# Normalisation of RNA seq data

Normalization is an important step while analysing RNA seq data, as it is used to make the data comparable within or across different samples (Bullard et al. 2010; Risso et al. 2014; Zhao et al. 2021). The main aim of normalization is to minimize the technical bias present in data (Robinson & Oshlack 2010).

## Biases present in RNA seq data

While comparing a gene expression between different samples, biases such as sequencing depth, batch effect exists, whereas, comparing different genes within a sample has inherent bias such as gene length and GC content (Bacher et al. 2017; Conesa et al. 2016). Sequencing depth or total read counts for a given sample, has no any optimal limit but a deep sequencing might improve the quantification and identification of transcripts, however it might also result in the detection of transcriptional noise and off-target transcripts (Conesa et al. 2016). Genes with shorter effective length have lower counts and higher rate of dropping out in comparison with loner genes (Phipson, Zappia & Oshlack 2017). If RNA sample has a small number of highly expressed genes, the reads representing those genes will take away substantial resources from the sample, which decreases the reads for genes causing under sampling, if not explicitly accounted for, this can cause the bias to the detection of differentially expressed. (Robinson & Oshlack 2010). Likewise, the batch effect might appear due to difference on experimental design, different sequencing platform, and sample handling person which might cause non biological variation in data. (Bullard et al. 2010; Leek et al. 2010). (Hansen, Irizarry & WU 2012) has found GC content of a gene also causes biases in a sample specific expression measurement.

## Normalization techniques

### Intra sample normalisation

1. **CPM (Counts per Million mapped reads)**

It is a basic gene expression unit which normalizes sequencing depth and useful for inter-sample comparison. It is implemented in edgeR. (Chen, Lun & Smyth 2016).

CPM =

1. **RPKM/FPKM (Reads/Fragments Ker kilo base of transcript per Million mapped reads)**

RPKM is a gene length normalised expression unit first described by (Mortazavi et al. 2008).

RPKM/FPKM =

Here, 103 normalizes for gene length and 106 for sequencing depth factor.

1. **TPM (Transcripts per million)**

TPM, is proposed as an alternative to RPKM, was first described by (Li et al. 2009).

TPM=A∗ ∗

Where, A =

Here, 106 for sequencing depth factor.

### Inter sample normalization

1. **TMM (Trimmed Mean of M-values)**

TMM, originally defined by (Robinson & Oshlack 2010), inter sample normalization technique. It is an expression unit which is recommended more than half of the genes in a data are believed not to be differentially expressed between any two samples. The TMM normalization method is implemented in the edgeR package. Normalization using the TMM method was performed on count data generated after CPM and filtering of lowly expressed genes are done (Law et al. 2018).

1. **GeTMM (Gene length corrected trimmed mean of M-values)**

This method allows both inter-sample and intra-sample comparison which is similar to TMM (Smid et al. 2018).

1. **DEseq2: median ratio**

The DESeq*2* normalization method like TMM assumes that most of the genes are not differentially expressed. In this method, first **a pseudo-reference sample is created by calculating row-wise geometric mean.** The second step follows the ratio calculation of each sample to the reference and normalization factor is calculated. lastly, count values are normalized using normalization factor (Love, Huber & Anders 2014).

### Batch effect and other artefacts

1. **Noiseq**

Noiseq consists of various functions for creating exploratory plots for various biases and correct them (Tarazona et al. 2012).

1. **RUV (Remove unwanted variation)**

RUV method was proposed to adjust for unknown technical variations by performing factor analysis on negative control genes namely housekeeping genes or spike-in controls, first proposed by (Risso et al. 2014)

1. **SVA (Surrogate Variable Analysis) seq**

SVA seq removes known batch effects and unknown artifacts present in data, In this approach, the unknown technical variations which is also called as surrogate variables are estimated by applying Singular Value Decomposition and artifacts are estimated by selecting singular vectors and technical bias was adjusted accordingly (Leek 2014).

### Log transformation

Log transformation of CPM normalised data is RNA seq pipeline is done for making the exploratory plots (Law et al. 2018).

# References

Bacher, R, Chu, L-F, Leng, N, Gasch, AP, Thomson, JA, Stewart, RM, Newton, M & Kendziorski, C 2017, 'SCnorm: robust normalization of single-cell RNA-seq data', *Nature Methods*, vol. 14, no. 6, pp. 584-586.

Bullard, JH, Purdom, E, Hansen, KD & Dudoit, S 2010, 'Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments', *BMC Bioinformatics*, vol. 11, no. 1, p. 94.

Conesa, A, Madrigal, P, Tarazona, S, Gomez-Cabrero, D, Cervera, A, McPherson, A, Szcześniak, MW, Gaffney, DJ, Elo, LL, Zhang, X & Mortazavi, A 2016, 'A survey of best practices for RNA-seq data analysis', *Genome Biology*, vol. 17, no. 1, p. 13.

Hansen, KD, Irizarry, RA & WU, Z 2012, 'Removing technical variability in RNA-seq data using conditional quantile normalization', *Biostatistics*, vol. 13, no. 2, pp. 204-216.

Leek, JT, Scharpf, RB, Bravo, HC, Simcha, D, Langmead, B, Johnson, WE, Geman, D, Baggerly, K & Irizarry, RA 2010, 'Tackling the widespread and critical impact of batch effects in high-throughput data', *Nature Reviews Genetics*, vol. 11, no. 10, pp. 733-739.

Phipson, B, Zappia, L & Oshlack, A 2017, 'Gene length and detection bias in single cell RNA sequencing protocols [version 1; peer review: 4 approved]', *F1000Research*, vol. 6, no. 595.

Risso, D, Ngai, J, Speed, TP & Dudoit, S 2014, 'Normalization of RNA-seq data using factor analysis of control genes or samples', *Nature Biotechnology*, vol. 32, no. 9, pp. 896-902.

Robinson, MD & Oshlack, A 2010, 'A scaling normalization method for differential expression analysis of RNA-seq data', *Genome Biology*, vol. 11, no. 3, p. R25.

Zhao, Y, Li, M-C, Konaté, MM, Chen, L, Das, B, Karlovich, C, Williams, PM, Evrard, YA, Doroshow, JH & McShane, LM 2021, 'TPM, FPKM, or Normalized Counts? A Comparative Study of Quantification Measures for the Analysis of RNA-seq Data from the NCI Patient-Derived Models Repository', *Journal of Translational Medicine*, vol. 19, no. 1, p. 269.