Normalization methods

Barbora Zwinsová

September 2023

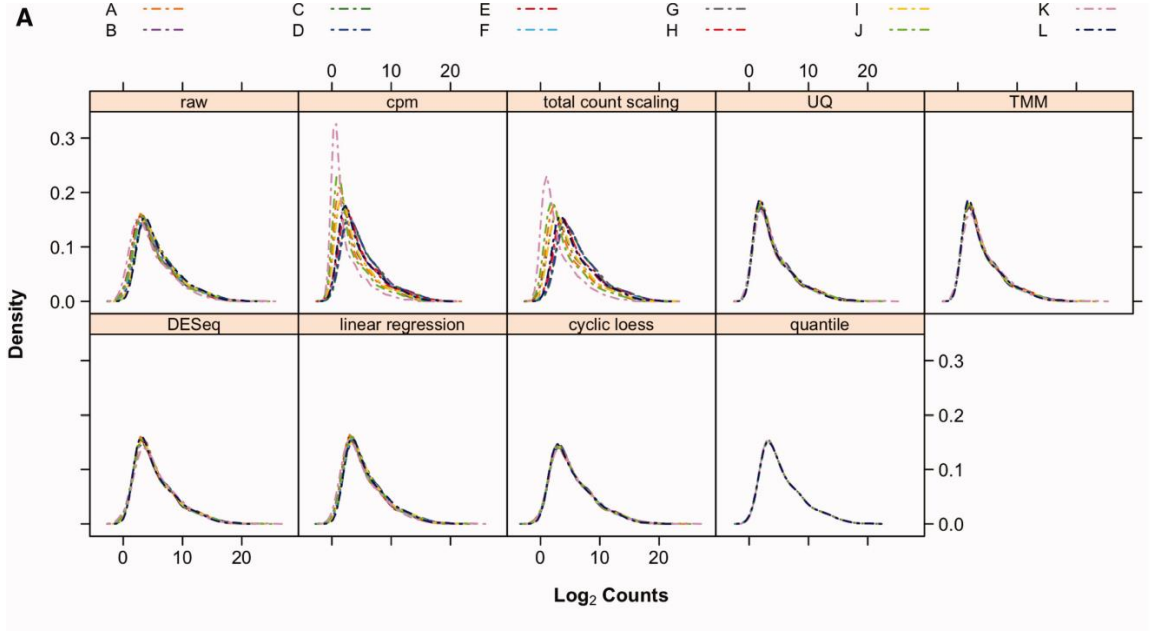
miRNA

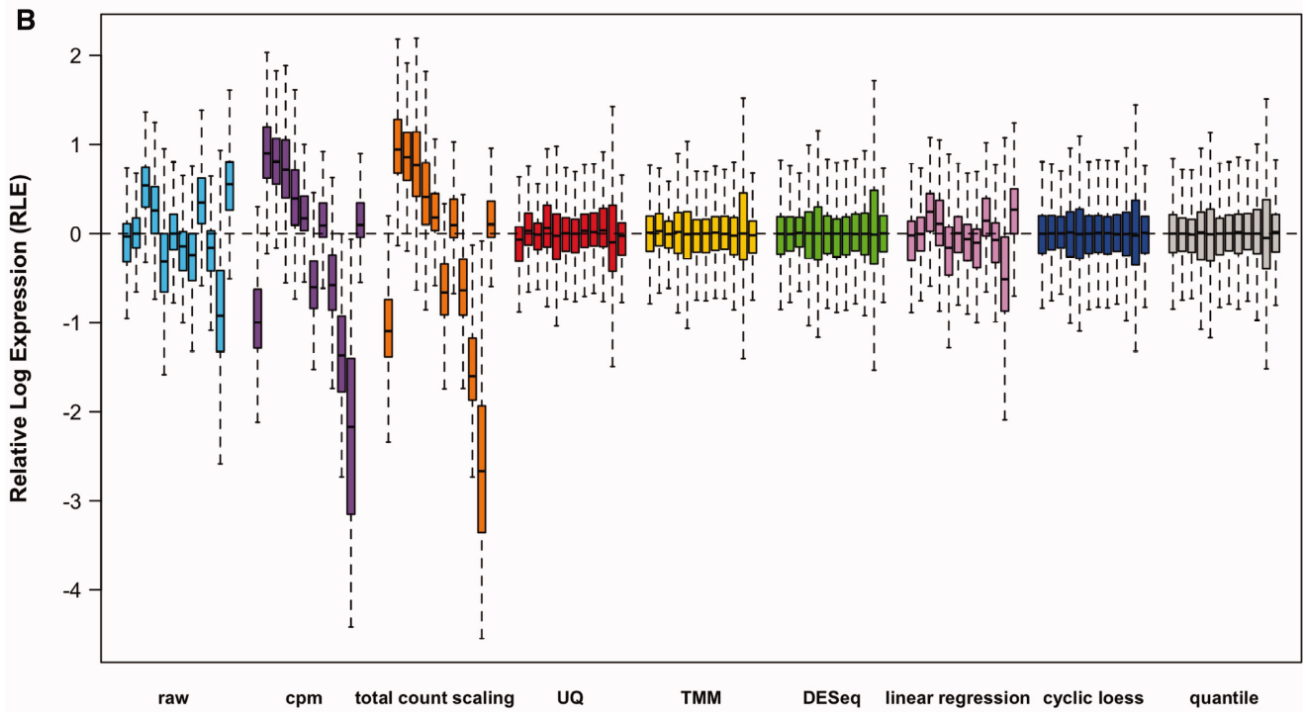
Overview of the methods (miRNA)

Tam et al., 2015, Briefings in Bioinformatics
https://doi.org/10.1093/bib/bbv019

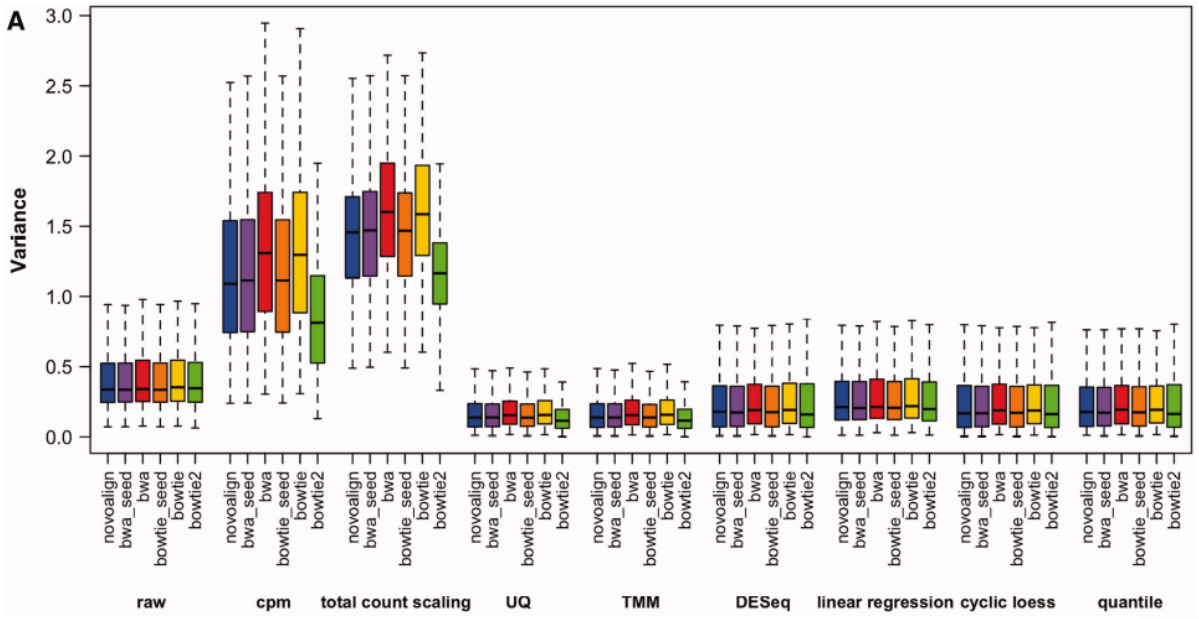
Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads *the simplest form of normalization*	sequencing depth	gene count comparisons between replicates of the same sample group; **NOT for within sample comparisons or DE analysis**
Total count scaling	after scaling each sample to its library size, they can be rescaled to a common value across all samples	sequencing depth and RNA composition	
Upper quantile scaling	modified quantile-normalization method: the upper quartile of expressed miRNAs is used instead as a linear scaling factor	sequencing depth	This method has been shown to yield better concordance with qPCR results than linear total counts scaling for RNA-seq data
Trimmed mean of M (edgeR)	calculates a linear scaling factor, di, for sample i, based on a weighted mean after trimming the data by log fold-changes (M) relative to a reference sample and by absolute intensity (A)	does not take into consideration the potentially different RNA composition across the samples	gene count comparisons between and within samples and for DE analysis
DESeq2's **median of ratios**	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for DE analysis; **NOT for within sample comparisons**
Linear regression	assumes that the systematic bias is linearly dependent on the count abundance	samples normalized to a baseline reference, which was defined as the median count of each element across the profiled samples	
Cyclic loess (nonlinear regression)		Baseline referece	
Quantile	non-scaling approach, forces the distribution of read counts in all samples across an experiment to be equivalent		assumes that most targets are not differentially expressed and that the true expression distribution is similar across all samples

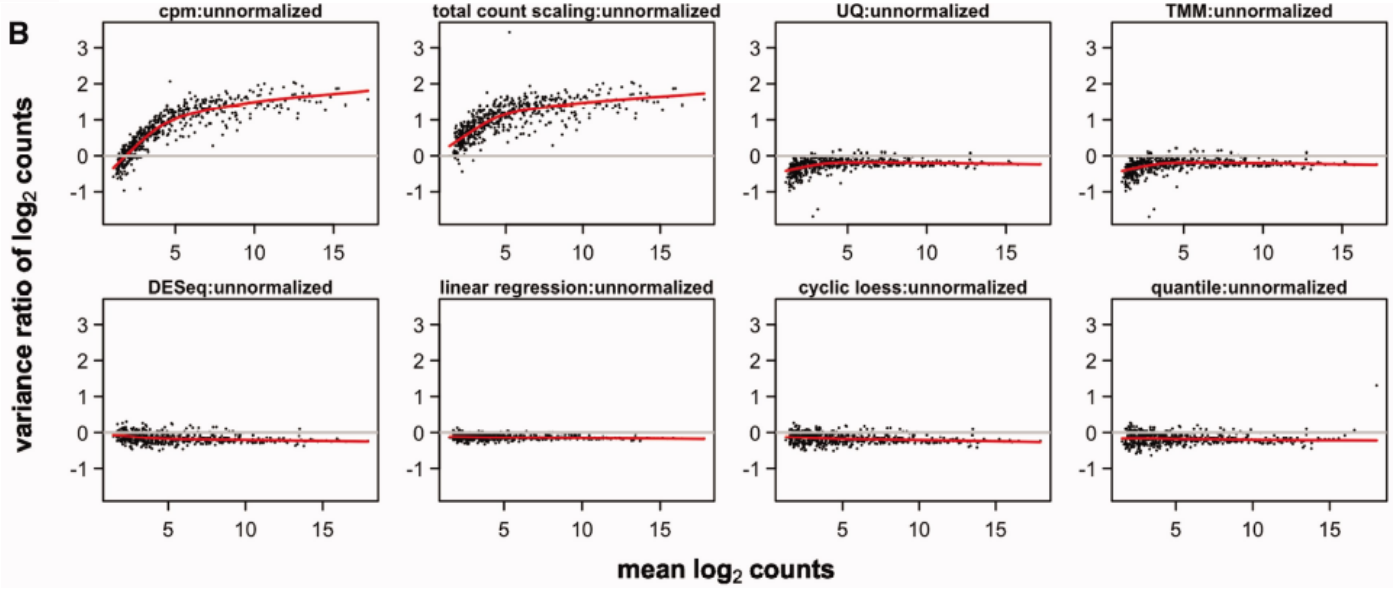
Comparison of data distribution (Tam et al., 2015)





Variance comparisons (Tam et al., 2015)





Conclusion (miRNA) (Tam et al., 2015)

- simply adjusting miRNA counts to the sequencing depth is inadequate
- the distinct number of miRNAs identified in replicate samples may differ because of the random sampling nature of the technology; normalizing to the library size ignores this.
- total count scaling introduces more variability by pushing all samples toward the same distribution
- UQ, TMM, DESeq, cyclic loess and quantile normalization are highly similar
- quantile and cyclic loess normalization may be too aggressive by forcing the distribution of the samples to be the same
- increased variability was noted in the lower abundance miRNAs compared with UQ and TMM normalized data
- Dillies et al. & Tam et al. support the use of TMM (and UQ) for the normalization of miRNA count data
- Tam et al. - **BWA with one mismatch across the entire read and UQ or TMM**, respectively, lead to **more accurate results in downstream analyses**

Transcriptome

Overview of the methods (transcriptomics)

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads the simplest form of normalization	sequencing depth	gene count comparisons between replicates of the same sample group; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios**	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for DE analysis; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values** (TMM)	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition, and gene length	gene count comparisons between and within samples and for DE analysis

CPM, RPKM and TPM

	Sample 1	Sample 2	Sample 3
Gene A (1.5kb)	50	25	85
Gene B (2kb)	75	50	90
Sequencing depth	**125**	**75**	**175**

CPM (Counts Per Million)

For the example I am scaling by 10 instead of 1000000

50/12.5 = 4

	Sample 1	Sample 2	Sample 3
Gene A (1.5kb)	4	3.33	4.85
Gene B (2kb)	6	6.66	5.14

RPKM (Reads Per Kilobase Million)

Step 1: Normalize for sequencing depth
For the example I am scaling by 10 instead of 1000000

50/12.5 = 4

	Sample 1	Sample 2	Sample 3
Gene A (1.5kb)	4	3.33	4.85
Gene B (2kb)	6	6.66	5.14

Step 2: Normalize for gene length

4/1.5 = 2.66

	Sample 1	Sample 2	Sample 3
Gene A (1.5kb)	2.66	2.22	3.23
Gene B (2kb)	3	3.33	2.57
Seq. depth	5.66	5.55	5.8

TPM (Transcripts Per Kilobase Million)

Step 1: Normalize for gene length

50/1.5 = 33.33

	Sample 1	Sample 2	Sample 3
Gene A (1.5kb)	33.33	16.66	56.66
Gene B (2kb)	37.5	25	45
Seq. depth	70.83	41.66	101.66

Step 2: Normalize for sequencing depth
For the example I am scaling by 10 instead of 1000000

33.33/7.083

	Sample 1	Sample 2	Sample 3
Gene A (1.5kb)	4.7	3.99	5.57
Gene B (2kb)	5.29	6	4.426
Seq. depth	9.99	9.99	9.99

RPKM/FPKM (not recommended)

- the normalized count values output by the RPKM/FPKM method are not comparable between samples
- the total number of RPKM/FPKM normalized counts for each sample will be different. Therefore, you cannot compare the normalized counts for each gene equally between samples.

DESeq2-normalized counts: Median of ratios method

- tools for differential expression analysis are comparing the counts between sample groups for the same gene, gene length does not need to be accounted for by the tool
- sequencing depth and RNA composition do need to be taken into account

Median of ratios normalization

Step 1: creates a pseudo-reference sample (row-wise geometric mean)

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
...	...	...	...

Step 2: calculates ratio of each sample to the reference

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

Step 3: calculate the normalization factor for each sample (size factor)

The median value (column-wise for the above table) of all ratios for a given sample is taken as the normalization factor (size factor) for that sample

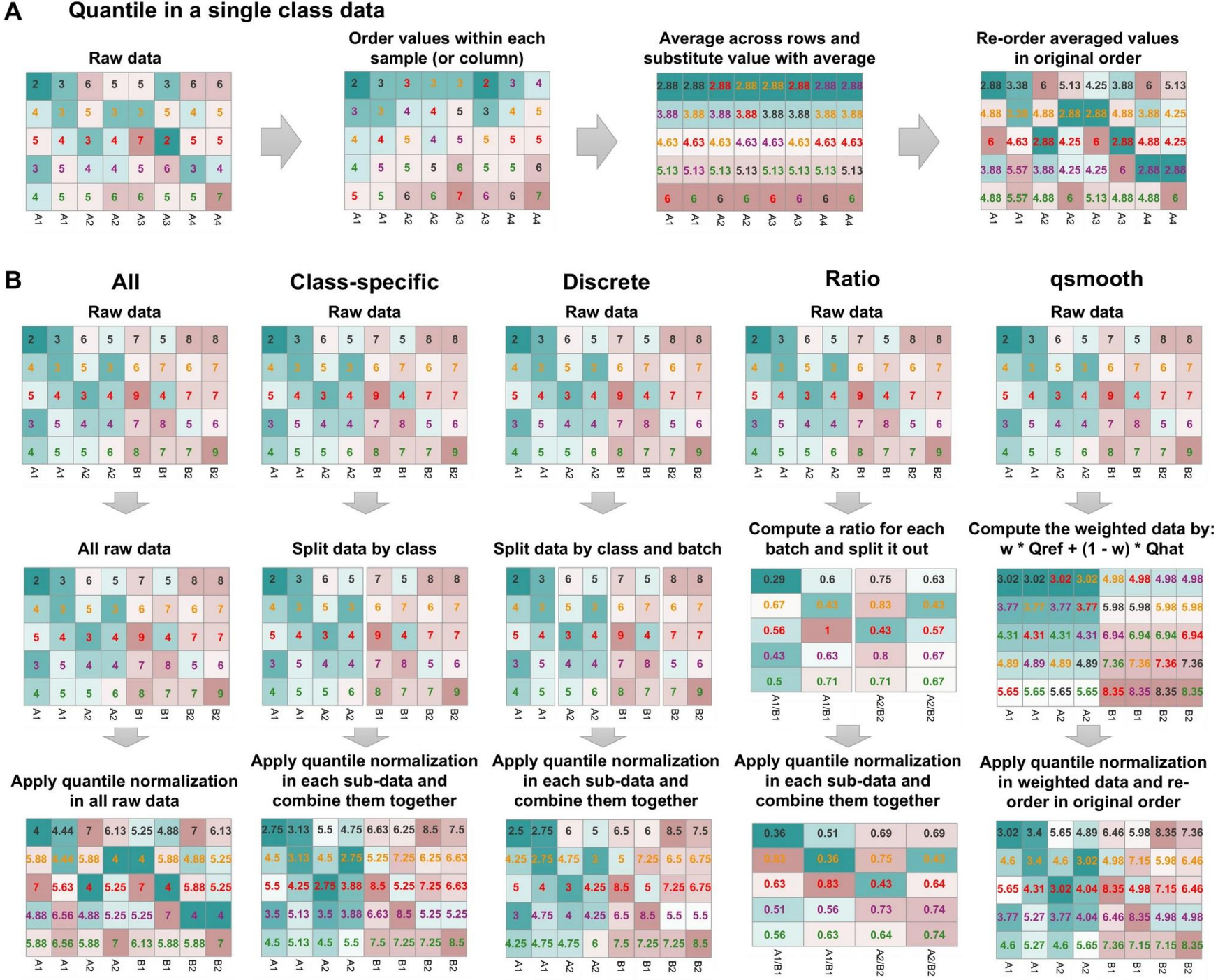
Step 4: calculate the normalized count values using the normalization factor

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

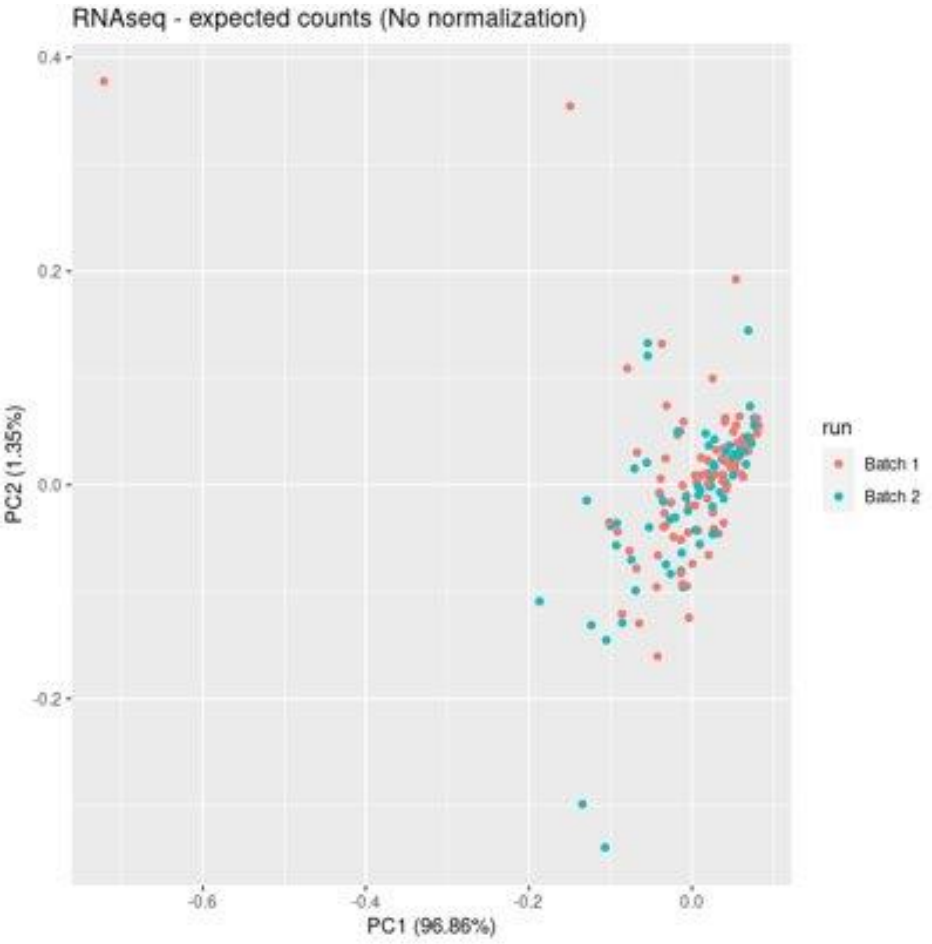
Quantile normalization

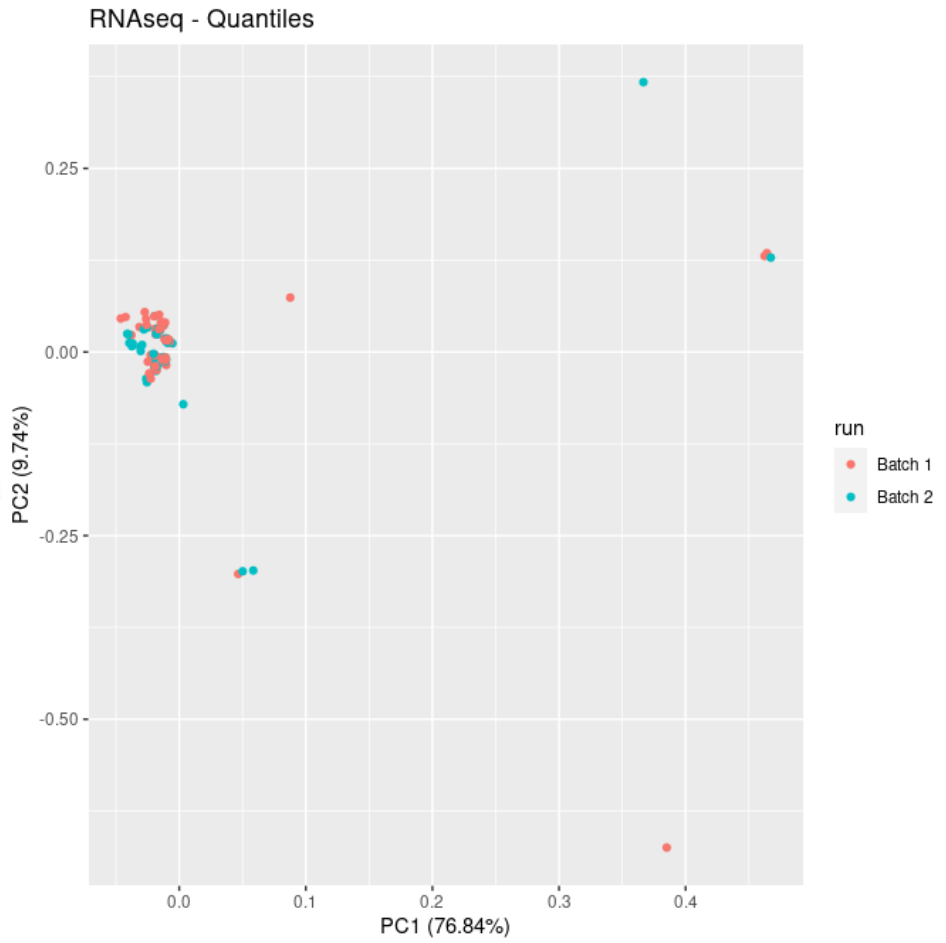
Smooth quantile normalization

Conditional quantile normalization

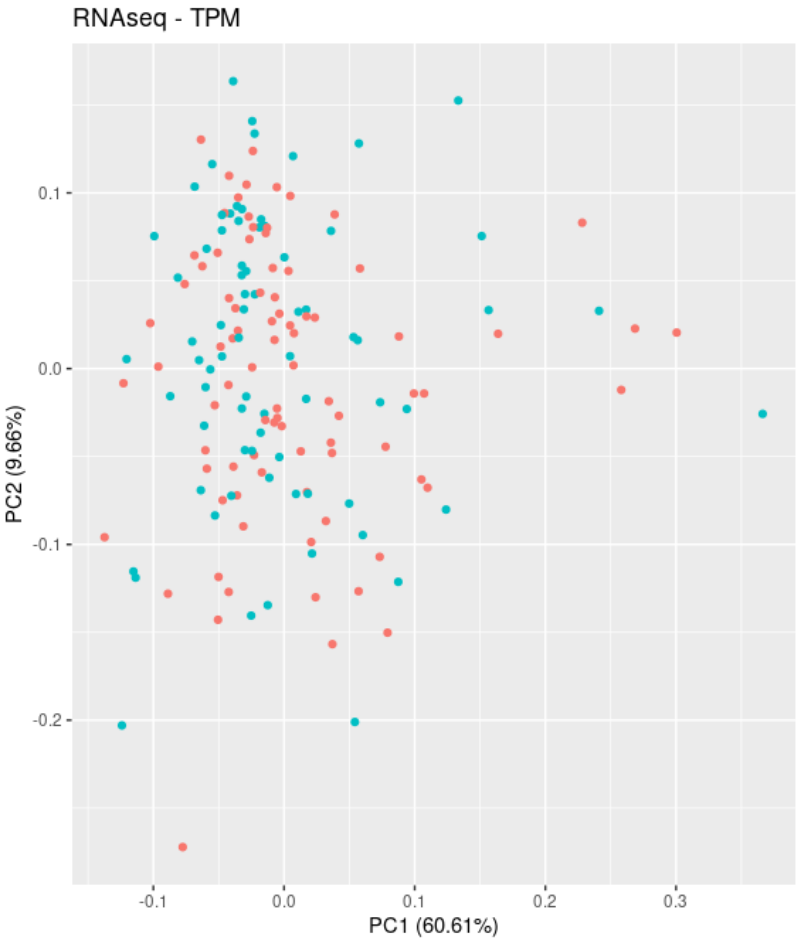


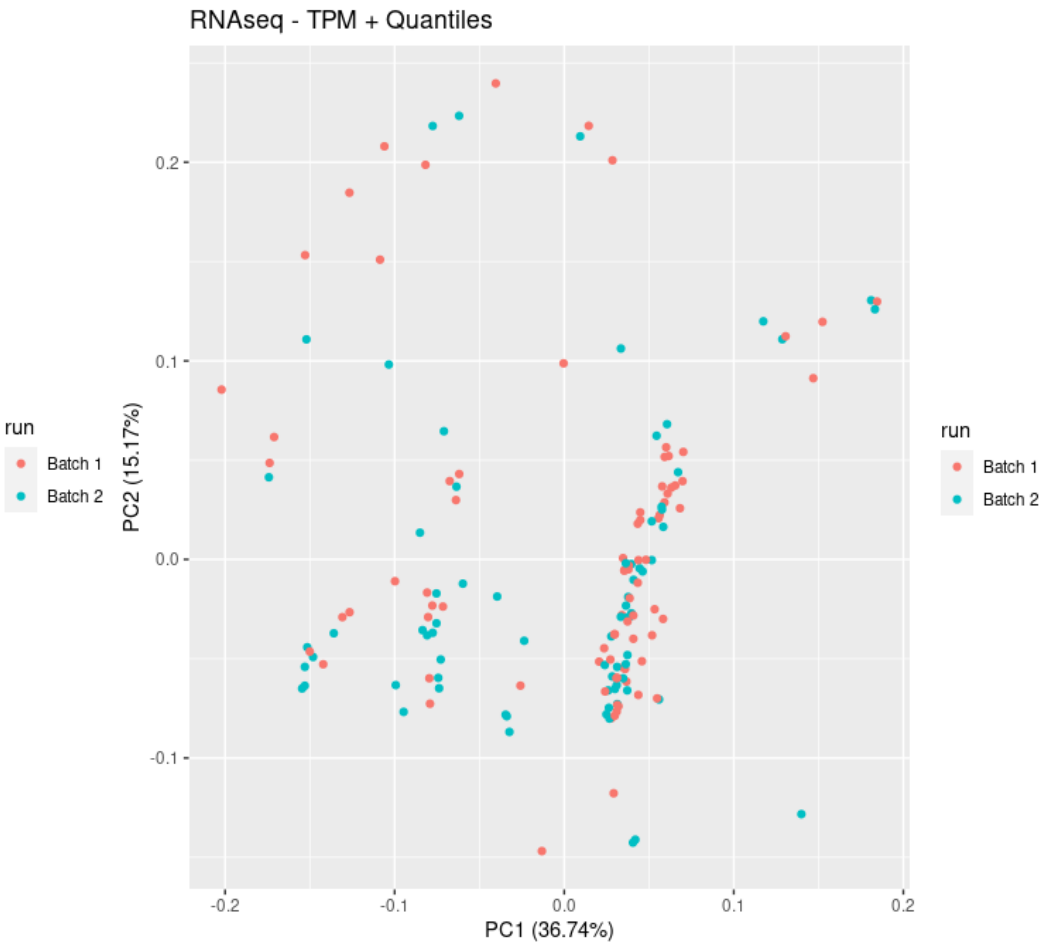
Projection of samples in 2D after PCA



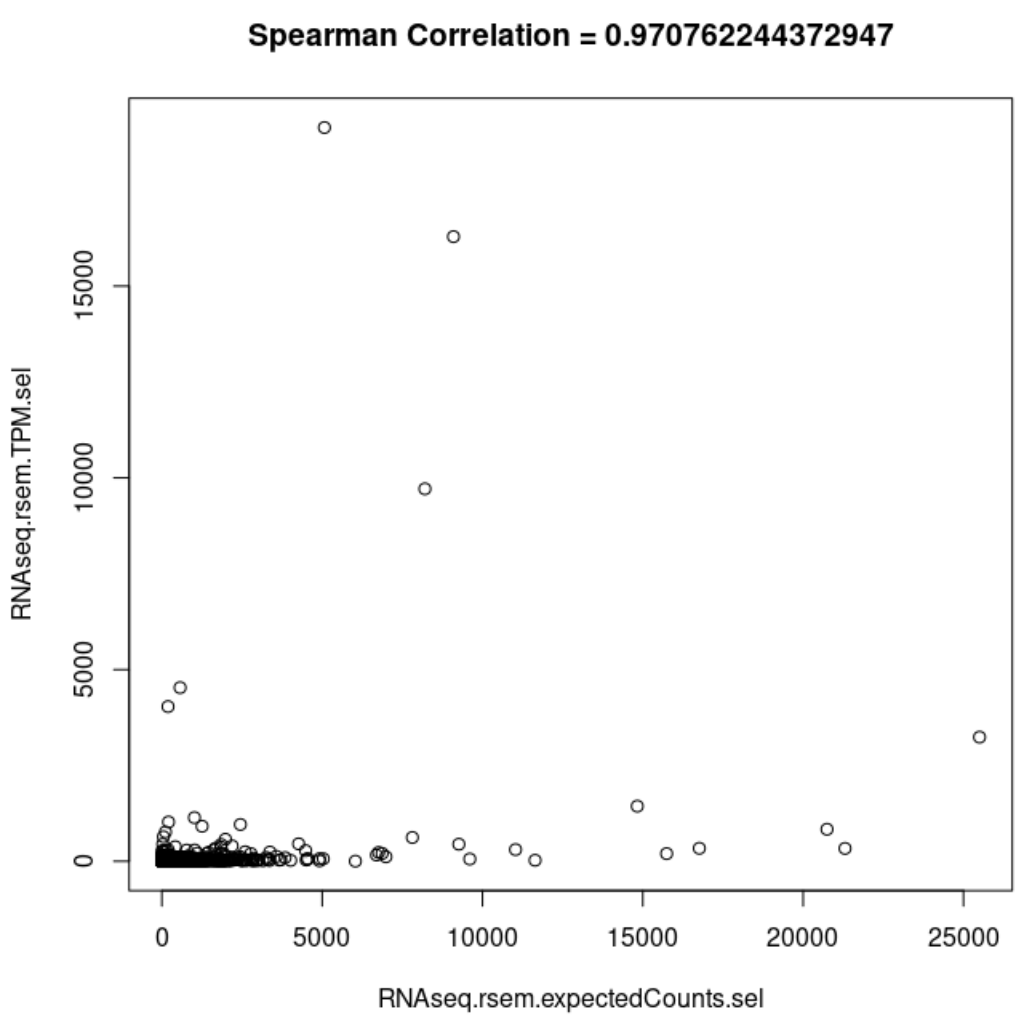


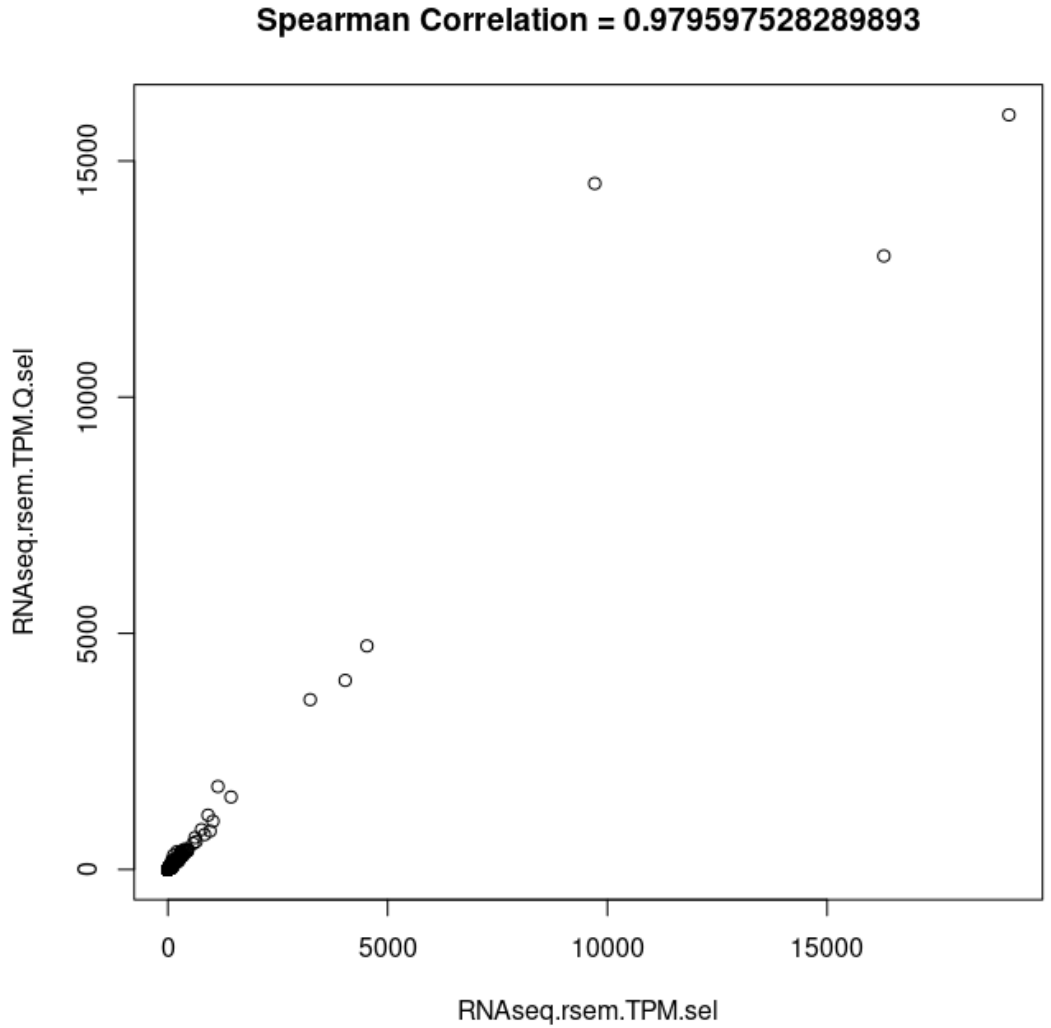
Projection of samples in 2D after PCA





Comparison of different approaches (sampling





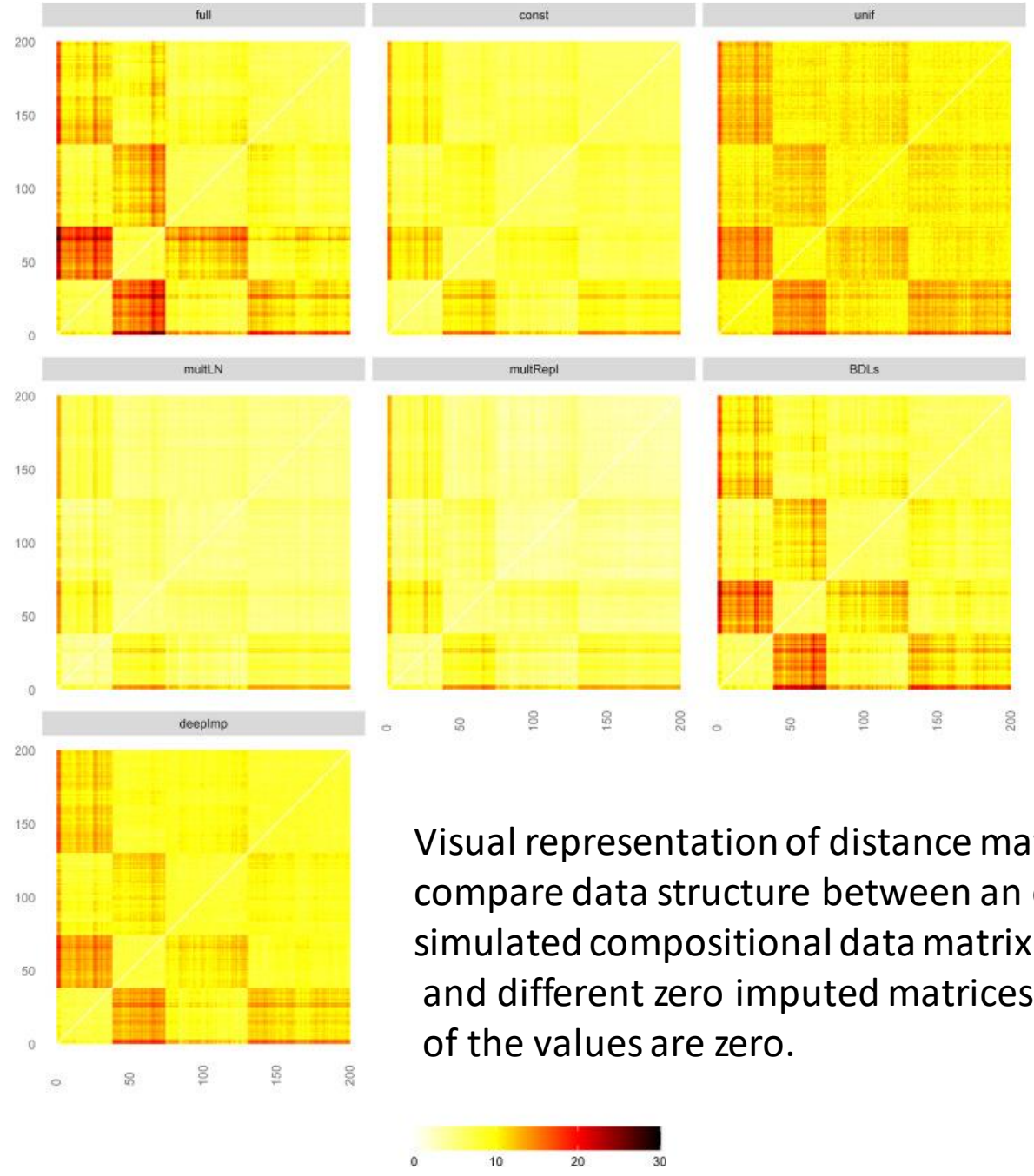
MetaTranscriptome

Microbiome

Sugnet Lubbe, Peter Filzmoser, Matthias Templ,
Comparison of zero replacement strategies for compositional data with large numbers of zeros. Chemometrics and Intelligent Laboratory Systems, 2021. https://doi.org/10.1016/j.chemolab.2021.104248

Overview of the methods for zero replacement

Normalization method	Description	
Add constant value	The simplest method is replacing all zeros with a constant value smaller than the detection limit. Martín-Fernández et al. (2003) found that 65% of the detection limit minimizes the distortion in the covariance structure. Using a constant value in the majority of cells leads to underestimation of the compositional variability.	0.65*detection limit = 0.65*1
Using uniform values between 0 and detection limit	Uniform values between 0 and the detection limit (DL) is often used, setting the first parameter at 0.1*DL prevents imputed values from being too close to zero.	runif(0.1*DL, DL)
Non-parametric multiplicative simple imputation	did not work if more than about half of the entries in the compositional data matrix were zero	ZComposition package in R
Model-based multiplicative lognormal imputation	The replacement is done in an iterative manner, and for that purpose the EM algorithm, Markov Chain Monte Carlo (MCMC) or multiple imputation are utilized.	ZComposition package in R
BDLs (Below detection limit)	iterative model-based procedure which performs regressions to replace the zeros (e.g. ordinary multiple linear regression, robust regression, and partial least-squares (PLS) regression), procedure is based on k-nearest-neighbour imputation for a large number of zeros there are too few neighbours with non-zeros available, which makes the algorithm not applicable in this context.	
deepImp	Imputation with deep learning methods, particularly using deep artificial neural networks in an EM-based approach	DeepImp package in R



Visual representation of distance matrices to compare data structure between an original simulated compositional data matrix and different zero imputed matrices when 50% of the values are zero.

Overview of the normalization methods

Normalization method	Description
CLR – centered log-ratio	divides each compositional part by the geometric mean of all parts CLR removes the value-range restriction (which is good for some applications), but does not remove the sum constraint
ILR – Isometric log-ratio	Instead of analyzing relative abundances, $y_{i,}$ of D different OTUs, the ILR transform produces D − 1 coordinates, x^*_i (called “balances”) Each balance corresponds to a single internal node of the tree and represents the averaged difference in relative abundance between the taxa in the two sister clades descending from that node
ALR – Additive log-ratio	One component is used as a baseline (reference), the proportion with the selected reference is logarithmized