Accepted Manuscript

It's all relative: analyzing microbiome data as compositions

Gregory B. Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, Juan José Egozcue

PII: S1047-2797(16)30073-4

DOI: 10.1016/j.annepidem.2016.03.003

Reference: AEP 7931

To appear in: *Annals of Epidemiology*

Received Date: 29 October 2015

Revised Date: 1 March 2016

Accepted Date: 23 March 2016

Please cite this article as: Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ, It's all relative: analyzing microbiome data as compositions, *Annals of Epidemiology* (2016), doi: 10.1016/j.annepidem.2016.03.003.

It's all relative: analyzing microbiome data as compositions

Gregory B. Gloor[a,*], Jia Rong Wu[a], Vera Pawlowsky-Glahn[b], Juan José Egozcue[c]

[a]*Dept. of Biochemistry, University of Western Ontario*

[b]*Dept. of Computer Science, Applied Mathematics, and Statistics, University of Girona, Spain*

[c]*Dept. of Civil and Environmental Engineering, Technical University of Catalonia, Spain*

[*]Corresponding author

Email address: ggloor@uwo.ca (Gregory B. Gloor)

Abstract

Purpose: The ability to properly analyze and interpret large microbiome datasets has lagged behind our ability to acquire such datasets from environmental or clinical samples. Sequencing instruments impose a structure on these data: the natural sample space of a 16S rRNA gene sequencing dataset is a simplex, which is a part of real space that is restricted to non-negative values with a constant sum. Such data are compositional, and should be analyzed using compositionally appropriate tools and approaches. However, the majority of the tools for 16S rRNA gene sequencing analysis assume these data are unrestricted.

Methods: We show that existing tools for compositional data (CoDa) analysis can be readily adapted to analyze high throughput sequencing datasets.

Results: The Human Microbiome Project tongue vs. buccal mucosa dataset shows how the CoDa approach can address the major elements of microbiome analysis. Reanalysis of a publicly available autism microbiome dataset shows that the CoDa approach in concert with multiple hypothesis test corrections prevent false positive identifications.

Conclusions: The CoDa approach is readily scalable to microbiome-sized analyses. We provide example code, and make recommendations to improve the analysis and reporting of microbiome datasets.

Keywords: compositional data, 16S rRNA gene sequencing, microbiome

2015 MSC: 09-311, 99-00 [*]

Introduction

High throughput sequencing has provided the laboratory tools needed for the large-scale culture-independent analysis of microbial communities. However, studies often fail to replicate earlier work even when similar technologies and strategies are used. For example, four recent papers indicating a link between autism and the gut microbiota have implicated many different associated genera. In mouse models, de Angelis et al. [1] identified 74 differential OTUs (operational taxonomic units) composing at least 90% of the sequence reads obtained, while de Theije et al [2] reported only three, and Hsiao et al. reported 67 [3]. While it is difficult to directly compare results across studies because of different sequencing and bioinformatic platforms, it is interesting to note that the three groups had very little overlap in taxonomically assigned genera, and the same named genus could exhibit statistically significant change in different directions in different experiments. In an additional study on humans, Kang et al. [4] identified 5 additional taxa that distinguished ASD (Autism Spectrum Disorder) from NT (neurotypical) control patients. However, as shown below, examination of these datasets suggests that their conclusions could be explained by chance alone. While these autism studies serve as a facile examples, the literature on the microbiome of other conditions are replete with similar experiments with the same shortcomings.

Hanage recently called for a skeptical re-examination of microbiome research results by posing five questions [5], with the first four being: Can experiments detect differences that matter? Are we examining correlation or causation? Is there a mechanism? Do the experiments reflect reality? These questions are beginning to be examined and addressed in detail by others.

Hanage's final question, "could anything else explain the results", was the most troubling because it is clear that the answer to the last question is a resounding yes! For example, work from the Microbiome

Quality Control Consortium (http://www.mbqc.org) shows that the wet lab, sequencing and computational approaches used can affect the quality, quantity and scope of the data and thus the conclusions reached. The consortium effort has already led to changes that make the analyses more reproducible (see for example [6, 7, 8]). A second example has been the realization that the reagents themselves often contribute contaminants to samples that contain low amounts of inputs [9, 10].

A further unappreciated problem is that of the sample space itself. We contend the data generated by high throughput sequencing are multivariate and constrained by an arbitrary constant sum. The constant sum constraint is imposed because all sequencing instruments have a fixed upper bound on the number of reads delivered. Data with this constraint are referred to as *compositional* data (CoDa) [11], and the only information that can be obtained from such a dataset is that of the ratios between the parts. We argue that it is more powerful and appropriate to examine these data using the existing tools designed for CoDa that have been used in the fields of geology and ecology [12]. Furthermore, we argue that sequencing data should be treated when possible in a Bayesian manner as probability distributions rather than as point estimates. Finally, we contend that proper statistical practice of correcting for multiple hypotheses is essential since the data are multivariate.

It may be helpful to draw an analogy between compositional data and proportional mortality. In both instances the true denominator is not known and only relative information is obtained. Additionally, variables with small numbers of events (OTUs represented by low counts, or deaths by rare causes) will be seen to be highly variable. Thus, a cautious interpretation of the dataset is required for compositional datasets as it is for proportional mortality datasets.

In this manuscript we demonstrate the utility of the CoDa frame of reference by comparing the buccal mucosa and tongue dorsum samples from the Human Microbiome Project. The CoDa approach almost completely separates samples from these adjacent sites, and many distinguishing OTUs can be identified. We then re-evaluate an available autism dataset with this framework and make recommendations for future work.

The origin of high throughput sequencing data

Datasets for 16S rRNA gene sequencing are generated from PCR (polymerase chain reaction) amplified random environmental samples of DNA molecules. We know that the total number of molecules varies by sample. For example, total bacterial load is quite different based on the environment (e.g. stool vs. mucous membranes for example), or across different phases of a cellular growth cycle. The data returned are random samples of the molecules in the environment, and each sample is subject to an arbitrary constant sum constraint imposed by the sequencer itself. Thus, the total number of reads assigned to an OTU can provide no information about the number of molecules in the original sample, and we can only investigate relative changes. This limitation is acknowledged when investigators treat 16S rRNA gene sequencing as proportions, percentages or 'relative abundances' in the data analysis, and by the RNA-seq convention of normalizing the counts across samples, an approach that has been advocated for microbiome studies [13]. However, these two approaches both effectively normalize all samples to a common denominator: relative abundance uses a constant denominator of 100, and the various count normalization approaches use an empirically determined denominator unique for each experiment [14].

Problems in the analysis of compositional data

A composition quantitatively describes parts of some whole. Parts are grouped in a vector of D positive components, $x = x_1, x_2, \ldots, x_D$. The composition is said to be closed when the sum of all components add up to a constant, for instance 1, 100, or a million. The major issue with these data is that the only relevant information is contained in the ratios between components [11, 15]. This property means that a composition can be multiplied by any positive constant without any change in its meaning. Thus, vectors with proportional positive components are equivalent from the compositional point of view [16, 17, 15]. Logarithmic transformation of the ratios ensures that the scale of ratios is symmetrical so that the permutation of numerator and denominator only causes a change of sign, and places values into an absolute scale. This so called log-ratio approach [11] allows inferences about compositional data to be performed on logarithms of ratios which do not change

when the composition is multiplied by a positive constant. Examples of these combinations are simple log-ratios as $\ln\ x_i / x_j$, or more complex log-ratios as $\ln\ \ x_1 x_2^{\ 1/2} /\ x_3 x_4 x_5^{\ 1/3}$; they are called log-contrasts. Remarkably, log-contrasts can be computed from non-closed compositions.

The analysis of compositions in a raw form has several inconsistencies. As first noted by Pearson in 1897 [18], the analysis of compositions with a constant sum results in spurious correlations. Several pitfalls can be detected even when the compositions are not closed.

Figure 1: Converting to proportions distorts the data. The Counts plot shows a scatter plot for samples A and B where each contain 50 OTUs. In the scatter plot the red points indicate 49 OTUs that differ only by a small random amount between two samples, and the blue point indicates one OTU that increased in abundance by 10-fold in sample B relative to sample A. The Proportions plot shows the same data after conversion to proportions by dividing the count values by the sample sum. This is formally identical to taking a random sample from the environment and subjecting it to DNA sequencing where the number of reads returned is determined by the instrument. Both plots are on a double logarithmic scale. The grey line shows were the values would fall if they were identical in both samples.

Figure 1 uses a simple example to illustrate one issue with the constant sum. Commonly an experimentalist wishes to count and compare the total number of molecules in two samples. Samples may contain many OTUs which values are independent and with similar abundance, and here one OTU has a 10-fold difference. A scatter plot of these two samples in the 'Counts' panel shows that we would infer by counting that the OTU represented by the blue circle has increased 10-fold in sample B compared to sample A, but that the rest are essentially unchanged.

The 'Proportions' panel shows the same data when these samples are subjected to DNA sequencing and random sampling in which case counting is no longer valid. The data are now constrained to a constant sum: note that *any* constant sum is exactly equivalent to a proportion and a constant scaling factor. All information regarding the total number of molecules in each sample is lost and only relative information remains. In the scatter plot, the distortion introduced by the constant sum is readily apparent: the invariant majority appear to have become *less* abundant, while the blue OTU appears to become more abundant: however it increases only from 6.8% in A to 41.9% in B. Note that this is *not* linearly related to the actual 10-fold difference. If we had many such samples, we would infer that the red OTUs are positively correlated (they all went down together!), and that they are negatively correlated with the blue OTU. This spurious correlation is caused only by restricting the data to have a constant sum.

Spurious correlation also appears when reducing a full composition to a composition with fewer parts: i.e., dealing with a so-called sub-composition. Note that 16S rRNA gene sequencing datasets are always sub-compositions because the constituent OTUs that are found in the samples depend on arbitrary decisions made during the data analysis pipelines (see the filtering and cutoff values recommended by QIIME [19] and mothur [20] for examples).

As an example of this problem, consider a vector of proportions like $x = \ x_1, x_2, x_3, x_4$, which is reduced to $y = C\ x_1, x_2, x_3$, where $C = 1/\ x_1 + x_2 + x_3$. Thus, *y* is a vector of proportions composed of only three first parts of *x*: clearly *y* is a sub-composition of *x*. Table 1 shows the sample covariance matrices for *x* and *y* for a synthetic data set of counts produced by simulating four log-normal variables with a given covariance structure (see the sub-composition code block in the supplement). If we are interested in evaluating the sample correlation between the first three components, we can see that these values are markedly different in the complete composition *x* than in the sub-compositon *y*. In fact none of the correlations involving the three initial components coincide, and differ substantially with no clear rule, thus deserving the name of spurious correlation. While not observed here, it is worth noting that a change of sign in the correlation is commonly found.

	x_1	x_2	x_3	x_4			y_1	y_2	y_3
x_1	1	-0.354	-0.154	-0.803		y_1	1	-0.835	-0.673
x_2	-0.354	1	0.343	-0.162		y_2	-0.835	1	0.155
x_3	-0.154	0.343	1	-0.358		y_3	-0.673	0.155	1
x_4	-0.803	-0.162	-0.358	1					

Table 1: Simulated example: sample size 100. Correlation matrices for the complete 4-part composition (left) and a 3-part sub-composition (right). The common 3×3 matrix differs substantially. Note the abundance of negative correlations related to the constant sum constrain for proportions.

It should be clear that any inference made on proportional data has the potential to be very misleading, and this is true whether we examine univariate differences [21, 22], correlations and methods that depend on correlation [11, 23, 24], or multivariate differences [11, 15].

The log-ratio solution

Aitchison realized that the underlying relationships between the parts can be captured by treating the data as ratios [11] and one widely used transformation is the centred log-ratio transformation of the D-part composition $x = x_1, x_2, \ldots, x_D$:

$$\mathrm{clr}\ x = \left(\ln \frac{x_1}{g_x}, \ln \frac{x_2}{g_x}, \ldots, \ln \frac{x_D}{g_x} \right) = z_1, z_2, \ldots, z_n \quad . \qquad (1)$$

Here g_x is the geometric mean of vector *x*. It should be noted that

$$z_i = \ln \frac{x_i}{g_x} = \ln x_i - \frac{1}{D} \ln x_1 + \ln x_2 + \cdots + \ln x_D \ ,$$

hence the sum of the components of $\mathrm{clr}\ x$ is always null. Moreover, each component z_i is a log-contrast where the value does not depend on whether *x* is closed to a constant or not. This is the first step to overcome the difficulties related to the constant sum constraint in compositional data analysis.

The clr transformation can be inverted to recover proportions from the clr-transformed data. The inverse is simply

$$x = \mathrm{clr}^{-1} = C \ \exp\ z_1 \ , \exp\ z_2 \ , \ldots, \exp\ z_D \quad , \quad C = \frac{1}{z_1 + z_2 + \cdots + z_D}.$$

Therefore, $\mathrm{clr}\ x$ is a representation of the composition *x* that is invariant when multiplied by positive constants and preserves all information conveyed by the ratios between its parts.

The elements of the log-ratio approach can be recast in a Euclidean space structure [15, 25]. This means that operations between compositions, namely perturbation and powering [11], behave as the sum and multiplication by real numbers in the ordinary Euclidean geometry. Furthermore, the Aitchison distance [26] between compositions is compatible with such operations. Perturbation of a composition can be interpreted as a multiplicative change in each part commonly reported as increase or decrease in percents. A way of computing the Aitchison distance between compositional vectors *x* and *y*, where $\mathrm{z} = \mathrm{clr}\ x$ and $\mathrm{w} = \mathrm{clr}\ y$ is

$$d_a\ x, y\ = \sqrt{z_1 - w_1\ ^2 + \ z_2 - w_2\ ^2 + \cdots + \ z_D - w_D\ ^2} \qquad (2)$$

The result is that *D*-part compositions, with perturbation, powering and Aitchison distance, have a $D-1$ -dimensional Euclidean space structure, called the Aitchison geometry of the simplex [27, 15]. As in the ordinary Euclidean space, compositions can be represented in Cartesian coordinates. These kind of coordinates are obtained by means of the so-called isometric log-ratio transformation (ilr) [28, 29]. These mathematical results are relevant as the standard multivariate statistical methods are applicable to ilr-coordinates without restrictions [30]. For instance, the Aitchison distance of equation 2 can be the base for cluster analysis of compositions and defining the variability of a compositional sample.

Compositional data and microbiome datasets

Microbiome analyses is impacted because the data are compositional, and below we show that compositional data originating from high throughput sequencing data can easily be evaluated in a framework that acknowledges the following principles [11, 17]:

• Permutation invariance: the order of the variables should not influence the results. In practice this is not difficult to achieve and is not discussed further. It should be noted that quantitative PCR is not permutation invariant since the choice of the internal standard affects the results.

• Scale invariance: multiplication of a composition by a positive constant must not change the information in the composition and, consequently, the conclusions.

• Sub-compositional coherence: any subset of the data should have distances between samples and variances that are equal to or less than those found for the full composition.

As a consequence of sub-compositional coherence we can add:

• Consistency of random sampling of variables: simple random sampling of variables should not lead to a contradictory change of the conclusions.

Unfortunately, many of the widely used practices for 16S rRNA gene sequencing do not adhere to these principles. For example, the unweighted Unifrac distance metric is not sub-compositionally coherent, since the values are strongly dependent on group membership and rarefaction instance [31]. Thus, apparently large differences between samples can arise by random chance. Other popular and useful metrics such as the weighted Unifrac distance and Bray-Curtis dissimilarity measure are also not sub-compositionally coherent, because these metrics have a common scale regardless of the number of OTUs included in the samples. No metric is consistent to random sampling of variables, and we have shown that ignoring random sampling can contribute to false positives [21, 22]. Furthermore, theses metrics are strongly affected by the relative abundances of the OTUs. In contrast, distances and variances determined after the centred log-ratio transformation conform to the first three principles, and we show below that this can be incorporated with Bayesian estimation to include the fourth principle as well. In addition, by recasting the analysis into a CoDa framework the data are now analyzed using variances and not abundances.

Results

In any given 16S rRNA gene sequencing experiment there are three main questions that typically are asked:

1. is there a difference between the samples in group A and group B? This is commonly known as β-diversity analysis and is a multivariate examination of the dataset.

2. which univariate differences in taxa abundances account for those differences? This is usually accomplished by simple statistical tests on each OTU, although a number of packages have been developed to account for specific experiments [32, 33, 21].

3. which taxa correlate?

These questions can (largely) be addressed using the tools of compositional data analysis.

Figure 2: Compositional biplot of the tongue vs. cheek HMP dataset. Samples are identified by a T or C for tongue or cheek (buccal mucosa), and coloured to make them distinguishable. The lowest named taxonomic rank name for each OTU is given in black. The arrows indicate the amount and direction of variation of the ratio of each OTU to all others in the dataset. The top and right axes indicate the values for variable loadings, and the bottom and left axes indicate the unit sum-of-squares values of the principle components. This is the so-called form biplot that preserves the relationships between samples. The proportion of variation explained by components 1 and 2 is also indicated.

β-Diversity: the compositional biplot

One major tool for CoDa analysis is the compositional biplot shown in Figure 2. This is based on a Singular Value Decomposition of the clr-transformed data displayed as a Principal Component Analysis (PCA) plot [34] that includes both samples and OTUs. This plot can replace Principle Co-ordinate plots based on Unifrac [35] or Bray-Curtis difference measures. Here we are examining the variation in the ratios of each part or sample on an absolute scale. These plots can be scaled to best project the variation in either the OTUs or the samples: the plot here is scaled to best show the relationships between samples, the so-called form biplot. The biplot is a rich resource that summarizes the dataset in a two-dimensional projection. In the example, only those named OTUs that are present in two-thirds of the samples are shown, however a biplot with OTUs found in at least 10% of the samples is given in the section 3.4 of the supplement, and shows little difference with the one displayed here. It is important to remember that the data has been converted to centred log-ratios, and so the length and direction of the arrow has no relationship to the abundance of the part.

Much information can be gleaned from these plots regarding the inter-relationships of all the parts, and Aitchison and Greenacre [34] give all the properties and rules for interpretation. First, Figure 2 shows that the two groups separate quite nicely upon PC1 which explains 25% of the variance in the dataset. At this point we could apply any standard multivariate method to discriminate between the groups, or to determine their separation, and section 3.3 of the supplement shows the results of k-means clustering applied to the HMP oral microbiome dataset. Second, the length and direction of the arrows gives a measure of the standard deviation of each OTU in the dataset, and so gives some idea regarding the contribution of each OTU clr-component to the observed separation between groups. However, we need to keep in mind that not all the OTUs are represented in this example dataset, and that this is a very rough approximation only. Nevertheless, it is interesting to observe that the variation vector for the majority of the OTUs are in the general direction of either the T or C samples, indicating that their variation is contributing to the observed split. Third, relationships between the OTUs can be observed: however, note that inferences about the OTU relationships must be performed on a covariance biplot projection. Nevertheless, for the purposes of explanation, pairs or groups of OTUs that have a short link between the arrowheads will probably exhibit a nearly constant ratio across all samples (we say they are compositionally associated [24]), and conversely those OTUs where the links between arrowheads are long will be found to have highly variable ratios and are said to be compositionally dissociated. Thus, in one simple plot, we can separate groups, identify species that exhibit high variation between groups and identify those species that may be strongly associated, up to the limit of the projection of the biplot. That the compositional biplot is much less resource intensive to make than a PCoA plot based on the Unifrac distance metric which requires the use of a standardized rooted tree is an added bonus [36, 31].

Univariate differences between groups

In theory, we could use the information in the compositional biplot to identify those taxa that are discriminatory between groups. However, in practice this is difficult when there are more than a few taxa under consideration, and when the projection is of poor quality. Two tools are available to examine univariate differences in OTU abundance that use the compositional data analysis approach, ALDEx2 [21, 22] and ANCOM [37]. Here we will illustrate the use of ALDEx2 since, rather than using a point estimate, it takes into account that the sequenced molecules are a random sample of the

DNA molecules collected. We have shown that point estimates can result in false positives in problematic datasets [22, 21].

Figure 3: An effect plot [38] summarizing the ALDEx2 output. In this plot each point represents an individual OTU from the dataset with the expected value of the log2 difference between groups on the y axis and the expected value of the maximum within-group dispersion on the x axis. Thus, the location each point in the plot provides a graphic summary of the standardized difference-dispersion relationship for each OTU. Points in red have an expected Benjamini-Hochberg adjusted p-value of 0.01 or less. Here we can observe that the majority of the OTUs have a dispersion of 4 or more (16-fold), and an absolute difference of 2 or less (4-fold). In other words the within-group dispersion greatly exceeds the between group difference of the vast majority of OTUs. Approximate lines of constant effect are shown where the solid lines represent approximate effect sizes of 2 and the dotted lines approximate effect sizes of 1.

ALDEx2 is a general purpose tool that can be used to identify univariate differences between groups within high throughput datasets that are count compositional data [22]. For each OTU, it uses Bayesian estimation to generate a posterior distribution of centred log-ratio values, conducts univariate analyses on the full distribution, and finally reports the expected value for both P and for the Benjamini-Hochberg corrected value of P. The between group difference is calculated as the median of the differences between random distributions of the clr values for groups A and B. The within group difference, called dispersion, is the median of the absolute differences between random distributions of the clr values within groups A or group B: the maximum value is reported and used (see [21] for a full technical description of how the distributions are generated and used). In essence, the ALDEx2 approach seeks to identify OTUs where the difference between groups is not likely to have occurred by random chance, and where the difference is robust to technical variation and multiple test correction. Note that other popular tools such as Metastats [33] or LefSe [32] examine point estimates of the simple proportions, which we see in Figure 1 can be distorted.

Figure 3 shows an 'effect plot' for the dataset, where the expected difference between groups is plotted vs. the expected maximum dispersion within either group ([21, 38]) plotted on a log scale. Each point in the plot represents an individual OTU, and they are coloured grey or red depending on their Benjamini-Hochberg adjusted p-value: those with a value < 0.01 are shown in red. The large number of red points is driven by the very large number of samples (366), and not necessarily by a large difference in relative abundance between groups. Note the concordance with Figure 2, where we see that the majority of OTUs display a standard deviation perpendicular to the axis separating the T and C groups. The large number of grey points with co-ordinates near [4,0] are low abundance OTUs that were filtered out prior to generating the biplot above.

However, we prefer to examine these datasets using an effect size metric, which for ALDEx2 is the expected value for the ratio between the between and within group differences [21, 39, 38]. Effect sizes are known to be more robust than p-value based measures [40], and Table 2 shows a much smaller set of OTUs that exceed an absolute effect size of 1. In other words, the OTUs generally are more variable within the groups than between the groups, despite having very low adjusted p-values. Thus, from this analysis we can see that the data are highly variable, and only a few OTUs are likely to be strongly diagnostic of these two body sites. Low effect sizes are typical for microbiome datasets.

	Diff	Disp	Effect	Overlap	E(p)	E(q)
22791	-4.691	4.008	-1.105	0.117	0.000	0.000
29014	-5.426	3.846	-1.327	0.081	0.000	0.000
30378	-4.750	3.710	-1.162	0.109	0.000	0.000
30902	-4.625	4.143	-1.058	0.114	0.000	0.000
38645	-4.843	3.707	-1.157	0.087	0.000	0.000
39103	-5.131	4.142	-1.143	0.087	0.000	0.000

Table 2: Table of significant OTUs. Row names indicate OTU identifier numbers. Column names are key ALDEx2 outputs and are midpoints of the Monte Carlo replicates. Diff: median difference between groups, Disp: maximum median difference within groups, Effect: median of Diff/Disp, Overlap: the overlapping proportion of the group A and B distributions, E(p): expected p-value of Wilcoxon Rank test: E(q): expected Benjamini-Hochberg corrected p-value

Examining correlations between OTUs

Finally, an increasing number of studies report correlation networks of microbiome datasets (see for example: [41, 42]), but as pointed out by others, the analysis of correlations in these datasets are problematic at best [24, 23, 37, 43, 44] and great care needs to be taken in their interpretation. As outlined above, and more completely in Lovell (2015), the essential problem is that whenever compositional data are analyzed the correlation matrix will be different if the OTUs making up the samples changes. Recall that the constituent OTUs in a sample varies because of arbitrary decisions made during data collection and analysis [19, 20]. Furthermore, it is now apparent that the way the OTUs are identified and grouped have significant implications for reproducibility of the data [8, 7]. Thus, microbiome datasets are necessarily only subsets of the true microbial diversity, and each subset can give a potentially unique result for correlation or clustering [24]. As noted by Lovell et al. (2015),

… in the absence of any other information or assumptions, correlation of relative abundances is just wrong.

The solution, is to identify pairs of OTUs with nearly invariant ratios across all samples [24]. Pairs of OTUs with this property are proportionally associated, and initial work has been done to develop this idea. The supplementary data of Lovell et al. [24] presents an excellent working walk-through with a real dataset.

Re-examining an existing dataset: what could be done better?

Figure 4: A compositional biplot summarizing control and *B. fragilis* treated microbiomes from the Hsiao et al. dataset. There were 10 control (IC) and 10 treated samples (Bf), and 703 OTUs in this dataset (black numbers). The control and treated samples separate poorly for such a small sample set, with 12% of the variation explained on the first principle component, 11% on the second, and 8% on the third.

The microbiome dataset of Hsiao et al. [3] was re-examined using a compositional data analysis paradigm. This group examined the effect of *Bacillus fragilis* supplementation on the gut microbiota of a mouse model of autism, and the conclusions are summarized in Figure 4 of that paper. Their examination of the multivariate dataset using any of a variety of multivariate measures indicated that the control and treatment groups were indistinguishable. An exploratory analysis with a compositional biplot leads to the same conclusion. Figure 4 shows that the control (IC) and treated (Bf) samples are intermingled, and that the percent variation explained suggested little more than random variance in this small dataset.

	Metastats p	LefSe p	Fig4 p	E(clr)	E(p)	E(q)	Count
R145	0.0010	0.0019	0.0022	0.1468	0.0622	0.6770	2.2000
L956	0.0040	0.0089	0.0145	0.7459	0.0618	0.7287	2.2000
L53	0.0110	0.0137	0.0201	1.3995	0.0527	0.7313	4.1500
S638	0.0277	0.0130	0.0149	-0.6421	0.4156	0.9020	0.2500
S836	0.0277	0.0130	0.0149	-0.7526	0.4349	0.9191	0.2500
L837	0.0080	0.0131	0.0147	0.1872	0.1956	0.8194	0.8000

Table 3: Table of highlighted taxa. The first four columns give the p-values associated with null hypothesis tests. The Hsiao paper relied primarily on Metastats and Lefse. The p-value in Figure 4 is

reported since this was a different value than from either of these tools. E(p) and E(q) are the ALDEx2 expected values for the Wilcoxon rank test and their associated Benjamini-Hochberg corrected values. E(clr) is the mean centered log-ratio abundance of the OTUs across all samples. Count is the mean count of the OTU across all samples.

The group went on to examine the univariate differences between groups, requiring that an OTU be identified by both the Metastats [33] and LefSe [32] tools to be considered significantly different. They concluded that there were 67 OTUs that were significantly different between the two conditions. A special emphasis was placed on 6 of the OTUs with the largest difference, and these are used in the example here. Table 3 shows the reported p-values from the original analysis where the p-values for Metastats and LefSe are taken from Table S2 of that paper, and the Fig4 p-values are from Figure 4 of the original paper [3]. These values can be compared to those from the Bayesian univariate compositional method implemented in ALDEx2. The final column reports the mean count of the OTUs across all samples.

Re-analysis of the dataset using an approach that accounts for random sampling and that uses the centred log-ratio transformed data indicates that none of the OTUs were different between the two conditions. There are three main reasons. First, a compositional approach was not used. Second, the values for OTUs S638, S836 and L837 are below a p-value cutoff of 0.05 with the point estimate methods (Metastats, LefSe), but not with ALDEx2. This is because these three OTUs are exceedingly rare in the dataset, having an average relative abundance across all samples, E(clr), that is approximately equal to or less than the geometric mean abundance. In point of fact, these three OTUs had mean counts of less than 1 in the dataset and thus were so rare that a point estimation of their true abundance was unreliable [21]. Third, while the authors used multiple different approaches to determine significance, they did not account for multiple hypothesis testing. This oversight is significant because there were 703 OTUs in this dataset, and so ~ 35 raw p-values less than 0.05 were expected by chance alone. The E(q) column of Table 3 shows that none of these OTUs were significant when the large number of hypothesis tests were properly accounted for, and indeed there are no OTUs with significantly different relative abundances as shown in section 4.2 of the supplement.

Reproducibility

We include as supplementary information a document with all the annotated R code needed to reproduce the figures and analyses reported here. In addition, we include several supplementary figures that more fully demonstrate and explain the point of view put forward in this perspective.

Conclusions

As outlined in the introduction, the origin of variability in 16S rRNA gene sequencing datasets include variation in protocols at all stages of collection and analysis including the wet lab, sequencing, bioinformatic and statistical protocols. Here we focussed on the statistical analysis and demonstrated that the datasets can be analyzed using techniques that have been developed for compositional data. We found that this approach shows strong separation of 16S rRNA gene sequencing survey samples from the tongue and buccal mucosa. A re-analysis of one dataset in the autism literature highlighted several significant shortcomings. However, it should be pointed out that all of the other autism microbiome papers cited in the introduction exhibited similar flaws of not using multiple test corrections, not accounting for sampling variation, and not treating their datasets as compositional. In addition, it should be pointed out that these shortcomings are rife in the microbiome literature, and editors, reviewers and readers must become aware of these problems.

In the future, we suggest that a compositional approach that accounts for sampling variation and that acknowledges the multivariate nature of the is likely to increase the reliability of microbiome analyses. We have four suggestions that will improve these analyses and their reporting:

1. All analyses must correct for multiple hypothesis tests. It is well known that the underlying distribution of p-values is random uniform, and the large number of OTUs in a typical experiment practically guarantee ‘significant’ OTUs regardless of the experiment.

2. Data should be treated as compositions. We have shown that the tools developed for other fields can be easily adapted for the multivariate comparison and analysis of a test HMP-derived dataset. We suggest that β-diversity analysis be performed with compositional biplots as an additional exploratory tool. Biplots have several advantages: 1) they be generated more rapidly and with fewer computational resources than can PCoA plots from weighted or unweighted Unifrac because they are not dependent on an underlying tree. This facilitates exploration, rather than an analysis that follows a check-box. 2) they are more easily interpreted than plots generated by Unifrac or Bray-Curtis dissimilarity measures because they are based on a mathematically rigorous underpinning that is less susceptible to the vagaries of filtering in the dataset. 3) both the samples and the OTUs can be displayed on a common plot with a scale that is directly interpretable as a variance, facilitating rapid identification of sample separation, the OTUs that are likely associated with that separation, and the identification of OTUs that are likely to be compositionally associated. This is quite unlike the biplots generated by QIIME, which display the separation of samples based on a distance or dissimilarity metric, but that displays the abundances of the taxa. The use of a compositional paradigm extends to the visualizations that depend on ordination, clustering, heat maps, etc. The supplementary information of Lovell et al. [24] illustrates the utility of the CoDa approach for this.

3. Data analysis can be greatly facilitated by using Bayesian approaches rather than point estimates. Investigators should realize that the counts observed are only one possible outcome of the sequencing run since sequencing the same library would give different results [21]. Investigators also need to realize that the greatest proportional difference is at the low count margin. Thus, incorporating this point of view into all analyses would prevent the rarest OTUs from being the 'most significant'. Furthermore, the Bayesian approach in combination with the centred log-ratio transformation removes the need to 'rarefy' or otherwise normalize for sequencing depth [21, 24] except when the difference in depth is egregious.

4. The original sequencing reads, the code used, and the raw tables of counts need to be made available as part of the review and publication process. While many journals require deposition of the raw data, many still do not require the exact code used to convert the raw reads to tables, nor do they require transparent reporting of all methods. Commendably, Hsiao et al. [3] are one of the few groups that made the raw table of counts available as part of their supplementary information. Others should be encouraged to do so. These raw tables of counts are required for Bayesian inference: tables of proportions, percentages or summary tables do not provide the level of detail required.

Glossary of technical terms

16S rRNA the highly conserved small ribosomal RNA gene

OTU Operational Taxonomic Unit, usually defined as encompassing all sequences of rRNA that are 97% identical

CoDa Compositional Data, datasets with an arbitrary sum in which only relative information is available

PCR Polymerase Chain Reaction, a method to amplify specific short segments of DNA from a genome

clr Centered log-ratio transformation, see Equation 1

QIIME Quantitative Insights Into Microbial Ecology: A comprehensive software suite for microbiome analysis

mothur A comprehensive software package for microbiome analysis: the name has no intrinsic meaning

Aitchison Simplex Cartesian co-ordinates for compositional data. It has one fewer dimension than there are components

Aitchison Distance Euclidian distance on the Aitchison Simplex

Unifrac distance A distance measure that incorporates relative relatedness of the OTUs between samples. Unweighted Unifrac for two samples is the sum of the unshared branch lengths in a phylogenetic tree divided by the sum of shared plus unshared branch lengths. Weighted Unifrac is the same measure scaled by the relative abundance of the OTUs on the phylogenetic tree

Bray-Curtis dissimilarity A dissimilarity measure commonly used in ecology. It is the sum of the minimum counts for species in common between groups A and B, divided by the total number of counts for all species in both groups.

Acknowledgements

Work in G.B. Gloor's lab has been supported by a Discovery Grant from the National Science and Engineering Research Council of Canada. J.R. Wu was supported by a CIHR grant to Dr. J. Allard and GBG. Drs J.J. Egozcue and V. Pawlowsky-Glahn have been supported by the *Spanish Ministry of Economy and Competitiveness* under the project *METRICS* (Ref. MTM2012-33236); and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under the project *COSDA* (Ref: 2014SGR551).

References

[1] M. De Angelis, M. Piccolo, L. Vannini, S. Siragusa, A. De Giacomo, D. I. Serrazzanetti, F. Cristofori, M. E. Guerzoni, M. Gobbetti, R. Francavilla, Fecal microbiota and metabolome of children with autism and pervasive developmental disorder not otherwise specified, PLoS One 8 (10) (2013) e76993. doi:10.1371/journal.pone.0076993.

[2] C. G. M. de Theije, H. Wopereis, M. Ramadan, T. van Eijndthoven, J. Lambert, J. Knol, J. Garssen, A. D. Kraneveld, R. Oozeer, Altered gut microbiota and activity in a murine model of autism spectrum disorders, Brain Behav Immun 37 (2014) 197–206. doi:10.1016/j.bbi.2013.12.005.

[3] E. Y. Hsiao, S. W. McBride, S. Hsien, G. Sharon, E. R. Hyde, T. McCue, J. A. Codelli, J. Chow, S. E. Reisman, J. F. Petrosino, P. H. Patterson, S. K. Mazmanian, Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders, Cell 155 (7) (2013) 1451–63. doi:10.1016/j.cell.2013.11.024.

[4] D.-W. Kang, J. G. Park, Z. E. Ilhan, G. Wallstrom, J. Labaer, J. B. Adams, R. Krajmalnik-Brown, Reduced incidence of Prevotella and other fermenters in intestinal microflora of autistic children, PLoS One 8 (7) (2013) e68322. doi:10.1371/journal.pone.0068322.

[5] W. P. Hanage, Microbiology: Microbiome science needs a healthy dose of scepticism, Nature 512 (7514) (2014) 247–8. doi:10.1038/512247a.

[6] R. Flores, J. Shi, G. Yu, B. Ma, J. Ravel, J. J. Goedert, R. Sinha, Collection media and delayed freezing effects on microbial composition of human stool, Microbiome 3 (2015) 33. doi:10.1186/s40168-015-0092-7.

[7] Y. He, J. G. Caporaso, X.-T. Jiang, H.-F. Sheng, S. M. Huse, J. R. Rideout, R. C. Edgar, E. Kopylova, W. A. Walters, R. Knight, H.-W. Zhou, Stability of operational taxonomic units: an

important but neglected property for analyzing microbial diversity, Microbiome 3 (2015) 20. doi:10.1186/s40168-015-0081-x.

[8] J. R. Rideout, Y. He, J. A. Navas-Molina, W. A. Walters, L. K. Ursell, S. M. Gibbons, J. Chase, D. McDonald, A. Gonzalez, A. Robbins-Pianka, J. C. Clemente, J. A. Gilbert, S. M. Huse, H.-W. Zhou, R. Knight, J. G. Caporaso, Subsampled open-reference clustering creates consistent, comprehensive otu definitions and scales to billions of sequences, PeerJ 2 (2014) e545. doi:10.7717/peerj.545.

[9] Salter, Reagent and laboratory contamination can critically impact sequence-based microbiome analyses, BMC Biology 12 (2014) 87.

[10] J. Jervis-Bardy, L. E. X. Leong, S. Marri, R. J. Smith, J. M. Choo, H. C. Smith-Vaughan, E. Nosworthy, P. S. Morris, S. O'Leary, G. B. Rogers, R. L. Marsh, Deriving accurate microbiota profiles from human samples with low bacterial content through post-sequencing processing of illumina miseq data, Microbiome 3 (2015) 19. doi:10.1186/s40168-015-0083-8.

[11] J. Aitchison, The Statistical Analysis of Compositional Data, Chapman & Hall, 1986.

[12] K. G. Van den Boogaart, R. Tolosana-Delgado, Analyzing compositional data with R, Springer, London, UK, 2013.

[13] P. J. McMurdie, S. Holmes, Waste not, want not: why rarefying microbiome data is inadmissible, PLoS Comput Biol 10 (4) (2014) e1003531. doi:10.1371/journal.pcbi.1003531.

[14] J. Sun, T. Nishiyama, K. Shimizu, K. Kadota, TCC: an R package for comparing tag count data with robust normalization strategies, BMC Bioinformatics 14 (2013) 219. doi:10.1186/1471-2105-14-219.

[15] V. Pawlowsky-Glahn, J. J. Egozcue, R. Tolosana-Delgado, Modeling and Analysis of Compositional Data, John Wiley & Sons, 2015.

[16] J. A. Martín-Fernández, C. Barceló-Vidal, V. Pawlowsky-Glahn, Dealing with zeros and missing values in compositional data sets using nonparametric imputation, Mathematical Geology 35 (3) (2003) 253–278.

[17] J. J. Egozcue, Reply to "on the harker variation diagrams;…" by ja cortés, Mathematical Geosciences 41 (7) (2009) 829–834.

[18] K. Pearson, 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 (1897) 489–498.

[19] J. G. Caporaso, J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman, E. K. Costello, N. Fierer, A. G. Peña, J. K. Goodrich, J. I. Gordon, G. A. Huttley, S. T. Kelley, D. Knights, J. E.

Koenig, R. E. Ley, C. A. Lozupone, D. McDonald, B. D. Muegge, M. Pirrung, J. Reeder, J. R. Sevinsky, P. J. Turnbaugh, W. A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld, R. Knight, QIIME allows analysis of high-throughput community sequencing data, Nat Methods 7 (5) (2010) 335–6. doi:10.1038/nmeth.f.303.

[20] P. D. Schloss, S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann, E. B. Hollister, R. A. Lesniewski, B. B. Oakley, D. H. Parks, C. J. Robinson, J. W. Sahl, B. Stres, G. G. Thallinger, D. J. Van Horn, C. F. Weber, Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities, Appl Environ Microbiol 75 (23) (2009) 7537–41. doi:10.1128/AEM.01541-09.

[21] A. D. Fernandes, J. M. Macklaim, T. Linn, G. Reid, G. B. Gloor, ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq, PLoS ONE 8 (7) (2013) e67019.

[22] A. D. Fernandes, J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, G. B. Gloor, 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 (2014) 15. doi:10.1186/2049-2618-2-15.

[23] J. Friedman, E. J. Alm, Inferring correlation networks from genomic survey data, PLoS Comput Biol 8 (9) (2012) e1002687. doi:10.1371/journal.pcbi.1002687.

[24] D. Lovell, V. Pawlowsky-Glahn, J. J. Egozcue, S. Marguerat, J. Bähler, Proportionality: a valid alternative to correlation for relative data, PLoS Comput Biol 11 (3) (2015) e1004075. doi:10.1371/journal.pcbi.1004075.

[25] D. Billheimer, P. Guttorp, W. F. Fagan, Statistical interpretation of species composition, Journal of the American statistical Association 96 (456) (2001) 1205–1214.

[26] J. Aitchison, Reducing the dimensionality of compositional data sets, Journal of the International Association for Mathematical Geology 16 (6) (1984) 617–635.

[27] V. Pawlowsky-Glahn, J. J. Egozcue, Geometric approach to statistical analysis on the simplex, Stochastic Environmental Research and Risk Assessment 15 (5) (2001) 384–398.

[28] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, C. Barcel O-Vidal, Isometric logratio transformations for compositional data analysis., Math. Geol. 35 (3) (2003) 279–300.

[29] J. Egozcue, V. Pawlowsky-Glahn, Groups of parts and their balances in compositional data analysis, Mathematical Geology 37 (7) (2005) 795–828.

[30] G. Mateu-Figueras, V. Pawlowsky-Glahn, J. J. Egozcue, V. Pawlowsky-Glahn, A. Buccianti, The principle of working on coordinates, Compositional data analysis: Theory and applications (2011) 31–42.

[31] C. Lozupone, R. Knight, Unifrac: a new phylogenetic method for comparing microbial communities, Applied and environmental microbiology 71 (12) (2005) 8228–8235.

[32] N. Segata, J. Izard, L. Waldron, D. Gevers, L. Miropolsky, W. S. Garrett, C. Huttenhower, Metagenomic biomarker discovery and explanation, Genome Biol 12 (6) (2011) R60. doi:10.1186/gb-2011-12-6-r60.

[33] J. R. White, N. Nagarajan, M. Pop, Statistical methods for detecting differentially abundant features in clinical metagenomic samples, PLoS Comput Biol 5 (4) (2009) e1000352. doi:10.1371/journal.pcbi.1000352.

[34] J. Aitchison, M. Greenacre, Biplots of compositional data, Journal of the Royal Statistical Society: Series C (Applied Statistics) 51 (4) (2002) 375–392.

[35] M. Hamady, C. Lozupone, R. Knight, Fast Unifrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and phylochip data, ISME J 4 (1) (2010) 17–27. doi:10.1038/ismej.2009.97.

[36] J. R. Long, V. Pittet, B. Trost, Q. Yan, D. Vickers, M. Haakensen, A. Kusalik, Equivalent input produces different output in the Unifrac significance test, BMC bioinformatics 15 (1) (2014) 278.

[37] S. Mandal, W. Van Treuren, R. A. White, M. Eggesbø, R. Knight, S. D. Peddada, Analysis of composition of microbiomes: a novel method for studying microbial composition, Microb Ecol Health Dis 26 (2015) 27663.

[38] G. B. Gloor, J. M. Macklaim, A. D. Fernandes, Displaying variation in large datasets: a visual summary of effect sizes, Journal of Computational and Graphical Statistics in press.

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

[40] L. G. Halsey, D. Curran-Everett, S. L. Vowler, G. B. Drummond, The fickle p value generates irreproducible results, Nat Methods 12 (3) (2015) 179–85. doi:10.1038/nmeth.3288.

[41] T. Kelder, J. H. M. Stroeve, S. Bijlsma, M. Radonjic, G. Roeselers, Correlation network analysis reveals relationships between diet-induced changes in human gut microbiota and metabolic health, Nutr Diabetes 4 (2014) e122. doi:10.1038/nutd.2014.18.

[42] K. Faust, J. Raes, Microbial interactions: from networks to models, Nat Rev Microbiol 10 (8) (2012) 538–50. doi:10.1038/nrmicro2832.

[43] K. Faust, J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers, J. Raes, C. Huttenhower, Microbial co-occurrence relationships in the human microbiome, PLoS Comput Biol 8 (7) (2012) e1002606. doi:10.1371/journal.pcbi.1002606.

[44] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, R. A. Bonneau, Sparse and compositionally robust inference of microbial ecological networks, PLoS Comput Biol 11 (5) (2015) e1004226. doi:10.1371/journal.pcbi.1004226.