# Normalization of RNA sequence data

RNA sequencing emerged 15 years ago (Marioni et al. 2008; Mortazavi et al. 2008; Nagalakshmi et al. 2008; Wilhelm et al. 2008), and has become preferred choice over other techniques such as microarray due its ability to find novel transcripts, provide enhanced resolution, wider detection capabilities (Wang, Z, Gerstein & Snyder 2009; Wilhelm & Landry 2009).

Normalization is a process of making data comparable within or across samples by minimising the technical variations and biases (Bullard et al. 2010; Risso et al. 2014; Robinson & Oshlack 2010; Zhao, Yingdong et al. 2021; Zwiener, Frisch & Binder 2014). It is one of the crucial steps which impacts the results of data analysis (Bullard et al. 2010; Evans, Hardin & Stoebel 2017; Risso et al. 2011; Robinson & Oshlack 2010; Zwiener, Frisch & Binder 2014). (Other crucial steps are statistical testing, alignment and abundance quantification which will be described on respective sections). Several benchmarking studies on normalization have been published however, the consensus which is the most accurate is lagging (Bullard et al. 2010; Dillies et al. 2013; Li, P et al. 2015; Lin et al. 2016; Maza et al. 2013; Zwiener, Frisch & Binder 2014; Zyprych-Walczak et al. 2015). The normalization techniques can be assessed by comparing the results with real-time qPCR which is well for benchmarking the true expression values (Li, P et al. 2015). Assessing the suitability of a selected normalization method can be challenging, but one approach is to employ multiple methods and evaluate the consistency of the outcomes.

# Intrasample normalization techniques

When comparing different genes within a sample, the inherent features of gene such as gene length and GC content causes bias. Genes with shorter effective length have lower counts and higher rate of dropouts (genes are expressed but not detected) (Oshlack & Wakefield 2009; Phipson, Zappia & Oshlack 2017). It has been found that GC content of a gene also causes variation in a sample specific expression measurement as genes with higher or lower GC content are underrepresented during sequencing (Benjamini & Speed 2012; Risso et al. 2011).

RPKM/FPKM (Reads/Fragments per kilo base of transcript per Million mapped reads) and TPM (Transcripts per million) are widely used techniques for intrasample comparison. RPKM is expression unit first described by (Mortazavi et al. 2008) for the single-end reads whereas FPKM is used for paired-end reads (Trapnell et al. 2010). RPKM corrects library size followed by gene length variance and was initially regarded as the best technique for normalization. However, evidence showed that while attempting to adjust for variations in gene length during a differential analysis, it lead to a biased impact on the variances per gene, especially for genes with low expression levels (Oshlack & Wakefield 2009).In addition, it corrects the library size first and then gene length making it less accurate and inconsistent among the samples (Wagner, Kin & Lynch 2012). To mitigate issues of RPKM, TPM (Transcripts per million) was developed (Li, B et al. 2009). Unlike RPKM, TPM corrects gene length followed by the library size. Hence, the resulting expression values reflects the true “relative molar concentration” (Wagner, Kin & Lynch 2012). The assumption of RPKM and TPM is that production and distribution of mRNAs across the samples is same. However, with the change in condition, amount of mRNA production also changes. This is main reason why RPKM and TPM can be used to compare genes only within sample or replicates of sample where the assumption of same amount and distribution of mRNA holds true (Zhao, S, Ye & Stanton 2020).

RPKM/FPKM =

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

TPM=A∗ ∗

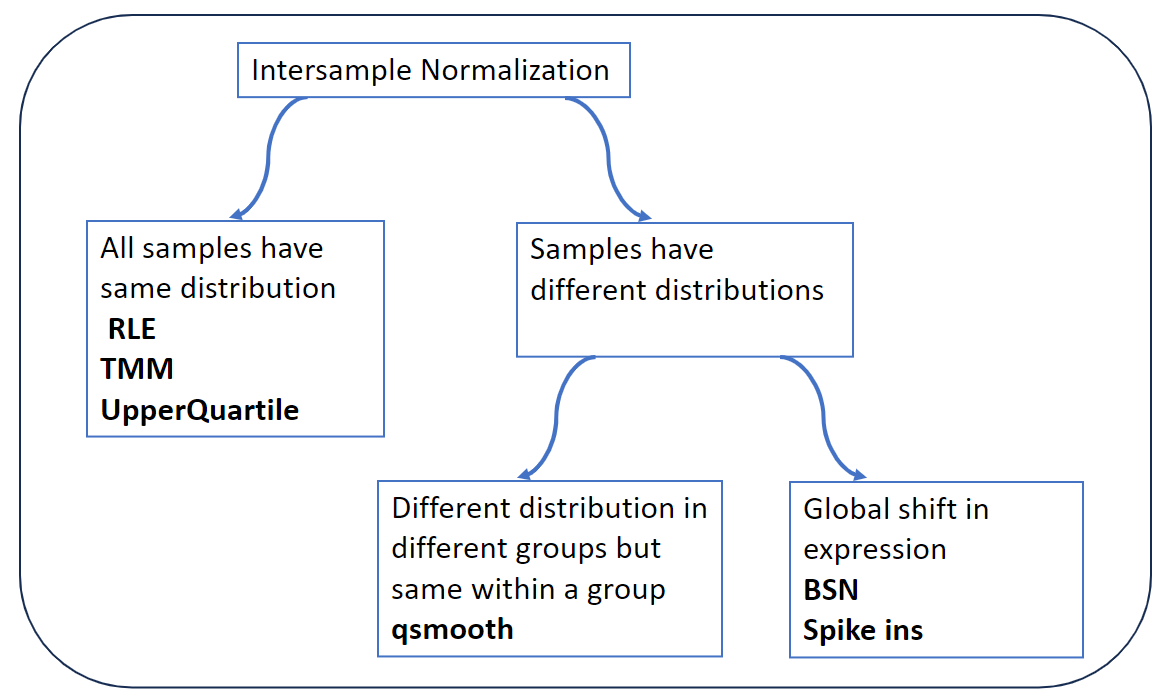
Where, A =

Here, 106 is for sequencing depth factor.

A GC content bias is observed partially due to PCR amplification while preparing library (Aird et al. 2011). GC content is confounding effect which is nonlinear and highly sample specific (Love, Hogenesch & Irizarry 2016). To account for GC bias, a within sample gene level approach is proposed in *EDASeq* package, three methods are implemented here such as regression normalization, global scaling normalization and quantile normalization and these techniques mainly stratify genes in same sized bins based on GC content, then match the count distributions across bins (Risso et al. 2011). GC effect is not consistent between repeated experiments making it difficult to account for (Benjamini & Speed 2012).

# Intersample normalization techniques

In intersample comparison sequencing depth variance, composition bias (Bullard et al. 2010; Robinson & Oshlack 2010), batch effects along with other known and unknown technical factors needs to be corrected. Sequencing depth, also known as, library size is total read counts mapped for a given sample. There is no optimal limit for sequencing depth as it depends on complexity of the organism and aim of the experiment (Conesa et al. 2016). A deep sequencing might improve the quantification and identification of transcripts leading to novel isoform detection, however it may also detect transcriptional noise and off-target transcripts (Conesa et al. 2016).While sequencing, different samples within a single experimental design may have different sequencing depths because total number of RNA molecules in cDNA library can vary (Dillies et al. 2013).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 composition bias to the detection of differentially expressed features (Robinson & Oshlack 2010). The batch effect appears due to difference on experimental design, different sequencing platforms, and researcher which causes unwanted noise in data. (Bullard et al. 2010; Leek et al. 2010).

Fig: Techniques for intersample normalization based on distribution of samples

## Normalization of sequencing depth and compositional bias

TMM (trimmed mean of M-values), RLE (relative log expression) and quantile normalization are widely used tools for intersample comparison which assumes all the samples have similar distribution and this assumption holds true when majority of genes are not dysregulated (Evans, Hardin & Stoebel 2017; Zhao, Yaxing, Wong & Goh 2020). TMM, RLE, and quantile normalization global are adjustment methods which force all the observed distribution of all the samples to be same and using these methods without understanding the actual distribution of all samples will distort the true biological signal (Evans, Hardin & Stoebel 2017; Wang, D et al. 2011; Zhao, Yaxing, Wong & Goh 2020). TMM (Robinson & Oshlack 2010), which is incorporated into the *edgeR* package, is widely used for differential gene expression. CPM (Counts per million mapped reads) is a basic gene expression unit (Chen, Lun & Smyth 2016; Law et al. 2018).

CPM =

After normalizing sequencing depth variation by CPM, lowly expressed genes are filtered and then TMM is used to handle the composition bias in count data. In the TMM method, a reference sample is chosen as a basis for comparison. The fold changes and absolute expression levels are then calculated for each gene relative to the reference sample. Genes that show significant differences in expression are trimmed off for robustness. To further normalize the data, the read counts are adjusted by dividing them by the product of the trimmed mean fold change and the total count of the corresponding sample. This process ensures that the expression values are scaled appropriately (Robinson & Oshlack 2010). TMMwsp (TMM with singleton pairing) is a variant of TMM to handle the zero inflated data. In TMM, gene having zero in either library while comparing a pair of libraries is ignored, but in TMMwsp it is not ignored. The makes the number of differentially expressed gene counts higher in TMMwsp and is more suitable approach to handle the inflated zero counts in the sample. The RLE in DESeq*2* Bioconductor package is similar to TMM, in which, first a **reference sample is created by calculating row-wise geometric mean.** Then, each count is divided by the reference and scaling factor iscalculated for each sample by calculating the median of each sample. lastly, count values are normalized using normalization factor (Love, Huber & Anders 2014). Quantile normalization was first proposed for microarray data. It is achieved by replacing each quantile in the data with the average or median value of that quantile, considering the values from all samples (Bolstad et al. 2003). This process aims to align the distributions and make them comparable across different samples. There are two types of quantile normalization, the upper quartile normalization which is determined by applying the 0.75 quantile to the gene counts from all runs and is applied in sequencing data (Bullard et al. 2010), whereas, the median quartile uses median value to the gene counts from all the runs and was first proposed for microarray data (Bolstad et al. 2003).

In large scale studies and studies involving different cells from different organs or cancer cells, the assumptions of TMM, RLE and Quantile methods might fail, in this type of data, all the samples may not have similar distributions and majority of genes might be upregulated in one biological condition relative to other condition. When cells produce equivalent amount of RNA per cell and yields of RNA remain same throughout the experiment, then normalised expression data provides the accurate representation of the each gene product (Lovén et al. 2012). But in cases where master regulator such as c-Myc is present, it can increase the mRNA expression by two to three folds bringing a global change (Lovén et al. 2012). For the normalization of global shift in expression and unbalanced number of upregulated and downregulated genes, either reference genes whose expression is constant, or spike ins can be used (Liu et al. 2019; Lovén et al. 2012). Another method for global shift in expression is BSN (Biological scaling normalization) which needs experimentally measured polyA+ RNA amounts. The measurement can also normalize the global shift in expression (Aanes et al. 2014). In some cases, distribution of samples with a group is same but across all the groups is different. For this situation, a normalization technique has been developed called smooth quantile normalization or *qsmooth* (Hicks et al. 2017). *qsmooth* is based on assumption that the distribution of each sample should be the same within a biological group or a condition, but the distribution may differ between different biological groups. It does not require control feature or external information to adjust the influence of unknown variation. It takes the group information and performs quantile normalization within a group. A care must be taken while using quantile normalization because a perfect graph obtained after quantile normalization might not be showing true biological signals, a completely different biological group might appears to be similar when applied blindly on the entire dataset (Wang, D et al. 2011; Zhao, Yaxing, Wong & Goh 2020).

## Normalization of batch effects and unwanted variations

A diagram of a process

Description automatically generatedAlthough increase in sample size gives more statistical power to analysis, the chances of getting batch effects and other unwanted variations also increases proportionally. There might be known and unknown batch effects, according to their nature, they can be accounted and corrected. There are mainly two ways, one way is correcting the batch effects first and separately modelling the corrected data with covariates of interest and the other way is accounting for the unknown batch effects during data modelling of both covariates of interest and batch variations together.

Fig: Techniques for accounting and correcting batch effects based on nature of unwanted variation

1. **Known batch effects**

BMC (Batch mean cantering) and Combat seq are techniques for correcting the known batch effects. These are practical and have broad applications (Wang, Y & LêCao 2019), here first unwanted batch effect is removed and corrected data is used for downstream analysis. Since, it does not consider correlation between covariates and batch, it does not detect correlation of variables associated with covariates (Wang, Y & LêCao 2019). BMC is simple and effective technique which centres the data within a batch by making the batch mean standardised to zero. It can be used for only normally distributed data. To use Combat seq batch effect must be known and should have a systematic pattern. Combat seq is revised version of Combat (developed for microarray data), can be used to correct known systematic batch effects. Combat seq is based on negative binomial regression model to estimate batch effect and also retains the integer nature of count data making it easier to interpret (Zhang, Parmigiani & Johnson 2020). The older version, Combat is based on Gaussian distribution model and was not able capture skewed and over dispersed nature of count data. Combat seq outperforms the Combat controlling the false positives (Zhang, Parmigiani & Johnson 2020). It uses the empirical Bayes method to estimate the parameters to adjust each batch. ComBat seq is implemented in *SVAseq* package.

1. **Unknown batch effects**

RUV (remove unwanted variation) and SVA (surrogate variable analysis) seq find the latent factors due to batch effect and use it as a covariate in the data modelling. RUV was developed to identify and adjust uncharacterized technical variations in data analysis using negative control variables. RUVseq follows generalised linear model (GLM) where both biological covariates and unknown factors due unwanted variation are regressed. Three approaches are used to estimate the unknown factors, they are, RUVg (negative control genes are used which have constant expression across samples), RUVs (negative control sample/replicate which is assumed to have covariates of interest are constant) and RUVr (residuals which are from first pass GLM regression on the known covariates). However, the use of spike ins as negative control is difficult, since it is challenging to maintain constant amount of spike ins in all samples in same ratio (Robinson & Oshlack 2010) and it is still unclear if the synthetic spike ins act in same manner as cellular transcripts (Grün & van Oudenaarden 2015). SVAseq builds a surrogate variable that can be used to adjust unknown variables due to technical noise. It is univariate method and based on Gaussian likelihood assumption. Data is modelled as the combination of known covariates of interest, known adjustment variables and unknown artifacts. Here, first a subset of data affected by artifact is identified and then matrix decomposition like is applied to estimate the artifacts, the resulted estimate is used to correct the analysis (Leek 2014). The genomic data affected by artefact is determined by identifying the variables which are not associated with biological variables if interest. This can be done by fitting a null model and full linear model to the data.

**Assessment of various normalization techniques**

Normalization, is not a single step process, is rather a layer of multiple processes to remove unwanted technical variations until data is clean with biological variation alone. First the lowly expressed genes and poor-quality samples are filtered. Then, data should be normalised for its varying sequencing depth, compositional bias and log transformed. Then, this normalized data is used to correct batch effect. If batch correction is done before normalization, then latent factors will have sequencing depth and compositional bias as a major batch difference. After correcting sequencing depth, compositional bias and batch effects, the data can be used for differential gene expression analysis. When most genes are not differentially expressed then methods such as TMM, RLE and Quantile normalisation can be used. But cases where this assumption does not hold true, external controls would be appropriate. It is crucial that a thorough evaluation of normalization techniques is required before conducting downstream statistical analysis. There are benchmarking schema are present to explore the accuracy and sensitivity of various normalization techniques (Abrams et al. 2019). The best application of the technique depends on research objective and nature of data.

# References

Aanes, H, Winata, C, Moen, LF, Østrup, O, Mathavan, S, Collas, P, Rognes, T & Aleström, P 2014, 'Normalization of RNA-Sequencing Data from Samples with Varying mRNA Levels', *PLOS ONE*, vol. 9, no. 2, p. e89158.

Abrams, ZB, Johnson, TS, Huang, K, Payne, PRO & Coombes, K 2019, 'A protocol to evaluate RNA sequencing normalization methods', *BMC Bioinformatics*, vol. 20, no. 24, p. 679.

Aird, D, Ross, MG, Chen, W-S, Danielsson, M, Fennell, T, Russ, C, Jaffe, DB, Nusbaum, C & Gnirke, A 2011, 'Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries', *Genome Biology*, vol. 12, no. 2, pp. 1-14.

Benjamini, Y & Speed, TP 2012, 'Summarizing and correcting the GC content bias in high-throughput sequencing', *Nucleic acids research*, vol. 40, no. 10, pp. e72-e72.

Bolstad, BM, Irizarry, RA, Åstrand, M & Speed, TP 2003, 'A comparison of normalization methods for high density oligonucleotide array data based on variance and bias', *Bioinformatics*, vol. 19, no. 2, pp. 185-193.

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.

Chen, Y, Lun, A & Smyth, G 2016, 'From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline [version 2; peer review: 5 approved]', *F1000Research*, vol. 5, no. 1438.

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.

Dillies, M-A, Rau, A, Aubert, J, Hennequet-Antier, C, Jeanmougin, M, Servant, N, Keime, C, Marot, G, Castel, D, Estelle, J, Guernec, G, Jagla, B, Jouneau, L, Laloë, D, Le Gall, C, Schaëffer, B, Le Crom, S, Guedj, M, Jaffrézic, F & on behalf of The French StatOmique, C 2013, 'A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis', *Briefings in Bioinformatics*, vol. 14, no. 6, pp. 671-683.

Evans, C, Hardin, J & Stoebel, DM 2017, 'Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions', *Briefings in Bioinformatics*, vol. 19, no. 5, pp. 776-792.

Grün, D & van Oudenaarden, A 2015, 'Design and Analysis of Single-Cell Sequencing Experiments', *Cell*, vol. 163, no. 4, pp. 799-810.

Hicks, SC, Okrah, K, Paulson, JN, Quackenbush, J, Irizarry, RA & Bravo, HC 2017, 'Smooth quantile normalization', *Biostatistics*, vol. 19, no. 2, pp. 185-198.

Law, C, Alhamdoosh, M, Su, S, Dong, X, Tian, L, Smyth, G & Ritchie, M 2018, 'RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR [version 3; peer review: 3 approved]', *F1000Research*, vol. 5, no. 1408.

Leek, JT 2014, 'svaseq: removing batch effects and other unwanted noise from sequencing data', *Nucleic acids research*, vol. 42, no. 21, pp. e161-e161.

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.

Li, B, Ruotti, V, Stewart, RM, Thomson, JA & Dewey, CN 2009, 'RNA-Seq gene expression estimation with read mapping uncertainty', *Bioinformatics*, vol. 26, no. 4, pp. 493-500.

Li, P, Piao, Y, Shon, HS & Ryu, KH 2015, 'Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data', *BMC Bioinformatics*, vol. 16, no. 1, p. 347.

Lin, Y, Golovnina, K, Chen, Z-X, Lee, HN, Negron, YLS, Sultana, H, Oliver, B & Harbison, ST 2016, 'Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster', *BMC Genomics*, vol. 17, no. 1, p. 28.

Liu, X, Li, N, Liu, S, Wang, J, Zhang, N, Zheng, X, Leung, K-S & Cheng, L 2019, 'Normalization methods for the analysis of unbalanced transcriptome data: a review', *Frontiers in bioengineering and biotechnology*, vol. 7, p. 358.

Love, MI, Hogenesch, JB & Irizarry, RA 2016, 'Modeling of RNA-seq fragment sequence bias reduces systematic errors in transcript abundance estimation', *Nature Biotechnology*, vol. 34, no. 12, pp. 1287-1291.

Love, MI, Huber, W & Anders, S 2014, 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, vol. 15, no. 12, p. 550.

Lovén, J, Orlando, David A, Sigova, Alla A, Lin, Charles Y, Rahl, Peter B, Burge, Christopher B, Levens, David L, Lee, Tong I & Young, Richard A 2012, 'Revisiting Global Gene Expression Analysis', *Cell*, vol. 151, no. 3, pp. 476-482.

Marioni, JC, Mason, CE, Mane, SM, Stephens, M & Gilad, Y 2008, 'RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays', *Genome Research*, vol. 18, no. 9, pp. 1509-1517.

Maza, E, Frasse, P, Senin, P, Bouzayen, M & Zouine, M 2013, 'Comparison of normalization methods for differential gene expression analysis in RNA-Seq experiments: a matter of relative size of studied transcriptomes', *Communicative & integrative biology*, vol. 6, no. 6, p. e25849.

Mortazavi, A, Williams, BA, McCue, K, Schaeffer, L & Wold, B 2008, 'Mapping and quantifying mammalian transcriptomes by RNA-Seq', *Nature Methods*, vol. 5, no. 7, pp. 621-628.

Nagalakshmi, U, Wang, Z, Waern, K, Shou, C, Raha, D, Gerstein, M & Snyder, M 2008, 'The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing', *Science*, vol. 320, no. 5881, pp. 1344-1349.

Oshlack, A & Wakefield, MJ 2009, 'Transcript length bias in RNA-seq data confounds systems biology', *Biology Direct*, vol. 4, no. 1, p. 14.

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.

Risso, D, Schwartz, K, Sherlock, G & Dudoit, S 2011, 'GC-Content Normalization for RNA-Seq Data', *BMC Bioinformatics*, vol. 12, no. 1, p. 480.

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.

Trapnell, C, Williams, BA, Pertea, G, Mortazavi, A, Kwan, G, van Baren, MJ, Salzberg, SL, Wold, BJ & Pachter, L 2010, 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation', *Nature Biotechnology*, vol. 28, no. 5, pp. 511-515.

Wagner, GP, Kin, K & Lynch, VJ 2012, 'Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples', *Theory in Biosciences*, vol. 131, no. 4, pp. 281-285.

Wang, D, Cheng, L, Wang, M, Wu, R, Li, P, Li, B, Zhang, Y, Gu, Y, Zhao, W, Wang, C & Guo, Z 2011, 'Extensive increase of microarray signals in cancers calls for novel normalization assumptions', *Computational Biology and Chemistry*, vol. 35, no. 3, pp. 126-130.

Wang, Y & LêCao, K-A 2019, 'Managing batch effects in microbiome data', *Briefings in Bioinformatics*, vol. 21, no. 6, pp. 1954-1970.

Wang, Z, Gerstein, M & Snyder, M 2009, 'RNA-Seq: a revolutionary tool for transcriptomics', *Nature Reviews Genetics*, vol. 10, no. 1, pp. 57-63.

Wilhelm, BT & Landry, J-R 2009, 'RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing', *Methods*, vol. 48, no. 3, pp. 249-257.

Wilhelm, BT, Marguerat, S, Watt, S, Schubert, F, Wood, V, Goodhead, I, Penkett, CJ, Rogers, J & Bähler, J 2008, 'Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution', *Nature*, vol. 453, no. 7199, pp. 1239-1243.

Zhang, Y, Parmigiani, G & Johnson, WE 2020, 'ComBat-seq: batch effect adjustment for RNA-seq count data', *NAR Genomics and Bioinformatics*, vol. 2, no. 3, p. lqaa078.

Zhao, S, Ye, Z & Stanton, R 2020, 'Misuse of RPKM or TPM normalization when comparing across samples and sequencing protocols', *Rna*, vol. 26, no. 8, pp. 903-909.

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.

Zhao, Y, Wong, L & Goh, WWB 2020, 'How to do quantile normalization correctly for gene expression data analyses', *Scientific Reports*, vol. 10, no. 1, p. 15534.

Zwiener, I, Frisch, B & Binder, H 2014, 'Transforming RNA-Seq Data to Improve the Performance of Prognostic Gene Signatures', *PLOS ONE*, vol. 9, no. 1, p. e85150.

Zyprych-Walczak, J, Szabelska, A, Handschuh, L, Górczak, K, Klamecka, K, Figlerowicz, M & Siatkowski, I 2015, 'The impact of normalization methods on RNA-Seq data analysis', *BioMed research international*, vol. 2015.