Austrian Journal of Statistics
July 2023, Volume 52, 180–197.
http://www.ajs.or.at/
doi:10.17713/ajs.v52i4.1617

amIcompositional: Simple Tests for Compositional Behaviour of High Throughput Data with Common Transformations

Gregory B. Gloor
The University of Western Ontario

Abstract

Compositional approaches are beginning to permeate high throughput biomedical sciences in the areas of microbiome, genomics, transcriptomics and proteomics. Yet non-compositional approaches are still commonly observed. Non-compositional approaches are particularly problematic in network analysis based on correlation, ordination and exploratory data analysis based on distance, and differential abundance analysis based on normalization. Here we describe the alc R package, a simple tool that answers the fundamental question: does the dataset or normalization exhibit compositional artefacts that will skew interpretations when analyzing high throughput biomedical data? The alc R package includes options for several of the most widely used normalizations and filtering methods. The R package includes tests for subcompositional dominance and coherence along with perturbation and scale invariance. Exploratory analysis is facilitated by an R Shiny app that makes the process simple for those not wishing to use an R console. This simple approach will allow research groups to acknowledge and account for potential artefacts in data analysis resulting in more robust and reliable inferences.

Keywords: compositional data, sub-compositions, data normalization, high throughput sequencing, R.

1. Introduction

In the last two decades, high throughput sequencing (HTS) has become the method of choice to characterize molecular events in cells, organisms and ecosystems. There are many experimental designs that interrogate gene expression (transcriptomics, metatranscriptomics, single-cell sequencing), chromatin structure (ChIP-seq, Hi-C, etc), protein function (SELEX, CRISPR-activity), and microbial community structure and function (16S rRNA tag sequencing, metagenomics). New methods are being developed that will examine spatial and temporal gene expression in cells, tissues and organisms. All these methods start with the design-specific production of a 'library' of short fragments of DNA, or RNA converted to DNA, which are then sequenced on an instrument. In each case, the instrument generates counts of the fragments in the input library up to the limit of the instrument which varies between hundreds of thousands and billions of fragment counts. However, each instrument has an upper bound.

For this reason the instruments generate 'count compositional' data; discrete fragment counts with a meaningless upper limit(Lovell, Müller, Taylor, Zwart, and Helliwell 2011).

Despite the uniform way that data is generated, the analytic tools for each experimental design developed organically and episodically. For example, the earliest approaches to characterize transcriptome data depended on a modified proportion, termed the RPKM, or reads per kilo base per million, where the total number of fragment counts per gene was normalized by both the gene length and the total counts in the sample and then multiplied by 1 million (Mortazavi, Williams, McCue, Schaeffer, and Wold 2008). Serious deficiencies in this approach were observed early on leading to the concept of 'count normalization' where a pseudo standard was chosen and the values in each sample were then divided by a normalization constant. The various normalization methods differed mainly in how they choose the pseudo standard. As another example, the microbiome field used rarefaction, or down sampling, to normalize the total number of counts in each sample (Hughes and Hellmann 2005). Anomalies in the results observed with rarefied data suggested that the count normalization approaches used in transcriptome analysis might be useful (McMurdie and Holmes 2014), and a comprehensive study showed that some microbiome datasets worked well with rarefaction, others with count normalization and others or with compositional approaches (Weiss, Xu, Peddada, Amir, Bittinger, Gonzalez, Lozupone, Zaneveld, Vázquez-Baeza, Birmingham, Hyde, and Knight 2017).

Multiple groups suggested that compositional data approaches should be universal tools for analyzing HTS datasets (Lovell *et al.* 2011; Fernandes, Reid, Macklaim, McMurrough, Edgell, and Gloor 2014; Quinn, Erb, Gloor, Notredame, Richardson, and Crowley 2019), but uptake has been sporadic. In part this is because of a lack of a standard of truth in many of the datasets. Once a result is in the public sphere any new approach, even if formally correct, may not be taken up by the community unless it reproduces at least some of the previous (erroneous) results.

Fundamentally, compositional data exists not in Euclidian space which contains the same number of dimensions as parts but on the Simplex which contains one less dimension than the the number of parts (Aitchison 1982). For example, a three part dataset in Euclidian space with co-ordinates (1,2,2) represents the locations of the data in the x, y and z dimensional space. The equivalent compositional dataset where the only the relative data is relevant can be represented by the co-ordinates 0.2, 0.4, 0.4 (all values divided by their sum). It should be clear that knowing the values of any two parts automatically defines the value of the third, and two dimensions of data are all that is necessary. This property is the fundamental constraint of compositional data, with only the ratios between the parts being relevant. This constraint has long been known by statisticians going back to Pearson (1897) but it took until 1982 for Aitchison (1982) to place the analysis of compositional data on a firm theoretical foundation through the use of log-ratios between the parts. Log-ratio analysis sets up an analysis approach that is parallel to the usual analyses done in Euclidian space, and with proper attention to detail almost all analytic tools can be adapted to analyze compositional data. The use of compositional approaches with a firm theoretical foundation is particularly useful when there is not a ground truth that can be used to evaluate outcomes.

Aitchison (1982, 1986, 1992) defined the criteria that the analysis of compositional data must fulfil based on analogy with the criteria for the analysis of data that exists in Euclidian space. Firstly, any analysis should be scale invariant because compositional data contain only relative information. In other words, a dataset of (1,2,2) is equivalent to a dataset of (10,20,20) because they can both be reduced to the unit dataset of (0.2, 0.4, 0.4) by dividing by the total. There is no non-compositional analogue with scale invariance because non-compositional data cannot be reduced to a unit dataset without loss of information. However, if such a transformation is done-say by converting a dataset of real numbers to proportions-then the dataset has become a composition. Secondly, any analysis should be sub-compositionally dominant; that is, the distances between samples in a sub-composition should be the same as or smaller than the distances between samples in the full composition.

This is the compositional analogue of subspace dominance in non-compositional data. Thirdly, the analysis should be perturbation invariant. This property refers to the change in distance between samples when the data has a systematic perturbation, and is the non-compositional analog of translation invariance. Fourthly, the data should be compositionally coherent, that is the correlation structure of the parts in a subset should be similar to the correlation structure of the whole set of parts. Non-compositional data adheres to this principle. Finally, any analysis should be permutation invariant; that is, the order of the parts should not affect the outcome. In practice, permutation invariance is easy to satisfy, as is compositional dominance, at least approximately. However, scale invariance, perturbation invariance and coherence can be difficult to achieve and can dramatically effect the interpretation of datasets making the results sensitive to particular ways of treating the data. The relative importance of each of these criteria can vary with the dataset and the type of analysis being conducted and scenarios where each criteria could result in a spurious result are given in the examples below.

It is not always apparent when the combination of a particular dataset and transformation fail to achieve the ideal properties laid out by Aitchison, and previous work showing that most approaches are not compositionally appropriate (Palarea-Albaladejo, Martìn-Fernàndez, and Soto 2012) can be difficult to conceptualize for the HTS community. The purpose of this report is to provide a simple toolbox that can be used to determine if a transformation is likely to give sensible and robust answers in a given HTS dataset. This toolbox is composed of the alc R package that can be used by those familiar with the command line and an R shiny app that can be called from within alc using the alc.runExample() command. The hope is that this toolkit can become part of a standard workflow that will be both educational and useful.

2. Methods and data

2.1. Data characteristics

Datasets were collected from a variety of sources and are all publicly available with the sources and availability given in the Table 1.

Table 1: Datasets and sources

Name	File	In aIc	Group size	ENA/SRA Accession	Source
transcriptome	transcriptome.tsv	yes	48, 48	PRJEB5348	Schurch 2016aa
single-cell	singleCell.tsv	yes	1000, 1000	N/A	Skinnider 2019
meta-transcriptome	mtsc.tsv	yes	8, 10	PRJEB31833	Wu 2021
16S rRNA	meta16S.tsv	yes	198, 161	SRP107602	Bian 2017
SELEX	SELEX.tsv	no	7, 7	N/A	McMurrough 2014

All datasets were generated from one of the Illumina HTS platforms using a variety of library preparation methods that are detailed in the references for each dataset and summarized below and in Table 1. Each dataset was used as an input for the aldex.effect() function from the ALDEx2 R package and the 'diff.win' and 'rab.all' values were plotted (R Development Core Team 2022; Fernandes *et al.* 2014; Gloor, Macklaim, and Fernandes 2016a) and are shown in Figure 1. The formula for calculating these two parameters is given by Fernandes *et al.* (2014). Simply put, the 'diff.win' is measuring the dispersion of the log-ratio of each part and differs from the standard deviation by a simple scaling factor in a Normal distribution, but provides sensible values even in skewed or multimodal data. The 'rab.all' value is the mean of the log-ratio of each part. Here it is clear that the relationship between dispersion and relative abundance varies greatly by dataset. This is important because the normalizations employed in analyzing high throughput sequencing datasets assume a dispersion v. abundance relationship like that observed for the transcriptome dataset (Anders and Huber 2010;

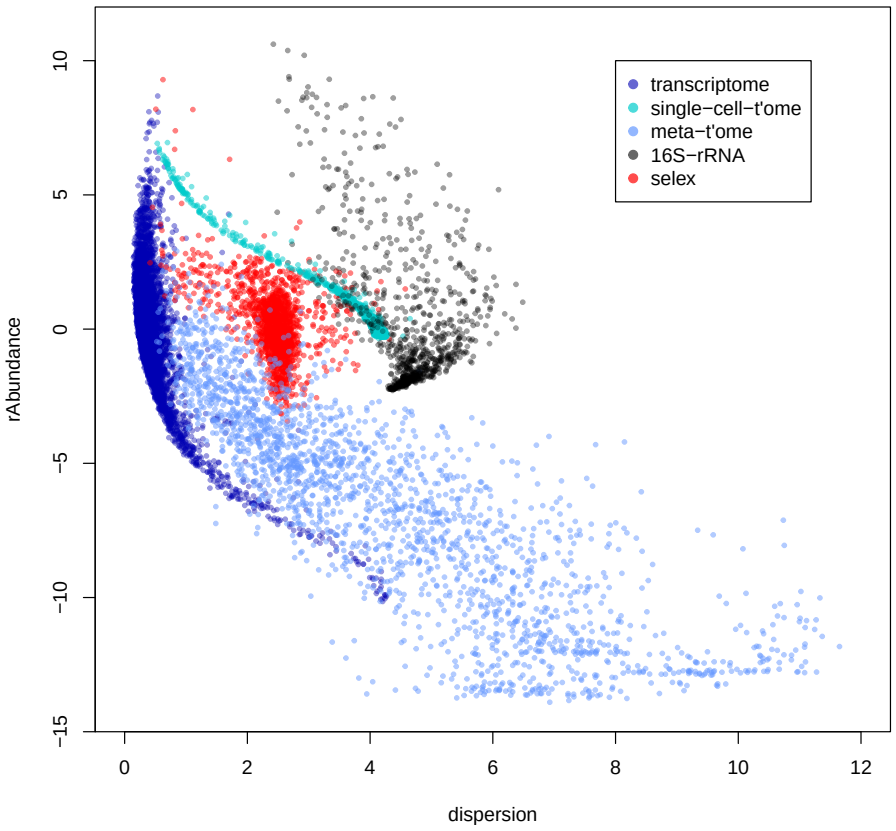

Figure 1: Dispersion vs. relative abundance characteristics of different high throughput sequencing datasets. Publicly available high throughput sequencing datasets were processed using the aldex.effect function from the ALDEx2 R package to estimate the within-group dispersion and relative abundance. These two variables were plotted on the x and y axes and are color-coded by their data type which is expanded on in the text. Abbreviations: t'ome-transcriptome; rRNA-ribosomal RNA; selex-in vitro selection experiment

Robinson and Oshlack 2010). It should be obvious that the application of a tool that expects such a predictable relationship will often fare poorly in datasets that have less dependence between the parameters such as is seen in the meta-transcriptome dataset or even has no discernible relationship as shown for the SELEX dataset.

The transcriptome data was generated by Schurch, Schofield, Gierliński, Cole, Sherstnev, Singh, Wrobel, Gharbi, Simpson, Owen-Hughes, Blaxter, and Barton (2016) and processed into a read table in house as described by Wu, Macklaim, Genge, and Gloor (2021). The raw reads for single-cell transcriptome dataset was from Zheng, Terry, Belgrader, Ryvkin, Bent, Wilson, Ziraldo, Wheeler, McDermott, Zhu, Gregory, Shuga, Montesclaros, Underwood, Masquelier, Nishimura, Schnall-Levin, Wyatt, Hindson, Bharadwaj, Wong, Ness, Beppu, Deeg, McFarland, Loeb, Valente, Ericson, Stevens, Radich, Mikkelsen, Hindson, and Bielas (2017), and processed into a read table by Skinnider, Squair, and Foster (2019). The data used here contrasts cells identified as cytotoxic T-cells and memory T-cells and includes only cells above the upper quartile of total reads, and includes only reads with an average read count greater than 0.11 across all samples. This was further reduced to 1000 cells at random for each group to keep the dataset manageable. Code for these filtering steps are in make-single-cell.R at http://github.com/ggloor/datasets. The meta-transcriptome dataset was described in (Macklaim and Gloor 2018; Wu *et al.* 2021). The 16S rRNA gene sequencing dataset is

a subset that compares pupils to centenarians from (Bian, Gloor, Gong, Jia, Zhang, Hu, Zhang, Zhang, Zhou, Zhang, Burton, Reid, Xiao, Zeng, Yang, and Li 2017). Finally, the SELEX dataset is the full dataset from (McMurrough, Dickson, Thibert, Gloor, and Edgell 2014) and is included as the test dataset in the ALDEx2 R package available from the Bioconductor repository (Gentleman, Carey, Bates, Bolstad, Dettling, Dudoit, Ellis, Gautier, Ge, Gentry, Hornik, Hothorn, Huber, Iacus, Irizarry, Leisch, Li, Maechler, Rossini, Sawitzki, Smith, Smyth, Tierney, Yang, and Zhang 2004). All these read count tables, excepting the SELEX dataset, are included as the alc R package as datasets, and the group sizes needed to determine normalizations per group are given in Table 1.

2.2. Data normalization

Data as collected from a high throughput sequencing instrument results from several processing steps that occur prior to sequencing, while on the instrument and following raw data collection. The data as collected are short sequence fragments of about 100 characters known as reads which are further processed by trimming, filtering, subsetting and binning. McLaren, Willis, and Callahan (2019) summarized how each of these steps can introduce bias in the detection, abundance and integrity of the parts (gene, function, species, etc) being measured.

The sequencing instrument itself imposes a limit on the total number of reads collected, and the total available reads are distributed amongst the number of samples being examined (Lovell *et al.* 2011). The random distribution of reads amongst samples means that even an identical replicate of a sample can have a substantially different total number distributed between the parts. This property of HTS led early on to the idea of 'library size' or normalization being needed to compare the read counts per part across samples and experimental replicates (Mortazavi *et al.* 2008). Broadly, there are two main ways of normalization. First, by normalizing only within a sample, that is each sample is normalized independently and the proportion, clr, iqlr and RPKM approaches are in this class. Second, by normalizing between samples, that is a reference is chosen external to the samples and all samples are normalized to a reference. The Trimmed mean of M values, hereafter the TMM (Robinson and Oshlack 2010), and the relative log expression scaling , hereafter the RLE (Anders and Huber 2010), methods being in this class.

Proportions and similar corrections: If we denote the raw read counts of a part p in sample j as y_{pj} with D being the total number of parts in a sample, then the total number of reads n for sample j is the sum across all parts

$$n_j = \sum_{p=1}^{D} y_{pj}. \tag{1}$$

The total counts observed across all samples is the sum of all n_j samples that were on the sequencing run, with the values of n_j being distributed around some location. The actual values of n_j have no meaning and are nuisance parameters. It was realized early on that in order to compare the values of y_{pj} with y_{pk} that it was necessary to scale or normalize these values in some way.

The simplest approach to normalization is the proportion

$$prop_{pj} = \frac{y_{pj}}{n_j}. \tag{2}$$

The realization that proportions were not real numbers and that the number of reads per part depended not just on its abundance in the input, but also its length led to further corrections such as reads per kilo base per million (RPKM) (Mortazavi *et al.* 2008), transcripts per million (TPM) (Wagner, Kin, and Lynch 2012) and others. Fundamentally, though these normalizations are simply proportions scaled by one or two constant values and are expected to have similar properties as the proportion.

Two commonly used normalizations (RLE, TMM) normalize each sample to a reference or pseudo-reference sample under the assumption that the majority of parts in the samples are invariant(Anders and Huber 2010; Robinson and Oshlack 2010). Thus, these transformations attempt to transform the read counts by using information about parts *across* samples.

The relative log expression (RLE) normalization: This normalization is widely used for both transcriptome and microbiome analysis and originated in the DESeq R package (Anders and Huber 2010). It was developed because obvious problems with proportions and other normalizations were observed. Thus, the focus changed to determining a size factor s with the goal of making data across samples comparable. The intention was that the common size would be relatable in some way to the actual values in the pre-sequencing samples. A key assumption was that the majority of parts in the underlying environment were invariant (or varied only by random effects). If this were true, then determining the set of invariant parts would allow the underlying non-compositional data to be exposed.

The RLE normalization determines a scaling factor by first determining the geometric mean of each part across all N samples $g_p = (\Pi_{j=1}^{N} y_{pj})^{\frac{1}{N}}$; this is a pseudo reference. The method proceeds by then calculating the ratio $r_{pj} = \frac{y_{pj}}{g_p}$ and for sample j determining the median value of $\mathbf{r}_{*j}$, and so on. This is the scaling factor for sample j and the final scaled counts of $\mathbf{s}_j$ is given by

$$s_{pj} = \frac{y_{pj}}{med(r_{*j})}. \tag{3}$$

By design, $med(r_{*j})$ is close to 1 so that the scaled values appear to be positive real numbers, but in fact the values are ratios between the count and the geometric mean value determined for each sample. Parts with 0 counts do not contribute to g_p in current versions of the software, but all non-zero counts are always scaled by this constant.

The trimmed mean of M values (TMM) normalization: A second widely used approach to determining scaling factors is the TMM method used in the edgeR R package for differential abundance (Robinson, McCarthy, and Smyth 2010; Robinson and Oshlack 2010). The TMM is also calculated in a multi-step process that first identifies a reference sample y_r as the one where the parts in the upper quartile of count values are the closest to the mean of all samples. Then parts that are between the 30^{th} and 70^{th} decile of log-ratio between the samples in the observation y_{pj} and reference group y_{pr}, and that are in between the bottom 5% and the top 95% abundance of the read count per sample. This set of parts $P*$ can then weighted by a variance function for each part $w_{pj}^{(r)}$ and the log-ratio sum is determined. The equation summarizing these steps from the supplement of Quinn, Erb, Richardson, and Crowley (2018) is reproduced below:

$$TMM_j = \sum_{P*} w_{pj}^{(r)} \log_2 \frac{y_{pj}}{y_{pr}}. \tag{4}$$

There is an alternative description in Maza (Maza, Frasse, Senin, Bouzayen, and Zouine 2013) that compares the steps involved in calculating both the RLE and TMM normalization. As with the RLE normalization, the count values values are then divided by the normalization factor, which again by design is close to 1, to output the scaled counts. The constraint that both y_{pj} and $y_{pr} > 0$ is enforced when calculating the scaling factor but all parts are scaled by TMM_j.

While both the RLE and TMM values are presented and used as counts for downstream processing of differential abundance it should be clear that they are actually ratios, and many analyses are done with the logarithm of those ratios. Quinn *et al.* (2018) argue that after a logarithm is taken the RLE and TMM normalized counts should have similar properties to centre log-ratio transformed values. However, the clr and derived transforms only use information about parts *within* samples and not *across* samples; this is a fundamental difference

between the clr and other widely used transformations. We shall see that these normalizations are not equivalent in practice.

The centred log-ratio transformation (clr) and similar: The centred log-ratio was introduced in the discussion following (Aitchison 1982) as the ratio between y_{pj} and the geometric mean of all parts of sample j

$$clr_{pj} = \frac{log(y_{pj})}{(\Pi_{p=1}^{D} y_{pj})^{\frac{1}{D}}}. \tag{5}$$

A related transform is the interquartile log-ratio (iqlr) that was introduced to help centre the data when the dataset contained an asymmetry between the groups (Wu *et al.* 2021). It differs from the clr in that only a subset of the parts are used to determine the geometric mean used as the denominator for the clr calculation. In the case of the iqlr, the parts are chosen as those that have a log-ratio variance between the first and third quartile of the data (the interquartile range of a standard boxplot), determined on a per-group basis and the intersect is take between the parts in the two groups.

2.3. Tests and parameters

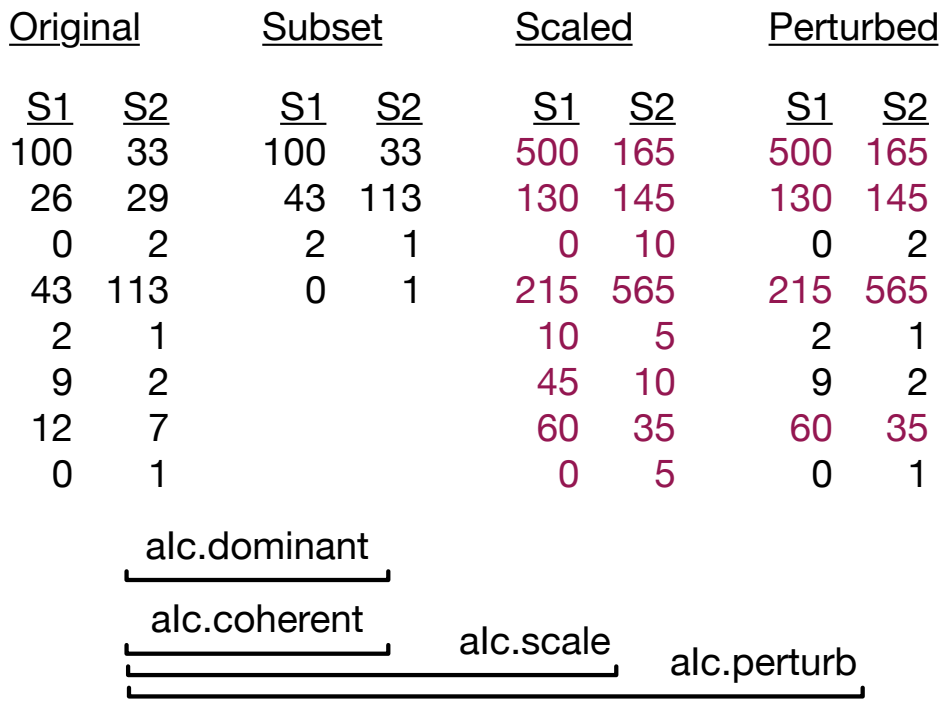

Figure 2: Graphical explanation of the data manipulations. Subsetting removes parts at random from each sample. Scaling multiplies each part by a constant in all samples. Perturbation multiplies a subset of data by a constant in all samples. The results of distance tests between sample 1 and sample 2 (S1, S2) are compared to the distances in the original dataset for dominance, scaling and perturbation tests. The results of the correlation for the parts in common between the original and subset are compared for the coherence test.

We employed the tests outlined in Palarea-Albaladejo *et al.* (2012) to examine the properties of each of the transforms in multiple datasets on distances between the samples and followed Lovell, Pawlowsky-Glahn, Egozcue, Marguerat, and Bähler (2015) to examine correlation. Figure 2 shows a summary of the data modifications and which tests were applied to each modification.

2.4. Availability of data and code

All data and code are available at https::/github.com/ggloor/amIcomp. The code, including the shiny app is available in the alc R package is available on the Comprehensive R Archive Network (Chang, Cheng, Allaire, Sievert, Schloerke, Xie, Allen, McPherson, Dipert, and Borges 2022). Datasets are included in either the alc or ALDEx2 R packages available on CRAN or Bioconductor.

3. Results

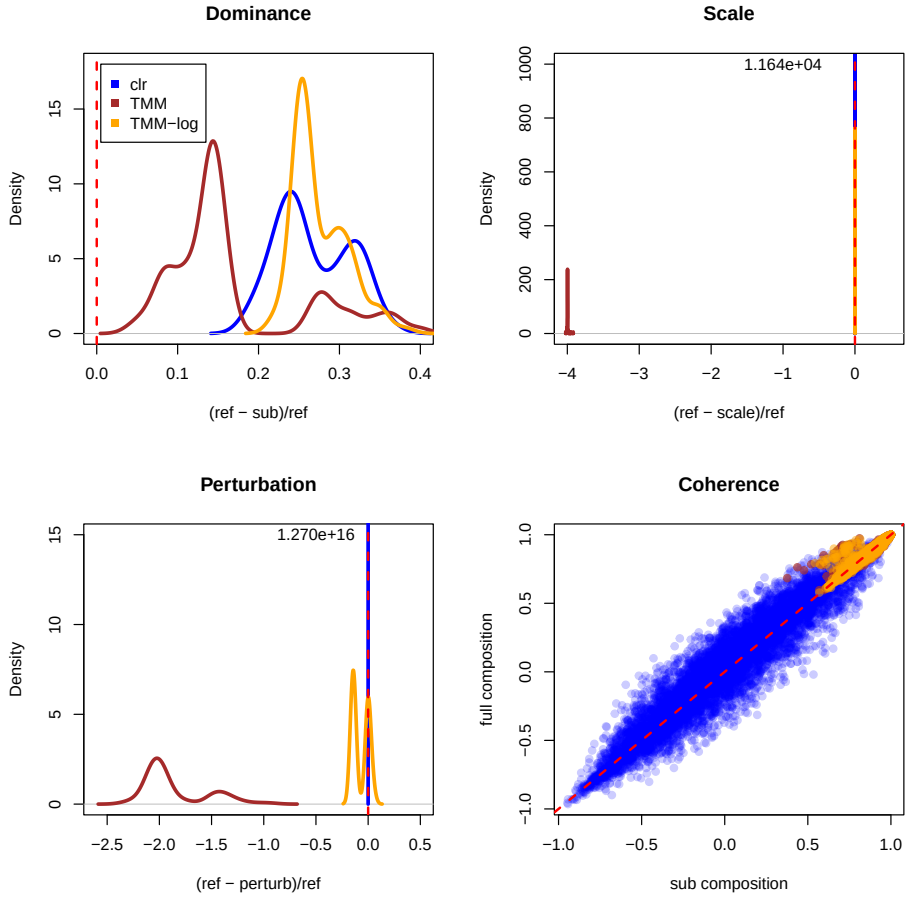

Figure 3: Graphical output from alc tests on the SELEX dataset. The outputs of the four main tests of alc are shown here with the location of ideal behaviour given in each panel as the dashed red line. The alc.dominant test determines if the distances in a full dataset are equal to or larger than the distances in a subset of the dataset. The alc.scale test determines if the distances between parts in a scaled version of the data is substantially similar to the unscaled version. The alc.perturbation test determines if distances between parts in the perturbed dataset is substantially similar to the unperturbed dataset. The alc.coherence test determines if the correlations between parts in common are similar in the full dataset or a subset of the dataset. In the ideal case, the correlations should be identical. Values in the upper part of the Scale and the Perturbation tests graphs represent the maximum density value for the clr transformed test. The extremely large maxima indicate that the distribution around 0 is very narrow.

For each dataset we examine the effect of each normalization on the outcomes of the four tests. The TMM and RLE normalizations are used to generate 'normalized counts' across samples for differential abundance analysis which uses a Negative Binomial fit to these data

(Anders and Huber 2010; Robinson *et al.* 2010), yet some uses of these outputs such as dimension reduction and clustering often use logarithms of the normalized values. Thus, we report results for both the original and after a logarithm is taken for each dataset.

Figure 3 shows the graphical output for each of the tests. Three of these tests examine the effect of data normalizations on distances between samples, and the fourth test examines the effect of data normalizations on correlations. These examples use the SELEX dataset generated by McMurrough *et al.* (2014). This dataset is useful for a number of reasons. First, by design, all samples and parts in the dataset are independent. Thus there should be no internal correlation structure in the data. Second, the dataset has only a single directional change in both relative and absolute abundance. Third, this dataset has a known standard of truth derived both from phylogenetic inference and from direct biochemical observation. For simplicity Figure 3 shows a graphical summary of the output of each test using only three transformations; the clr, TMM and the logarithm of the TMM transform.

Subcompositional dominance determines the effect of subsetting the data on the distances between samples, and is the compositional analog of subspace dominance in Euclidian space. Changing the data by subsetting corresponds to reducing the number of dimensions in the data. In order to pass this test Euclidian distances between samples in the subset should always be equal to or less than the distance between the samples in the complete dataset (Aitchison 1992). More formally, if we have the full composition X and a sub composition x distances $d(X_i, X_j) >= d(x_i, x_j)$. All high throughput datasets are subsets either derived by subsetting at the initial data collection step–rRNA depletion for transcriptomes, amplification bias for 16S rRNA gene sequencing–or by subsetting during the computational pipeline or both. It is also common to remove parts that are present in only a small percentage of the samples. The test employed here removes 50% of the parts at random and compares distances between samples in both the original and the subset data. It is expected that Euclidian distance based on clr transformed data, i.e, the Aitchison distance, will always be sub compositionally dominant and so serves as a sanity check on the methods.

The 'Dominance' test in Figure 3 outputs the distance between samples of the subset relative to the starting distances between samples. Success in this test can be summarized as the proportion of pairwise distances between samples that are larger in the full set than in the subset. A single number summary of this graphic is the proportion of reference distances that are greater than the distances in the sub composition for parts in common. In the example shown all three transforms pass the dominance test. Table 2 shows this summary statistic for the five different datasets. Sub-compositional dominance seems to be relatively easy to achieve for most datasets except for the 16S rRNA dataset when the RLE, TMM normalizations are used. In general, raw proportions are not sub-compositionally dominant as expected (Palarea-Albaladejo *et al.* 2012).

Table 2: Tests of compositional dominance

Dataset	prop	prop-l	clr	iqlr	RLE	TMM	RLE-l	TMM-l
SELEX	0.09	1	1	1	1	1	1	1
16S	0.05	1	1	1	0.81	0.53	1	1
tsc-ome	0.04	1	1	1	1	1	1	1
single cell	0	1	1	1	1	1	1	1
meta-tsc	0.37	1	1	1	1	1	1	1

Scale invariance determines the effect of changing the scale of the data (Aitchison 1992). In the context of high throughput sequencing this property is desirable because we would like to make congruent inferences regardless of the sequencing depth. Indeed, the need for scale invariance is one of the pillars underling the logic of all transforms used to analyze HTS data(Robinson and Oshlack 2010). Formally, if we have the original composition X and a composition x that X multiplied by a constant (5), then distances are scale invariant if

$d(X_i, X_j) = d(x_i, x_j)$. The 'Scale' test that is graphically shown in Figure 3 measures the distances between samples for the dataset and a scaled version of the same dataset. This corresponds to situations where samples have wildly divergent library sizes or where the data was pooled across different sequencing runs. In this test success occurs when there is no significant difference between the distances; graphically this is represented by the density of the output being tightly grouped around 0 difference. This test is passed by both the clr as expected, and the logarithm of the TMM, but is failed by the raw TMM values which show an approximate 4-fold inflation of sample distances in the scaled dataset. This is not surprising because the raw TMM is merely an adjustment of the count values and a 5-fold change in scale is expected. A single number summary of this test is the maximum deviation from 0 value on the graph, and Table 3 shows this summary for all the tests and datasets. In practice, the scale invariance test is more difficult to pass than is the dominance test. The raw TMM, RLE and proportion consistently failed this test. Unexpectedly the TMM showed a 17-fold change in distance for the 16S rRNA dataset. Note that even the logarithm of the TMM normalization showed less than ideal behaviour with some samples having up to 32% difference from the desired outcome of no deviation in some datasets. This has obvious implications for distance-based analyses such as clustering and ordination.

Table 3: Tests of scale invariance

Dataset	prop	prop-l	clr	iqlr	RLE	TMM	RLE-l	TMM-l
SELEX	0.01	0	0	0.02	4.03	4.02	0	0
16S	0.01	0	0	0.03	4.02	17.2	0	0.32
tsc-ome	0.02	0	0	0	4.04	4.05	0	0
single cell	0	0	0	0.01	4.02	4.84	0	0.03
meta-tsc	0	0	0	0.01	4.01	4.05	0	0

Perturbation invariance determines the effect of a systematic change on the distances between samples when only a subset of the parts have scaled values; this is the compositional analog of translation invariance in Euclidian space (Aitchison 1992). In the context of HTS this corresponds to some parts being more (or less) easily observable because of upstream collection and processing steps and could correspond to an amplification bias or selection bias for some parts (McLaren *et al.* 2019). The test employed here compares the original dataset and a dataset where the most abundant 50% of the parts are perturbed by a factor of 5.

The 'Perturbation' test shown graphically in Figure 3 measures the distances between samples for the whole dataset and for the perturbed version of the same dataset. The ideal situation is that the distances between samples is not affected by perturbation since this corresponds to a simple translation of the samples from one location in space to another without changing the relationship between the samples. Again the ideal behaviour is no change in distance between the samples in the original and the perturbed datasets. Formally, if we have the original composition X and a perturbed composition x that has arbitrary values of X multiplied by a second composition, then distances are perturbation invariant if $d(X_i, X_j) = d(x_i, x_j)$. As can be seen in Table 4 this test is very problematic for almost all normalizations in almost all datasets. Here we can see that only the clr and the closely related iqlr transforms are reliably perturbation invariant. All other transformations can have arbitrarily large changes in the distances between samples upon perturbation of the data ranging from a 1% change to over a 6-fold change in distance. The raw RLE, TMM transforms behave very poorly, and log transformation improves the outcome somewhat from 2- to 4-fold changes in distances to double or single digit percentage differences in distance depending on the dataset and transformation. The actual variance from ideal behaviour appears to be unpredictable regarding the transformation and the dataset. Again, such a pathology will have obvious implications for distance-based analyses such as clustering and ordination.

Sub-compositional consistency (coherence) determines the effect of subsetting the data on the

Table 4: Tests of perturbation invariance

Dataset	prop	prop-l	clr	iqlr	RLE	TMM	RLE-l	TMM-l
SELEX	0.51	0.33	0	0	2.3	2.34	0.16	0.16
16S	1.76	0.31	0	0	3	6.15	0.33	0.52
tsc-ome	0	0	0	0	4.01	4.02	0	0
single cell	1.54	0.05	0	0	2.04	2.04	0.05	0.05
meta-tsc	0.2	0.02	0	0	4.26	5.08	0.02	0.04

Table 5: Tests of correlation coherence

Dataset	prop	prop-l	clr	iqlr	RLE	TMM	RLE-l	TMM-l
SELEX	0.98	0.98	0.96	0.92	0.97	0.97	0.97	0.97
16S	1	1	0.99	0.99	1	0.98	1	0.98
tsc-ome	1	1	1	1	1	1	1	1
single cell	1	1	1	1	1	1	1	1
meta-tsc	0.98	0.98	1	0.99	0.99	0.98	0.99	0.98

correlation between parts using the same subsetting as the dominance test. In the context of HTS it is common to filter parts on either relative or count abundance to remove those near the low count margin, and to filter by occurrence to identify parts that are in a plurality or majority of the samples. The test employed here removes 50% of the parts at random and compares Pearson or Spearman correlation coefficients between variables in both the supplied and the data subset; the example shown here uses Pearson correlation between the parts. An alternative test by (Greenacre 2011) can be used, but it is not as efficient for large datasets. Note that this is *not testing* the correlation between the pre- and post-sequencing data; i.e. between the underlying counts and what the instrument returns which is known to be non-reproducible (Friedman and Alm 2012; Lovell *et al.* 2015). Here, the ideal situation would be a correlation coefficient of the correlations calculated between the parts in common to be 1. Again Figure 3 shows a graphical example. Interestingly, the correlation between the parts in the SELEX dataset the TMM and the logarithm of the TMM look extremely similar and are clustered around strong positive correlation. In contrast, the correlation between the parts in the SELEX data are dispersed and the clr transformed analysis shows a marginal distribution that is non-skewed and strongly platykurtic. In practice, the correlation between correlation coefficients in the complete and subset data is very high for all tests, but the investigator should characterize the underlying correlation graphically to guard against the spurious observation that was seen in the SELEX dataset with the TMM normalization. Examination of the correlation plots of the other datasets shows that the uniform high correlation observed for TMM in the SELEX dataset is unusual and that the correlation plots more closely resemble that observed for the clr transform for all other datasets and transforms. This suggests that the TMM normalization is introducing some unwanted correlation structure in this transformed dataset.

Skinnider *et al.* (2019) noted that correlations in HTS data are problematic because of inconsistency and surprisingly low sensitivity for grouping known biological associations. Erb (2020) suggested that partial correlation coefficients of clr transformed data provide both consistency and higher accuracy, although this was not tested here. Thus, given the broad agreement between the subset and the full dataset for correlation within each sample and normalization, we examined the correlations within each dataset across normalizations. The results are summarized in Figure 4. We observed that the consistency of the correlation coefficients varies widely by dataset. Surprisingly, inter-normalization correlation was not associated with intra-dataset variance as the 16S rRNA dataset has high dispersion while the single cell transcriptome has relatively low dispersion, yet both datasets show high inter-

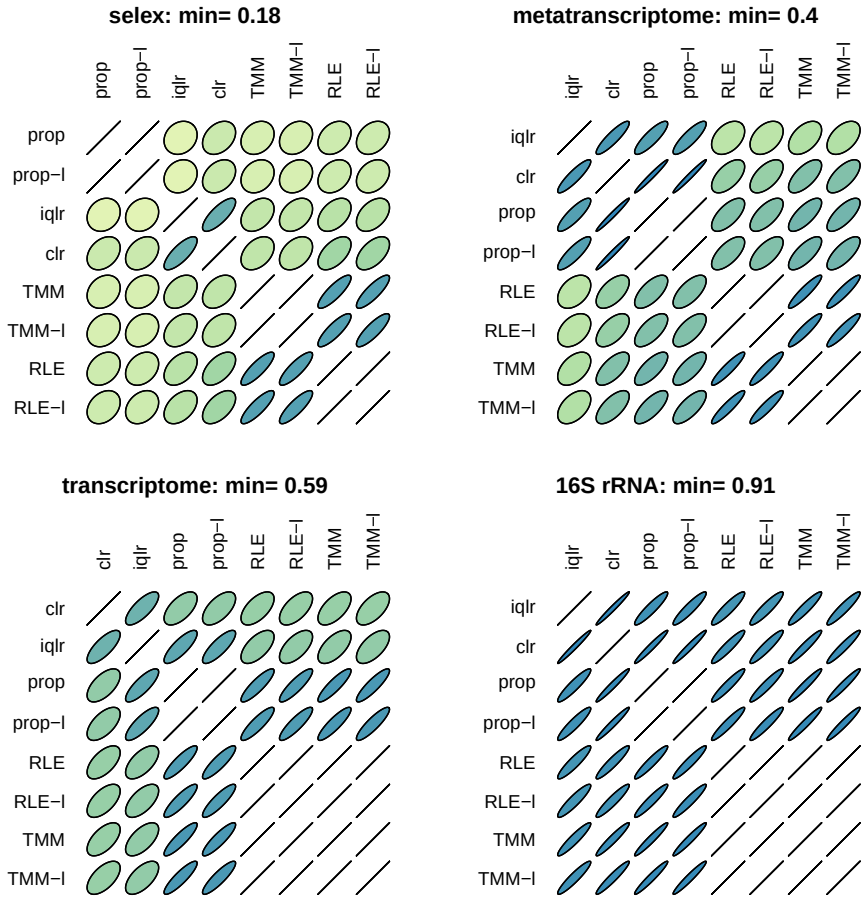

Figure 4: Correlation within datasets between normalizations. The correlation of correlation coefficients for pairs of features in the noted datasets are shown using the ellipse R package (Murdoch and Chow 2022). These range from a low of 0.18 within the SELEX dataset to very consistent correlations in the 16S rRNA and single cell transcriptome datasets. The datasets are arranged from lowest to largest number of samples. Sample sizes given in Table 1

normalization correlation. On the other hand, consistency seemed to be associated with group size. Consistency was lowest for the SELEX and metatranscriptome datasets which have small sample sizes; SELEX contains 7 samples in each group, and metatranscriptome has group sizes of 7 and 10, while the moderately sized transcriptome dataset with 48 samples per group had modest correlation, and the large 16S rRNA dataset had very high inter-normalization correlation. In keeping with the sample size being the primary driver of this correlation, the very large single cell transcriptome dataset with a sample size of 1000 in each group had an even higher minimum inter-normalization correlation of 0.92, but the results are not shown. We conclude that precision, but not accuracy can be achieved with very large sample sizes when estimating correlation.

Finally, we examined the effect of zero replacement strategies on the robustness of this analysis. All analyses to this point used a uniform prior; that is adding 0.5 to all parts in all samples as this is known to minimally distort the data with and is very efficient (Fernandes, Macklaim, Linn, Reid, and Gloor 2013; Gloor, Macklaim, Vu, and Fernandes 2016b). For this comparison, we chose the 16S rRNA dataset because this was the only dataset that did not show compositional dominance with all normalization strategies. Thus, we would have the opportunity to observe if one of the advanced zero replacement strategies used by the zCompositions R package performed differently than the uniform prior approach. The results are shown in Table 6. As expected, none of the approaches gave meaningful results with no

zero replacement when taking the logarithm of the ratios. The log-ratio between parts with count 0 and any non-zero part will be ± infinity. All the non log ratio-approaches performed poorly with no zero replacement. However surprisingly neither the GBM nor the CZM zero imputation approach performed appreciably differently than did the uniform prior method; all three methods of zero replacement gave very similar results. This is likely because of the wide, sparse and high-dimensional nature of the data where the parts with 0 values are found in all samples and because of the large difference between the log-ratio of any imputed value and infinity.

Table 6: Tests of zero replacement

Comp Test	0 replace	prop	prop-l	clr	iqlr	RLE	TMM	RLE-l	TMM-l
scale	none	0	NA	NA	NA	4.04	4.44	NA	NA
scale	prior	0.01	0	0	0.03	4.03	17.2	0	0.32
scale	GBM	0.01	0	0	0.04	4.02	4.3	0	0.26
scale	CZM	0	0	0	0.04	4.06	16	0	0.02
dominance	none	0.05	NA	NA	NA	0.75	0.60	NA	NA
dominance	prior	0.05	1	1	1	0.81	0.53	1	1
dominance	GBM	0.05	1	1	1	0.71	0.52	1	1
dominance	CZM	0.05	1	1	1	0.78	0.53	1	1
perturbation	none	1.9	NA	NA	NA	2.2	4.0	NA	NA
perturbation	prior	1.76	0.31	0	0	3	6.15	0.33	0.52
perturbation	GBM	1.86	0.24	0	0	2.8	3.9	0.26	0.57
perturbation	CZM	1.86	0.26	0	0	2.8	3.95	0.3	0.49
coherence	none	0.99	NA	NA	NA	0.99	0.96	NA	NA
coherence	prior	1	1	0.99	0.99	0.99	0.96	1	0.98
coherence	GBM	1	1	1	0.99	0.99	0.97	1	0.99
coherence	CZM	1	1	0.99	0.99	0.99	0.97	1	0.99

4. Discussion and conclusions

High throughput sequencing comes with the constraint that the platform imposes a limit on the number of reads obtained, thus ensuring that the data behave as compositions (Lovell *et al.* 2011; Friedman and Alm 2012; Fernandes *et al.* 2014). Aitchison (1982) identified rational and reproducible approaches to dealing with well-known limitations of compositional data that arise because these data have one less dimension than expected. These limitations are revealed by the four tests (Palarea-Albaladejo *et al.* 2012) used here and the pathologies noted here have caused confusion in the literature. In the context of HTS these data pathologies may or may not manifest depending on the dataset. Several groups have shown that in some instances non-compositional approaches can provide appropriate answers, but in other cases independent non-compositional approaches may give wildly divergent answers (Weiss *et al.* 2017; Weiss, Van Treuren, Lozupone, Faust, Friedman, Deng, Xia, Xu, Ursell, Alm, Birmingham, Cram, Fuhrman, Raes, Sun, Zhou, and Knight 2016; Nearing, Douglas, Hayes, MacDonald, Desai, Allward, Jones, Wright, Dhanani, Comeau, and Langille 2022). Several recent reports show that compositional approaches provide more consistent and more accurate answers, in general, than do non-compositional or partially compositional approaches (Skinnider *et al.* 2019; Nearing *et al.* 2022; Armstrong, Martino, Rahman, Gonzalez, Vázquez-Baeza, Mishne, and Knight 2021).

HTS analysis has several standard steps, and while the actual tools vary between data types, in concept the steps all aim to achieve a dataset and analysis that conforms to the standards of the field. The first step is to collect the data, and regardless of data type, it is usual to collect only a subset of the available data. In the case of RNA-seq (including transcriptome and single cell), this means collecting only the mRNA or other sub-population of RNA from the

cell and discarding the majority. In the case of metagenomics (whether amplified or not) this involves isolating DNA from different species where the DNA will have different efficiencies of isolation, and different propensities to be amplified or further processed downstream. It is standard practice for all HTS data to filter the parts (reads) to remove those that are near the low count margin (frequency filtering) or that occur in only a very small subset of the samples (occurrence filtering). These types of biases result in subcompositions which affect both the distances and the correlations observed meaning that subcompositional dominance and correlation coherence are desirable properties. A major limitation of current workflows is the apparent inability to identify when non-compositional approaches will fail.

The issue of systematic bias in HTS is well-studied across multiple data types. McLaren *et al.* (2019) identified systematic biases in metagenomic datasets and provided a mechanism to adjust for these biases based on logratios between the parts using internal standards. Nearing, Comeau, and Langille (2021) recently reviewed the full suite of systematic biases that can occur in metagenomic datasets and outlined the open challenges in addressing those biases. The early RNA-seq literature contains many studies and reviews that outlined systemic problems that have been addressed by multiple corrections for nucleotide content and multiple normalizations. In short, most of these biases result in compositional perturbations whereby a subset of the parts are systematically observed more or less frequently than expected.

The choice of sequencing platform determines the nominal scale of the sequencing data; for example, the same library can be run on an Illumina MiSeq (20M reads) or an Illumina NovaSeq (2 B reads). Here it is clearly desirable that the only variation should be that rare parts from the former should be estimated with higher precision on the latter (Gloor *et al.* 2016b), and merely changing the read depth should not materially affect the observations for parts that are observed with differing precision. This underlines the importance of the scale invariance property.

The work outlined here shows that HTS data can be relatively predictable when testing for sub-compositional dominance or correlation coherence. For the most part, it may be safe to assume that distances observed will be similar when examining the full and the sub-composition regardless of the data transformation. However, not all datasets and transforms give smaller distances in sub-compositions with non-compositional transforms and this could lead to false inferences unless the analyst specifically tests for sub-compositional dominance.

It may also be safe to assume that methods such as network analysis will give similar results for the parts in common when only a subset of the parts are used if the sample sizes are large. However, the caveat here is that the observed correlations between transforms are not themselves necessarily consistent between normalization methods, and again the investigator should determine if the correlations observed with different normalization methods are congruent. Inspection of these datasets and transformations indicates that the correct answer is sometimes, as shown in Figure 4. Indeed, two recent studies demonstrated just how problematic correlation, and thus network analysis, can be in HTS datasets (Skinnider *et al.* 2019; Erb 2020). Thus, correlation within HTS datasets should be considered an open problem with Erb (2020) suggesting that partial correlation should be explored as the appropriate tool.

In summary, this work provides a framework to examine four different tests that can be applied to datasets to determine if compositional approaches are likely to be more appropriate than non-compositional approaches. The work is supported by the alc R package and amlcomp shiny app that facilitate these tests.

5. Acknowledgements

The author thanks the reviewers for their kind and useful suggestions to improve this work. Early versions of this work benefitted from discussions with participants at CoDaWork2022. The author has no specific funding for this work, and has no conflicts of interest.

References

Aitchison J (1982). "The Statistical Analysis of Compositional Data, London. Reprinted 2003 with additional material by The Blackburn Press, London, UK." *Journal of the Royal Statistical Society: Series B (Methodological)*, **44**(2), 139–160.

Aitchison J (1986). *The Statistical Analysis of Compositional Data*. Chapman & Hall.

Aitchison J (1992). "On Criteria for Measures of Compositional Difference." *Mathematical Geology*, **24**(4), 365–379.

Anders S, Huber W (2010). "Differential Expression Analysis for Sequence Count Data." *Genome Biology*, **11**(10), R106. doi:10.1186/gb-2010-11-10-r106.

Armstrong G, Martino C, Rahman G, Gonzalez A, Vázquez-Baeza Y, Mishne G, Knight R (2021). "Uniform Manifold Approximation and Projection (UMAP) Reveals Composite Patterns and Resolves Visualization Artifacts in Microbiome Data." *mSystems*, **6**(5), e0069121. doi:10.1128/mSystems.00691-21.

Bian G, Gloor GB, Gong A, Jia C, Zhang W, Hu J, Zhang H, Zhang Y, Zhou Z, Zhang J, Burton JP, Reid G, Xiao Y, Zeng Q, Yang K, Li J (2017). "The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young." *mSphere*, **2**(5), e00327–17. doi:10.1128/mSphere.00327-17.

Chang W, Cheng J, Allaire J, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2022). *shiny: Web Application Framework for R*. R package version 1.7.2, URL https://CRAN.R-project.org/package=shiny.

Erb I (2020). "Partial Correlations in Compositional Data Analysis." *Applied Computing and Geosciences*, **6**(6), 100026. doi:doi.org/10.1016/j.acags.2020.100026.

Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB (2013). "Anova-like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." *PLoS One*, **8**(7), e67019. doi:10.1371/journal.pone.0067019.

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB (2014). "Unifying the Analysis of High-Throughput Sequencing Datasets: Characterizing RNA-Seq, 16S RRNA Gene Sequencing and Selective Growth Experiments by Compositional Data Analysis." *Microbiome*, **2**, 15.1–15.13. doi:10.1186/2049-2618-2-15.

Friedman J, Alm EJ (2012). "Inferring Correlation Networks From Genomic Survey Data." *PLoS Computational Biology*, **8**(9), e1002687. doi:10.1371/journal.pcbi.1002687.

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, Tierney L, Yang JYH, Zhang J (2004). "Bioconductor: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**(10), R80. doi:10.1186/gb-2004-5-10-r80.

Gloor GB, Macklaim JM, Fernandes AD (2016a). "Displaying Variation in Large Datasets: Plotting a Visual Summary of Effect Sizes." *Journal of Computational and Graphical Statistics*, **25**(3C), 971–979. doi:10.1080/10618600.2015.1131161. http://dx.doi.org/10.1080/10618600.2015.1131161, URL http://dx.doi.org/10.1080/10618600.2015.1131161.

Gloor GB, Macklaim JM, Vu M, Fernandes AD (2016b). "Compositional Uncertainty Should Not Be Ignored in High-Throughput Sequencing Data Analysis." *Austrian Journal of Statistics*, **45**, 73–87. doi:doi:10.17713/ajs.v45i4.122.

Greenacre M (2011). "Measuring Subcompositional Incoherence." *Mathematical Geosciences*, **43**(6), 681–693. ISSN 1874-8961. doi:10.1007/s11004-011-9338-5. URL http://ws.isiknowledge.com/cps/openurl/service?url_ver=Z39.88-2004&rft_id=info:ut/WOS:000293473100005.

Hughes JB, Hellmann JJ (2005). "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity." *Methods in Enzymology*, **397**, 292–308. doi:10.1016/S0076-6879(05)97017-1.

Lovell D, Müller W, Taylor J, Zwart A, Helliwell C (2011). "Proportions, Percentages, Ppm: Do the Molecular Biosciences Treat Compositional Data Right?" In V Pawlowsky-Glahn, A Buccianti (eds.), *Compositional Data Analysis: Theory and Applications*, chapter 14, pp. 193–207. John Wiley and Sons New York, NY, London. doi:doi.org/10.1002/9781119976462.ch14.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015). "Proportionality: A Valid Alternative to Correlation for Relative Data." *PLoS Computational Biology*, **11**(3), e1004075. doi:https://doi.org/10.1371/journal.pcbi.1004075.

Macklaim JM, Gloor GB (2018). "From RNA-seq to Biological Inference: Using Compositional Data Analysis in Meta-Transcriptomics." *Methods in Molecular Biology*, **1849**, 193–213. doi:10.1007/978-1-4939-8728-3_13.

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*, **6**(6), e25849. doi:10.4161/cib.25849.

McLaren MR, Willis AD, Callahan BJ (2019). "Consistent and Correctable Bias in Metagenomic Sequencing Experiments." *ELife*, **8**. doi:10.7554/eLife.46923.

McMurdie PJ, Holmes S (2014). "Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible." *PLoS Computational Biology*, **10**(4), e1003531. doi:10.1371/journal.pcbi.1003531.

McMurrough TA, Dickson RJ, Thibert SMF, Gloor GB, Edgell DR (2014). "Control of Catalytic Efficiency by a Coevolving Network of Catalytic and Noncatalytic Residues." *Proceedings of the National Academy of Sciences of the United States of America*, **111**(23), E2376–83. doi:10.1073/pnas.1322352111.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008). "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nature Methods*, **5**(7), 621–8. doi:10.1038/nmeth.1226.

Murdoch D, Chow ED (2022). *ellipse: Functions for Drawing Ellipses and Ellipse-Like Confidence Regions*. Western University. R package version 0.4.3, URL https://CRAN.R-project.org/package=ellipse.

Nearing JT, Comeau AM, Langille MGI (2021). "Identifying Biases and Their Potential Solutions in Human Microbiome Studies." *Microbiome*, **9**(1), 113. doi:10.1186/s40168-021-01059-0.

Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N, Jones CMA, Wright RJ, Dhanani AS, Comeau AM, Langille MGI (2022). "Microbiome Differential Abundance Methods Produce Different Results Across 38 Datasets." *Nature Communications*, **13**(1), 342. doi:10.1038/s41467-022-28034-z.

Palarea-Albaladejo J, Martìn-Fernàndez JA, Soto JA (2012). "Dealing with Distances and Transformations for Fuzzy C-Means Clustering of Compositional Data." *Journal of Classification*, **29**, 144–169. doi:https://doi.org/10.1007/s00357-012-9105-4.

Pearson K (1897). "Mathematical Contributions to the Theory of Evolution. – on a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs." *Proceedings of the Royal Society of London*, **60**, 489–498.

Quinn TP, Erb I, Gloor G, Notredame C, Richardson MF, Crowley TM (2019). "A Field Guide for the Compositional Analysis of Any-Omics Data." *GigaScience*, **8**(9). doi:10.1093/gigascience/giz107.

Quinn TP, Erb I, Richardson MF, Crowley TM (2018). "Understanding Sequencing Data as Compositions: An Outlook and Review." *Bioinformatics*, **34**(16), 2870–2878. doi:10.1093/bioinformatics/bty175.

R Development Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Robinson MD, McCarthy DJ, Smyth GK (2010). "edgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data." *Bioinformatics*, **26**(1), 139–40. doi:10.1093/bioinformatics/btp616.

Robinson MD, Oshlack A (2010). "A Scaling Normalization Method for Differential Expression Analysis of RNA-Seq Data." *Genome Biology*, **11**(3), R25.1–R25.9. doi:10.1186/gb-2010-11-3-r25.

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ (2016). "How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?" *RNA*, **22**(6), 839–51. doi:10.1261/rna.053959.115.

Skinnider MA, Squair JW, Foster LJ (2019). "Evaluating Measures of Association for Single-Cell Transcriptomics." *Nature Methods*, **16**(5), 381–386. doi:10.1038/s41592-019-0372-4.

Wagner GP, Kin K, Lynch VJ (2012). "Measurement of mRNA Abundance Using RNA-Seq Data: RPKM Measure Is Inconsistent Among Samples." *Theory in Biosciences = Theorie in den Biowissenschaften*, **131**(4), 281–5. doi:10.1007/s12064-012-0162-3.

Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ, Birmingham A, Cram JA, Fuhrman JA, Raes J, Sun F, Zhou J, Knight R (2016). "Correlation Detection Strategies in Microbial Data Sets Vary Widely in Sensitivity and Precision." *ISME J*, **10**(7), 1669–81. doi:10.1038/ismej.2015.235.

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R (2017). "Normalization and Microbial Differential Abundance Strategies Depend Upon Data Characteristics." *Microbiome*, **5**(1), 27. doi:10.1186/s40168-017-0237-y.

Wu JR, Macklaim JM, Genge BL, Gloor GB (2021). "Finding the Centre: Compositional Asymmetry in High-Throughput Sequencing Datasets." In P Filzmoser, K Hron, JA Martìn-Fernàndez, J Palarea-Albaladejo (eds.), *Advances in Compositional Data Analysis: Festschrift in Honour of Vera Pawlowsky-Glahn*, pp. 329–346. Springer International Publishing, Cham. ISBN 978-3-030-71175-7. doi:10.1007/978-3-030-71175-7_17.

Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, Ziraldo SB, Wheeler TD, McDermott GP, Zhu J, Gregory MT, Shuga J, Montesclaros L, Underwood JG, Masquelier DA, Nishimura SY, Schnall-Levin M, Wyatt PW, Hindson CM, Bharadwaj R, Wong A, Ness KD, Beppu LW, Deeg HJ, McFarland C, Loeb KR, Valente WJ, Ericson NG, Stevens EA, Radich JP, Mikkelsen TS, Hindson BJ, Bielas JH (2017). "Massively Parallel Digital Transcriptional Profiling of Single Cells." *Nature Communications*, **8**, 14049. doi:10.1038/ncomms14049.

Affiliation:

Gregory B. Gloor
Department of Biochemistry
Schulich School of Medicine & Dentistry
The University of Western Ontrio
London, Ontario, Canada
E-mail: ggloor@uwo.ca
URL: https://ggloor.github.io

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 52
July 2023
http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2022-09-22
Accepted: 2023-02-27