust version of normalisation by distribution.
- A set of non-DE genes are detected through hypothesis testing
- Tolerates a larger difference in number of over and under expressed genes between conditions
- Methods include PoissonSeq, DEGES

## Normalisation using Controls

- Assumes controls are not affected by experimental condition and technical effects are similar to all other genes
- Useful in conditions with global shift in expression
- Controls could be house-keeping genes or spike-ins
- Methods include RUV, CLS

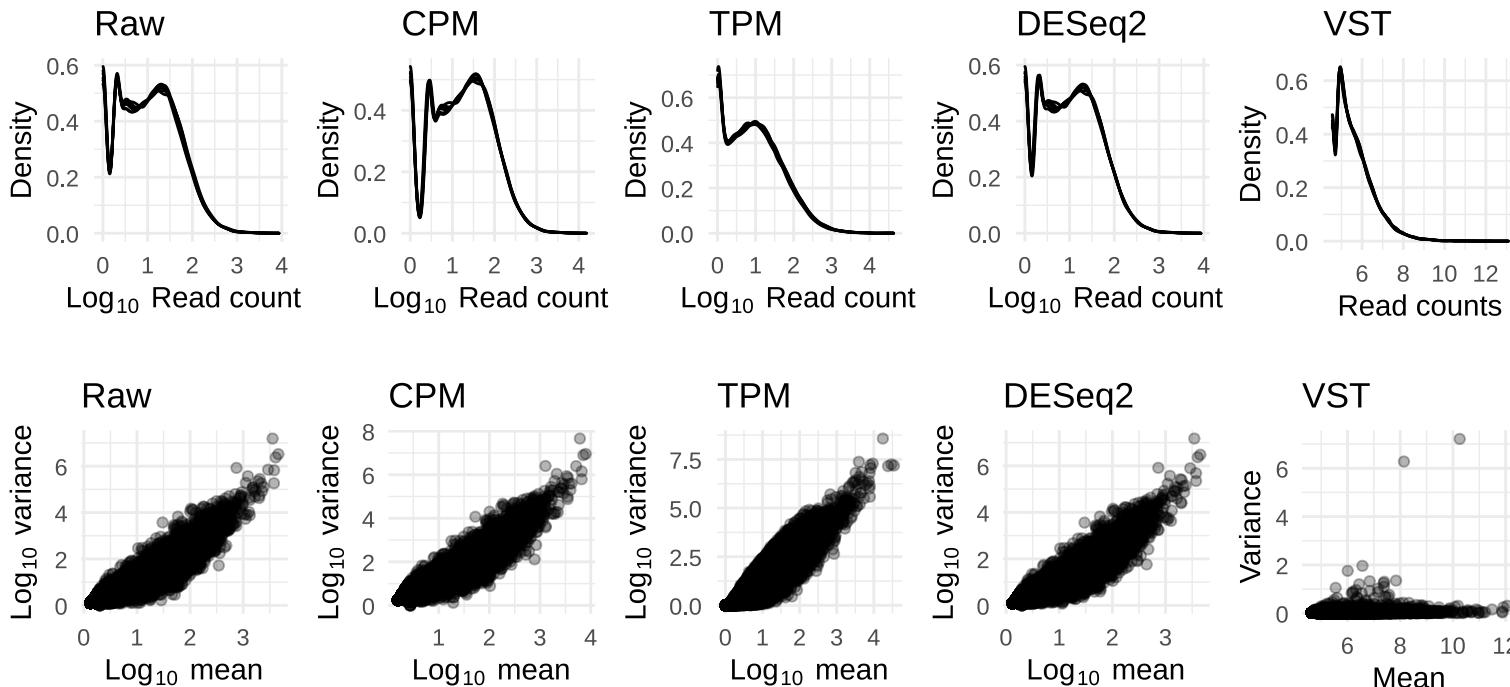
## Stabilizing variance

- Variance is stabilised across the range of mean values
- Methods include VST, RLOG, VOOM
- For use in exploratory analyses. Not for DE.
- `vst()` and `rlog()` functions from *DESeq2*
- `voom()` function from *Limma* converts data to normal distribution

# Normalisation

## Recommendations

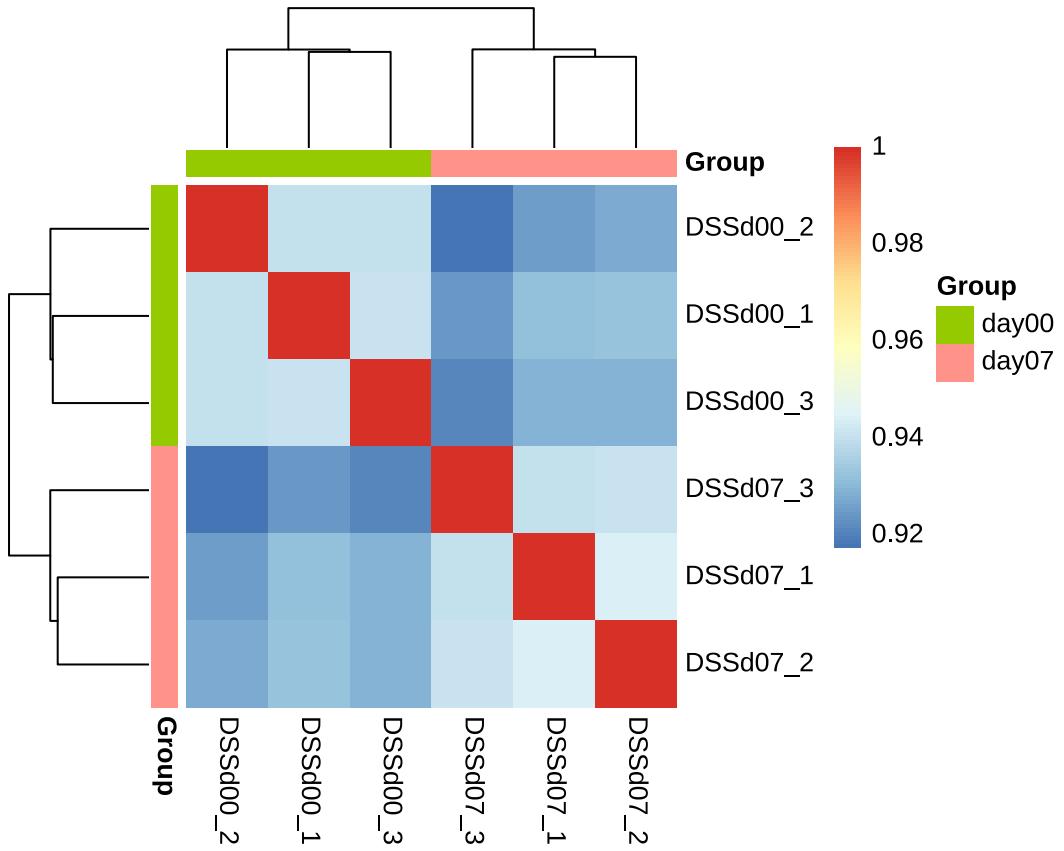
- Most tools use a mix of many different normalisations
- For DGE using DGE R packages (DESeq2, edgeR, Limma etc), use raw counts
- For visualisation (PCA, clustering, heatmaps etc), use VST or RLOG
- For own analysis with gene length correction, use TPM (maybe geTMM?)
- Custom solutions: spike-ins/house-keeping genes



# Exploratory | Correlation

- Correlation between samples

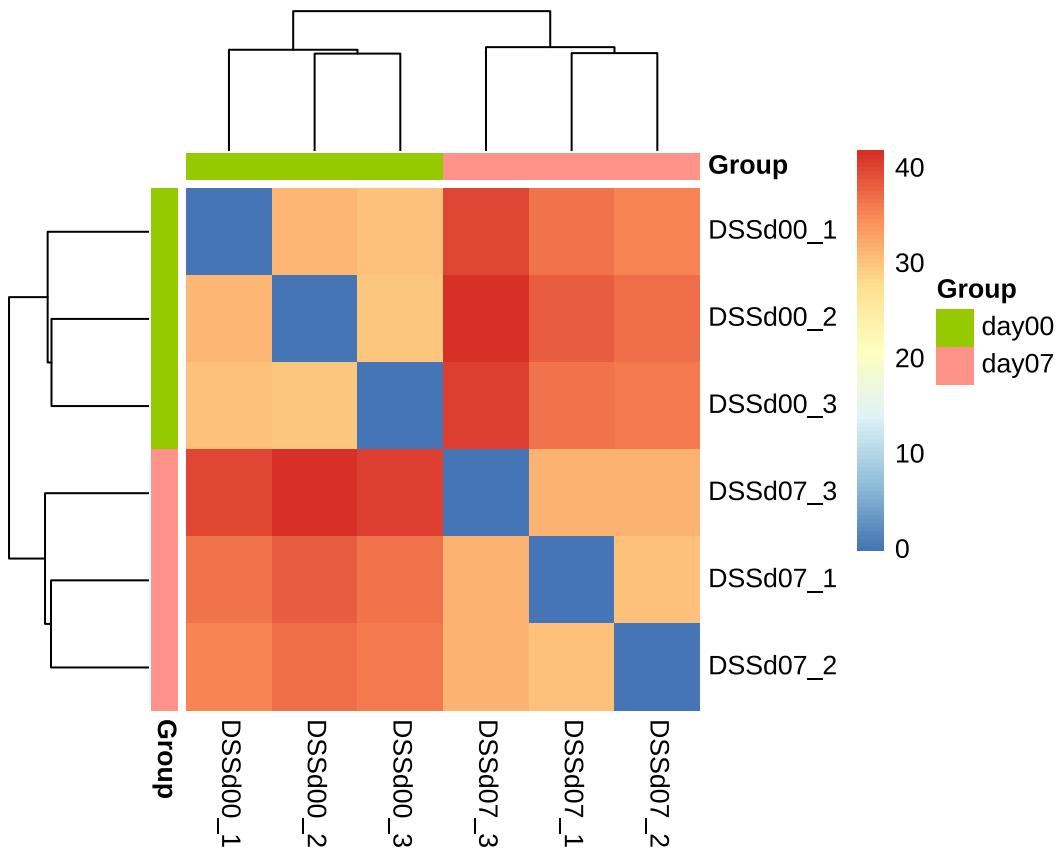
```
dmat <- as.matrix(cor(cv,method="spearman"))
pheatmap::pheatmap(dmat,border_color=NA,annotation_col=mr[, "Group",drop=F],
                   annotation_row=mr[, "Group",drop=F],annotation_legend=T)
```



# Exploratory | Distance

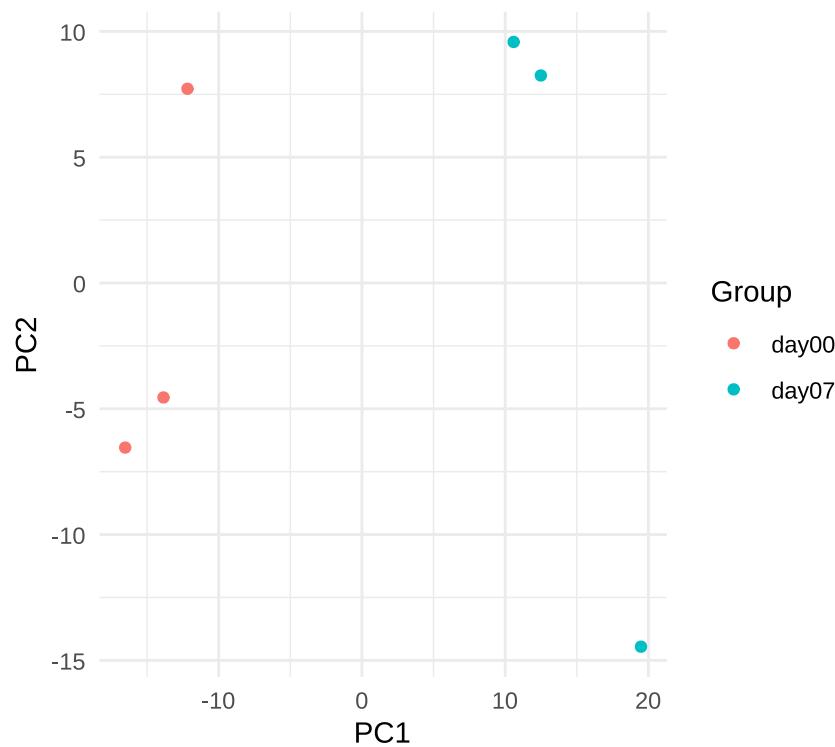
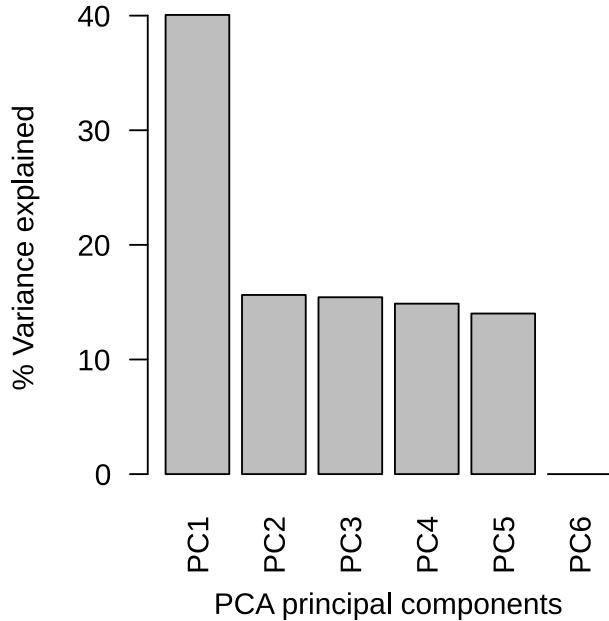
- Similarity between samples

```
dmat <- as.matrix(dist(t(cv)))
pheatmap(dmat,border_color=NA,annotation_col=mr[,"Group",drop=F],
          annotation_row=mr[,,"Group",drop=F],annotation_legend=T)
```



# Exploratory | PCA

- Relationship between samples



# 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 : 📅 27-Nov-2020 at ⏱ 13:40:04

2020 • SciLifeLab • NBIS