Data Preprocessing

Workshop on RNA-Seq

NBIS | 23-Nov-2020

NBIS, SciLifeLab

Raw data

- Raw count table

```
##                 Sample_1 Sample_2 Sample_3 Sample_4 Sample_5 Sample_6
## ENSG00000000003      321      303      204      492      455      359
## ENSG00000000005        0        0        0        0        0        0
## ENSG00000000419      696      660      472      951      963      689
## ENSG00000000457       59       54       44      109       73       66
## ENSG00000000460      399      405      236      445      454      374
## ENSG00000000938        0        0        0        0        0        1
```

- Metadata

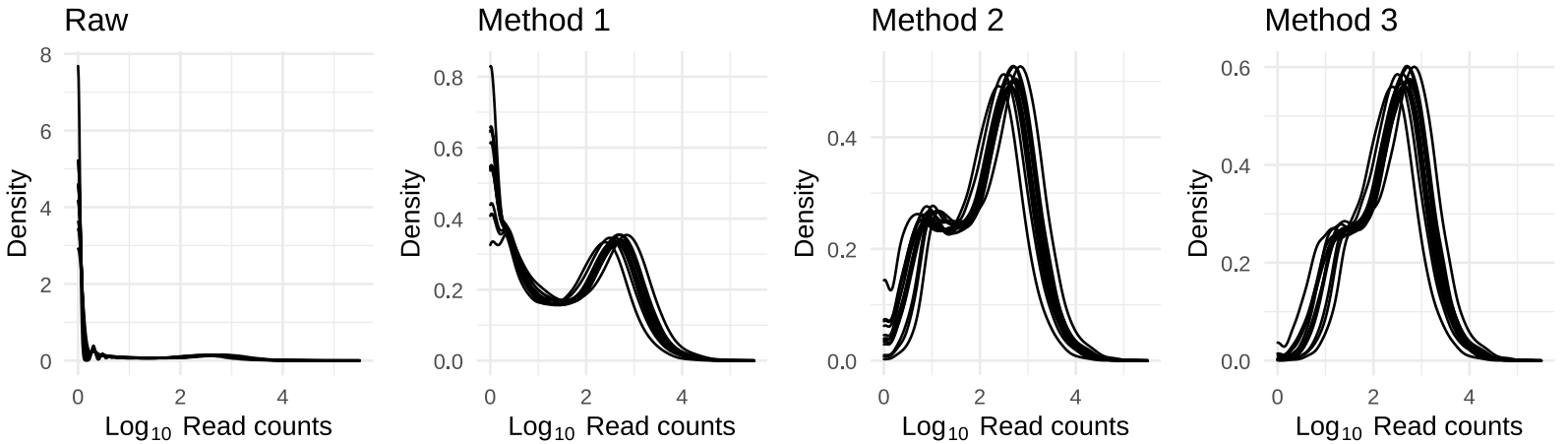
```
##          Sample_ID Sample_Name Time Replicate Cell
## Sample_1  Sample_1        t0_A   t0         A A431
## Sample_2  Sample_2        t0_B   t0         B A431
## Sample_3  Sample_3        t0_C   t0         C A431
## Sample_4  Sample_4        t2_A   t2         A A431
## Sample_5  Sample_5        t2_B   t2         B A431
## Sample_6  Sample_6        t2_C   t2         C A431
```

Preprocessing

- Remove genes and samples with low counts

```
cf1 <- cr[rowSums(cr>0) >= 2, ]
cf2 <- cr[rowSums(cr>5) >= 2, ]
cf3 <- cr[rowSums(edgeR::cpm(cr)>1) >= 2, ]
```

- Inspect distribution

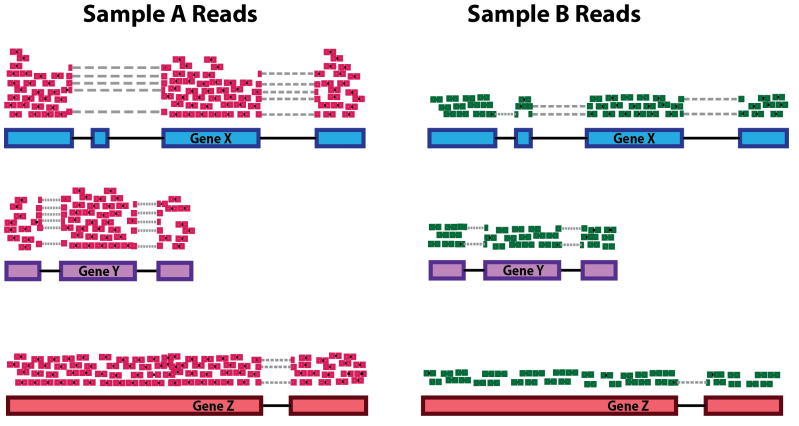


- Inspect the number of rows (genes)

```
## Raw: 59573, Method 1: 24194, Method 2: 16519, Method 3: 14578
```

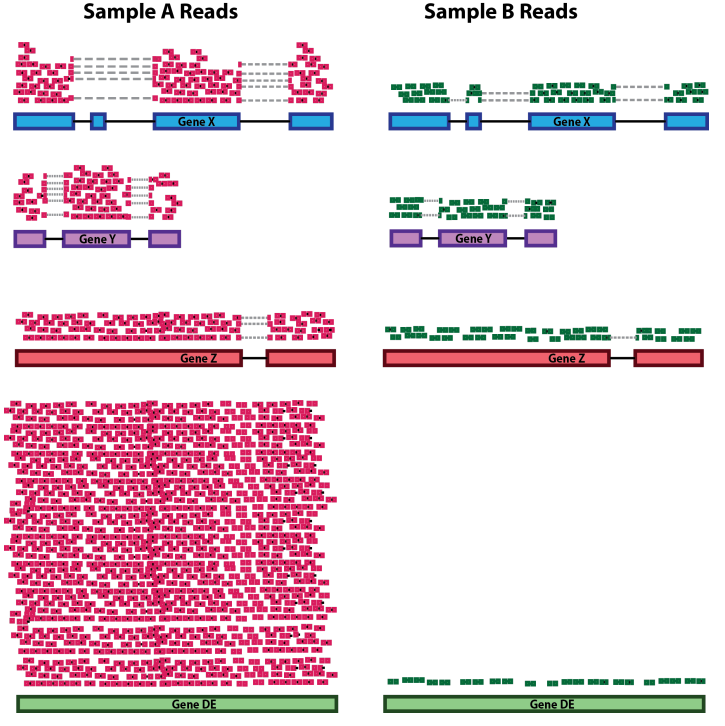
Normalisation

- 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
```

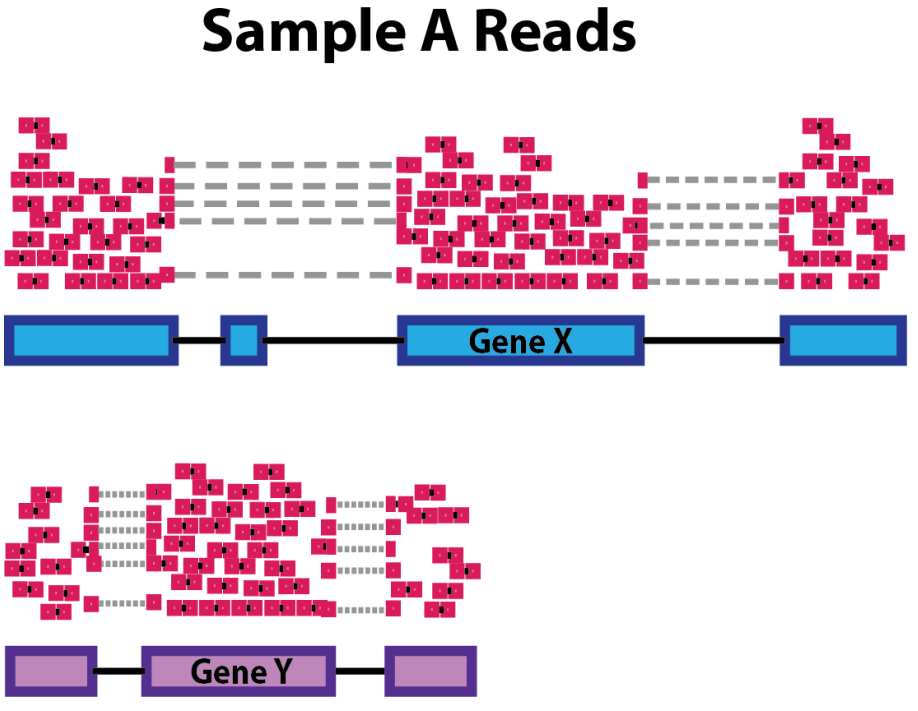
- Control for compositional bias




```
##      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)




```
##   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

Evans, Ciaran, Johanna Hardin, and Daniel M. Stoebel. "Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions." Briefings in bioinformatics (2017)

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

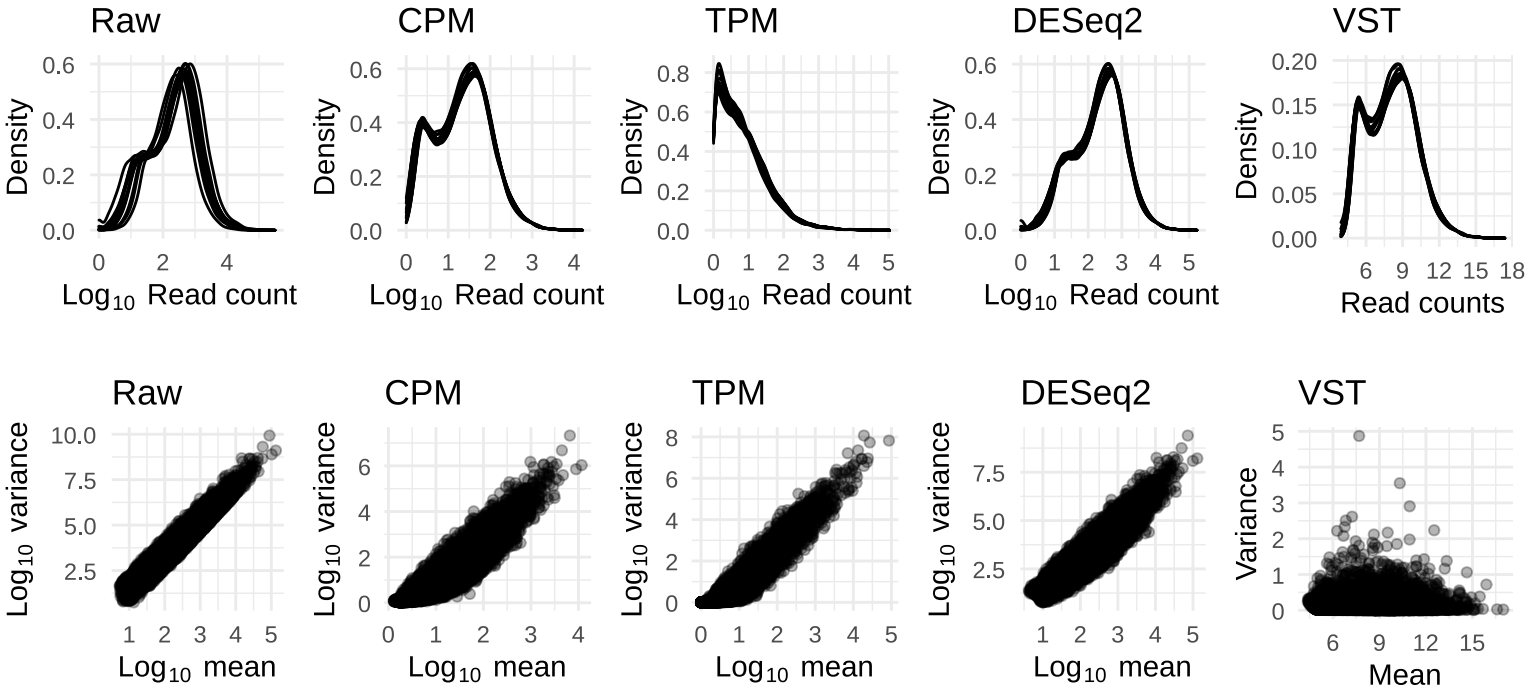
Stabilizing variance

- Variance is stabilised across the range of mean values
- Mwthods 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

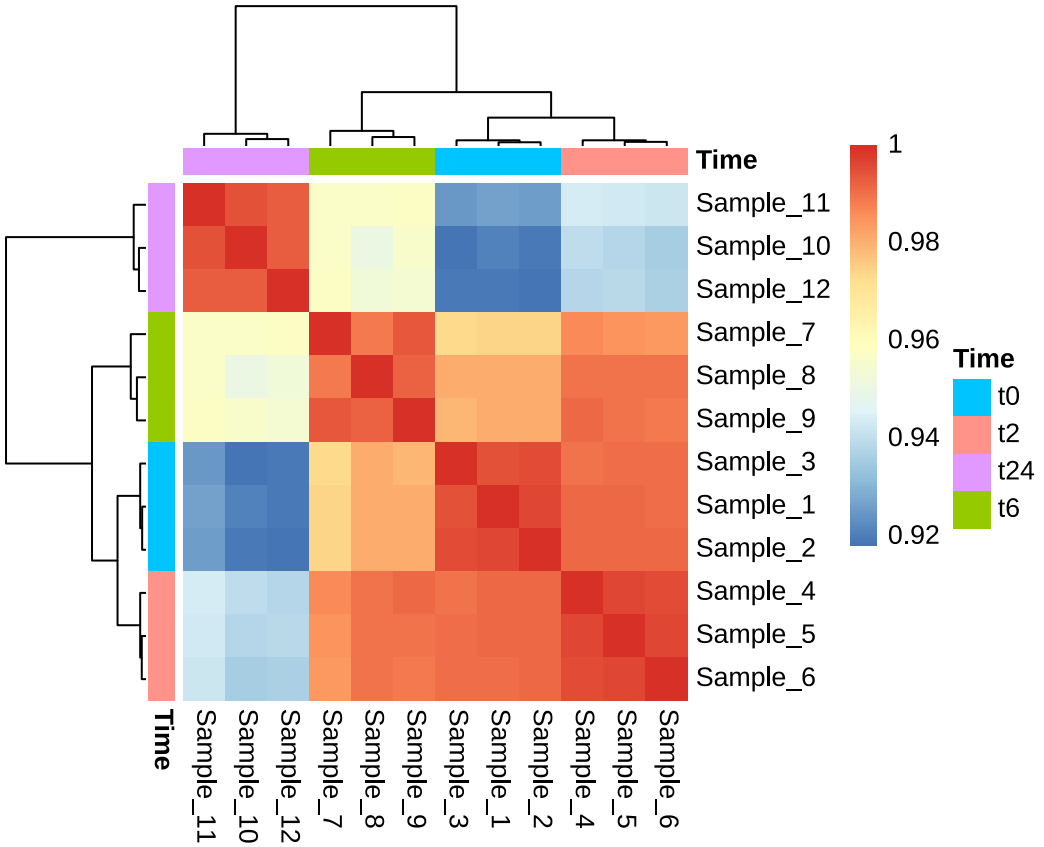


Dillies, Marie-Agnes, *et al.* "A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis." Briefings in bioinformatics 14.6 (2013): 671-683

Exploratory | Correlation

- Correlation between samples

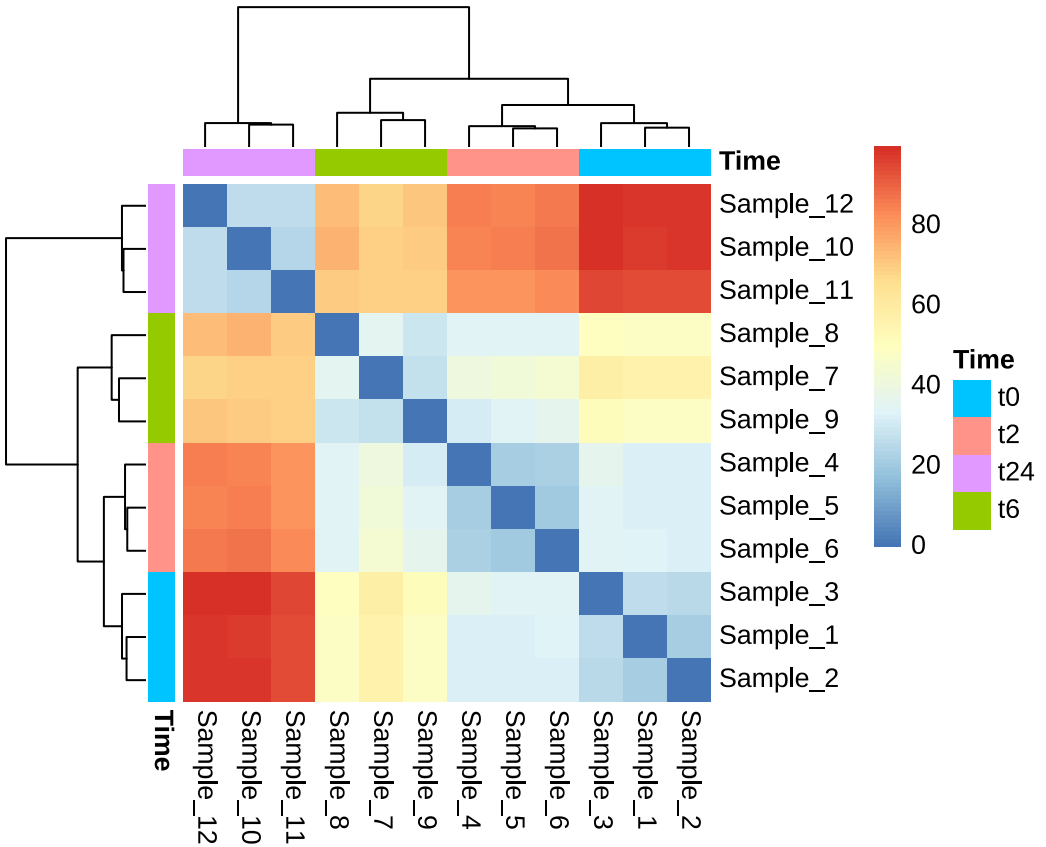
```
dmat <- as.matrix(cor(cv,method="spearman"))
pheatmap::pheatmap(dmat,border_color=NA,annotation_col=mr[,"Time",drop=F],
         annotation_row=mr[,"Time",drop=F],annotation_legend=T)
```



Exploratory | Distance

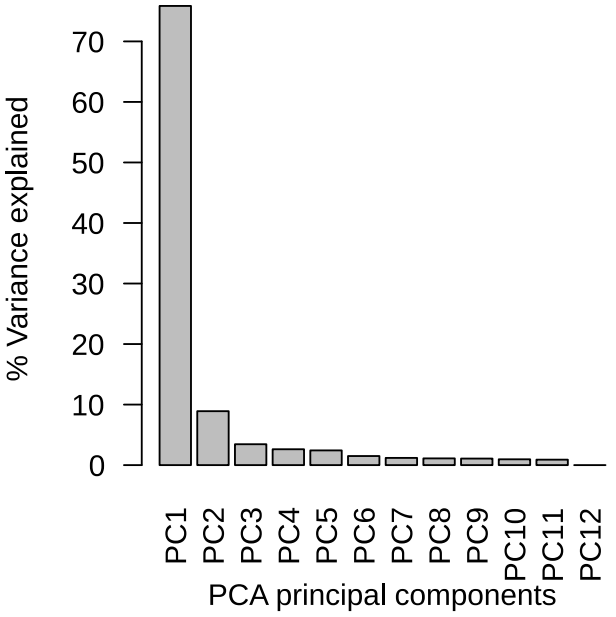
- Similarity between samples

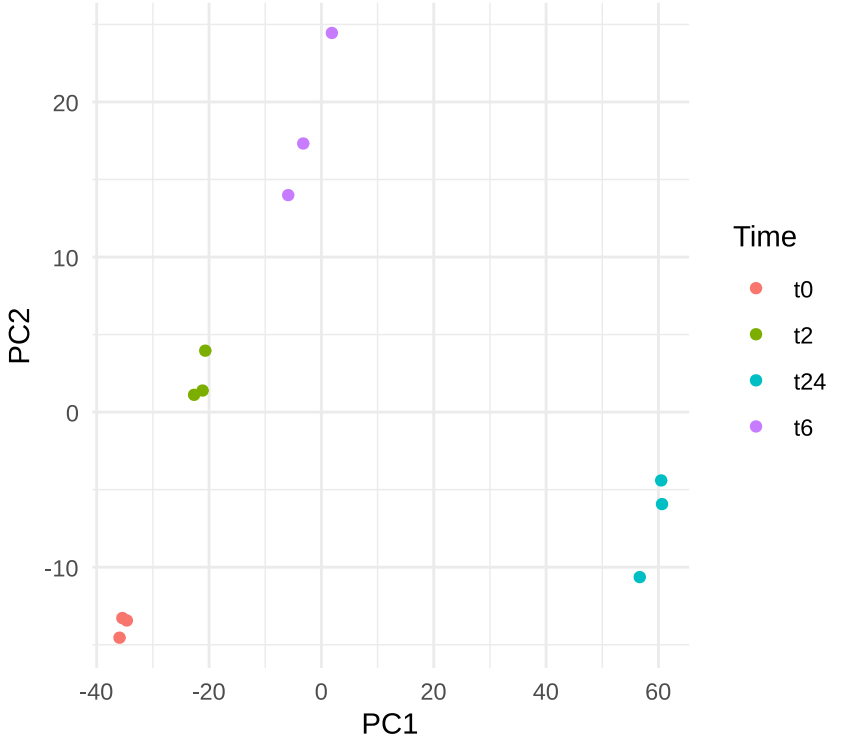
```
dmat <- as.matrix(dist(t(cv)))
pheatmap(dmat,border_color=NA,annotation_col=mr[,"Time",drop=F],
         annotation_row=mr[,"Time",drop=F],annotation_legend=T)
```



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 : 23-Nov-2020 at 09:29:53