NRC Research Press

ARTICLE

Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data

Gregory B. Gloor and Gregor Reid

Abstract: A workshop held at the 2015 annual meeting of the Canadian Society of Microbiologists highlighted compositional data analysis methods and the importance of exploratory data analysis for the analysis of microbiome data sets generated by high-throughput DNA sequencing. A summary of the content of that workshop, a review of new methods of analysis, and information on the importance of careful analyses are presented herein. The workshop focussed on explaining the rationale behind the use of compositional data analysis, and a demonstration of these methods for the examination of 2 microbiome data sets. A clear understanding of bioinformatics methodologies and the type of data being analyzed is essential, given the growing number of studies uncovering the critical role of the microbiome in health and disease and the need to understand alterations to its composition and function following intervention with fecal transplant, probiotics, diet, and pharmaceutical agents.

Key words: microbiome, compositional data, correlation, multiple test correction.

Résumé : Un atelier tenu dans le cadre du colloque annuel de 2015 de la Société Canadienne des Microbiologistes a traité des méthodes d'analyse de données et de l'importance de l'analyse de données exploratoire dans l'analyse d'ensembles de données générées par séquençage d'ADN à haut débit. Dans le présent article, on fait la synthèse du contenu de l'atelier, des nouvelles méthodes d'analyse et des éléments signalant l'importance d'une analyse réfléchie. L'atelier a abordé la logique justifiant l'usage de l'analyse des données sur la composition et a fait la démonstration de ces méthodes afin d'examiner 2 ensembles de données de microbiomes. Il est essentiel de se doter d'une conception claire des méthodes de bioinformatique et du type de données à analyser, au vu du nombre croissant d'études révélant le rôle prépondérant du microbiome pour la santé et dans les pathologies, et de la nécessité de comprendre les changements dans sa composition et son fonctionnement des suites d'interventions mettant en jeu la transplantation de matières fécales, des probiotiques ou des produits pharmaceutiques. [Traduit par la Rédaction]

Mots-clés : microbiome, données sur la composition, corrélation, correction en raison de tests multiples.

Introduction

Human microbiome studies have shown a major link between microbial composition and health and disease and dysbiosis (Frémont et al. 2013; Lourenço et al. 2014; Urbaniak et al. 2014). High-throughput DNA sequencing methodologies have made this possible, along with breakthroughs in culturing techniques. The former has used approaches such as 16S rRNA gene sequencing, metagenomics, transcriptomics, and meta-transcriptomics, leading to vast data sets that must be simplified and analyzed (Di Bella et al. 2013). Indeed, each sample may have tens of thousands to millions of sequence reads associated with it, and the entire data set across all samples can easily exceed many hundreds of millions of reads. Such has been the rapidity of these developments that some studies appear to have been published using methods that are potentially flawed. The result can be papers with serious deficiencies that are publicized as major advances or breakthroughs (Reardon 2013), when in some cases the data are far from sufficient for such claims. We will examine the evidence for one of these papers below (Hsiao et al. 2013).

Data for microbiome analysis are collected by the following general workflow. The sample (swab, stool, saliva, urine, or other type) is collected; the DNA is isolated and used in a polymerase chain reaction with primers specific to one or more variable regions of the 16S rRNA gene. It is also possible to target other conserved genes

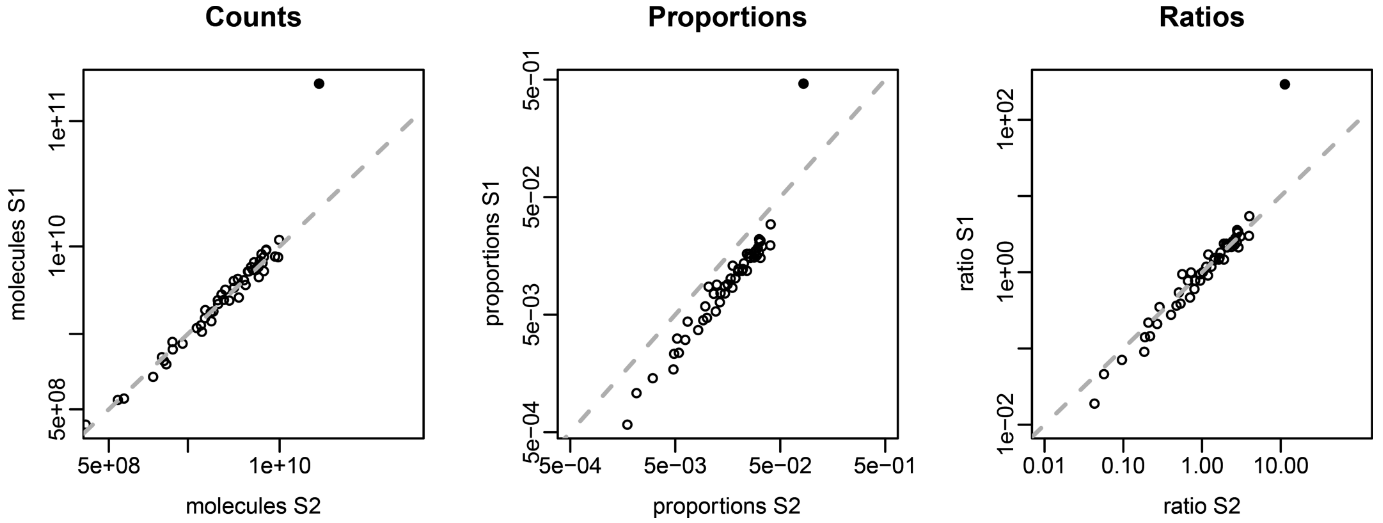
Received 8 December 2015. Revision received 30 March 2016. Accepted 11 April 2016.

G.B. Gloor. Department of Biochemistry, Western University, London, Ontario, Canada; Canadian Center for Human Microbiome and Probiotic Research, Lawson Health Research Institute, London, Ontario, Canada.
G. Reid. Canadian Center for Human Microbiome and Probiotic Research, Lawson Health Research Institute, London, Ontario, Canada; Department of Microbiology and Immunology, and Department of Surgery, Western University, London, Ontario, Canada.
Corresponding author: Gregory B. Gloor (email: ggloor@uwo.ca).

Can. J. Microbiol. **62**: 692–703 (2016) dx.doi.org/10.1139/cjm-2015-0821

Published at www.nrcresearchpress.com/cjm on 12 April 2016.

Fig. 1. The difference between counts, proportions, and ratios. The "Counts" panel shows a scatterplot of a simulated data set with 2 samples composed of 49 invariant taxa in open circles, and 1 taxon that changes in count 10-fold (black-filled circle). This is the type of data that most current analysis tools in the microbiome field expect is being analyzed. The "Proportions" panel shows the same samples after they have been sequenced and so constrained to have a constant sum. With such a constraint, their representation is the same whether the sum is 1 (as shown here) or an arbitrarily larger number (such as would be obtained from a sequencing instrument). The distortion in the data is obvious: the black-filled circle still appears to be more abundant, but the open circles appear to have become less abundant! It is obvious that we would draw incorrect inferences regarding abundance changes in these data, yet these are the data as used by existing tools. The third panel shows that much of this distortion can be removed using a ratio transformation where each count (or proportion) is divided by the geometric mean of the 50 taxa in the sample. Examination of the data after this transformation can thus provide more robust inferences.



such as the *cpn60* gene (Schellenberg et al. 2009). However, analysis problems are the same regardless of the amplification target chosen, and Walker et al. (2015) present a good summary of how choices taken upstream of data analysis affect the results. Following amplification, a random sample of the product is used to make a sequencing library, and it is common to multiplex many samples in the library. A small aliquot of the library is processed on the high-throughput DNA sequencing instrument. As outlined below, this workflow imposes constraints on the resulting data.

It should be recognized that the investigator is sequencing a random sample of the DNA in the library, which is itself a random sample of the DNA in the environment. Thus, it is important to ensure that any analysis takes this random component into account (Fernandes et al. 2013).

Perhaps less obvious is that the number of sequencing reads obtained for a sample bears no relationship to the number of molecules of DNA in the environment, because the number of reads obtained for a sample is determined by the capacity of the instrument. For example, the same library sequenced on an Illumina MiSeq or HiSeq would return approximately 20 million or 200 million reads. That there is no information in the actual read numbers per sample is implicitly acknowledged by the common use of "relative abundance" values for analysis of microbiome data sets. Such data sets are referred to as compositional, and there is a long history of the development of proper analysis techniques for such data in other fields (Pawlowsky-Glahn et al. 2015).

Compositional data are a term used to describe a data set in which the parts in each sample have an arbitrary or noninformative sum (Aitchison 1986), such as data obtained from high-throughput DNA sequencing (Friedman and Alm 2012; Fernandes et al. 2013, 2014). These data have long been known to be problematic (Pearson 1896), and we now understand that multivariate data analysis approaches such as ordination and clustering and univariate methods that measure differential abundance are invalid (Aitchison 1986; Warton et al. 2012; Friedman and Alm 2012; Fernandes et al. 2013; Pawlowsky-Glahn et al. 2015).

The essential problem is illustrated in Fig. 1, in which we set up an artificial example and count the number of molecules in the environment. We allow one part (shown as solid black circle) to increase 10-fold between samples 1 and 2, while the abundance of the other 49 parts (in open circles) remains unchanged. The proportion panel shows how the data are distorted when we convert it to relative abundances or proportions, or as happens when the sequencing instrument imposes a constant sum. The black part still appears to become more abundant, although it is less than a 10-fold change. However, the 49 other parts appear to become less abundant. This property leads to the *negative correlation bias* observed in compositional data, and renders invalid any type of correlation- or covariance-based analysis such as correlation networks, principal component analysis, and others (Pearson 1896; Aitchison 1986). Note that this distortion will also lead to false univariate inferences as well (Fernandes et al. 2013, 2014).

The original issue with compositional data identified by Pearson (1896) was that of spurious correlation. That is, 2 or more variables can appear to be correlated simply because the data are transformed to have a constant sum. Spurious correlation also causes the correlations observed in these data to depend on the membership of the sample. For example, consider the simple case of 3 samples (a, b, and c) with 4 taxonomic variables measured to have the following absolute counts in 3 environmental samples (i.e., samples are in rows, taxa are in columns):

$$\mathrm{abc} = \begin{bmatrix} 470 & 66 & 839 & 751 \\ 541 & 569 & 787 & 512 \\ 167 & 906 & 959 & 504 \end{bmatrix},$$

$$\mathrm{cor(abc)} = \begin{bmatrix} & -0.68 & -\mathbf{0.99} & 0.36 \\ -0.77 & & \mathbf{0.59} & -0.93 \\ -\mathbf{0.30} & -\mathbf{0.37} & & -0.25 \\ 0.55 & -0.95 & 0.62 & \end{bmatrix}.$$

The Pearson correlation for the numerical values is in the upper triangle of the right-hand matrix, and we see that taxon 1 and taxon 3 have a near perfect negative correlation of –0.99 (shown in bold), and taxon 2 and taxon 3 have a positive correlation of 0.59. The lower triangle on the right-hand matrix shows the Pearson correlation values that are found when these are converted to relative abundances by dividing by the total sum of counts in each sample. Now, the correlations between the same taxa have changed. The correlation between 1 and 3 is now moderately negative at –0.30, and that between 2 and 3 is now –0.37. Thus, the correlations observed in compositional data are not the same as the correlations for the counts, and the correlations measured can even change sign.

There is a further complication: the correlations observed in compositional data depend on the membership in the sample. So, for example, when the last value is dropped from each sample, the correlation between taxa 1 and 2 is positive (0.43), and the correlation between taxa 2 and 3 is even more strongly negative at –0.79. Thus, a correlation determined from compositional data has the potential to be wildly wrong, and normal approaches to determine correlation cannot be used (Friedman and Alm 2012; Lovell et al. 2015; Kurtz et al. 2015). It is worth noting that any method of determining correlation (including Spearman, Kendall, etc.) will suffer from the same problems. Thus, the current tools used to examine the analysis goals give results that may be inconsistent, difficult to interpret, and in many cases completely wrong (Filzmoser et al. 2009; Friedman and Alm 2012; Fernandes et al. 2013, 2014, Lovell et al. 2015; Kurtz et al. 2015).

The essential first step of proper compositional data analysis (CoDa) is to convert the relative abundances of each part, or the values in the table of counts for each part, to ratios between all parts. This can be accomplished in several ways (Aitchison 1986), but the most widely used and most convenient for our purposes is to convert the data using the centred log-ratio (clr) transformation. So if X is a vector of numbers that contains D parts

$$X = [x_1, x_2, \ldots x_D]$$

the clr of X can be computed as

$$X_{\mathrm{clr}} = [\log(x_1/g_x), \log(x_2/g_x), \ldots \log(x_D/g_x)]$$

where g_x is the geometric mean of all values in vector X (Aitchison 1986). This simple transformation renders valid all standard multivariate analysis techniques (Aitchison 1986; van den Boogaart and Tolosana-Delgado 2013; Pawlowsky-Glahn et al. 2015) and, as shown in the Ratios panel of Fig. 1, can reconstitute the shape of the data so that univariate analyses are also more likely to be valid. This transformation is also the starting point for essentially all CoDa-based assessments of the data sets.

A CoDa approach would be robust if microbiome data sets were not sparse, that is, they did not contain any 0 values. However a frequent criticism of the CoDa approach is that the geometric mean cannot be computed if any of the values in the vector are 0. It is here we reiterate that our data represent the counts per taxon through the process of random sampling (Fernandes et al. 2013, 2014). Thus, some 0 values could arise simply by random chance, while others arise because of true absence of the taxon in the environment. Fortunately, we can couple Bayesian approaches to estimate the likelihood of 0 values with the compositional analysis approach (Fernandes et al. 2013, 2014; Gloor et al. 2016*b*). With this paradigm we dispose of taxa with 0 counts in all or most samples (Palarea-Albaladejo and Martín-Fernández 2015), and assign an estimate of the likelihood of the 0 being a sampling artifact to the remainder. When performing univariate tests or correlation analyses, it is often convenient to keep many such estimates of 0 and to determine the expected value of test statistics to reduce false-positive inferences (Friedman and Alm 2012; Fernandes et al. 2013, 2014).

Microbiome analysis tools that account for compositional data

Fortunately, the CoDa problem of microbiome data sets is starting to be examined by several groups and there are now an increasing number of tools available as outlined below.

These tools can be applied to address 3 major objectives of many microbiome analyses:

1. Do the data show any structure? That is, do the data partition into groups?
2. What is the difference between groups? This can be between groups identified beforehand or following the exploratory data analysis.
3. What is the correlation structure of the taxonomic groups? Do any of these taxa correlate with the metadata?

These analyses are usually done using either the mothur (Schloss et al. 2009) or the QIIME (Kuczynski et al. 2012) aggregated toolsets, containing approaches adapted from the field of ecology. However, the use of an analysis paradigm based on CoDa (Aitchison 1986) offers a number of advantages over these tools, as explained below.

The first objective is to determine if there is structure in the data set. In the microbiome field this is generally described as β-diversity analysis. β-Diversity as currently used requires a distance or dissimilarity measure, and popular ones include the unweighted or weighted Unifrac distance metrics (Lozopone and Knight 2005) or the Bray–Curtis dissimilarity metric. These methods are included in both the mothur and QIIME toolkits. The distance metrics from these tools can be used to generate Principal Co-ordinate (PCoA) plots that can be used to assess similarities and differences between samples and groups. Unfortunately, distance-based tools can confuse location (difference) and dispersion (variance) effects (Warton et al. 2012), and so additional approaches based on a compositional paradigm should be used for exploratory data analysis.

The CoDa analysis analog to PCoA is a principal component analysis (PCA) of clr-transformed data that has been modified to either remove taxa with 0 observed counts or to adjust 0 values to an estimated value (Palarea-Albaladejo and Martín-Fernández 2015). PCA has the advantage of being a more interpretable metric than PCoA, since it directly assesses the variance in the data and because both the locations of the samples and the contribution of each taxon to the total variance can be shown on the so-called compositional biplot (Aitchison and Greenacre 2002). The ability to examine variation of both the samples and the taxa on the same plot provides powerful insights into which taxa are compositionally associated and which taxa are driving (or not) the location of particular samples. Thus, the biplot can serve as a summary of the entire data set, and it is up to the investigator to attach numerical significance to the qualitative results observed. The example usage of compositional biplots is explained in detail below.

The second major objective is often to determine which taxa are driving the difference observed between groups. Several methods are in widespread use to assess the difference in abundance of taxa between groups. These include microbiome-specific methods such as Metastats (White et al. 2009) or LEfSe (Segata et al. 2011), and more general *t* tests or nonparametric tests. However, all use as input a table of proportional abundances. As shown in Fig. 1, examination of proportions can result in a gross distortion of the data, such that some taxa can appear to change in abundance when measured by proportion, when in fact, their true abundance in the environment may be unchanged. This effect can be ameliorated by the clr transformation.

There are 2 approaches that assess differential abundance in a CoDa framework. The simplest approach is the ANCOM tool (Mandal et al. 2015), which assesses statistical significance on log-ratio-transformed data. This is more robust than both traditional *t* tests and more sophisticated approaches such as zero-inflated Gaussian methods. It should be noted that the 0-replacement value used is fixed in the software.

A slightly more complex approach is used by the ALDEx2 (Fernandes et al. 2013, 2014) package, available from Bioconductor (Gentleman et al. 2004). Like ANCOM, ALDEx2 clr transforms the data prior to the assessment of statistical significance; however, ALDEx2 differs greatly in how values of 0 are handled. ALDEx2 estimates a large number of possible values for 0 (and any other count for a taxon in a sample), conducts significance tests on all estimated values, and takes the average significance test value as the most representative for that taxon. In essence, ALDEx2 determines which taxa are significantly different between groups after accounting for the random sampling that occurs when the DNA is extracted and loaded onto the sequencing instrument. In either case, both ANCOM and ALDEx2 explicitly acknowledge the multivariate compositional nature of the data and control for false-positive identifications much better than do the usual approaches.

The third objective is to determine if there are taxa in the data set with correlated abundances. As noted above, spurious correlation is a very large problem in microbiome data sets. Therefore, analyses that report correlations using traditional methods, such as Pearson's or Spearman's correlations, Kendall's τ, or Partial correlations are likely to be wrong (Friedman and Alm 2012; Lovell et al. 2015, Kurtz et al. 2015). However, there are a number of approaches that use a compositional data analytic approach to correlation. In a compositional approach, the variance between ratios of 2 taxa should be 0 or nearly so for the 2 taxa to be counted as correlated (Aitchison 1986; Lovell et al. 2015). The difficulty comes when placing this approach into a familiar null hypothesis test framework or when applying a consistent scale to the measure. The simplest approach is to calculate the ϕ statistic for 2 taxa X and Y, which is the var(log(X/Y))/var(log(X)) (Lovell et al. 2015), where log() is meant to imply the clr values of X or Y. This measure has the advantage of being easily calculated and of strictly enforcing the CoDa approach. The SparCC method (Friedman and Alm 2012) uses Bayesian estimates of the value of *X* and *Y* but calculates a mean value of a measure similar to the concordance correlation coefficient. The SPIEC-EASI approach (Kurtz et al. 2015) uses clr-transformed values and infers a graphical model under the assumption of a sparse correlation network. Both of the latter approaches make strong assumptions about the sparsity of the data, and so are less rigorous for estimating correlations in compositional data than is the calculation of ϕ. However, they both offer the advantage of using a full or partial Bayesian approach, which is generally more powerful than point-estimate based approaches.

Application of CoDa to 2 case studies

Having introduced the issue of CoDa, we now present the results of 2 worked examples presented at the Bioinformatics Workshop was held on 16 June 2015 in Regina, Saskatchewan, at the Annual Scientific Meeting of the Canadian Society of Microbiologists. This illustrates how these approaches can be applied to 2 different 16S rRNA gene sequencing data sets from the recent literature. A full description of the methodology, the data sets, and the code used to generate the figures is given in the Supplementary file workshop.Rnw (Gloor 2016). Downloading and running this file in R (R Core Team 2015) or RStudio will generate the associated workshop.pdf (see supplementary data[1]). The .Rnw document contains both the code and annotation for the code, and the .pdf document contains the code and the resulting figures.

The first worked example is a vaginal microbiome data set. This data set is from an experiment that examined the effect of treating women suffering from bacterial vaginosis (BV) with antibiotics and placebo or antibiotics plus a probiotic supplement (Macklaim et al., 2015). For this example, we extracted only the "before" (samples labelled as BXXX) and "after" (AXXX) treatment samples, which were further identified by their Nugent status, a Gram stain scoring system that acts as a rough indicator of whether the subject had BV or was healthy (normal, n), or whose status was indeterminate (labelled as "i" for intermediate). In addition, individual taxa were aggregated to genus level using QIIME (Kuczynski et al. 2012), except for *Lactobacillus iners* and *Lactobacillus crispatus*, which remained as separate species in the tables. This relatively simple data set will be used to introduce and explain the CoDa analysis methods.

The compositional biplot is the essential initial tool for exploratory CoDa and replaces ordinations based on Unifrac or Bray–Curtis metrics. Compositional biplots are principal component plots of the singular value decomposition of the data. This approach displays the major axes of variance (or change) in a data set (Aitchison and Greenacre 2002). Properly made and interpreted, these plots summarize all the essential results of an experiment. However, it should be remembered that they are descriptive and exploratory, not quantitative. Quantitative tools can be applied later to support the conclusions derived from the biplot.

For simplicity, we filtered the data set to include only those taxa that were at least 0.1% abundant in any sample. One of the desirable properties of CoDa is that subsets of the data set are expected to give essentially the same answer as the entire data set *for the taxa in common* between the whole and the subset data set (Aitchison 1986).

Figure 2 shows the compositional biplot for this data set along with the associated scree plot that displays the percentage of variance explained by each sample or component. The sample names (labelled in red for BV, blue for Normal, or purple for Intermediate) illustrate the variance of the samples, and the taxa values (represented by the black rays) illustrate the variance between the taxa. In fact, the length of the arrow for each taxon is proportional to the standard deviation of the ratio of each taxon to all other taxa. There are many interpretation rules for biplots of compositional data (Aitchison and Greenacre 2002), but these rules are dependent on remembering that only the *ratios* between taxa can be examined. Thus, the links between the tips of the rays or between samples contain the most information. Keeping this in mind, we can see the following:

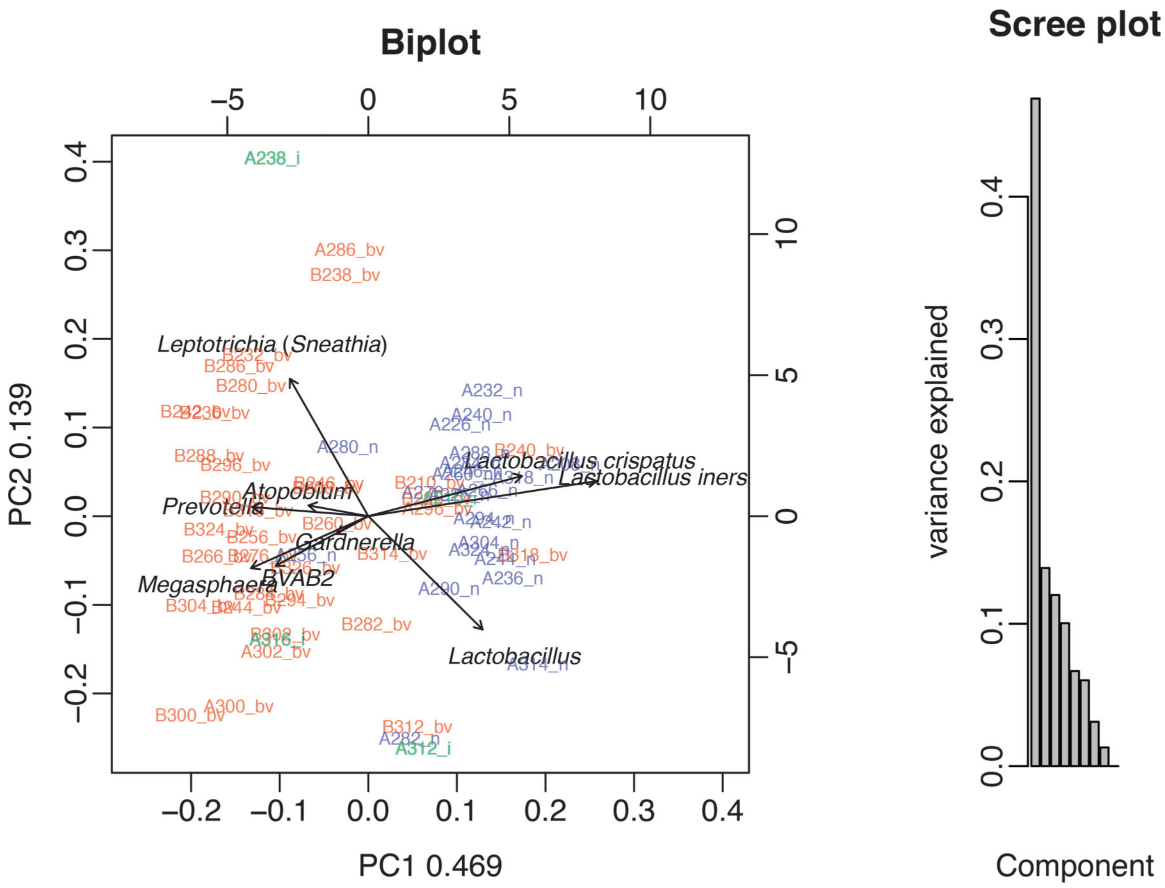
First, the proportion of variance explained in the first component is very good, being 47%, then falling to 13% on component 2, and decreasing rapidly thereafter. This indicates that the major difference between samples can be captured in essentially one direction along component 1. While the amount of variance explained on the first component is relatively large in this data set, a rule of thumb is that PCA plots that display less than 80% of the variance on the first 2 components are not necessarily accurate projections of the data. Thus, some of the quantitative results are expected to be somewhat different than is displayed in the qualitative PCA projection.

Second, the longest link from the center to a taxon is the one to *L. iners*. This indicates that the ratio of this taxon to all others is the most variable across all samples. Likewise, the shortest link is to *Gardnerella*, implying that the ratio of this taxon to all others is the least variable.

Third, the longest link is between *L. iners* and *Leptotrichia* (*Sneathia*). This means we can infer that these 2 taxa likely have the strongest reciprocal ratio relationship. That is, when one becomes more abundant relative to everything else, the other becomes less abundant relative to everything else.

[1] Supplementary data are available with the article through the journal Web site at http://nrcresearchpress.com/doi/suppl/10.1139/cjm-2015-0821.

Fig. 2. The left figure shows a covariance biplot of the abundance-filtered data set, the right figure shows a scree plot of the same data. This exploratory analysis is encouraging, but not definitive, because the amount of variance explained is substantial with 0.469 of the variance being explained by component 1, and 0.139 being explained by component 2. The numbers on the left and bottom indicated unit-scaled variance of the taxa, the numbers on the top and right indicate unit scaled variances of the samples. Samples are colored in red if diagnosed as bacterial vaginosis, blue if healthy, and green if intermediate. The scree plot also shows that the majority of the variability is on component 1. We can interpret this biplot with some confidence, although it is likely that any associations will be found to have large variation.



Fourth, the shortest link observed in the plot is between *Megasphaera* and BVAB2. From this we conclude that the ratio of these 2 taxa is relatively constant across all samples. That is, their ratio abundance is highly correlated. These 2 taxa should be seen to have a low value of ϕ, but we must keep in mind the limit of the projection of the data.

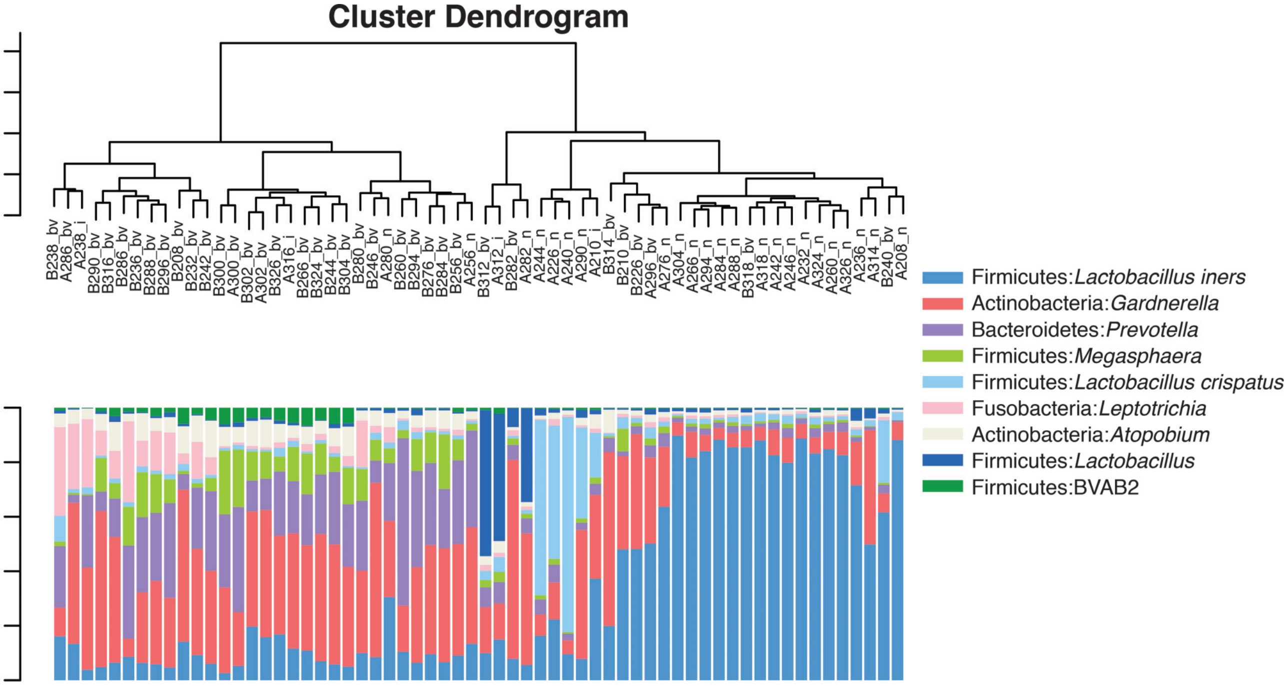
Fifth, the link between *Prevotella* and *L. crispatus* passes directly through *Atopobium*. This indicates that these 3 taxa are linearly related. In this case, it is clear when *L. crispatus* increases, the other 2 will decrease. Likewise, this property can be extended to any linear relationships containing 3 or more links.

Sixth, the link between *L. iners* and *Megasphaera*, and the link between *Leptotrichia* (*Sneathia*) and *Lactobacillus* cross at approximately 90°. The cosine of the angle approximates the correlation between the connected log ratios. Thus, we can conclude that the abundance relationship between the former pair of taxa is poorly correlated with that of the latter 2 taxa. In other words, these 2 pairs vary independently in the data set.

Some samples (A312_bv, B312_i, A282_n at the bottom) are tightly grouped, indicating that they contain similar sets of taxa at similar ratio abundances. We can see from the biplot that these samples contain an abundance of *Lactobacillus* and are depleted in *Leptotrichia* (*Sneathia*). Furthermore, we can see that the samples divide into 2 fairly clear groups, with most of the before or "B" samples on the left, and most of the after or "A" samples on the right. We further observe that the majority of the B samples are colored red, indicating a diagnosis of BV, and the majority of the A samples are colored blue, indicating a diagnosis of non-BV.

The result of the biplot suggested that there were 2 main groups that could be defined with this set of data. With a few exceptions, there appears to be a fairly strong separation between the samples containing a majority of *Lactobacillus* spp. and those lacking them. We can explore this by performing an unsupervised cluster analysis on the log-ratio-transformed data. In traditional microbiome evaluation methodologies, clustering is based on the weighted or unweighted Unifrac distances or on the

Fig. 3. Unsupervised clustering of the reduced data set. The top figure shows a dendrogram of relatedness generated by unsupervised clustering of the Aitchison distances, which is a distance that is robust to perturbations and subcompositions of the data (Aitchison 1986). The bottom figure shows a stacked bar plot of the samples in the same order. The legend indicating the colour scheme for the taxa is on the right side.



Bray–Curtis dissimilarity metric, for example, see the standard workflow in QIIME (Kuczynski et al. 2012). These metrics are much more sensitive to the abundance of community members than is the Aitchison distance used in CoDa (Martín-Fernández et al. 1998). Thus, here we used the Aitchison distance metric that fulfills the criteria required for compositional data. In particular, by using a compositional approach, it is appropriate to examine a defined subcomposition of the data (i.e., a subset of the taxa).

The results of unsupervised clustering of the data set are shown in Fig. 3. Again, it is important to remember that all distances are calculated from the ratios between taxa and not on the taxa abundances themselves. For this figure, we used the ward.D2 method, which clusters groups together by their squared distance from the geometric mean distance of the group. There are many other options, and the user should choose one that best represents the data, although ward.D and ward.D2 are usually the most appropriate (Martín-Fernández et al. 1998).

The cluster analysis supports the results of the biplot and shows the split between 2 types of samples rather clearly. Samples containing an abundance of *Lactobacillus* spp. are grouped together on the right, and samples with an abundance of other taxa are grouped together on the left. The cluster analysis helps explain and clarify the compositional biplot. For example, the 4 samples in the middle lower part of the biplot in Fig. 2, labelled A/B312 and A/B282, group together in both the biplot and the cluster plot. These samples are atypical for both the N and BV groups, containing substantially more of the *Lactobacillus* taxon, and somewhat more of the taxa normally found in BV than in the other N samples. Based on these 2 results it would be appropriate to exclude these 4 samples from further analysis because of their atypical makeup.

Next, a univariate comparison between the B and A groups was performed. For simplicity of coding, we kept the outlier samples, but the reader is encouraged to remove them and see how the results change. For this, we used the ALDEx2 tool (Fernandes et al. 2013, 2014) that incorporates a Bayesian estimate of taxon abundance into a compositional framework, with the results shown in Table 1 and the effect plot (Gloor et al. 2016*a*) shown in Fig. 4. Of note, ALDEx2 examines differential abundance by estimating the measurement error inherent in high-throughput DNA sequencing experiments, including the measurement error associated with 0 count taxa, and uses the assumptions of CoDa to normalize the data for the differing number of reads in each sample (Fernandes et al. 2013; Lovell et al. 2015).

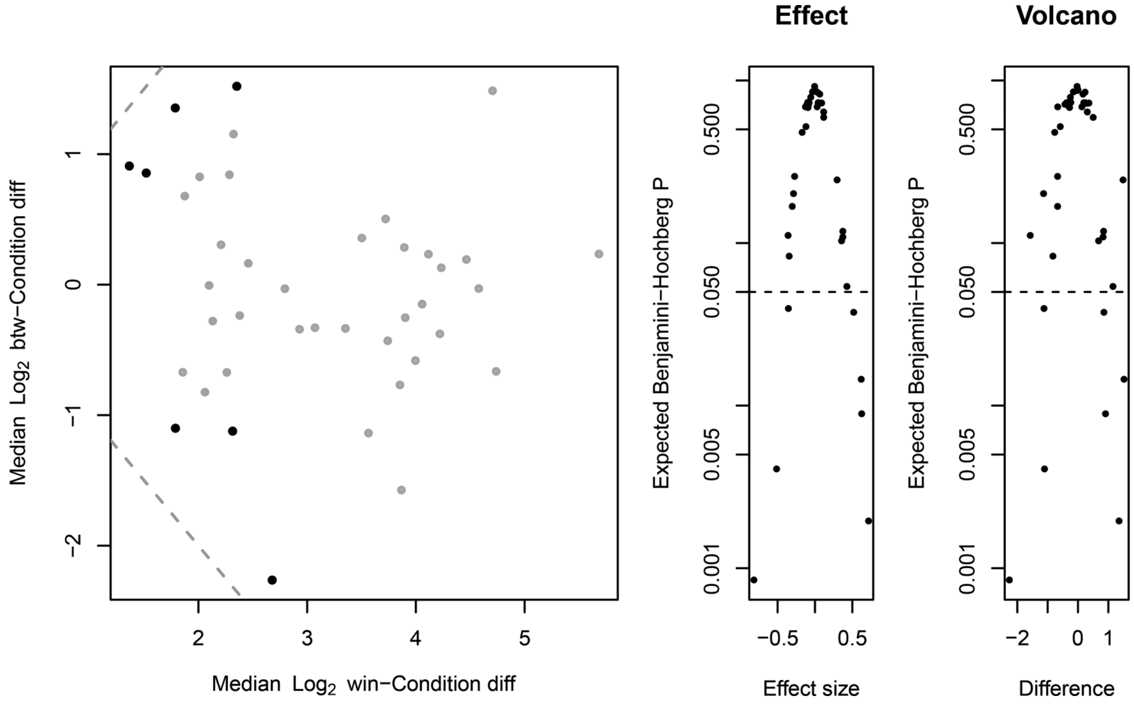
When interpreting these results, it is important to remember that we are actually examining ratios between values, rather than abundances. Thus, we are examining the change in abundance of a taxon *relative to all others* in the data set. The user should also remember that all values reported are the means or medians over the number of Dirichlet instances, as given by the mc.samples variable in the aldex.clr function and explained more

Table 1. List of significantly different taxa.

Taxon	diff.btw	diff.win	effect	overlap	wi.ep	wi.eBH
Atopobium	0.86	1.51	0.53	0.30	0.007	0.037
Prevotella	1.41	1.77	0.75	0.22	0.000	0.002
Lactobacillus crispatus	–1.07	1.78	–0.49	0.23	0.000	0.004
Lactobacillus iners	–2.25	2.68	–0.79	0.20	0.000	0.001
Streptococcus	–1.14	2.38	–0.37	0.30	0.008	0.041
Dialister	0.89	1.38	0.59	0.25	0.001	0.009
Megasphaera	1.56	2.31	0.63	0.28	0.002	0.015

Note: diff.btw, median difference between groups on a log base 2 scale; diff.win, largest median variation within group H (healthy) or BV (bacterial vaginosis); effect, effect size of the difference, median of diff.btw/diff.win; overlap, confusion in assigning an observation to H or BV group, smaller is better; wi.ep, expected value of the Wilcoxon Rank Test *P* value; wi.eBH, expected value of the Benjamini–Hochberg corrected *P* value.

Fig. 4. An effect plot showing the univariate differences between groups (Gloor et al. 2016*a*). The left plot shows a plot of the maximum variance within the B or A group vs the difference between groups. The black points indicate those that have a mean Benjamini–Hochberg adjusted *P* value of 0.05 or less using *P* values calculated with the Wilcoxon rank test. The middle plot shows a plot of the effect size vs the adjusted *P* value. In general, effect size measures are more robust than are *P* values and are preferred. For a large sample size such as this one, an effect size of 0.5 or greater will likely correspond to biological relevance. The right plot shows a volcano plot where the difference between groups is plotted vs the adjusted *P* value.



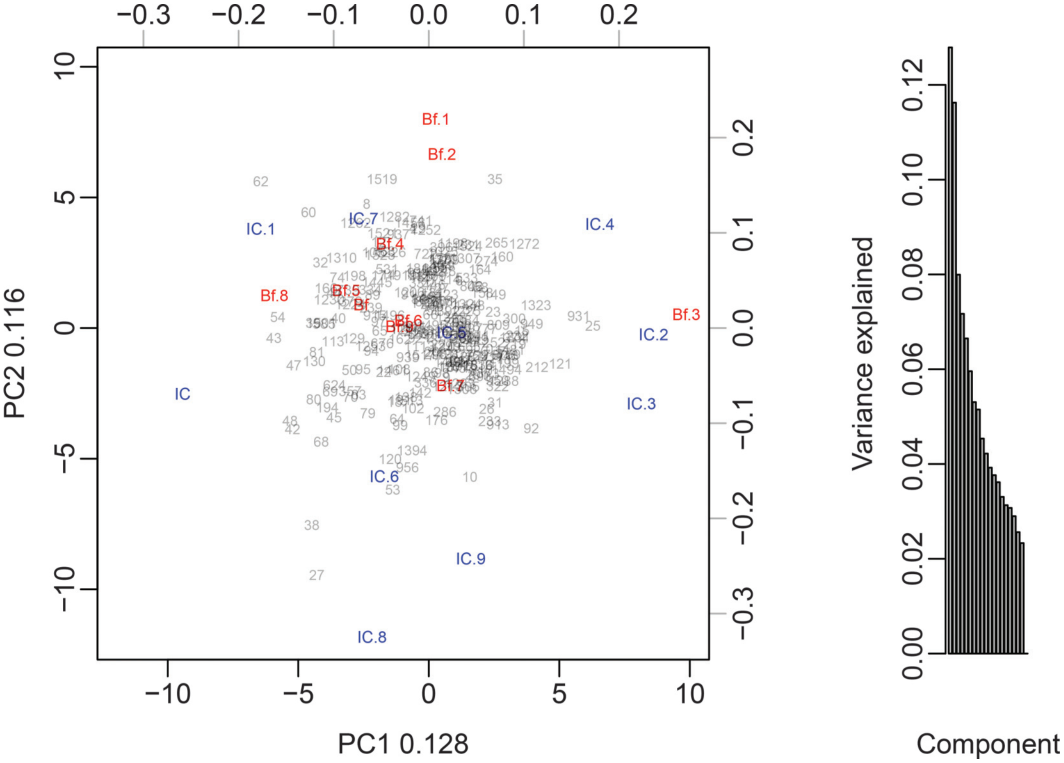
fully in the supplementary material[1] and the original papers (Fernandes et al. 2013, 2014).

In the examples given in Table 1, we filtered to show only those taxa for which the expected Benjamini and Hochberg (1995) adjusted *P* value was less than 0.05, meaning that the expected likelihood of a false-positive identification per taxon is less than 5%, with the actual value per taxon given in the wi.eBH column. Using *L. iners*, we note that the absolute difference between groups can be up to –2.25. Thus, the absolute fold change in the ratio between *L. iners* and all other taxa between groups for this organism is on average 4.76-fold ($1/2^{-2.25}$) — being more abundant in the A samples than in the B samples. However, the difference within the groups (roughly equivalent to the standard deviation) is even larger, giving an effect size of –0.79. Thus, the difference between groups is less than the variability within a group, a result that is typical for microbiome studies.

These quantitative results are largely congruent with the biplot, which showed that the taxa represented here were the ones that best explained the variation between groups, and that the *Leptotrichia* (*Sneathia*) and *Lactobacillus* taxa were not contributing to the separation of the 2 large groups and so would not be expected to be significantly different, despite being highly variable.

The left panel of Fig. 4 shows a plot of the within (diff.win) to between (diff.btw) condition differences, with the large black dots representing those that have a

Fig. 5. A form biplot of the Hsiao et al. (2013) data set that best represents the distances between samples. Here we can see that the control and experimental samples are intermingled, suggesting no separation between the groups. Furthermore, in the scree plot the proportion of variance explained in the first component is not large when compared with the other components. The evidence of structure within this data set is thus weak.



Benjamini–Hochberg adjusted *P* value of 0.05 or less. Taxa that are more abundant than the mean in the B samples have positive *y* values, and those that are more abundant than the mean in the A samples have negative *y* values. These are referred to as "effect size" plots, and they summarize the data in an intuitive way (Gloor et al. 2016*a*). The grey lines represent the line of equivalence for the within and between group values. Small black dots represent taxa that are less abundant than the mean taxon abundance: here it is clear that the abundance of rare taxa, are generally difficult to estimate with any precision.

The middle plot in Fig. 4 shows a plot of the effect size vs. the Benjamini–Hochberg adjusted *P* value, with a strong correspondence between these 2 measures. In general, an effect size cutoff is preferred because it is more robust than *P* values. The right plot in this figure shows a volcano plot for reference.

Finally, we can determine which taxa are most correlated or compositionally associated. As noted above, correlation is especially problematic, and the only way to avoid false-positive associations is to identify those taxa that have constant or nearly constant ratios in all samples: this is the underlying basis of the ϕ measure (Lovell et al. 2015). In the example shown in the supplementary material, we calculate the mean ϕ using the same philosophy as outlined above for univariate statistical tests.

In the context of microbiome data sets, the ϕ metric (Lovell et al. 2015) seeks to identify those pairs of taxa that have a near constant ratio abundance across all samples. Applying this approach to the data set shows that the 2 most compositionally associated taxa are *Prevotella* sp. and *Megasphaera* sp. Note, these taxa do not have the shortest links in the compositional biplot, indicating that the amount of variance explained is not high enough to provide an accurate projection of the data set.

For the second worked example, we include in the workshop.Rnw document a second example based on the data of Hsiao et al. (2013) who examined the effect of *Bacteroides fragilus* supplementation on the microbiome composition of a mouse model of autism. That paper determined that there was a strong functional association between *B. fragilus* supplementation and mouse behavior. One of the major conclusions was that this functional change in behavior was associated with changes in abundance of a number of bacteria that composed the mouse gut microbiome. We will focus our analysis only on the conclusions derived from the analysis of the microbiome data that were presented in figure 4 of Hsiao et al. (2013).

Figure 5 shows a compositional biplot of this data set, and it is obvious that there is little evidence of difference between the poly-IC treated control (IC) and poly-IC treated mice supplemented with *B. fragilus* groups when analyzed using this approach. This is in accordance with their conclusions when analyzing the data using an unweighted Unifrac distance-based approach. Interestingly, the compositional biplot shows that the *B. fragilus* samples are generally closer to the origin of the plot than are the IC samples, suggesting that the *B. fragilus* samples have lower dispersion than the IC samples.

Since the authors concluded that there was no evidence for multivariate differences between groups, and the CoDa approach agrees, it is generally not advised to conduct a univariate analysis, since it is likely that only false-positive results would be obtained (Hubert and Wainer 2012).

However, these authors went on to identify a number of univariate differences in taxon abundance between groups using the LEfSe and Metastats tools that are standard in the field (White et al. 2009; Segata et al. 2011), but that do not assume the data are multivariate compositions. When examining univariate differences with the ALDEx2 tool, we found that none of the univariate differences reported in the original paper were supported by subsequent analysis. In particular, the authors indicated that the largest differences between groups were found for 6 taxa labelled as 53, 145, 638, 836, 837, and 956 in figure 4 of Hsiao et al. (2013). The reason for this discrepancy is that inspection of the original paper reveals that raw *P* values were reported, not Benjamini–Hochberg adjusted *P* values. Thus, it is likely that the majority, if not all, of the taxa different between the control and treatment groups are false-positive identifications. This result is congruent with the multivariate results found in both the original paper and by the compositional biplot. Finally, in support of this assertion, we observe that all of these predicted differences become insignificant following a multiple test correction using either the *P* values reported in the paper or *P* values calculated using the ALDEx2 software.

While we have been critical of the microbiome analysis methods used in this paper, we must acknowledge that other published papers exhibit many of the same flaws, namely, an over-reliance on tools that do not treat the data as compositions, the identification of extremely rare taxa as the most “significantly different” taxa between groups, and a general lack of corrections for multiple hypothesis testing.

Summary

Because the total number of reads is uninformative in high-throughput DNA sequencing data sets, the only information available is the ratio of abundances between components, thus, these data are compositional. Using two 16S rRNA gene sequencing data sets, we have illustrated that microbiome data can be examined using a multivariate CoDa approach, where the data are ratios between the operational taxonomic unit count in a sample and the geometric mean for that sample. Dirichlet Monte-Carlo replicates coupled with the clr transformation can ameliorate the sparse data problem inherent in microbiome data sets.

In essence, we argue here that 16S rRNA gene sequencing data sets are not special and do not need their own unique statistical analysis approaches. The data generated can be examined by a general multivariate approach after accounting for the compositional nature of the data, and such an analysis is comparable or superior to domain-specific approaches, such as those used in the second example paper (Hsiao et al. 2013).

With the human body associated with a large number and diversity of bacteria, we need to understand the evolution of this association and how and when this intimate association develops. Such understanding will in turn lead us to robust approaches focussed on when and how to influence the microbiome by probiotic supplementation or by nutrient or antimicrobial means. More and more studies are exploring how the microbiome can predict outcomes, including following fecal transplant, probiotic, dietary, and drug treatment (David et al. 2014; Kwak et al. 2014; Rajca et al. 2014; Seekatz et al. 2014). Such work will require carefully designed studies with high quality clinical documentation, and samples that are processed using some of the methods described herein. As the compositional toolkit for microbiome analysis evolves, these studies will reveal aspects of human life not previously envisaged. To have confidence in such findings, data sets must be interrogated with rigour. The public is thirsty for knowledge and the media anxious to attract attention. Our sole reliance on pharmaceutical agents is no longer acceptable, and the ability to manipulate the microbiome is not only appealing but also actually feasible. Thus, studies that help to understand how such manipulations occur and what communication is taking place between microbes and the host will allow for more precisely targeted interventions, even to some extent personalized, particularly for the latter, since precise knowledge of microbiome components and activity will be critical.

Interested readers wishing to progress beyond this demonstration should consult the compositional data literature but in particular the original book by Aitchison (1986) and a comprehensive book by Pawlowsky-Glahn et al. (2015), which outlines the essential geometric problem of compositional data as it is understood at present. For a guide that goes beyond the introduction given here and in the supplementary material, a book outlining how to use the compositions R package by Van den Boogaart and Tolosana-Delgado (2013) is particularly helpful, although none of the examples are drawn from the biological literature. For others wishing to understand bioinformatics and data analysis of sequencing data in general terms, hopefully this paper will prove helpful and encourage people to enroll in specialized courses. The temptation may be to rely on proprietary third party systems, even at a cost, but the "devil is in the details", and for thoroughness, we recommend developing the highest level of skill possible, especially to continue to create new analytical tools.

We hope that this report will help researchers to better understand their data and thereby conduct analyses that are more likely to be robust and, more importantly, bring badly needed breakthroughs in prevention, treatment, and cure of disease.

Funding

Financial support for this study was provided by a Canadian Institutes of Health Research (CIHR) Emerging Team Grant and a Genome British Columbia (GBC) Grant awarded to GR and Debra Money as co-PI, and GG was a co-investigator (grant No. 108030). Work in the lab of GG is also supported by an NSERC Discovery Grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aitchison, J. 1986. The statistical analysis of compositional data. Chapman and Hall, London, UK.

Aitchison, J., and Greenacre, M. 2002. Biplots of compositional data. J. R. Stat. Soc. Ser. C Appl. Stat. **51**: 375–92. doi:10.1111/1467-9876.00275.

Benjamini, Y., and Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol. **57**: 289–300. Available from http://www.jstor.org/stable/2346101.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., et al. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature, **505**(7484): 559–563. doi:10.1038/nature12820. PMID:24336217.

Di, Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., and Reid, G. 2013. High throughput sequencing methods and analysis for microbiome research. J. Microbiol. Methods, **95**(3): 401–414. doi:10.1016/j.mimet.2013.08.011. PMID:24029734.

Fernandes, A.D., Macklaim, J.M., Linn, T.G., Reid, G., and Gloor, G.B. 2013. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One, **8**(7): e67019. doi:10.1371/journal.pone.0067019. PMID:23843979.

Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., and Gloor, G.B. 2014. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome, **2**: 15. doi:10.1186/2049-2618-2-15. PMID:24910773.

Filzmoser, P., Hron, K., and Reimann, C. 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci. Total Environ. **407**: 6100–6108. doi:10.1016/j.scitotenv.2009.08.008. PMID:19740525.

Frémont, M., Coomans, D., Massart, S., and De, Meirleir, K. 2013. High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. Anaerobe, **22**: 50–56. doi:10.1016/j.anaerobe.2013.06.002. PMID:23791918.

Friedman, J., and Alm, E.J. 2012. Inferring correlation networks from genomic survey data. PLoS Comput. Biol. **8**(9): e1002687. doi:10.1371/journal.pcbi.1002687. PMID:23028285.

Gentleman, R.C., Carey, V.J., Bates, D.M., Bolstad, B., Dettling, M., Dudoit, S., et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. **5**(10): R80. doi:10.1186/gb-2004-5-10-r80. PMID:15461798.

Gloor, G.B. 2016. Compositional data analysis for high throughput sequencing: an example from 16S rRNA gene sequencing. Available from https://github.com/ggloor/CJM_Supplement. doi:10.5281/zenodo.49579.

Gloor, G.B., Macklaim, J.M., and Fernandes, A.F. 2016*a*. Displaying variation in large datasets: a visual summary of effect sizes. J. Comput. Graph. Stat. In press. doi:10.1080/10618600.2015.1131161.

Gloor, G.B., Macklaim, J.M., Vu, M., and Fernandes, A.F. 2016*b*. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. Austrian Journal of Statistics. In press.

Hsiao, E.Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., et al. 2013. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell, **155**(7): 1451–1463. doi:10.1016/j.cell.2013.11.024. PMID:24315484.

Hubert, L., and Wainer, H. 2012. A statistical guide for the ethically perplexed. CRC Press, London, UK.

Kuczynski, J., Stombaugh, J., Walters, W.A., González, A., Caporaso, J.G., and Knight, R. 2012. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Chapter 1, Unit 1E-5. Curr. Protoc. Microbiol. doi:10.1002/9780471729259.mc01e05s27. PMID:23184592.

Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., and Bonneau, R.A. 2015. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput. Biol. **11**: e1004226. doi:10.1371/journal.pcbi.1004226. PMID:25950956.

Kwak, D.S., Jun, D.W., Seo, J.G., Chung, W.S., Park, S.E., Lee, K.N., et al. 2014. Short-term probiotic therapy alleviates small intestinal bacterial overgrowth, but does not improve intestinal permeability in chronic liver disease. Eur. J. Gastroenterol. Hepatol. **26**(12): 1353–1359. doi:10.1097/MEG.0000000000000214. PMID:25244414.

Lourenço, T.G., Heller, D., Silva-Boghossian, C.M., Cotton, S.L., Paster, B.J., and Colombo, A.P. 2014. Microbial signature profiles of periodontally healthy and diseased patients. J. Clin. Periodontol. **41**(11): 1027–1036. doi:10.1111/jcpe.12302. PMID:25139407.

Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., and Bähler, J. 2015. Proportionality: a valid alternative to correlation for relative data. PLoS Comput. Biol. **11**: e1004075. doi:10.1371/journal.pcbi.1004075. PMID:25775355.

Lozopone, C., and Knight, R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. **71**: 8228–8235. doi:10.1128/AEM.71.12.8228-8235.2005. PMID:16332807.

Macklaim, J.M., Clemente, J.C., Knight, R., Gloor, G.B., and Reid, G. 2015. Changes in vaginal microbiota following antimicrobial and probiotic therapy. Microb. Ecol. Health Dis. **26**: 27799. doi:10.3402/mehd.v26.27799. PMID:26282697.

Mandal, S., Van Treuren, W., White, R.A., Eggesbø, M., Knight, R., and Peddada, S.D. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb. Ecol. Health Dis. **26**: 27663. doi:10.3402/mehd.v26.27663. PMID:26028277.

Martín-Fernández, J.A., Barceló-Vidal, C., and Pawlowsky-Glahn, V. 1998. Measures of difference for compositional data and hierarchical clustering methods. *In* Proceedings of the IAMG'98, 1998 Annual Conference of the International Association for Mathematical Geology. *Edited by* A. Buccianti, G. Nardi, and R. Potenza. pp. 526–531.

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

Pawlowsky-Glahn, V., Egozcue, J.J., and Tolosana-Delgado, R. 2015. Modeling and analysis of compositional data. John Wiley & Sons, Springer, London, UK.

Pearson, K. 1896. 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. **60**: 489–498. doi:10.1098/rspl.1896.0076.

R Core Team. 2015. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from https://www.R-project.org/.

Rajca, S., Grondin, V., Louis, E., Vernier-Massouille, G., Grimaud, J.C., Bouhnik, Y., et al. 2014. Alterations in the intestinal microbiome (dysbiosis) as a predictor of relapse after infliximab withdrawal in Crohn's disease. Inflamm. Bowel Dis. **20**(6): 978–986. doi:10.1097/MIB.0000000000000036. PMID:24788220.

Reardon, S. 2013, Bacterium can reverse autism-like behaviour in mice. Nature. doi:10.1038/nature.2013.14308.

Schellenberg, J., Links, M.G., Hill, J.E., Dumonceaux, T.J., Peters, G.A., Tyler, S., et al. 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. Appl. Environ. Microbiol. **75**: 2889–2898. doi:10.1128/AEM.01640-08. PMID:19270139.

Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., et al. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microbiol. **75**: 7537–7541. doi:10.1128/AEM.01541-09. PMID:19801464.

Seekatz, A.M., Aas, J., Gessert, C.E., Rubin, T.A., Saman, D.M., Bakken, J.S., and Young, V.B. 2014. Recovery of the gut microbiome following fecal microbiota transplantation. MBio, **5**(3): e00893–14. doi:10.1128/mBio.00893-14. PMID:24939885.

Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., and Huttenhower, C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. **12**: R60. doi:10.1186/gb-2011-12-6-r60. PMID:21702898.

Urbaniak, C., Cummins, J., Brackstone, M., Macklaim, J.M., Gloor, G.B., Baban, C.K., et al. 2014. Microbiota of human breast tissue. Appl. Environ. Microbiol. **80**(10): 3007–3014. doi:10.1128/AEM.00242-14. PMID:24610844.

Van den Boogaart, K.G., and Tolosana-Delgado, R. 2013. Analyzing compositional data with R. Springer, Berlin Heidelberg, Germany.

Walker, A.W., Martin, J.C., Scott, P., Parkhill, J., Flint, H.J., and Scott, K.P. 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome, **3**: 26. doi:10.1186/s40168-015-0087-4. PMID:26120470.

Warton, D.I., Wright, S.T., and Wang, Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods Ecol. Evol. **3**: 89–101. doi:10.1111/j.2041-210X.2011.00127.x.

White, J.R., Nagarajan, N., and Pop, M. 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. **5**: e1000352. doi:10.1371/journal.pcbi.1000352. PMID:19360128.