Annals of Epidemiology 26 (2016) 330–335

Contents lists available at ScienceDirect

Annals of Epidemiology

journal homepage: www.annalsofepidemiology.org

The Microbiome and Epidemiology

Compositional data analysis of the microbiome: fundamentals, tools, and challenges

Matthew C.B. Tsilimigras, Anthony A. Fodor PhD *

Department of Bioinformatics and Genomics, UNC Charlotte, Bioinformatics Building, The University of North Carolina, Charlotte 9201, University City Blvd, Charlotte

ARTICLE INFO

Article history:
Received 1 December 2015
Accepted 23 March 2016
Available online 31 March 2016

Keywords:
Metagenomics
Data interpretation, statistical
Microbiota
RNA, Ribosomal
16S
Statistics as topic
Selection bias
High-throughput nucleotide sequencing

ABSTRACT

Purpose: Human microbiome studies are within the realm of compositional data with the absolute abundances of microbes not recoverable from sequence data alone. In compositional data analysis, each sample consists of proportions of various organisms with a sum constrained to a constant. This simple feature can lead traditional statistical treatments when naively applied to produce errant results and spurious correlations.
Methods: We review the origins of compositionality in microbiome data, the theory and usage of compositional data analysis in this setting and some recent attempts at solutions to these problems.
Results: Microbiome sequence data sets are typically high dimensional, with the number of taxa much greater than the number of samples, and sparse as most taxa are only observed in a small number of samples. These features of microbiome sequence data interact with compositionality to produce additional challenges in analysis.
Conclusions: Despite sophisticated approaches to statistical transformation, the analysis of compositional data may remain a partially intractable problem, limiting inference. We suggest that current research needs include better generation of simulated data and further study of how the severity of compositional effects changes when sampling microbial communities of widely differing diversity.

© 2016 Elsevier Inc. All rights reserved.

Introduction

Compositional data are vectors of nonnegative elements constrained to sum to a constant. This simple feature of compositional data can have surprisingly adverse effects when traditional methods of multivariate statistics are naively used [1]. The dangers of ignoring the effects of compositionality were noted by Pearson, who recognized more than a century ago, that “spurious correlations” would result, should values constructed as proportions be compared haphazardly [2]. Compositional data is subject to the “closure problem” that occurs when components necessarily compete to make up the constant sum constraint [3]. This can cause large changes in the absolute abundance of one component to drive apparent changes in the measured abundance of others, violating the assumption of sample independence and creating inevitable errors in covariance estimates that can lead to bias and flawed inference. Diverse academic disciplines have begun to appreciate the complexity of the analysis of compositional data, ranging from forensics [4,5] and psychology [6] to the assessment of antibiotic use [7] and nutritional epidemiology [8].

In the case of the microbiome sequencing surveys, the compositional nature of the data comes from the fact that a correction must be made for different samples having different numbers of sequences while the total absolute abundance of all bacteria in each sample is unknown. These complications arise from sample collection, polymerase chain reaction (PCR) amplification, and the sequencing technology itself from which the absolute abundances of bacteria are not recoverable from sequence counts, but the proportions of different taxa are still relevant. Numerous schemes are used in the literature to convert the number of sequences for each taxon within each sample to relative abundance with popular techniques, including proportional abundance and rarefying, the latter being the default choice in the popular Quantitative Insights Into Microbial Ecology pipeline [9,10]. Neither of these approaches corrects for compositionality and it has been argued that this lack of correction has led to erroneous analyses that fail to discriminate between true and spurious correlations between taxa [11,12]. However, it remains unclear whether these sorts of normalization schemes routinely produce spurious correlations in the study of complex microbial communities, like the gut, or whether errors due to compositionality are instead restricted to analysis of microbial communities where only a few taxa dominate, such as the vaginal microbiome.

* Corresponding author. Department of Bioinformatics and Genomics, The University of North Carolina; Charlotte 9201, University City Blvd, Charlotte, NC 28223-0001. Tel.: 704-687-8214.
E-mail address: anthony.fodor@gmail.com (A.A. Fodor).

http://dx.doi.org/10.1016/j.annepidem.2016.03.002
1047-2797/© 2016 Elsevier Inc. All rights reserved.

In this review, we examine the historical literature on the compositionality problem and some modern approaches to its solution that have been proposed for the analysis of next-generation sequencing data sets. We track recent progress and indicate where we think more research is needed. We also emphasize that the analysis of compositional data will always be at least a partially intractable problem despite the development of sophisticated statistical transformations as the absolute abundances of microbes before sequencing can never be recovered from sequence data alone, and this will inevitably color inference based on compositional samples.

Compositional data sets are best analyzed after a log-ratio transformation

The initial literature on compositional data analysis has largely been attributed to a pioneering author, John Aitchison, whose classic treatise, "The Statistical Analysis of Compositional Data," has remained enormously influential for nearly 3 decades [3]. However, Aitchison, developing his theory in the 1980s, was analyzing data sets considerably smaller than those of current next-generation sequencing. His examples were often sourced from geology and usually featured problems such as how different mineral components were used to categorize variability in rock specimens. Despite the relative simplicity of the data sets he analyzed, the theory Aitchison developed was surprisingly complex. His work eventually led to the realization that the unit-sum constraint yielded a new geometrical space requiring a substantial background in advanced multivariate linear algebra to fully appreciate. A central challenge for researchers wanting to apply these elegant mathematical formalisms to modern genomics data is the complexity of sequencing data sets, which, unlike simple geology data sets, can have tens of thousands of different categories (high dimensionality), have zeros dominating all other values (sparsity), and have a number of samples substantially fewer than the number of variables (underdetermination) [13,14]. Aitchison recognized these problems but does not offer complete solutions to them in his treatise, and attempts to satisfactorily address these difficult compositional data sets continue to the current day.

Aitchison argued that taking the logarithm of ratios is a transformation of compositional data that restores much of the utility of traditional statistical analyses in situations such as relative abundance. This transformation is structured so that the constant sum constraint does not distort the underlying covariance or correlation structure originating from the natural interaction of the components [3]. A natural problem in using a ratio-based transformation is that one has to choose what will be in the denominator; that is to say, which value to use to normalize all the values in a sample. Aitchison considered two possible transformations in his text, both of which are still in use. The simplest transformation is to choose one component as a reference. For example, in a metagenomics experiment analyzed at the phyla level, one could choose as a reference the phyla "Firmicutes." Then all other taxa would be reported as a ratio of each taxa to Firmicutes. Although compositionality was not the motivation, this was in fact the transformation that was used in an early landmark study of the human microbiome, which reported that the ratio of Bacteroidetes to Firmicutes was associated with obesity in a human population (interestingly this observation has proven to be difficult to replicate[15]). Choosing reference taxa has the advantage of simplicity, but there may not always be an obvious reference to choose and results may vary substantially dependent on the choice of reference [13]. One solution might be to systematically perform inference on every possible pair of taxa, but performing N^2 analyses, and then correcting for N^2 multiple hypotheses is not usually feasible given the large numbers of distinct taxa in many metagenomic analysis pathways. Aitchison called this simple choice of using one reference taxon and taking the logarithm an "additive log-ratio" (alr). As an alternative, Aitchison recommended transforming each taxon within a sample by taking the log-ratio of the counts for that taxon divided by the geometric mean of the counts of all taxa, called the centered log-ratio (clr). This approach is necessarily more robust than the additive log-ratio as it does not depend on the choice of an arbitrary reference. This algorithm has found use in the current microbial literature [16] where it was argued that this transformation could be used to successfully analyze microbiome data as well as RNA-seq data and, indeed, any next-generation sequence data set. Egozcue et al. [17] later defined a third isometric log-ratio transformation (ilr), which is the product of the clr and the transpose of a matrix which consists of elements that are clr-transformed components of an orthonormal basis. This ilr transformation is an orthonormal isometry that addresses certain difficulties of alr and clr, but its interpretability is subject to the selection of its basis, which has somewhat limited its adoption [17].

Although the centered log-ratio has mathematical elegance and has found sophisticated champions in the current metagenomic literature, it has potential problems when applied to metagenomic data sets. This difficulty arises from extreme variability of library sizes and the great sparsity of metagenomic data sets. In a highly sparse data set, the geometric mean of all taxa can often be zero or near zero. Obviously, if it is zero, a transformation that involves dividing by the geometric mean is undefined. One can of course correct for this by adding a pseudo-count to each cell, but it is not immediately clear what the value of this pseudo-count should be. For example, if the value 1 is chosen for the pseudo-count, then dividing by the geometric mean in a highly sparse data set is equivalent to simply not normalizing the data (because you are dividing by 1 before the log transformation). Performing statistical inference on unnormalized data will often lead to results that do not reveal biological variability, but merely reflect differences in sequencing depth [18]. For example, Weiss et al. [19] has shown that the first principal coordinate analysis axes of data sets are often well correlated to the number of sequences per sample. This problem is not ameliorated by a transformation such as taking geometric mean while using a small pseudo-count (Fodor lab, unpublished data).

One could choose some other value for the normalizing counts other than the geometric mean. Packages made for RNA-seq data, DESeq, for example, use values based on medians or certain percentiles in the denominator [20,21]. This offers some of the advantages of the geometric mean, but there is still no guarantee that even very high percentiles of a metagenomics data set do not yield zeros subject to the routinely encountered sparsity. One article [18] recommends the use of RNA-seq pipelines for analysis of metagenomic data but does not offer much guidance on how best to set the normalizing threshold to avoid normalizing by zero or the pseudo-count.

Another problem related to sequencing depth in metagenomic experiments is the difficult decision of when to remove samples that have few sequences [18,22]. In general, these samples tend to be outliers. The low number of sequences in such samples may reflect a PCR error or indicate a sample in which there was no input microbial DNA and the sequences reflect kit microbes or other artifacts [23]. However, it is not clear how to define the cutoff value that indicates that a sample has so few sequences that it should be removed from downstream analysis. This difficult decision of sequence count thresholds impacts the corrections for compositional data described previously in ways that are not fully appreciated as the compositionality corrections work in relative space, but the decision to threshold is in absolute space, and the interaction of making decisions in these two spaces is unclear.

It should be stressed that even with all the algorithms that have been developed to appropriately analyze compositional data

Table 1
Overview and brief comparison of methods for compositional analysis described in this review

Software	Method	Utility
CCREPE	• Correlation based • Establishes corrections based on a null distribution	• Corrects for spurious correlations • Improved similarity measure
MetagenomeSeq	• Gaussian mixed models • Advanced normalization routine	• Addresses undersampling problem • Differential abundance capabilities outperformed tools used in differential expression analyses
SPARCC	• Correlation based • Sparse network assumptions are used to refine variance dependencies	• Corrects for spurious correlations • Identifies true associations missed by Pearson correlation
SPIEC-EASI	• Neighborhood • Covariance selection • Graphical model • Sparse network assumptions	• Addresses underpowering issues • Advanced simulations may more accurately reflect microbiome interaction networks • Corrects for indirect taxa-taxa associations • More reproducible results compared with SPARCC and CCREPE

(summarized in Table 1), there are fundamental limits to the types of information that can be derived from compositional data. The Aitchison transformations can prevent variance-covariance matrices from containing artifacts caused by the closure problem; they cannot recover the unknown absolute abundances present before sequencing [3,13]. Aitchison's classical example of these limits in relative data compared with absolute data involves the overnight changes to the composition of an outdoor planter containing soil, water, and seed. He argues that based only on relative data, one cannot tell whether changes in the composition are due to rain increasing the amount of water, wind blowing away topsoil, or some other more complicated combination of elements [24]. In a 16S rRNA experiment, the proportional increase in one particular taxon in a differential abundance study cannot be discriminated between the taxon potentially taking advantage of a new resource and growing on its own while other populations remain stable in absolute numbers, or its proportion growing as a result of its competition with the other taxa present which would drive down their absolute numbers and hence their relative proportions. Ultimately, in a relative abundance, experiment where one taxon goes up while another goes down, and these two taxa make up most of the observed sequences, we have no way of knowing whether that has occurred because in absolute space one taxon has increased or the other has decreased or both. If information about absolute abundance is required for biological inference, experiments besides just sequencing, such as qPCR measuring the total abundance of the 16S rRNA gene, are required so that conclusions initially arising from sequencing experimentation alone can ultimately be supported by multiple lines of evidence [25].

Sparsity is a central challenge in the analysis of 16S rRNA-sequence data

Characteristic of, but not unique to, the microbiome is a tendency toward pronounced sparsity seen as the absence of many taxa across samples. Some of this sparseness is due to the true discovery of low-abundance taxa that are only in a few samples, but in most experiments, much sparsity derives from sequencing artifacts and the highly variable sequencing depth between samples [26,27]. This sparsity undermines the ability of tools and methodologies developed for compositionality before the advent of next-generation sequencing to be used without caution or correction. Sparsity also poses general numerical challenges for many traditional tools of statistical analysis [28,29]. Parametric models must make accurate estimates of variance for meaningful inference and such estimates are essentially impossible on samples that consist mostly of zeros. There is an additional layer of complication as the zero values have multiple causes and so no generalized treatment strategy exists [30]. Attempts to correct for zeros generally consider two categories: zeros that result from undersampling (called "rounded" zeros) and zeros that truly represent the absence of taxa from a particular sample (called "essential" or "structural" zeros; [31]). One solution to rounded zeros is to replace them with a small, nonzero value [32], often termed a pseudo-count. The process of adding such pseudo-counts is called imputation. This step is mandatory for the log-ratio transformations mentioned earlier to avoid taking logarithms of zero. The theoretical justification for adding pseudo-counts is that these correspond to some value below the detection limit. However, numerous studies have indicated that making all analyses robust to varying imputation values is difficult if not impossible, especially if the degree of sparsity changes dramatically [32–34]. Structural zeros in continuous data have been modeled with some success by using a binomial conditional logistic normal model [35], whereas progress with such zeros in discrete data has been achieved by using models based on the Poisson-Log Normal distribution [36].

Nonparametric methods are generally insensitive to such factors as how pseudo-counts are added to an experiment [37] and avoid making variance estimates that can be skewed by sparse samples. Often in metagenomic experiments, however, there are many taxa but few samples and nonparametric methods will lack power to perform inference on the many low-abundance taxa that make up a metagenomic sample. In such cases, parametric models can be used, but this comes with the risk of the assumptions of parametric models being violated. An increasingly popular method for explicitly modeling zero counts using parametric models is the mixture linear model wherein rounded zeros are explicitly modeled with a distribution, such as the binomial distribution, and the essential zeros are modeled as the probability of seeing a zero under a distribution such as the negative binomial or the Poisson distribution, the nonzero values of which are used to model the rest of the data [38]. In addition, the negative binomial addresses the problem of overdispersion often seen in these studies as it removes the mean-variance constraint of the Poisson distribution. Although these models can substantially boost power in comparison to nonparametric methods, they make an additional set of assumptions over and above the assumption of normality made by simple nonmixture linear models. These assumptions include (1) that the rounded zeros indeed follow the chosen distribution, (2) that there is a single parameter that explains the fraction of rounded zero observed and that this parameter is the same in all samples or can be adequately modeled by the model's independent variables, and (3) that the presence of essential and rounded zeros are independent in each sample. In practice, these assumptions can be difficult to verify, especially since we do not have a general theoretical framework under which to explain or predict when zeros are essential or rounded. Results from these models, therefore, must be treated with due caution.

No matter which statistical model is ultimately used, care must be taken during normalization to account for the sparsity of the data set. As described previously, if a centered log-ratio transformation is used on a highly sparse data set, the normalizing variable used in the denominator will be close to 1, and this transformation will therefore fail to correct for differences in sequencing depth between samples. One tool designed to address the problem of normalization in a sparse environment, metagenomeSeq, examines the data to find a suitable percentile for each data set for normalization to avoid normalizing by zero or near zero values. This process, called "cumulative-sum scaling," generates an appropriate percentile for normalization for the data set which can then be used for normalization before application of Gaussian mixed models which are used for inference in the metagenomeSeq pipeline [39].

Although the pipeline used by metagenomeSeq has many attractive features, normalizing data by rarefying or simple proportion—pipelines long used in the metagenomics field—are also not subject to the problem of normalizing by a zero or pseudo-count value, and as a consequence these pipelines can show less dependence on sequence depth and its influence on sparsity than pipelines that perform Aitchison-style transformations [19] (Fodor lab, unpublished data). However, unlike the ratio transformation used in the metagenomeSeq pipeline, relative abundance and rarefying do not eliminate the closure problem from compositionality. This is clearly problematic for low-diversity communities, such as the vaginal microbiome, where one or a few taxa dominate the entire community. It is less clear that not correcting for compositionality is necessarily a problem in complex communities like the gut, where there can be many hundreds of taxa. It may be that in high complexity environments, although it is technically true that the taxa must sum to 1, that there are so many taxa that changes to a single taxon are effectively independent of effects on all other taxa. That is, in a complex microbial community where there is no single dominant taxa, it may be that the relative abundance of taxa can be considered effectively independent of each other, whereas this will never be the case in a low-diversity community in which most of the reads are associated with a single taxon and therefore large changes in that taxon have to be directly reflected in the number of reads for the few other taxa. This is a crucial area for future research. In particular, it will be essential to discover whether analyses at lower phylogenetic levels (like the operational taxonomic unit level) are more insensitive to compositional artifacts than analyses at higher taxonomic levels (such as phyla) where the number of distinct categories is much smaller and Aitchison-style corrections therefore potentially more useful. An Aitchison-style correction introduces some complexity to the analysis path, requiring us to make a choice about how we will transform the data to work in some type of ratio space. Future research may tell us when this additional complexity is needed to guard against compositionality artifacts and when it is not.

Regardless of the method used to try to resolve the zero problem, the fact remains that certain rare taxa with unique functional properties can have profound influences on microbial ecology. Therefore, a strategy of simply ignoring low-abundance taxa, while perhaps statistically sound, may miss vital biology [40,41]. In cases where inference suggests that low-abundance taxa play an important role in an ecosystem, confirmation by methods other than just 16S rRNA sequencing is crucial, especially if that inference depended on difficult to verify parametric assumptions.

Compositional data analysis in practice

Ordination and dimensionality reduction of compositional data requires several important considerations with distance metrics being chief among them. The Aitchison distance, formed by the sum of log-ratio differences over all taxa, is one such means of working within the restrictions of the Aitchison geometry to retain metric properties [42]. In the metagenomics literature, however, distance measures and dissimilarities like Bray–Curtis and UniFrac are much more commonly used. It remains an open research question as to how much the use of an Aitchison transformation would beneficially impact results when compared with these more commonly used metrics [13].

Aitchison describes inference and regression on compositional data as largely proceeding following standard multivariate techniques after the transformation of the data [3,42]. Therefore, once the microbiome sequence data has been transformed appropriately, model building and selection can proceed largely as the case with noncompositional data, subject to sparsity considerations described previously. Indeed, many usual checks for the suitability of standard multivariate techniques, like goodness-of-fit tests, have been made usable for compositional data [3].

After the log-ratio transformation, the *P*-values produced by standard linear models will not be artifacts of a variance-covariance matrix that has been corrupted by the closure requirement of compositionality data. However, that does not mean that models will then be entirely free of the influence of compositionality. For example, consider a simple microbial community with only two taxa: A and B. If we are comparing samples across two conditions, say cancer patient versus noncancer patients, a *t* test that produces a significant value for A changing will very likely also produce a significant value for B changing. This will occur even if biologically only A goes up in response to cancer, only B goes down in response to cancer, or both. We can reliably detect that there was a change caused by the cancer variable, but even after an Aitchison transformation, we are much less certain as to which taxa actually changed even if we get significant *P*-values from both taxa. This is not necessarily fatal to a research program as there are other ways to confirm the hypotheses that A and B have changed independently, such as qPCR. And, it is not clear that these complications in inference would occur in complex ecosystems such as the gut where the total number of counts is not dependent on only a handful of abundant taxa. However, the possible confounding of one taxon for another should be a consideration in the interpretation of metagenomic data from low-diversity ecosystems such as vaginal samples or cystic fibrosis sputum samples.

In addition to doing taxa-by-taxa inference, there has been a good deal of interest in being able to determine when two taxa are correlated [43–46]. The building of correlated networks is especially sensitive to compositionality artifacts as two taxa can appear to be biologically correlated when in fact they are just both responding to a large change in the number of sequences from another taxa. Again, Atchison-inspired transformations do not fully solve this problem.

One method has approached the basic problem by trying to establish a null distribution of apparent correlations caused by compositionality and comparing that to the actual distributions of observed correlations. Compositionality Corrected by Renormalization and Permutation (CCREPE) uses permutation-based methods to estimate the null distribution [11]. It has been argued, however, that such permutation approaches will fail to adequately control for compositional effects and lead to "false confidence" in the observed correlations [47]. As an alternative, Sparse Correlation for Compositional data (SparCC) approaches the association network problem by using the standard log-ratio transformation and an iterated refinement of the variance matrix describing the dependencies of the components, which is then used to compute the correlations [47]. The model assumes that there are a sufficiently large number of taxa, and that these taxa on average are uncorrelated with each other leading to a sparse network (though the authors demonstrate that the predictions are in fact robust subject to deviations from sparseness). The model then finds pairs of taxa that are outliers to this background network of weak correlations at some threshold.

An appreciation for problems of potential overestimation, and corrections to the underlying association networks in methods like SparCC and CCREPE has recently been investigated in Sparse Inverse Covariance Estimation for Ecological Association Inference (SPIEC-EASI; [14]). This approach uses graphical network models in place of the commonly used correlation as evidence of ecological association. While CCREPE and SparCC estimate a correlation for each pair of taxa, and then use these estimates to build a correlation network, SPIEC-EASI attempts to directly infer the entire correlation network. SPIEC-EASI is designed to further address the problems of underpowering in taxa-taxa associations through incorporating additional evidence of interactions or equivalent model assumptions. The graphical model it uses to represent interactions stresses conditional independence, so that indirectly connected, but still correlated, taxa have this degree of separation appropriately indicated. However, complete network recovery is not likely given the much larger number of taxa when compared with the number of samples. The authors also provide advanced data simulation tools that more accurately reflect the variety of network motifs simultaneously present in the microbiome. SPIEC-EASI was shown to perform well on the American Gut Project [48] data set, and it constructed more highly reproducible association networks when compared against SparCC and CCREPE. It additionally was able to depict sparser interactions compared with the other two methods.

These tools address different difficulties arising from the compositional nature of microbiome analyses. However, they are all predicated on assumptions about the underlying data. Many of these assumptions and general features of such data sets have yet to be compared with an established "gold standard" [16]. Some of the tools indicated when specific biological features of microbial communities limit its utility, but others are less transparent in this regard. In addition, the methods demonstrate the increasing trend of sophistication in the simulation of microbiome data and the selection of biological test data sets.

Further discussion

There is a growing appreciation for the importance of the compositional nature of data in microbiome studies. Tools and methods must properly account for both the biological and the statistical implications from the structure of such data. This includes the nonnormality of the experimental distribution of taxa [49], the sparsity of many data sets, problems related to underdetermined nature of the data, and lack of a "gold standard" for testing the validity of predicted interactions on the community scale. Work in the production of a "gold standard" [14] that explore sample-processing and sequencing artifacts that detract from true biological differences come from studies like the Microbiome Quality Control project for human-associated studies and Xiao et al.'s mouse gut metagenome catalogue [50]. Such concerns exist alongside initiatives to set reproducible and robust standards that do not stifle innovation and allow for the incorporation of necessary revisions. The ongoing dramatic decrease in sequencing costs does alleviate some power limitations present when using more appropriate nonparametric tests [51,52] as a lower cost per sample can allow for larger sample sizes. However, samples sequenced to increasing depth become increasingly sparse. Where they can be used, nonparametric tests should be favored as they are not dependent on the parametric assumptions that can complicate the construction and interpretation of linear models. The use of nonparametric tests can reduce the benefits associated with log-ratio transformations since nonparametric statistics have no variance-covariance matrix to be corrupted by the closure problem. However, neither the use of nonparametric tests nor Aitchison-style log-ratio data transformations will prevent detection of spurious correlations in co-occurrence networks or results from one taxa influencing another in a spurious manner when comparing relative abundance across different experimental conditions. The prevention of spurious correlations in compositional data has been a very active area of research and we expect this to continue in the near future.

Beyond the previously mentioned difficulties future methods must address, it should be reiterated that library size differences do also influence the decision to accept or reject a given sample, but such variability may not reflect biological differences. One means of understanding this variability in a biological context is the use of repli data that might allow acceptance of conclusions from numerous samples even if each sample has a low sequencing depth. In addition, qPCR could provide numerical estimates indicating directionality of population shifts, which remain ambiguous when restricted to being seen using proportional data alone. Similarly, those methods elucidating functional characteristics and ecological niches could provide mechanistic insight into such changes in community structure. Current limitations in compositional data analysis theory mean that important methods like rank correlations and mutual information have yet to be implemented [47]. It bears repeating that Aitchison and other developers of compositional data analysis recognized the fundamental limitations present in inferences produced by purely compositional data.

In conclusion, compositional data analysis provides a means to recover the utility of variance-covariance relationships for data whose components are subject to the unit-sum constraint. However, it is a potentially intimidating body of literature, especially for biologists who may have limited backgrounds in linear algebra. Indeed, the application of compositional data analysis techniques to new fields has largely been driven by statisticians taking interest in a particular field rather than area specialists reaching into the compositional data analysis literature. Despite considerable recent progress, the question of the accurate detection of all spurious correlations subject to the complexity of sequencing data is still unresolved. Improvements to the generation of simulated data have begun to provide a more appropriate background with which to compare various methods. In addition, methods or studies providing detailed tracking of the severity of compositional effects across a range of diversity communities are currently an area in need of further research.

References

[1] Bacon-Shone J. A short history of compositional data analysis. In: Pawlowsky-Glahn V, Buccianti A, editors. Compos. Data Anal. Theory Appl. 1st ed. West Sussex, United Kingdom: Wiley; 2011. p. 3–11.

[2] Pearson K. 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 R Soc Lond 1897;60:489–98 http://www.jstor.org/stable/115879.

[3] Aitchison J. The Statistical Analysis of Compositional Data. 2003rd ed. Caldwell, New Jersey: The Blackburn Press; 1986.

[4] Campbell GP, Curran JM, Miskelly GM, Coulson S, Yaxley GM, Grunsky EC, et al. Compositional data analysis for elemental data in forensic science. Forensic Sci Int 2009;188:81–90.

[5] Neocleous T, Aitken C, Zadora G. Transformations for compositional data with zeros with an application to forensic evidence evaluation. Chemometer Intell Lab 2011;109:77–85.

[6] Pennington L, James P, McNally R, Pay H, McConachie H. Analysis of compositional data in communication disorders research. J Commun Disord 2009;42:18–28.

[7] Faes C, Molenberghs G, Hens N, Muller A, Goossens H, Coenen S. Analysing the composition of outpatient antibiotic use: a tutorial on compositional data analysis. J Antimicrob Chemother 2011;66:vi89–94.

[8] Leite ML. Applying compositional data methodology to nutritional epidemiology. Stat Methods Med Res 2014 [Epub ahead of print].

[9] Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 2010;7:335–6.

[10] Kuczynski J, Stombaugh J, Walters WA, González A, Caporaso JG, Knight R. Using QIIME to Analyze 16S rRNA Gene Sequences from Microbial Communities. Curr Protoc Microbiol 2012;Chapter 1:1–30.

[11] Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the Human Microbiome. PLoS Comput Biol 2012;8:e1002606.

[12] Jackson DA. Compositional data in community ecology: the paradigm or peril of proportions? Ecology 1997;78:929–40.

[13] Li H. Microbiome, Metagenomics and High-Dimensional Compositional Data Analysis. Annu Rev Stat Its Appl 2015;2:73–94.

[14] Kurtz ZD, Mueller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol 2015;11:e1004226.

[15] Finucane MM, Sharpton TJ, Laurent TJ, Pollard KS. A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter. PLoS One 2014;9:e84689.

[16] Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome 2014;2:15.

[17] Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric logratio transformations for compositional data analysis. Math Geol 2003;35:279–300.

[18] McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 2014;10:e1003531.

[19] Weiss SJ, Xu Z, Amir A, Peddada S, Bittinger K, Gonzalez A, et al. Effects of library size variance, sparsity, and compositionality on the analysis of microbiome data. PeerJ Prepr 2015;3:e1408.

[20] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol 2014;15:550.

[21] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol 2010;11:R106.

[22] Kumar R, Eipers P, Little RB, Crowley M, Crossman DK, Lefkowitz EJ, et al. Getting started with microbiome analysis: sample acquisition to bioinformatics. Curr Protoc Hum Genet 2014;82:1–41.

[23] Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 2014;12:87.

[24] Aitchison J. A concise guide to compositional data analysis, CDA work. Girona 2003;24:73–81.

[25] Jespers V, Menten J, Smet H, Poradosú S, Abdellati S, Verhelst R, et al. Quantification of bacterial species of the vaginal microbiome in different groups of women, using nucleic acid amplification tests. BMC Microbiol 2012;12:83.

[26] Gloor GB, Hummelen R, Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R, et al. Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. PLoS One 2010;5:e15406.

[27] Poretsky R, Rodriguez-R LM, Luo C, Tsementzi D, Konstantinidis KT. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. PLoS One 2014;9:e93827.

[28] Lucas J, Carvalho C, Wang Q, Bild A. Sparse statistical modelling in gene expression genomics. In: Bayesian Inference for Gene Expression and Proteomics. 1st ed. New York: Cambridge University Press; 2006. p. 1–25.

[29] Little RJA, Rubin DB. Statistical Analysis with Missing Data. 2nd ed. Hoboken: Wiley; 2002.

[30] Martín-Fernández JA, Palarea-Albaladejo J, Olea RA. Dealing with Zeros. In: Pawlowsky-Glahn V, Buccianti A, editors. Compos. Data Anal. Theory Appl. 1st ed. West Sussex, United Kingdom: John Wiley & Sons, Ltd; 2011.

[31] van den Boogaart KG, Tolosana-Delgado R. Analyzing compositional data with R. Springer-Verlag: Berlin Heidelberg; 2013.

[32] Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. Comput Stat Data Anal 2012;56:2688–704.

[33] Filzmoser P, Hron K, Reimann C. Interpretation of multivariate outliers for compositional data. Comput Geosci 2012;39:77–85.

[34] Palarea-Albaladejo J, Martín-Fernández JA. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. Chemometer Intell Lab 2015;143:85–96.

[35] Aitchison J, Kay JW. Possible solutions of some essential zero problems in compositional data analysis. Compos Data Anal Work Girona 2003; 2003:6.

[36] Bacon-Shone J. Discrete and continuous compositions. In: Daunis-i-Estadella J, Martínez-Fernández J, editors. Proc. CoDaWork '08, 3rd Compos. Data Anal. Work. Girona, Spain: University of Girona; 2008.

[37] Martín-Fernández J, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Math Geol 2003;35:253–78.

[38] Zuur AF, Ieno EN, Walker NJ, Saveliev AA, Smith GM. Mixed Effects Models and Extensions in Ecology with R. New York, NY: Springer Science & Business Media, LLC; 2009.

[39] Paulson JN, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods 2013;10:1200–2.

[40] Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, et al. Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nat Biotechnol 2013;31:814–21.

[41] Pärtel M. Community ecology of absent species: hidden and dark diversity. J Veg Sci 2014;25:1154–9.

[42] Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R. Modeling and Analysis of Compositional Data. 1st ed. Chennai, India: Wiley; 2015.

[43] Carr R, Shen-Orr SS, Borenstein E. Reconstructing the genomic content of microbiome taxa through shotgun metagenomic deconvolution. PLoS Comput Biol 2013;9:e1003292.

[44] Brown RE, Ghannoum MA, Mukherjee PK, Gillevet PM, Sikaroodi M. Quorum-sensing dysbiotic shifts in the HIV-infected oral metabiome. PLoS One 2015;10:e0123880.

[45] Duran-Pinedo AE, Paster B, Teles R, Frias-Lopez J. Correlation network analysis applied to complex biofilm communities. PLoS One 2011;6: e28438.

[46] Fisher CK, Mehta P. Identifying keystone species in the human gut microbiome from metagenomic timeseries using sparse linear regression. PLoS One 2014;9:e102451.

[47] Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. PLoS Comput Biol 2012;8:e1002687.

[48] Mcdonald D, Birmingham A, Knight R. Context and the human microbiome. Microbiome 2015;3:52.

[49] Wagner BD, Robertson CE, Harris JK. Application of two-part statistics for comparison of sequence variant counts. PLoS One 2011;6:e20296.

[50] Xiao L, Feng Q, Liang S, Sonne SB, Xia Z, Qiu X, et al. A catalog of the mouse gut metagenome. Nat Biotechnol 2015;33:1103–8.

[51] La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS One 2012;7:e52078.

[52] Sinclair L, Osman OA, Bertilsson S, Eiler A. Microbial community composition and diversity via 16S rRNA gene amplicons: evaluating the illumina platform. PLoS One 2015;10:e0116955.