frontiers in Microbiology

MINI REVIEW
published: 15 November 2017
doi: 10.3389/fmicb.2017.02224

Microbiome Datasets Are Compositional: And This Is Not Optional

Gregory B. Gloor[1]*, *Jean M. Macklaim*[1], *Vera Pawlowsky-Glahn*[2] and *Juan J. Egozcue*[3]

[1] Department of Biochemistry, University of Western Ontario, London, ON, Canada, [2] Departments of Computer Science, Applied Mathematics, and Statistics, Universitat de Girona, Girona, Spain, [3] Department of Applied Mathematics, Universitat Politècnica de Catalunya, Barcelona, Spain

Datasets collected by high-throughput sequencing (HTS) of 16S rRNA gene amplimers, metagenomes or metatranscriptomes are commonplace and being used to study human disease states, ecological differences between sites, and the built environment. There is increasing awareness that microbiome datasets generated by HTS are compositional because they have an arbitrary total imposed by the instrument. However, many investigators are either unaware of this or assume specific properties of the compositional data. The purpose of this review is to alert investigators to the dangers inherent in ignoring the compositional nature of the data, and point out that HTS datasets derived from microbiome studies can and should be treated as compositions at all stages of analysis. We briefly introduce compositional data, illustrate the pathologies that occur when compositional data are analyzed inappropriately, and finally give guidance and point to resources and examples for the analysis of microbiome datasets using compositional data analysis.

Keywords: microbiota, compositional data, high-throughput sequencing, correlation, Bayesian estimation, count normalization, relative abundance

OPEN ACCESS

Edited by:
Jessica Galloway-Pena,
University of Texas MD Anderson Cancer Center, United States

Reviewed by:
Ionas Erb,
Centre for Genomic Regulation, Spain
Jennifer Stearns,
McMaster University, Canada

****Correspondence:***
Gregory B. Gloor
ggloor@uwo.com

Specialty section:
This article was submitted to Systems Microbiology, a section of the journal Frontiers in Microbiology

Received: *10 July 2017*
Accepted: *30 October 2017*
Published: *15 November 2017*

Citation:
Gloor GB, Macklaim JM, Pawlowsky-Glahn V and Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. 8:2224. doi: 10.3389/fmicb.2017.02224

1. INTRODUCTION

The collection and analysis of microbiome datasets presents many challenges in the study design, sample collection, storage, and sequencing phases, and these have been well reviewed (Robinson et al., 2016). Many methods for the analysis of microbiome datasets assume that sequencing data are equivalent to ecological data where the counts of reads assigned to organisms are often normalized to a constant area or volume. Methods applied include count-based strategies such as Bray-Curtis dissimilarity, zero-inflated Gaussian models and negative binomial models (McMurdie and Holmes, 2014; Weiss et al., 2017).

In an ecological study it is possible for many different species to co-exist, and their absolute abundance may be important. For example, in an area containing only tigers, it is important to know if the population size is sufficient to maintain needed genetic diversity for long-term survival (Shaffer, 1981). However, the abundance of one species may not influence the abundance of another; the area may contain both tigers and ladybugs, and the migration of several ladybugs into the area would not be expected to affect the number of tigers.

The assumption of true independence can not hold in high-throughput sequencing (HTS) experiments because the sequencing instruments can deliver reads only up to the capacity of the instrument. Thus, it is proper to think of these instruments as containing a fixed number of slots which must be filled. Returning to our tiger and ladybug analogy, the migration of ladybugs into an area containing a fixed number of slots that are already filled must displace either tigers or ladybugs from the occupied slots. This analogy extents, without restriction, to any number of taxa, and to any fixed capacity instrument (Aitchison, 1986; Lovell et al., 2011; Friedman and Alm, 2012; Fernandes et al., 2013, 2014; Lovell et al., 2015; Mandal et al., 2015; Gloor et al., 2016a,b; Gloor and Reid, 2016; Tsilimigras and Fodor, 2016). Thus, the total read count observed in a HTS run is a fixed-size, random sample of the relative abundance of the molecules in the underlying ecosystem. Moreover, the count can not be related to the absolute number of molecules in the input sample as shown in **Figure 1**. This is implicitly acknowledged when microbiome datasets are converted to relative abundance values, or normalized counts, or are rarefied (McMurdie and Holmes, 2014; Weiss et al., 2017) prior to analysis. Thus the number of reads obtained is irrelevant, and contains only information on the precision of the estimate (Fernandes et al., 2013). Data that are naturally described as proportions or probabilities, or with a constant or irrelevant sum, are referred to as compositional data. Compositional data contains information about the relationships between the parts (Aitchison, 1986; Pawlowsky-Glahn et al., 2015).

Data about a microbiome collected by high throughput sequencing are often examined under the assumption that sequencing is, in some way, *counting the number of molecules associated with the bacteria in the population*, as illustrated by the top barplot in **Figure 1B**. We can see the difference between counts and compositions by comparing the data for the actual counts for three samples in the top barplot with their proportions in the bottom barplot. Note, that samples 2 and 3 in **Figure 1B** have the same proportional abundances even though they have different absolute counts prior to sequencing. The difference in apparent direction of change is shown in **Figure 1C** and we can observe that the relationship between absolute abundance in the environment and the relative abundance after sequencing is not predictable.

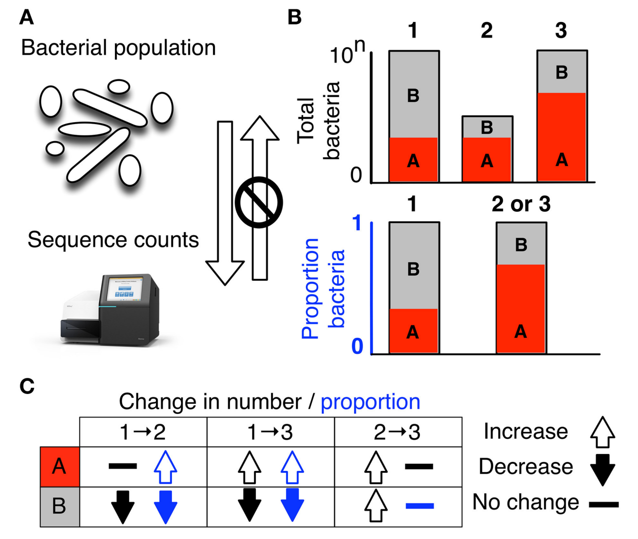


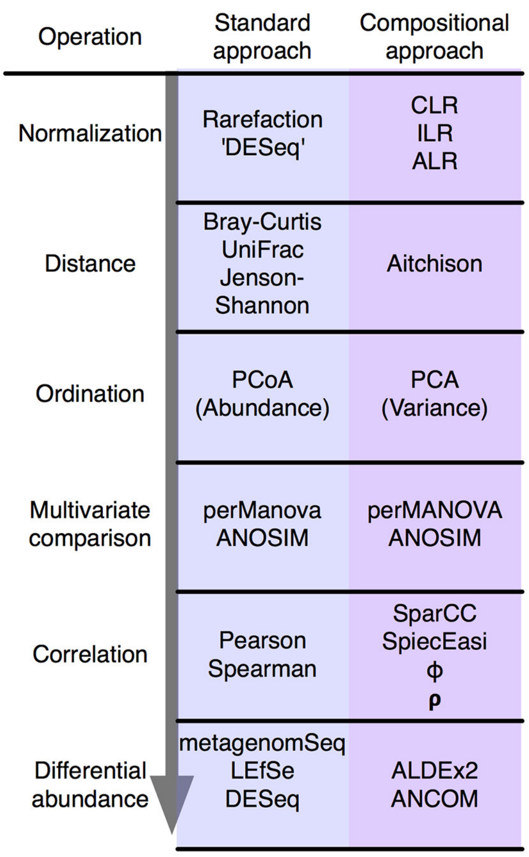
FIGURE 1 | High-throughput sequencing data are compositional. **(A)** illustrates that the data observed after sequencing a set of nucleic acids from a bacterial population cannot inform on the absolute abundance of molecules. The number of counts in a high throughput sequencing (HTS) dataset reflect the proportion of counts per feature (OTU, gene, etc.) per sample, multiplied by the sequencing depth. Therefore, only the relative abundances are available. The bar plots in **(B)** show the difference between the count of molecules and the proportion of molecules for two features, A (red) and B (gray) in three samples. The top bar graphs show the total counts for three samples, and the height of the color illustrates the total count of the feature. When the three samples are sequenced we lose the absolute count information and only have relative abundances, proportions, or "normalized counts" as shown in the bottom bar graph. Note that features A and B in samples 2 and 3 appear with the same relative abundances, even though the counts in the environment are different. The table below in **(C)** shows real and perceived changes for each sample if we transition from one sample to another.

2. PROBLEMS WITH CURRENT METHODS OF ANALYSIS

We will briefly outline the problems that arise when compositional data are examined using a non-compositional paradigm, stepping through the usual stages of analysis shown in **Figure 2**. All these issues have been extensively reviewed and debated in both the older and the more recent literature in fields as diverse as economics, geology and ecology. Thus, rather than present an exhaustive explanation of the problems, we will outline the major issue and cite a few useful resources.

It is very difficult to collect exactly the same number of sequence reads for each sample. This can be because of differences in platform (e.g., MiSeq vs. HiSeq) or because of technical difficulties in loading the same molar amounts of the sequencing libraries on the instrument, or because of random variation. The total number of counts observed (often referred to as read depth) is a major confounder for distance or dissimilarity calculations for multivariate ordinations derived from these distances (McMurdie and Holmes, 2014). Initial attempts in the microbiome field used "rarefaction" or subsampling of the read counts of each sample to a common read depth to attempt to correct this problem (Lozupone et al., 2011; Wong et al., 2016). The use of subsampling has been questioned since it results in a loss of information and precision (McMurdie and Holmes, 2014), and the practice of count normalization from the RNA-seq field has been advocated instead. There are a number of count normalization methods used and two, the trimmed mean of *M* values (TMM) (Robinson and Oshlack, 2010), and the median method (Anders and Huber, 2010) are similar to a log-ratio transformations, but are less suitable in highly asymmetrical or sparse datasets (Fernandes et al., 2013; Gloor et al., 2016a). These transformations are further undesirable since the number of counts observed by the instrument, by design, can not contain any information on the actual number of molecules in the environment, and because the investigator naturally interprets the results as counts instead of log-ratios.

One of the first analysis steps in a traditional analysis, following rarefaction or count normalization, is the calculation of a distance or dissimilarity (DD) matrix from the data that is used



Operation	Standard approach	Compositional approach
Normalization	Rarefaction 'DESeq'	CLR ILR ALR
Distance	Bray-Curtis UniFrac Jenson-Shannon	Aitchison
Ordination	PCoA (Abundance)	PCA (Variance)
Multivariate comparison	perManova ANOSIM	perMANOVA ANOSIM
Correlation	Pearson Spearman	SparCC SpiecEasi ϕ ρ
Differential abundance	metagenomSeq LEfSe DESeq	ALDEx2 ANCOM

FIGURE 2 | The standard microbiome analysis tool kit and the compositional replacements. A simplified standard microbiome computational workflow is illustrated. The initial normalization steps are not formally equivalent since compositional data are inherently "normalized", and read count normalization is unnecessary. The other steps are functionally equivalent and substitute a compositionally appropriate approach for one that is not.

for downstream analyses such as ordination, and discrimination. Distances between features are non-linear when examined from a Euclidian perspective (Martín-Fernández et al., 1998; Aitchison et al., 2000) and many DD matrices are used that partially address this problem. As noted above the total number of reads in a sample is a strong confounding variable on all these methods, indicating that the composition of the sample is not the primary property being measured. However, apparently useful DD matrices can be generated after normalization. Three DD matrices dominate the literature; UniFrac (both the weighted and unweighted variants) (Lozupone et al., 2011), Bray-Curtis and Jensen-Shannon divergence, and while all have their uses, they do not account for the compositional nature of the data. It should be noted that the weighted UniFrac distance approach captures important phylogenetic information, and a recent compositional replacement has been developed (Silverman et al., 2017).

The major uses for the DD matrices are ordination and clustering. Here, the shortcomings of these DD methods become apparent. In addition to being sensitive to the total read depth of a sample, DD methods largely discriminate between samples based on the most relatively abundant features in the samples, not on the features that are necessarily the most variable between samples (Gorvitovskaia et al., 2016; Wong et al., 2016). This can lead to the location of samples in an ordination changing dramatically when different features are included or excluded from the dataset, and to a lack of sensitivity in identifying outlier samples (Wong et al., 2016).

Severe problems with correlation in compositional data were first noted at the dawn of statistical practice by Pearson (1897) and rediscovered in the context of microbiome studies (Lovell et al., 2011; Friedman and Alm, 2012; Lovell et al., 2015; Kurtz et al., 2015; Morton et al., 2017). Unfortunately, the effect cannot be diluted away as has been recommended (Weiss et al., 2016). Understanding that there is a correlation problem is crucial, since unconstrained correlation or covariation are key concepts for ordination, clustering, network analysis and differential (relative) abundance determination. Compositional data have a negative correlation bias and a different correlation structure than the underlying count data. Even worse, compositional data exhibit spurious correlation upon subsetting or aggregation. The "Correlation" section in the Supplement shows that correlation is not a reliable or a reproducible indicator of the underlying data when dealing with compositional data.

Finally, differential (relative) abundance measures do not account for compositionality (Fernandes et al., 2013; Mandal et al., 2015; Gloor et al., 2016a). Large scale tool benchmarking has revealed that differential (relative) abundance tools in common use are sensitive to sparsity (Thorsen et al., 2016) and consequently exhibit unacceptably high false positive identification rates (Hawinkel et al., 2017).

In summary the analysis of compositional data using current protocols has several challenges. However, as shown below these issues can be addressed in a satisfactory way using tools that account for the compositional nature of the data.

3. ANALYSIS OF HTS USING CODA METHODS

Compositional datasets from HTS can be analyzed in a rigorous manner by adapting tools from other fields (Van den Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn et al., 2015) and using new tools based on the same underlying foundations (Fernandes et al., 2013; Erb and Notredame, 2016; Silverman et al., 2017; Quinn et al., 2017). There are now examples in the literature that provide guidance on how to do some or all of these analyses on HTS datasets, including meta-transcriptomics (Macklaim et al., 2013) and tag-sequencing (McMurrough et al., 2014; Bian et al., 2017). We briefly review the approaches below.

The starting point for any compositional analyses is a ratio transformation of the data. Ratio transformations capture the relationships between the features in the dataset and these ratios are the same whether the data are counts or proportions. Taking the logarithm of these ratios, thus log-ratios, makes the data symmetric and linearly related, and places the data in a log-ratio coordinate space (Pawlowsky-Glahn et al., 2015). Thus, we can obtain information about the log-ratio abundances of features *relative to other features* in the dataset, and this information is directly relatable to the environment. We cannot get information about the absolute abundances since this information is lost during the sequencing process as explained in **Figure 1**. However, log-ratios have the nice mathematical property that their sample space is real numbers, and this represents a major advantage for the application of standard statistical methods that have been developed for real random variables.

Often the centered log-ratio (clr) transformation introduced by Aitchison (1986) is used. Given an observation vector of D "counted" features (taxa, operational taxonomic units or OTUs, genes, etc.) in a sample, $\boldsymbol{x} = [x_1, x_2, ...x_D]$, the clr transformation for the sample can be obtained as follows:

$$\begin{aligned} \boldsymbol{x}_{clr} &= [log(x_1/G(\boldsymbol{x})), log(x_2/G(\boldsymbol{x})) \ldots log(x_D/G(\boldsymbol{x}))], \\ G(\boldsymbol{x}) &= \sqrt[D]{x_1 \cdot x_2 \cdot ... \cdot x_D} \end{aligned} \quad (1)$$

$G(\boldsymbol{x})$ is the geometric mean of $\boldsymbol{x}$. The clr transformed values can be used as inputs for multivariate hypothesis testing using tools such as MANOVA, regression etc. (Van den Boogaart and Tolosana-Delgado, 2013) and for model building. The clr-transformed values are scale-invariant; that is the same ratio is expected to be obtained in a sample with few read counts or an identical sample with many read counts, only the precision of the clr estimate is affected. This is elaborated in the "Probability" and "Log-ratio transformations" section in the Supplement, but the consequence is that count normalization is unnecessary and indeed, undesirable since information on precision is lost.

The $G(\boldsymbol{x})$ cannot be determined for sparse data without deleting, replacing or estimating the 0 count values. Fortunately, there are acceptable methods of dealing with 0 count values as both point estimates using `zCompositions` `R` package (Palarea-Albaladejo and Martín-Fernández, 2015), and as a probability distribution using `ALDEx2` available on Bioconductor. Converting the single estimate to a probability vector prior to clr transformation produces a scale-invariant measure since this accounts for the precision of the estimate of the probabilities for each feature; we refer advanced readers to the more technical literature (Jaynes and Bretthorst, 2003; Fernandes et al., 2013; Gloor et al., 2016a) and the "Probability" section of the Supplement for more information.

There are compositional replacements for distance determination that is used for clustering and ordination. The first is the `philr` phylogenetic transform (and `R` package) based on balances (binary partitions) along an evolutionary tree (Silverman et al., 2017) that is a replacement for the familiar UniFrac distance metric. Distances determined by phylogenetic transforms have the advantage that the binary partitions chosen have a simple interpretation and the correlation structure of the data is fully accounted for. However, the disadvantage is that only the relationships between the chosen partitions can be examined. A second distance metric is the Aitchison distance, which is simply the Euclian distance between samples after clr transformation, and the distances between samples are the same as the phylogenetic ilr. The Aitchison distance is superior to both the widely used Jensen-Shannon divergence and the Bray-Curtis dissimilarity metrics, being more stable to subsetting and aggregating of the data, and being a true linear distance (Aitchison et al., 2000).

The replacement for β-diversity exploration of microbiome data is the variance-based compositional principal component (PCA) biplot (Aitchison, 1983; Aitchison and Greenacre, 2002) where the relationship between inter-OTU variance and sample distance can be observed (Gloor et al., 2016b). The compositional biplot has several advantages over the principal co-ordinate (PCoA) plots for β-diversity analysis. The results obtained are very stable when the data are subset (Bian et al., 2017), meaning that exploratory analysis is not driven simply by the presence absence relationships in the data nor by excessive sparsity (Wong et al., 2016; Morton et al., 2017). PCA plots can be substantially more reproducible, since they do not depend upon an presumed underlying tree that may need to be regenerated with each data subset, or when new taxa need to be incorporated. This simplicity facilitates exploratory data analysis. Compositional PCA biplots display the relationships between OTUs and the distances between samples on a common plot. It is possible to glean substantial qualitative information regarding the quality of the dataset and the relationships between groups with this tool (Aitchison and Greenacre, 2002; Gloor et al., 2016b), and examples are shown in the "Biplot" section of the Supplement.

As noted above, the correlation is unreliable in compositional datasets because of the negative correlation bias and the instability of correlation to subsetting the data. This is explained more fully in the supplement (Pearson, 1897; Aitchison, 1986) but these problems are observed with all non-compositional correlation methods (Ortego and Egozcue, 2013). Unfortunately, correlation cannot be subjected to a principled process to determine the optimal method as has been advocated recently (Weiss et al., 2016).

There are several more rigorous approaches that can be applied to analyze correlation in microbiome datasets, including SPARCC (Friedman and Alm, 2012) and SPieCeasi (Kurtz et al., 2015), both of which assume a sparse data matrix, and the ϕ (Lovell et al., 2015) and ρ (Erb and Notredame, 2016) metrics (the published versions of which required a non-sparse matrix). These latter two metrics have been incorporated into the `R` package `propr`, that includes an adaptation allowing the calculation of the metrics with sparse data that gives an expected value of ρ (E(ρ)), that approaches 1 if the two features have exactly constant ratios in the data (Lovell et al., 2015; Quinn et al., 2017). Supplementary Figure 2 shows that the expected value of ρ is much more stable to subsetting than are familiar correlation metrics, and becomes more reproducible as the value of E(ρ) approaches 1, thus indicating greater precision in estimation as correlation becomes stronger. However, determining an optimal and general approach for correlation in compositional datasets is an open research problem. Supplementary Figures 2–5 have a more extended explanation of the correlation problem and the use of E(ρ) as a proposed solution.

Differential (relative) abundance of OTUs between groups in compositional data is often examined using purpose-built tools that compare the difference in relative abundance across samples, and recently tools adapted from the domain of RNA-seq have been suggested. Unfortunately, these approaches do not account for the compositional nature of the data, and so can be particularly sensitive to the negative correlation bias and large variability of such datasets (Fernandes et al., 2013). Indeed benchmarking suggests that traditional tools exhibit different false positive rates with different levels of sparsity (Thorsen et al., 2016), and that the false positive rates can be up to 20× higher than expected (Hawinkel et al., 2017).

Tools based on an approximate compositional foundation are available. The ANCOM tool performs statistical tests on point estimates of data transformed by an additive log ratio, where (presumed) invariant taxa are chosen as the denominator (Mandal et al., 2015). ANCOM is being incorporated into the popular QIIME suite of microbiome analysis tools (Weiss et al., 2017). The ALDEx2 tool performs statistical tests on the clr values from a modelled probability distribution of the dataset (Supplementary data Equations 1–4), and reports the expected values of parametric and non-parametric statistical tests along with effect-size estimates. This approach reduces the false-positive identification problem to near 0 in real and modelled microbiome datasets with little effect of sensitivity (Thorsen et al., 2016) and is observed to be relatively insensitive to change when the data are subset (Fernandes et al., 2014). There are many examples in the literature on its use (Macklaim et al., 2013; McMurrough et al., 2014; Bian et al., 2017) and in the Supplementary.

In summary, the analysis of compositional data by traditional methods can appear to give satisfactory results. However, these results can be misleading and unpredictable. Compositionally-appropriate tools exist as drop-in replacements at each stage of the analysis as shown in **Figure 2**, and interested readers are directed to the supplementary and to other published examples (Macklaim et al., 2013; Fernandes et al., 2014; McMurrough et al., 2014; Lovell et al., 2015; Mandal et al., 2015; McMillan et al., 2015; Gloor and Reid, 2016; Gloor et al., 2016b; Bian et al., 2017; Silverman et al., 2017; Quinn et al., 2017), and the similar correspondence analysis implemented in the phyloseq package (McMurdie and Holmes, 2013).

AUTHOR CONTRIBUTIONS

GG conceived and wrote the initial draft of the manuscript. JM conceived and made **Figures 1**, **2**. JM, JE, and VP-G edited the draft. All authors agreed with the contents of the final version.

FUNDING

Work by GG and JM were supported by grants from the National Sciences and Engineering Research Council grant RGPIN-03878-2015, and by funding from the Ministry of Agriculture and Agri-Food AgriInnovation Program. VP-G and JE acknowledge financial support by the Spanish Ministry of Education and Science under project 'CODA-RETOS' (Ref. MTM2015-65016-C2-1 (2)-R (MINECO/FEDER,UE)).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224/full#supplementary-material

REFERENCES

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika* 70, 57–65. doi: 10.1093/biomet/70.1.57

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V. (2000). Logratio analysis and compositional distance. *Math. Geol.* 32, 271–275. doi: 10.1023/A:1007529726302

Aitchison, J., and Greenacre, M. (2002). Biplots of compositional data. *J. Roy. Stat. Soc. Ser. C* 51, 375–392. doi: 10.1111/1467-9876.00275

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biol.* 11:R106. doi: 10.1186/gb-2010-11-10-r106

Bian, G., Gloor, G. B., Gong, A., Jia, C., Zhang, W., Hu, J., et al. (2017). The gut microbiota of healthy aged chinese is similar to that of the healthy young. *mSphere* 2:e00327-17. doi: 10.1128/mSphere.00327-17

Erb, I., and Notredame, C. (2016). How should we measure proportionality on relative gene expression data? *Theory Biosci.* 135, 21–36. doi: 10.1007/s12064-015-0220-8

Fernandes, A. D., Macklaim, J. M., Linn, T., Reid, G., and Gloor, G. B. (2013). ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS ONE* 8:e67019.

Fernandes, A. D., Reid, J. N., Macklaim, J. M., McMurrough, T. A., Edgell, D. R., and Gloor, G. B. (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., and Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* 8:e1002687. doi: 10.1371/journal.pcbi.1002687

Gloor, G. B., Macklaim, J. M., Vu, M., and Fernandes, A. D. (2016a). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Aust. J. Stat.* 45, 73–87. doi: 10.17713/ajs.v45i4.122

Gloor, G. B., and Reid, G. (2016). Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* 62, 692–703. doi: 10.1139/cjm-2015-0821

Gloor, G. B., Wu, J. R., Pawlowsky-Glahn, V., and Egozcue, J. J. (2016b). It's all relative: analyzing microbiome data as compositions. *Ann. Epidemiol.* 26, 322–329. doi: 10.1016/j.annepidem.2016.03.003

Gorvitovskaia, A., Holmes, S. P., and Huse, S. M. (2016). Interpreting prevotella and bacteroides as biomarkers of diet and lifestyle. *Microbiome* 4:15. doi: 10.1186/s40168-016-0160-7

Hawinkel, S., Mattiello, F., Bijnens, L., and Thas, O. (2017). A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief. Bioinf.* bbx104. doi: 10.1093/bib/bbx104

Jaynes, E. T., and Bretthorst, G. L. (2003). *Probability Theory: The Logic of Science*. Cambridge, UK: Cambridge University Press

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput. Biol.* 11:e1004226. doi: 10.1371/journal.pcbi.1004226

Lovell, D., Müller, W., Taylor, J., Zwart, A., and Helliwell, C. (2011). "Proportions, percentages, ppm: do the molecular biosciences treat compositional data right," in *Compositional Data Analysis: Theory and Applications,* eds V. Pawlowsky-Glahn and A. Buccianti (London: John Wiley & Sons, Ltd.), 193–207.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J., Marguerat, S., and Bähler, J. (2015). Proportionality: a valid alternative to correlation for relative data. *PLoS Comput. Biol.* 11:e1004075. doi: 10.1371/journal.pcbi.1004075

Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J., and Knight, R. (2011). Unifrac: an effective distance metric for microbial community comparison. *ISME J.* 5, 169–172. doi: 10.1038/ismej.2010.133

Macklaim, M. J., Fernandes, D. A., Di Bella, M. J., Hammond, J.-A., Reid, G., and Gloor, G. B. (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome* 1:15. doi: doi: 10.1186/2049-2618-1-12

Mandal, S., Van Treuren, W., White, R. A., Eggesbø, M., Knight, R., and Peddada, S. D. (2015). Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* 26:27663. doi: 10.3402/mehd.v26.27663

Martín-Fernández, J., Barceló-Vidal, C., Pawlowsky-Glahn, V., Buccianti, A., Nardi, G., and Potenza, R. (1998). Measures of difference for compositional data and hierarchical clustering methods. *Proc. IAMG*. 98, 526–531.

McMillan, A., Rulisa, S., Sumarah, M., Macklaim, J. M., Renaud, J., Bisanz, J. E., et al. (2015). A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. *Sci. Rep.* 5, 14174. doi: 10.1038/srep14174

McMurdie, P. J., and Holmes, S. (2013). phyloseq: an r package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE* 8:e61217. doi: 10.1371/journal.pone.0061217

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

McMurrough, T. A., Dickson, R. J., Thibert, S. M. F., Gloor, G. B., and Edgell, D. R. (2014). Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. *Proc. Natl. Acad. Sci. U.S.A.* 111, E2376–E2383. doi: 10.1073/pnas.1322352111

Morton, J. T., Toran, L., Edlund, A., Metcalf, J. L., Lauber, C., and Knight, R. (2017). Uncovering the horseshoe effect in microbial analyses. *mSystems* 2:e00166-16. doi: 10.1128/mSystems.00166-16

Ortego, M. I., and Egozcue, J. J. (2013). "Spurious copulas," in *Proceedings of the 5th Workshop on Compositional Data Analysis, CoDaWork 2013* (Vorau).

Palarea-Albaladejo, J., and Martín-Fernández, J. A. (2015). zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr. Intel. Lab. Syst.* 143, 85–96. doi: 10.1016/j.chemolab.2015.02.019

Pawlowsky-Glahn, V., Egozcue, J. J., and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data.* London: John Wiley & Sons.

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. *Proc. Roy. Soc. Lond.* 60, 489–498.

Quinn, T., Richardson, M. F., Lovell, D., and Crowley, T. (2017). propr: An R-package for identifying proportionally abundant features using compositional data analysis. *bioRxiv*. doi: 10.1101/104935

Robinson, C. K., Brotman, R. M., and Ravel, J. (2016). Intricacies of assessing the human microbiome in epidemiologic studies. *Ann. Epidemiol.* 26, 311–321. doi: 10.1016/j.annepidem.2016.04.005

Robinson, M. D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* 11, R25.1–R25.9. doi: 10.1186/gb-2010-11-3-r25

Shaffer, M. L. (1981). Minimum population sizes for species conservation. *BioScience* 31, 131–134.

Silverman, J. D., Washburne, A. D., Mukherjee, S., and David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *Elife* 6:21887. doi: 10.7554/eLife.21887

Thorsen, J., Brejnrod, A., Mortensen, M., Rasmussen, M. A., Stokholm, J., Al-Soud, W. A., et al. (2016). Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 4, 62. doi: 10.1186/s40168-016-0208-8

Tsilimigras, M. C. B., and Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26, 330–335. doi: 10.1016/j.annepidem.2016.03.002

Van den Boogaart, K. G., and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*, London, UK: Springer.

Weiss, S., Van Treuren, W., Lozupone, C., Faust, K., Friedman, J., Deng, Y., et al. (2016). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J.* 10, 1669–1681. doi: 10.1038/ismej.2015.235

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5, 27. doi: 10.1186/s40168-017-0237-y

Wong, R. G., Wu, J. R., and Gloor, G. B. (2016). Expanding the UniFrac toolbox. *PLoS ONE* 11:e0161196. doi: 10.1371/journal.pone.0161196

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.