Canadian Science Publishing
NRC Research Press journals | Publishing Services

Canadian Journal of Microbiology
Revue canadienne de de microbiologie

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

Journal:	*Canadian Journal of Microbiology*
Manuscript ID	cjm-2015-0821.R2
Manuscript Type:	Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Gloor, Gregory; The University of Western Ontario, Biochemistry Reid, Gregor; The Lawson Research Institute
Keyword:	microbiome, compositional data, correlation, multivariate statistics, multiple test correction

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

Gregory B. Gloor (1,2)*, Gregor Reid (2, 3)

1. Department of Biochemistry, Western University, London, Ontario, Canada
2. Canadian Center for Human Microbiome and Probiotic Research, Lawson Health Research Institute, London, Ontario, Canada
3. Departments of Microbiology and Immunology, and Surgery, Western University, London, Ontario, Canada

* Address for Correspondence: Gregory B. Gloor, E-mail: ggloor@uwo.ca

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 datasets 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 two microbiome datasets. 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

Introduction

Human microbiome studies have shown a major link between microbial composition and health and disease and dysbiosis (Fremont 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 datasets 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 dataset 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. 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 such as the *cpn60* gene (Schellenburg 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 datasets. Such datasets 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 is a term used to describe a dataset in which the parts in each sample have an arbitrary or non-informative 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 Figure 1 where we set up an artificial example and count the number of molecules in the environment. We allow one part (shown in blue) to increase 10-fold between samples 1 and 2, while the abundance of the other 49 parts (in red) remain 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 blue part still appears to become more abundant, although it is less than a 10-fold change. However, the 49 red 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, principle 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).

Indeed, the original issue with compositional data identified by Pearson (1896) was that of spurious correlation. That is, two 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 three samples (a, b and c) with four taxonomic variables measured to have the following absolute counts in three environmental samples (i.e., samples are in rows, taxa are in columns):

$$\text{abc} = \begin{bmatrix} 470 & 66 & 839 & 751 \\ 541 & 569 & 787 & 512 \\ 167 & 906 & 959 & 504 \end{bmatrix}, \text{cor(abc)} = \begin{bmatrix} & -0.68 & \mathbf{-0.99} & 0.36 \\ -0.77 & & \mathbf{0.59} & -0.93 \\ \mathbf{-0.30} & \mathbf{-0.37} & & -0.25 \\ 0.55 & -0.95 & 0.62 & \end{bmatrix}.$$

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 between 2 and 3 is now -0.37. Thus, the correlation observed in compositional data is not the same as the correlation 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 correlations between taxa 1 and 2 is positive (0.43), and the correlation between 2 and 3 is even more strongly negative at -0.79. Thus, 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 (Filmoser et al.

2009, Friedman and Alm 2012, Fernandes et al 2013, Fernandes et al. 2014, Lovell et al. 2015, Kurtz et al. 2015).

The essential first step of proper compositional data analysis 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 centered log-ratio of *X* can be computed as:

$X_{clr} = [\log[x_1/g_X], \log[x_2/ g_X], \ldots \log[x_D/ g_X]$,

where g_X is the geometric mean of all values in vector *X* (Aichison 1986). This simple transformation renders valid all standard multivariate analysis techniques (Aitchison 1986, van den Boogaart 2013, Pawlowsky-Glahn et al. 2015), and as shown in the Ratios panel of Figure 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 compositional data analysis (CoDa) based assessments of the datasets.

A CoDa approach would be robust if microbiome datasets 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). With this paradigm we dispose of taxa with 0 counts in all or most samples (Palarea-Albaladejo and Martin-Fernandez 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, Fernandes et al. 2014).

Microbiome analysis tools that account for compositional data

Fortunately, the compositional data analysis problem of microbiome datasets 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 three 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 compositional data analysis (Aitchison 1986), or CoDa, offers a number of advantages over these tools, as explained below.

The first objective is to determine if there is structure in the dataset. In the microbiome field this is generally described as beta-diversity analysis. Beta-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 Principle 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 principle component analysis (PCA) of center-log ratio 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 Martin-Fernandez 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

206 driving (or not) the location of particular samples. Thus, the biplot can serve as a
207 summary of the entire dataset, and it is up to the investigator to attach numerical
208 significance to the qualitative results observed. The example usage of compositional
209 biplots is explained in detail below.

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

219 There are two approaches that assess differential abundance in a compositional
220 data analysis framework. The simplest approach is the ANCOM tool (Mandal et al.
221 2015), which assesses statistical significance on log-ratio transformed data. This is
222 more robust than both traditional t-tests and more sophisticated approaches such as
223 zero-inflated Gaussian methods. It should be noted that the software is not deposited
224 into a public repository, and that the 0-replacement value used is fixed in the software.

225 A slightly more complex approach is used by the ALDEx2 package, available
226 from Bioconductor (Fernandes et al 2013, Fernandes et al 2014). Like ANCOM,
227 ALDEx2 centre log-ratio transforms the data prior to the assessment of statistical
228 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 significant 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 dataset with correlated abundances. As noted above, spurious correlation is a very large problem in microbiome datasets. Therefore, analyses that report correlations using traditional methods, such as Pearson's or Spearman's correlations, Kendall's Tau 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 two taxa should be 0 or nearly so for two 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 phi statistic for two 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 compositional data analysis approach. The SparCC method (Friedman and Alm, 2012) uses Bayesian estimates of the value of X and Y but

252 calculates a mean value of a measure similar to the concordance correlation coefficient.
253 The SPIEC-EASI approach (Kurtz et al. 2015) uses clr-transformed values and infers a
254 graphical model under the assumption of a sparse correlation network. Both of the latter
255 approaches make strong assumptions about the sparsity of the data, and so are less
256 rigorous for estimating correlations in compositional data than is the calculation of phi.
257 However, they both offer the advantage of using a full or partial Bayesian approach,
258 which is generally more powerful than point-estimate based approaches.

259 **Application of CoDa to Two Case Studies**

260 Having introduced the issue of compositional data analysis, we now present the
261 results of two worked examples presented at the Bioinformatics Workshop was held on
262 June 16, 2015 in Regina at the Annual Scientific Meeting of the Canadian Society of
263 Microbiologists. This illustrates how these approaches can be applied to two different
264 16S rRNA gene sequencing datasets from the recent literature. A full description of the
265 methodology, the datasets and the code used to generate the figures is given in the
266 Supplementary file workshop.Rnw. Downloading and running this file in R (R Core
267 Team 2015) or RStudio will generate the associated workshop.pdf. The .Rnw document
268 contains both the code and annotation for the code, and the .pdf document contains the
269 code and the resulting figures.

270 The first worked example is a vaginal microbiome dataset. This dataset is from
271 an experiment that examined the effect of treating women suffering from bacterial
272 vaginosis (BV) with antibiotics and placebo or antibiotics plus a probiotic supplement
273 (Macklaim et.al, 2015). For this example, we extracted only the ‘before’ (samples
274 labeled 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 (labeled 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 dataset will be used to introduce and explain the CoDa analysis methods.

The compositional biplot is the essential initial tool for exploratory compositional data analysis and replaces ordinations based on Unifrac or Bray-Curtis metrics. Compositional biplots are principle component plots of the singular value decomposition of the data. This approach displays the major axes of variance (or change) in a dataset (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 dataset to include only those taxa that were at least 0.1% abundant in any sample. One of the desirable properties of compositional data analysis is that subsets of the dataset are expected to give essentially the same answer as the entire dataset *for the taxa in common* between the whole and the subset dataset (Aitchison 1986).

Figure 2 shows the compositional biplot for this dataset along with the associated scree plot that displays the percentage of variance explained by each sample or component. The sample names (labeled 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 dataset, a rule of thumb is that PCA plots that display less than 80% of the variance on the first two 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 *Lactobacillus 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 two 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.

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

Fifth, the link between *Prevotella* and *Lactobacillus crispatus* passes directly through *Atopobium*. This indicates that these three taxa are linearly related. In this case, it is clear when *L. crispatus* increases, the other two will decrease. Likewise, this property can be extended to any linear relationships containing three 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 two taxa. In other words, these two pairs vary independently in the dataset.

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 be depleted in *Leptotrichia* (*Sneathia*). Furthermore, we can see that the samples divide into two 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 two 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* sp., 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 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 compositional data analysis (Martin Fernandez 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 sub-composition of the data (i.e., a subset of the taxa).

The results of unsupervised clustering of the dataset are shown in Figure 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 (Martin-Fernandez 1998).

The cluster analysis supports the results of the biplot and shows the split between two types of samples rather clearly. Samples containing an abundance of *Lactobacillus* sp. 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 four samples in the middle lower part of the biplot in Figure 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 two results it would be appropriate to exclude these four 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) shown in Figure 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 compositional data analysis 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

the change in abundance of a taxon *relative to all others* in the dataset. 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 fully in the supplementary material and the original papers (Fernandes et al. 2013, 2014).

In the examples given in Table 1, we filtered to show only those taxa where the expected Bejamini-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 two large groups.

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
L. crispatus	-1.07	1.78	-0.49	0.23	0.000	0.004
L. 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

diff.btw: median difference between groups on a log base 2 scale

diff.win: largest median variation within group H or BV

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

The left panel of Figure 4 shows a plot of the within (diff.win) to between (diff.btw) condition differences, with the red dots representing those that have a BH 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. 2015). The grey lines represent the line of equivalence for the within and between group values. Black dots are 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 Figure 4 shows a plot of the effect size vs. the BH adjusted P value, with a strong correspondence between these two 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 phi measure (Lovell et al. 2015). In the example shown in the supplementary material, we calculate the mean phi using the same philosophy as outlined above for univariate statistical tests.

In the context of microbiome datasets, the phi 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 dataset shows that the two most compositionally associated taxa are *Prevotella* sp. and *Megasphaera* sp. Note, that 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 dataset.

For the second worked example we include in the workshop.Rnw document a second example based on the data of Hsiao et al. (2013) that examined the effect of *Bacteriodes fragilus* supplementation on the microbiome composition of a mouse model of autism. This 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 and 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 the paper.

Figure 5 shows a compositional biplot of this dataset, and it is obvious that there is little evidence of difference between the poly-IC treated control (IC) and poly-IC treated mice supplemented witt *B. fragilus* (Bf) 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 Bf samples are generally closer to the origin of the plot than are the IC samples, suggesting that the Bf 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 agree, 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. 2012), but that do not assume the data are 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 six taxa labeled as 53, 145, 638, 836, 837, and 956 in Figure 4 of the paper. The reason for this discrepancy is that inspection of the original paper reveals that raw, and not Benjamini-Hochberg adjusted P values were reported. 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 datasets, the only information available is the ratio of abundances between components: thus these data are compositional. Using two 16S rRNA gene sequencing datasets, we have illustrated that microbiome data can be examined using a multivariate CoDa approach that the data as ratios where the denominator is the geometric mean for a sample. Dirichlet Monte-Carlo replicates coupled with the centered log-ratio transformation can ameliorate the sparse data problem inherent in microbiome datasets.

In essence, we argue here that 16S rRNA gene sequencing datasets 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; Seekatz et al. 2014; Rajca 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. In order to have confidence in such findings, datasets must be interrogated with rigour. The public is thirsty for knowledge and the media anxious to attract attention. Reliance on pharmaceutical agents is longer acceptable, and the ability to manipulate the microbiome is not only appealing but actually feasible. Thus, studies that help to understand how such manipulations occur, what communication is taking place between microbes and the host, will allow for more precisely targeted interventions, even to some extent personalized. In particular for the latter, as 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) that outlines the essential geometric problem of compositional data as it is understood at present. For a guide that

goes beyond the introduction given 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 to bring badly needed breakthroughs in prevention, treatment and cure of disease.

Figure Legends

Figure 1: The difference between counting, proportions and ratios. The 'Counts' panel shows a scatter plot of a simulated dataset with two 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 the 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.

Figure 2: The left figure shows a covariance biplot of the abundance-filtered dataset, 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 right indicated unit-scaled variance of the taxa, the numbers on the top and right indicate unit scaled variances of the samples. 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.

Figure 3: Unsupervised clustering of the reduced dataset. 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 sub-compositions 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.

Figure 4: An effect plot showing the univariate differences between groups (Gloor et al. 2015). The left plot shows a plot of the maximum variance within the B or A group vs. the difference between groups. Red 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.

Figure 5: A form biplot of the Hsiao et al. (2013) dataset 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, the proportion of variance explained in the first component is not large when compared to the other components. The evidence of structure within this dataset is thus weak.

Funding: Financial support for this study was provided by a joint Canadian Institutes of Health Research (CIHR) Emerging Team Grant and a Genome British Columbia (GBC) grant awarded on which GR was a co-PI and GG and ML were co-investigators (grant reference #108030). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Aitchison, J. 1986. The statistical analysis of compositional data, Chapman and Hall, London England. ISBN 1-930665-78-4
2. Aitchison, J and Greenacre, M. 2002. Biplots of compositional data. J. Royal Stat. Soc: Series C. 51:375-92

3. Benjamini, Y., Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. . Royal Stat. Soc: Series B (Methodological), 289-300.
4. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., Biddinger, S.B., Dutton, R.J., Turnbaugh, P.J. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 505(7484):559-63.
5. Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., Reid, G. 2013. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013 Dec;95(3):401-14.
6. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., Gloor, G. B. 2013. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PloS One, 8(7), e67019.
7. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., 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.
8. Filzmoser, P., Hron, K., Reimann, C. 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci tTtal Environ. 407:6100-8.
9. Frémont, M., Coomans, D., Massart, S., De Meirleir, K.. 2013. High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. Anaerobe. 22:50-6.

10. Friedman, J., Alm, E. J. 2012. Inferring correlation networks from genomic survey data. PLoS Comput. Biol. 8(9): e1002687

11. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Gen. Biol. 5 (10): R80.

12. Gloor, G.B., Macklaim, J.M., Fernandes, A.F. 2016. Displaying variation in large datasets: a visual summary of effect sizes. J. Comput. Graph. Stat. (in press)

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

14. Hsiao, E. Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F., Patterson, P.H., Mazmanian,S.K. 2013. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 155(7):1451-63

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

16. Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., Knight, R. 2012. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr. Prot. Microbiol. 1E-5.

17. Kurtz, Zachary D and Müller, Christian L and Miraldi, Emily R and Littman, Dan R and Blaser, Martin J and Bonneau, Richard A 2015. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comp. Bio. 11:e1004226

18. Kwak, D.S., Jun, D.W., Seo, J.G., Chung, W.S., Park, S.E., Lee, K.N., Khalid-Saeed, W., Lee, H.L., Lee, O.Y., Yoon, B.C., Choi, H.S. 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):1353-9.

19. Lourenço, T.G., Heller, D., Silva-Boghossian, C.M., Cotton, S.L., Paster, B.J., Colombo, A.P. 2014. Microbial signature profiles of periodontally healthy and diseased patients. J Clin Periodontol. 41(11):1027-36.

20. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J. Marguerat, S., Bähler, J. 2015. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol 11:e1004075.

21. Lozopone, C., Knight, R. 2005. Unifrac: a new phylogenetic method for comparing microbial communities. Applied Env. Micro. 71:8228-8235.

22. Macklaim, J.M., Clemente, J.C., Knight, R., Gloor, G.B., Reid, G. 2015. Changes in vaginal microbiota following antimicrobial and probiotic therapy. Microb Ecol Health Dis. 26:27799.

23. Mandal, S., Van Treuren, W., White, RA., and Eggesbø, M., Knight, R., Peddada, S. D. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microl. Ecol. Health Dis. 26:27663.

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

25. Palarea-Albaladejo J., Antoni Martín-Fernández, J. 2015. zCompositions --- R package for multivariate imputation of left-censored data under a compositional approach. Chemometrics and Intelligent Laboratory Systems. 143:85-96
26. Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons. Springer. 258 pg, London, UK.
27. 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. Royal Soc. Lond. 60:489-498
28. R Core Team 2015. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
29. Rajca, S., Grondin, V., Louis, E., Vernier-Massouille, G., Grimaud, J.C., Bouhnik, Y., Laharie, D., Dupas, J.L., Pillant, H., Picon, L., Veyrac, M., Flamant, M., Savoye, G., Jian, R., Devos, M., Paintaud, G., Piver, E., Allez, M., Mary, J.Y., Sokol, H., Colombel, J.F., Seksik, P. 2014. Alterations in the intestinal microbiome (dysbiosis) as a predictor of relapse after infliximab withdrawal in Crohn's disease. Inflamm Bowel Dis. 20(6):978-86.
30. Reardon, S. 2013, Bacterium can reverse autism-like behaviour in mice. Nature. doi:10.1038/nature.2013.14308.
31. Schellenberg, J., Links, M. G., Hill, J. E., Dumonceaux, T. J., Peters, G. A., Tyler, S., Ball, T. B., Severini, A., Plummer, F. A. 2009. Pyrosequencing of the

chaperonin-60 universal target as a tool for determining microbial community composition. Appl Environ Microbiol. 75: 2889-98.

32. Schloss, P.D, Westcott, S.L, Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., and Van Horn, D.J., Weber, C.F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities

33. Seekatz, A.M., Aas, J., Gessert, C.E., Rubin, T.A., Saman, D.M., Bakken, J.S., Young, V.B. 2014. Recovery of the gut microbiome following fecal microbiota transplantation. MBio. 5(3):e00893-14.

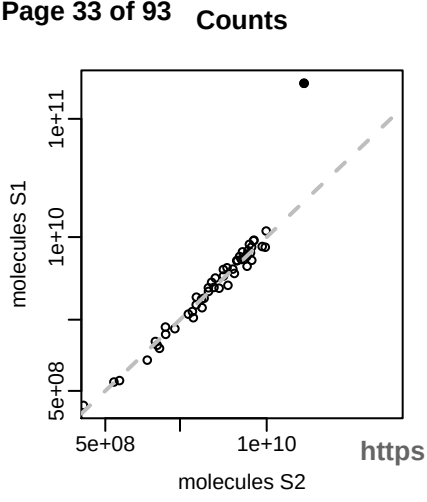
34. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12:R60

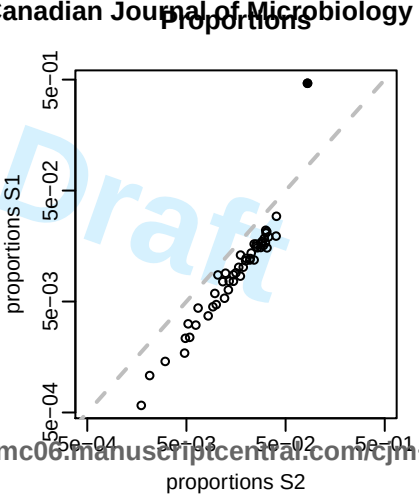
35. Urbaniak, C., Cummins, J., Brackstone, M., Macklaim, J.M., Gloor, G.B., Baban, C.K., Scott, L., O'Hanlon, D.M., Burton, J.P., Francis, K.P., Tangney, M., Reid, G. 2014. Microbiota of human breast tissue. Appl Environ Microbiol. 80(10):3007-14.

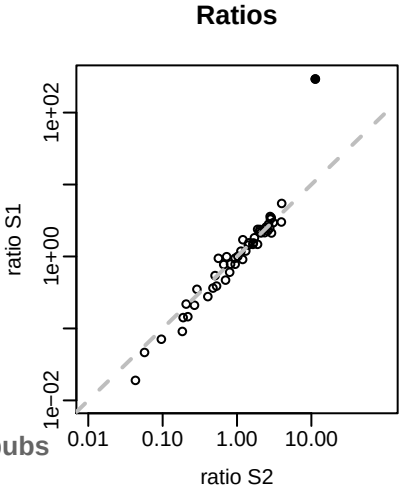
Van den Boogaart, K. G., Tolosana-Delgado, R. 2013. Analyzing compositional data with R. Heidelberg: Springer. Heidelberg 258 pages.

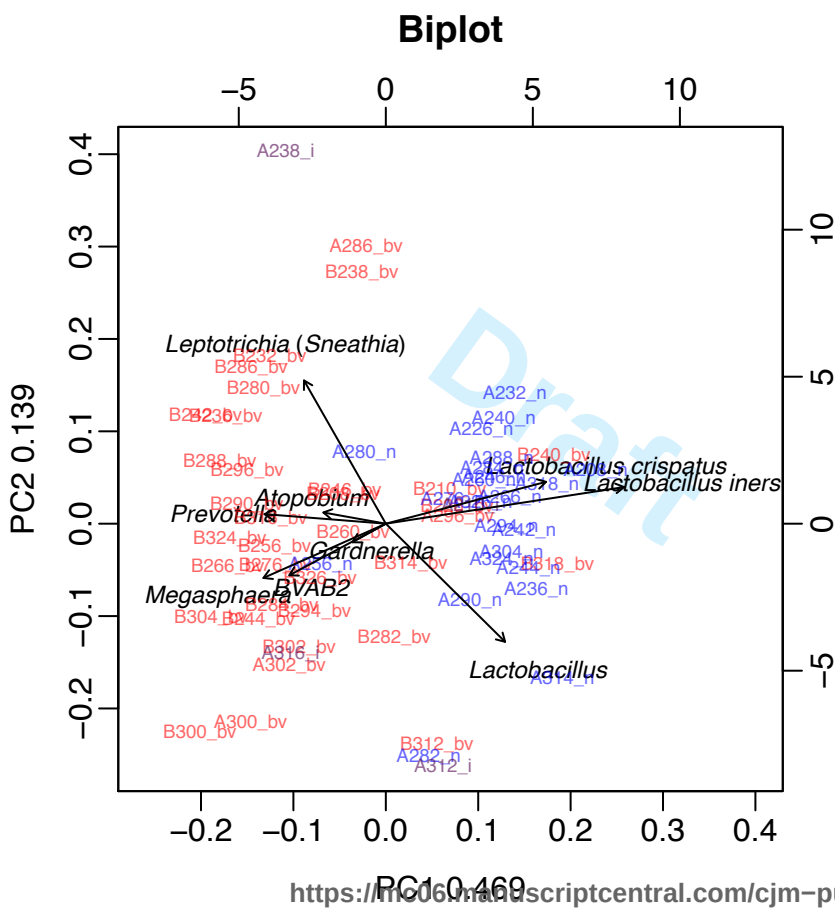
36. Walker, A. W., and Martin, J.C., Scott, P., Parkhill, J., Flint, H. J. Scott, K. P. 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome. 3:26

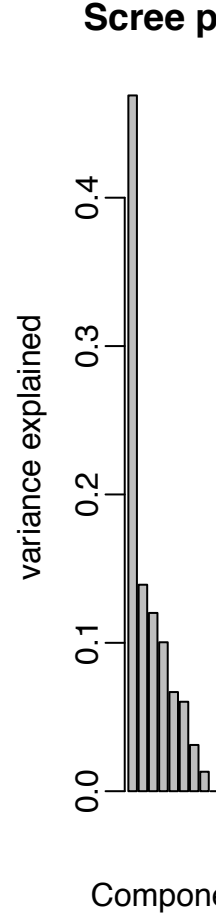
698 37. Warton, D.I., Wrigth, S.T., Wang, Y. 2012. Distance-based multivariate analyses
699 confound location and dispersion effects. Methods Ecol. Evol. 3:89-101.\
700 38. White, J.R., Nagarajan, N., Pop, M. 2009. PLoS Comput. Biol. 5:e1000352
701

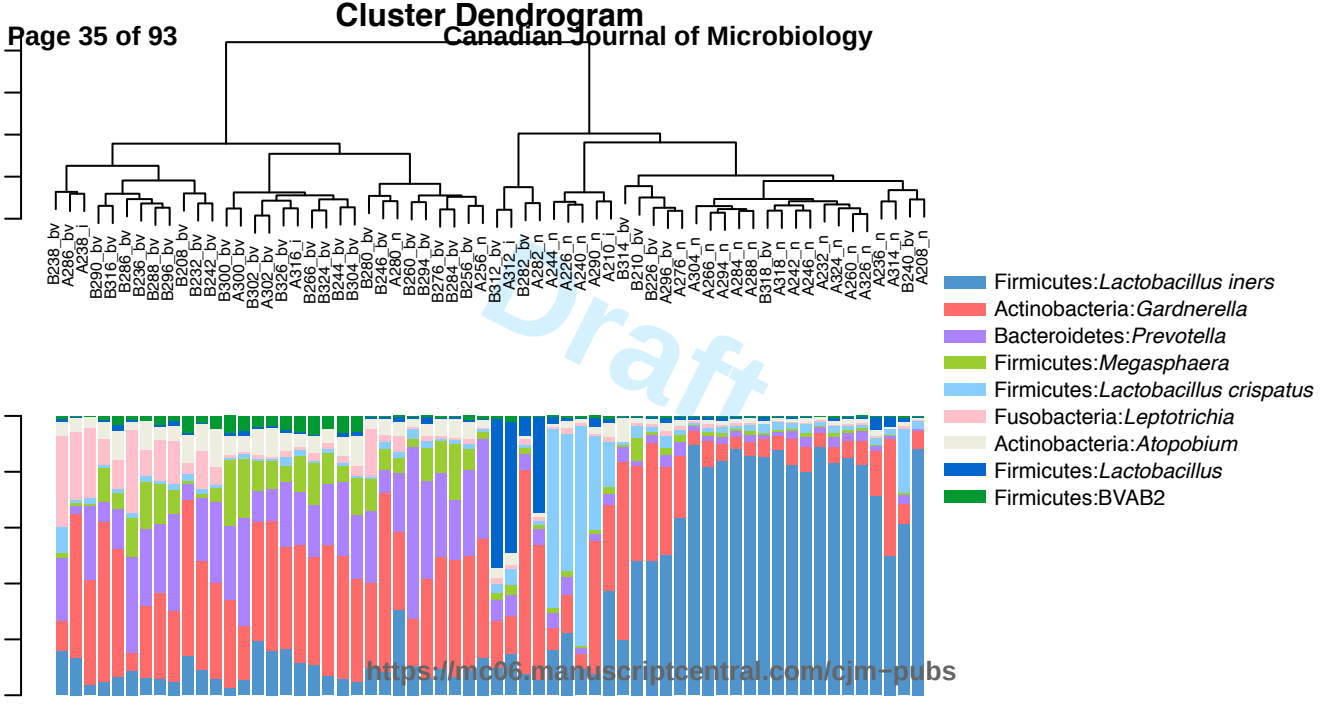


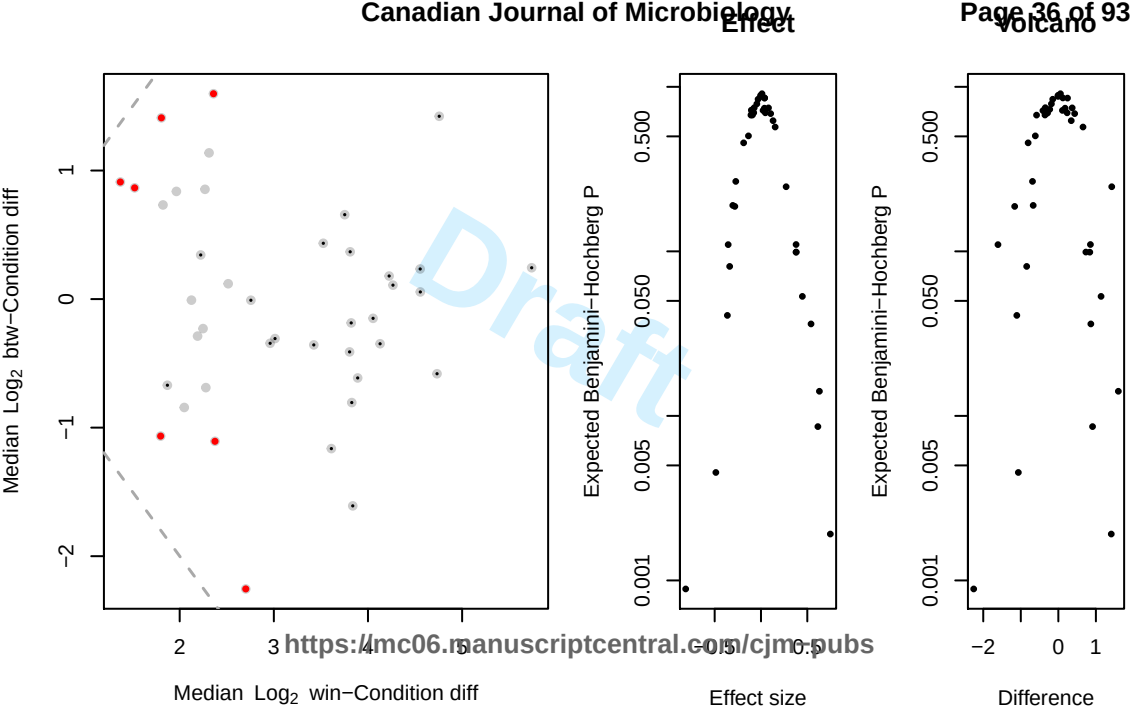


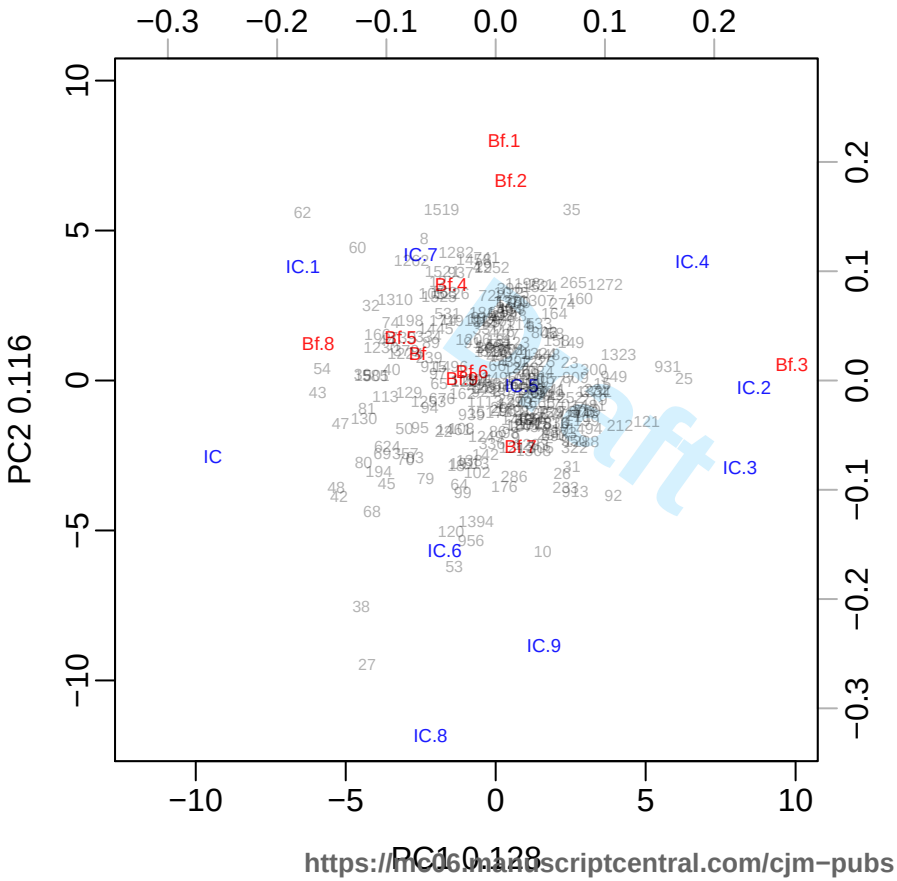


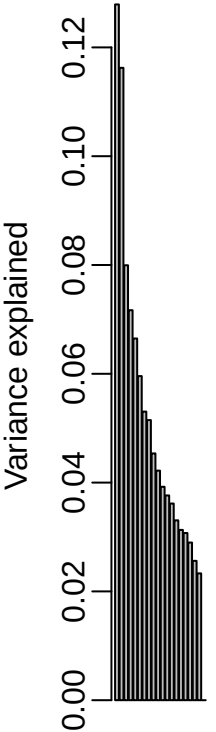

Biplot
PC2 0.139
PC1 0.469
Leptotrichia (Sneathia)
Atopobium
Prevotella
Gardnerella
BVAB2
Megasphaera
Lactobacillus crispatus
Lactobacillus iners
Lactobacillus
A238_i
A286_bv
B238_bv
B286_bv
B280_bv
A232_n
A240_n
A226_n
A280_n
B296_bv
B324_bv
B256_bv
B266_bv
A242_n
A304_n
A236_n
A290_n
B282_bv
A302_bv
A300_bv
B300_bv
B312_bv
A312_i
Draft


Scree plot
variance explained
Component









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

Gregory B. Gloor (1,2)*, Gregor Reid (2, 3) and Matthew Links (4)

1. Department of Biochemistry, Western University, London, Ontario, Canada
2. Canadian Center for Human Microbiome and Probiotic Research, Lawson Health Research Institute, London, Ontario, Canada
3. Departments of Microbiology and Immunology, and Surgery, Western University, London, Ontario, Canada
4. Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, SK, Canada; Agriculture and AgriFood Canada, Saskatoon, SK, Canada

* Address for Correspondence: Gregory B. Gloor, E-mail: ggloor@uwo.ca

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 datasets 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 focusssed on explaining the rationale behind the use of compositional data analysis, and a demonstration of these methods for the examination of two microbiome datasets ~~using published datasests as examples~~. 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 clusters of microbe~~of the microbiome~~s~~ in health and disease, and the need to understand alterations to ~~their~~ its composition and function following intervention with fecal transplant, probiotics, diet and pharmaceutical agents.

Key Words: microbiome, compositional data, correlation, multiple test correction

Introduction

Human microbiome studies have shown a major link between microbial composition and health and disease and dysbiosis (Fremont et al. 2013; Lourenço et al. 2014; Urbaniak et al. 2014). High throughput DNA sequencing methodologies have made this possible, along with breakthroughs in ~~culture~~ culturing techniques. The former has used approaches such as 16S rRNA gene sequencing, metagenomics, transcriptomics and meta-transcriptomics, leading to vast datasets 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 dataset across all samples can easily ~~have~~ exceed many hundreds of millions ~~to billions~~ of reads. Such has been the rapidity of these developments that some studies appear to have been published using methods that are ~~at worst~~potentially ~~flawed and at best not sufficiently well utilized~~. The result can be papers with serious deficiencies that ~~journals and the lay media~~ 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~~; Mazmanian 2015~~).

~~A Bioinformatics Workshop was held on June 16, 2015 in Regina at the Annual Meeting of the Canadian Society of Microbiologists with the goal of demonstrating how~~

to analyse large microbiome datasets, the importance of individual and group exploratory data analysis, and on how to increase the accuracy and dependability of the analyses. Participants were introduced to the concept and importance of multivariate compositional data and why datasets of this type can give misleading results when analyzed by current methods. Examples were used from the literature that outlined the dangers of improper understanding of data analysis in this area. This manuscript provides an expanded rationale for using compositional data analysis methods, a brief review of methods for compositional data analysis of microbiome datasets and a complete description of the approach. A supplementary document is included that contains all the R code and further explanation for two worked examples.

The goals of analysis: Any analysis of a microbiome dataset usually has three major goals:

1. Does the data show any structure? That is, does the data partition into groups?
2. What is the difference between groups? This can be between groups identified beforehand, or by exploratory analysis of the data.
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) of aggregated toolsets, containing approaches adapted from ecology. However, the use of an analysis paradigm based on compositional data analysis (Aitchison 1986), or CoDa, offers a number of advantages over these tools, as explained below.

~~Brief explanation of the data type:~~ Data for microbiome analysis are collected by the following general workflow. The sample (swab, stool, saliva, urine or other type) is collected, ~~and~~ the DNA is isolated and ~~a small amount amplified using the~~ 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 such as the *cpn60* gene (Schellenburg et al. 2009). However, analysis problems are the same regardless of ~~the~~ the amplification target chosen, and Walker et al. (2015) present a good summary of how choices taken upstream of data analysis ~~the effect~~ affect the results~~choices taken upstream of data analysis can 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, t~~T~~his 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~~.~~ ~~This is~~, because the number of reads obtained for a sample is determined by the capacity of the instrument~~: f~~. 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 datasets. Such datasets 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).

~~Why worry about compositional data? Data constrained to relative abundances are subject to the constant arbitrary sum constraint that generates significant unanticipated problems for analysis and is the main reason to use a compositional data approach (Aitchison 1986).~~ Compositional data is a term used to describe a dataset in which the parts in each sample have an arbitrary or non-informative sum (Aitchison 1986), such as data obtained from high throughput DNA sequencing ~~data~~ (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~~when analyzed using standard approaches~~ (Aitchison 1986, Warton et al. 2012, Friedman and Alm 2012, Fernandes et al. 2013 Pawlowsky-Glahn et al. 2015) ~~(Pearson 1896)~~.

The essential problem is illustrated in Figure 1 where we set up an artificial example and count the number of molecules in the environment~~, and~~. We allow one part (shown in blue) to increase~~s~~ 10-fold between samples 1 and 2, while the abundance of the other 49 ~~other~~ parts (in red) remain unchanged. The proportion panel shows how the data are distorted when we convert it to relative abundances or proportions, or as happens when a constant sum is imposed by the sequencing instrument. The blue part still appears to become more abundant, although it is less than a 10-fold change.

However, the 49 red 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, principle 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).

Indeed, the original issue with compositional data identified by Pearson (1896) was that of spurious correlation. That is, two 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 three samples (a, b and c) with four taxonomic variables measured to have the following absolute counts in three environmental samples (i.e., samples are in rows, taxa are in columns):

$$\text{abc} = \begin{bmatrix} 470 & 66 & 839 & 751 \\ 541 & 569 & 787 & 512 \\ 167 & 906 & 959 & 504 \end{bmatrix}, \text{cor(abc)} = \begin{bmatrix} & -0.68 & \mathbf{-0.99} & 0.36 \\ -0.77 & & \mathbf{0.59} & -0.93 \\ \mathbf{-0.30} & \mathbf{-0.37} & & -0.25 \\ 0.55 & -0.95 & 0.62 & \end{bmatrix}.$$

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 between 2 and 3 is now -0.37. Thus, the correlation observed in compositional data

is not the same as the correlation for the counts, and the correlations measured can even change sign.

~~However, t~~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 correlations between taxa 1 and 2 is positive (0.43), and the correlation between 2 and 3 is even more strongly negative at -0.79. Thus, correlation determined from ~~proportional or~~ 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 ~~also~~ 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 (Filmoser et al. 2009, Friedman and Alm 2012, Fernandes et al 2013, Fernandes et al. 2014, Lovell et al. 2015, Kurtz et al. 2015).

The essential first step of proper compositional data analysis 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~~,~~ ~~Aitchison and Greenacre 2002~~), 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 ~~we~~ *X* is~~have~~ a vector of numbers ~~*X*~~ that contains *D* parts:

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

the centered log-ratio of *X* can be computed as:

Formatted: Indent: First line: 0.5"

X_{clr} = [log[x_1/g_X], log[x_2/ g_X], … log[x_D/ g_X],

where g_X is the geometric mean of all values in vector X (Aichison 1986). This simple transformation renders valid all standard multivariate analysis techniques (Aitchison 1986, van den Boogaart 2013, Pawlowsky-Glahn et al. 2015), and as shown in the Ratios panel of Figure 1 ~~Ratios~~, ~~reconstitutes~~ 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 ~~correlation~~essentially all- compositional data analysis (CoDa) based assessments of the datasets.

~~This~~ A CoDa approach would be ~~ideal~~ robust if microbiome datasets were not sparse, that is, they did not contain any 0 values~~,~~. H~~h~~owever 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~~a~~ 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). With this paradigm we dispose of taxa with ~~very~~ 0 counts in all or most samples (Palarea-Albaladejo and Martin-Fernandez 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, Fernandes et al.~~,~~ 2014).

Microbiome analysis tools that account for compositional data

Fortunately, the compositional data analysis problem of microbiome datasets 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 three 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 compositional data analysis (Aitchison 1986), or CoDa, offers a number of advantages over these tools, as explained below.

~~1: is there structure in the dataset?~~The first objective is to determine if there is structure in the dataset. In the microbiome field ~~This~~this is generally ~~handled~~ described as~~by~~ beta-diversity analysis ~~in the microbiome field~~. Beta-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 Principle Co-ordinate

Formatted: Indent: First line: 0", Tab stops: 0.64", Left + 1.27", Left + 1.91", Left + 2.54", Left + 3.18", Left + 3.82", Left + 4.45", Left + 5.09", Left + 5.73", Left + 6.36", Left + 7", Left + 7.63", Left + 8.27", Left + 8.91", Left + 9.54", Left + 10.18", Left

Formatted: Font: English (Canada)

(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 ~~compositional data~~CoDa analysis analog to PCoA is a principle component analysis (PCA) of center-log ratio 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 Martin-Fernandez 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 dataset, 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.

~~2: What is the difference between groups?~~ 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 Figure 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 ~~is~~ may be unchanged. This effect can be ameliorated by the center-log ratio transformation.

There are two approaches that assess differential abundance in a compositional data analysis 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 software is not deposited into a public repository, and that the 0-replacement value used is fixed in the software.

A slightly more complex approach is used by the ALDEx2 package, available from Bioconductor (Fernandes et al 2013, Fernandes et al 2014). Like ANCOM, ALDEx2 centre log-ratio 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 significant 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.

~~3: What is the correlation structure of the taxonomic groups?~~ The third objective is to determine if there are taxa in the dataset with correlated abundances. As noted above, spurious correlation is a very large problem in microbiome datasets~~, and~~. Therefore, ~~any~~ analys~~i~~es that report~~s~~ correlations using traditional methods, such as Pearson's or Spearman's correlations, Kendall's Tau or Partial correlations ~~is~~ 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 two taxa should be 0 or nearly so for two 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 phi statistic for two 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 compositional data analysis 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 phi~~, and make strong assumptions about the scarcity of the data.~~. 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 Two ~~Results and Use Cases~~Case Studies

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

The fi~~F~~irst worked example is (a vaginal microbiome dataset. This~~): We first use a~~ dataset 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 labeled 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 (labeled as ' i' for intermediate). In addition, individual taxa were aggregated to genus level using QIIME (Kuczynski et al. 2012), except for *Lactobacillus*

Formatted: Font: Not Bold

Formatted: Tab stops: 0.64", Left + 1.27", Left + 1.91", Left + 2.54", Left + 3.18", Left + 3.82", Left + 4.45", Left + 5.09", Left + 5.73", Left + 6.36", Left + 7", Left + 7.63", Left + 8.27", Left + 8.91", Left + 9.54", Left + 10.18", Left

Formatted: Font: Not Bold

Formatted: No underline

Formatted: No underline

Formatted: No underline

Formatted: No underline

iners and *Lactobacillus crispatus*, which remained as separate species in the tables. This relatively simple dataset will be used to introduce and explain the CoDa analysis methods.

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

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

Figure 2 shows the compositional biplot for this dataset along with the associated scree plot ~~showing~~ that displays the percentage of variance explained by each sample or component. The sample names (labeled in red for BV, blue for Normal or ~~cyan~~ purple for Intermediate) illustrate the variance ~~between~~ 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

Formatted: Font: Not Bold

ratio of each taxon to all other taxa. There are many interpretation rules for biplots of compositional data (Aitchison and Greenacre 2002), but these ~~boil down to~~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~~s~~ 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 dataset, a rule of thumb is that PCA plots that display less than 80% of the variance on the first two 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 *Lactobacillus 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 two 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.

Fourth, the shortest link observed in the plot is between *Megasphaera* and BVAB2. From this we conclude that the ratio of these two taxa is relatively constant

Formatted: Font: Italic

365 across all samples. That is, their ratio abundance is highly correlated. These two taxa
366 should be seen to have a low value of phi, but we must keep in mind the limit of the
367 projection of the data.
368 Fifth, the link between *Prevotella* and *Lactobacillus*. *crispatus* passes directly
369 through *Atopobium*. This indicates that these three taxa are linearly related. In this case,
370 it is clear when *L. crispatus* increases, the other two will decrease. Likewise, this
371 property can be extended to any linear relationships containing three or more links.
372 Sixth, the link between *L. iners* and *Megasphaera*, and the link between
373 *Leptotrichia* (*Sneathia*) and *Lactobacillus* cross at approximately 90°. The cosine of the
374 angle approximates the correlation between the connected log ratios. Thus, we can
375 conclude that the abundance relationship between the former pair of taxa is poorly
376 correlated with that of the latter two taxa. In other words, these two pairs vary
377 independently in the dataset.
378 Some samples (A312_bv, B312_~~bv~~i, A282_n at the bottom), are tightly grouped,
379 indicating that they contain similar sets of taxa at similar ratio abundances. We ~~would~~
380 can see from the biplot ~~expect~~ that these samples contain an abundance of
381 *Lactobacillus* and be depleted in *Leptotrichia* (*Sneathia*). Furthermore, we can see that
382 the samples divide into two fairly clear groups, with most of the before or “B” samples
383 on the left, and most of the after or “A” samples on the right. We further observe that~~,~~
384 ~~and that~~ the majority of the B samples are colored red indicating a diagnosis of BV, and
385 the majority of the A samples are colored blue indicating a diagnosis of non-BV.
386 The result of the biplot suggested that there were two main groups that could be
387 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* sp., 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 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 compositional data analysis (Martin Fernandez 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 sub-composition of the data (i.e., a subset of the taxa).

Formatted: Font: Not Bold

The results of unsupervised clustering of the dataset are shown in Figure 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 (Martin-Fernandez 1998).

The cluster analysis supports the results of the biplot and shows the split between two types of samples rather clearly. Samples containing an abundance of *Lactobacillus* sp. 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 four samples in the middle lower part

of the biplot in Figure 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 two results it would be appropriate to exclude these four 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 ~~four~~ 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) shown in Figure 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 compositional data analysis to normalize the data for ~~sequencing effort~~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 the change in abundance of a taxon *relative to all others* in the dataset. 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

Formatted: Font: Not Bold

explained more fully in the ~~Supplement~~supplementary material and the original papers (Fernandes et al. 2013, 2014).

In the examples given in Table 1, we filtered to show only those taxa where the expected Bejamini-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 ~~uncorrelated~~ not contributing to ~~in these samples with~~ the separation of the two large groups.

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
L. crispatus	-1.07	1.78	-0.49	0.23	0.000	0.004
L. 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

diff.btw: median difference between groups on a log base 2 scale

diff.win: largest median variation within group H or BV

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

The left panel of Figure 4 shows a plot of the within (diff.win) to between (diff.btw) condition differences, with the red dots representing those that have a BH 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. 2015). The grey lines represent the line of equivalence for the within and between group values. Black dots are taxa that are less abundant than the mean taxon abundance: here it is clear that the abundance of ~~these~~ rare taxa, ~~in general,~~ are generally difficult to estimate with any precision.

The middle plot in Figure 4 shows a plot of the effect size vs. the BH adjusted P value, with a strong correspondence between these two 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 phi measure (Lovell et al. 2015). In the example shown in the ~~S~~supplementary material, we calculate the mean phi using the same philosophy as outlined above for univariate statistical tests.

~~When examining compositional data it is important to remember that the ratio between taxa is the only information available.~~ In the context of microbiome datasets, the phi metric (Lovell et al. 2015) seeks to identify those pairs of taxa that have ~~the mosta~~ near constant ratio abundance across all samples~~s~~. Applying this approach to the dataset shows that the two most compositionally associated taxa are *Prevotella* sp. and *Megasphaera* sp. Note, that 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 dataset.

~~Second worked example:Examining the Hsiao Dataset using compositional approaches:~~For the second worked example w~~W~~e include in the workshop.Rnw document a second example based on the data of Hsiao et al. (2013) that examined the effect of ~~*Bacillus*~~ *Bacteriodes fragilus* supplementation on the microbiome composition of a mouse model of autism. This 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 and changes in abundance of a number of bacteria that composed the mouse gut

Formatted: Font: Not Bold

Formatted: Font: Italic

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 the paper.

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

Formatted: Indent: First line: 0"

Since the~~multivariate way~~ authors concluded that there was no evidence for multivariate differences between groups, and the CoDa approach agree, 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 ~~identified~~ 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. 2012), but that do not assume the data are 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 six taxa labeled as 53, 145, 638, 836, 837, and 956 in Figure 4 ~~(Supplementary Table 3)~~ of the paper. The reason for this discrepancy is that inspection of the original paper reveals that raw, and not Benjamini-

Hochberg adjusted P values were reported. 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 beenbeing critical of this 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 inappropriate use or lack of corrections for multiple hypothesis testing.

Formatted: Indent: First line: 0.5"

Discussion and Summary

Because the total number of reads is uninformative in high throughput DNA sequencing datasets, the only information available is the ratio of abundances between components: thus these data are compositional. Using twoa 16S rRNA gene sequencing datasets, we have illustrated that microbiome data are can be examined using a multivariate CoDa approachcompositional and so best treated as ratios that because the total number of reads is uninformative. By treating the data as ratios where the denominator is the geometric mean for a sample. , we can accurately recapitulate the shape and the error profile of the input data. Dirichlet Monte-Carlo replicates

coupled with the centered log-ratio transformation can ameliorate the sparse data problem inherent in microbiome datasets. ~~show that point estimates of statistical significance in a real dataset can substantially inflate the observed P value because of random partitioning of low count values across datasets.~~

In essence, we argue here that 16S rRNA gene sequencing datasets~~, RNA-seq datasets, and many other -seq datasets~~ are not special and do not ~~each~~ 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 ~~the~~ domain~~-~~specific approaches, ~~(~~such as those used in the ~~Hsieh et al~~ second example~~.~~ paper (Hsiao et al. 2013).

With the human body associated with a large number and diversity of bacteria, we need to understand ~~why it evolved like this~~the evolution of this association and~~,~~ how and when ~~programing~~ this intimate association develops. Such understanding will in turn lead us to robust approaches focussed on when and how~~between microbes and human cells happens, and how and when~~ to influence ~~we~~ the microbiome ~~can influence it~~ by ~~either~~ 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; Seekatz et al. 2014; Rajca et al. 2014). ~~This~~ 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. In order to have confidence in such findings, datasets must be

interrogated with rigour. The public is thirsty for knowledge and the media anxious to attract attention. Reliance on pharmaceutical agents is longer acceptable, and the ability to manipulate the microbiome is not only appealing but actually feasible. Thus, studies that help to understand how such manipulations occur, what communication is taking place between microbes and the host, will allow for more precisely targeted interventions, even to some extent personalized. In particular for the latter, as 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) that outlines the essential geometric problem of compositional data as it is understood at present ~~by Pawlowsky-Glahn et al. (2015)~~. For a ~~step-by-step~~ guide that goes beyond the introduction given in the ~~Supplement~~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, ~~avoid making claims that are~~

later disproven, and more importantly to bring badly needed breakthroughs in prevention, treatment and cure of disease.

Figure Legends

Figure 1: The difference between counting, proportions and ratios. The 'Counts' panel shows a scatter plot of a simulated dataset with two 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 filed 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, there 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 cirle 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 the a ratio transformation where each count (or proportion) is divided by the geometric mean of the 50 taxa in the sample a. Examination of the data after this transformation can thus provide more robust inferences.

Figure 2: The left figure shows a covariance biplot of the abundance--filtered dataset, 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 rather 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 right indicated unit-scaled variance of the taxa, the numbers on the top and right indicate unit scaled variances of the samples. 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.

Figure 3: Unsupervised clustering of the reduced dataset. 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 sub-compositions 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.

Figure 4: An effect plot showing the univariate differences between groups (Gloor et al. 2015). The left plot shows a plot of the maximum variance within the B or A group vs. the difference between groups. Red 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.

Figure 5: A form biplot of the Hsiao et al. (2013) dataset 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,

the ~~and that the~~ proportion of variance explained in the first component is not ~~as~~ large when compared to the other components ~~is not as obvious as in the biplot in Figure 2~~. The evidence of structure within this dataset is thus weak.

Funding: Financial support for this study was provided by a joint Canadian Institutes of Health Research (CIHR) Emerging Team Grant and a Genome British Columbia (GBC) grant awarded on which GR was a co-PI and GG and ML were co-investigators (grant reference #108030). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Aitchison, J. 1986. The statistical analysis of compositional data, Chapman and Hall, London England. ISBN 1-930665-78-4
2. Aitchison, J and Greenacre, M. 2002. Biplots of compositional data. J. Royal Stat. Soc: Series C. 51:375-92
3. Benjamini, Y., Hochberg, Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. . Royal Stat. Soc: Series B (Methodological), 289-300.
4. David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., Biddinger, S.B., Dutton, R.J., Turnbaugh, P.J. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature. 505(7484):559-63.

5. Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., Reid, G. 2013. High throughput sequencing methods and analysis for microbiome research. J Microbiol Methods. 2013 Dec;95(3):401-14.
6. Fernandes, A. D., Macklaim, J. M., Linn, T. G., Reid, G., Gloor, G. B. 2013. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PloS One, 8(7), e67019.
7. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., 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.
8. Filzmoser, P., Hron, K., Reimann, C. 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci tTtal Environ. 407:6100-8.
9. Frémont, M., Coomans, D., Massart, S., De Meirleir, K.. 2013. High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients. Anaerobe. 22:50-6.
10. Friedman, J., Alm, E. J. 2012. Inferring correlation networks from genomic survey data. PLoS Comput. Biol. 8(9): e1002687
11. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B. et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. Gen. Biol. 5 (10): R80.
12. Gloor, G.B., Macklaim, J.M., Fernandes, A.F. 2016. Displaying variation in large datasets: a visual summary of effect sizes. J. Comput. Graph. Stat. (in press)

~~12.~~13. Gloor, G.B., Macklaim, J.M., Vu, M, Fernandes, A.F. 2016. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. Austrian Journal of Statistics (in press).

14. Hsiao, E. Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A., Chow, J., Reisman, S.E., Petrosino, J.F., Patterson, P.H., Mazmanian,S.K. 2013. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. Cell. 155(7):1451-63

~~13.~~15. Hubert, L., Wainer, H. 2012. A statistical guide for the ethically perplexed. CRC Press, London, UK.

~~14.~~16. Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., Knight, R. 2012. Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr. Prot. Microbiol. 1E-5.

~~15.~~17. Kurtz, Zachary D and Müller, Christian L and Miraldi, Emily R and Littman, Dan R and Blaser, Martin J and Bonneau, Richard A 2015. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comp. Bio. 11:e1004226

~~16.~~18. Kwak, D.S., Jun, D.W., Seo, J.G., Chung, W.S., Park, S.E., Lee, K.N., Khalid-Saeed, W., Lee, H.L., Lee, O.Y., Yoon, B.C., Choi, H.S. 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):1353-9.

Formatted: Font: (Default) Arial, Font color: Text 1, English (U.S.)

~~17.~~19. Lourenço, T.G., Heller, D., Silva-Boghossian, C.M., Cotton, S.L., Paster, B.J., Colombo, A.P. 2014. Microbial signature profiles of periodontally healthy and diseased patients. J Clin Periodontol. 41(11):1027-36.

20. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J. Marguerat, S., Bähler, J. 2015. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol 11:e1004075.

~~18.~~21. Lozopone, C., Knight, R. 2005. Unifrac: a new phylogenetic method for comparing microbial communities. Applied Env. Micro. 71:8228-8235.

~~19.~~22. Macklaim, J.M., Clemente, J.C., Knight, R., Gloor, G.B., Reid, G. 2015. Changes in vaginal microbiota following antimicrobial and probiotic therapy. Microb Ecol Health Dis. 26:27799.

~~20.~~23. Mandal, S., Van Treuren, W., White, RA., and Eggesbø, M., Knight, R., Peddada, S. D. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microl. Ecol. Health Dis. 26:27663.

~~21.~~24. Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. 1998. Measures of difference for compositional data and hierarchical clustering methods. In A. Buccianti, G. Nardi, & R. Potenza (Eds.), Proc. IAMG (Vol. 98, pp. 526-531).

~~22. Mazmanian, S. 2015. https://sfari.org/funding/grants/abstracts/a-probiotic-therapy-for-autism.~~

~~23.~~25. Palarea-Albaladejo J., Antoni Martín-Fernández, J. 2015. zCompositions -- R package for multivariate imputation of left-censored data under a

compositional approach. Chemometrics and Intelligent Laboratory Systems. 143:85-96

24.26. Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. 2015. Modeling and Analysis of Compositional Data. John Wiley & Sons. Springer. 258 pg, London, UK.

25.27. 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. Royal Soc. Lond. 60:489-498

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

29. Rajca, S., Grondin, V., Louis, E., Vernier-Massouille, G., Grimaud, J.C., Bouhnik, Y., Laharie, D., Dupas, J.L., Pillant, H., Picon, L., Veyrac, M., Flamant, M., Savoye, G., Jian, R., Devos, M., Paintaud, G., Piver, E., Allez, M., Mary, J.Y., Sokol, H., Colombel, J.F., Seksik, P. 2014. Alterations in the intestinal microbiome (dysbiosis) as a predictor of relapse after infliximab withdrawal in Crohn's disease. Inflamm Bowel Dis. 20(6):978-86.

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

28.31. Schellenberg, J., Links, M. G., Hill, J. E., Dumonceaux, T. J., Peters, G. A., Tyler, S., Ball, T. B., Severini, A., Plummer, F. A. 2009. Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. Appl Environ Microbiol. 75: 2889-98.

~~29.~~32. Schloss, P.D, Westcott, S.L, Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B., Lesniewski, R.A., Oakley, B.B., Parks, D.H., Robinson, C.J., Sahl, J.W., Stres, B., Thallinger, G.G., and Van Horn, D.J., Weber, C.F. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities

33. Seekatz, A.M., Aas, J., Gessert, C.E., Rubin, T.A., Saman, D.M., Bakken, J.S., Young, V.B. 2014. Recovery of the gut microbiome following fecal microbiota transplantation. MBio. 5(3):e00893-14.

~~30.~~34. Segata, N., Izard, J., Waldron, L., Gevers, D., Miropolsky, L., Garrett, W.S., Huttenhower, C. 2011. Metagenomic biomarker discovery and explanation. Genome Biol. 12:R60

~~31.~~35. Urbaniak, C., Cummins, J., Brackstone, M., Macklaim, J.M., Gloor, G.B., Baban, C.K., Scott, L., O'Hanlon, D.M., Burton, J.P., Francis, K.P., Tangney, M., Reid, G. 2014. Microbiota of human breast tissue. Appl Environ Microbiol. 80(10):3007-14. Van den Boogaart, K. G., Tolosana-Delgado, R. 2013. Analyzing compositional data with R. Heidelberg: Springer. Heidelberg 258 pages.

~~32.~~36. Walker, A. W., and Martin, J.C., Scott, P., Parkhill, J., Flint, H. J. 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

37. Warton, D.I., Wrigth, S.T., Wang, Y. 2012. Distance-based multivariate analyses confound location and dispersion effects. Methods Ecol. Evol. 3:89-101.\

~~33.~~38. White, J.R., Nagarajan, N., Pop, M. 2009. PLoS Comput. Biol. 5:e1000352

763

Compositional data analysis for high throughput sequencing: an example from 16S rRNA gene sequencing.

Greg Gloor, ggloor@uwo.ca
CSM Workshop: 2015

March 4, 2016

Contents

1 What is this?

This is an document that contains intermingled LaTeXand *R* information. It is saved with the extension .Rnw, and if you have these two programs loaded on your machine you can regenerate this document on most platforms by running the `build_workshop.sh` bash script after you load in the `bbv_probiotic_samples.txt` file. If this is gibberish to you, don't worry, all the code to generate the outputs are in this pdf document. You can copy and paste them into an RStudio window, or equivalent, to make the figures.

2 Counts vs proportions

First, load in required libraries and functions.

```
# load the required R packages
require(compositions) # exploratory data analysis of compositional data
require(zCompositions) # used for 0 substitution
require(ALDEx2) # used for per-OTU comparisons
require(xtable) # used to generate tables from datasets
library(igraph) # used to generate graphs from phi data
library(car) # used to generate graphs from phi data
# you will need to download this directly from github
# https://github.com/DavidRLovell/propr
source("~/git/proprBayes/R/propr-functions.R") # rename proprBayes to propr
source("Rfunctions/functions.R") # rename proprBayes to propr
```

Not that everything is loaded, we can get to work. Make sure that all the packages are available in your R installation.

```
# make sure we get the same random numbers every time
set.seed(10000)

# generate the data, it is just random uniform
s1 <- c(runif(50, min=1e8, max=1e10))
s1[50] <- 2e10
s2 <- s1 * runif(length(s1), min=0.8, max=1.2)
s2[50] <- 2e11
s <- cbind(s1,as.vector(s2))


# proportions with a little bit of multivariate Poisson
# randomness added in
s.p <- apply(s,2,function(x){rdirichlet(1,(x/(0.0001*sum(x))))})

# clr
s.clr <- apply(s.p,2, function(x){log2(x) - mean(log2(x))})

# define variables for use later in the text
# number of the variable that changes
n.start <- s1[50]
n.end <- s2[50]
p.start <- round(s.p[50,1] * 100, 1)
p.end <- round(s.p[50,2] * 100, 1)
sum.s1 <- round(sum(s1)/1e9,1)
sum.s2 <- round(sum(s2)/1e9,1)
```

This section reconstitutes Figure 1 in the paper.

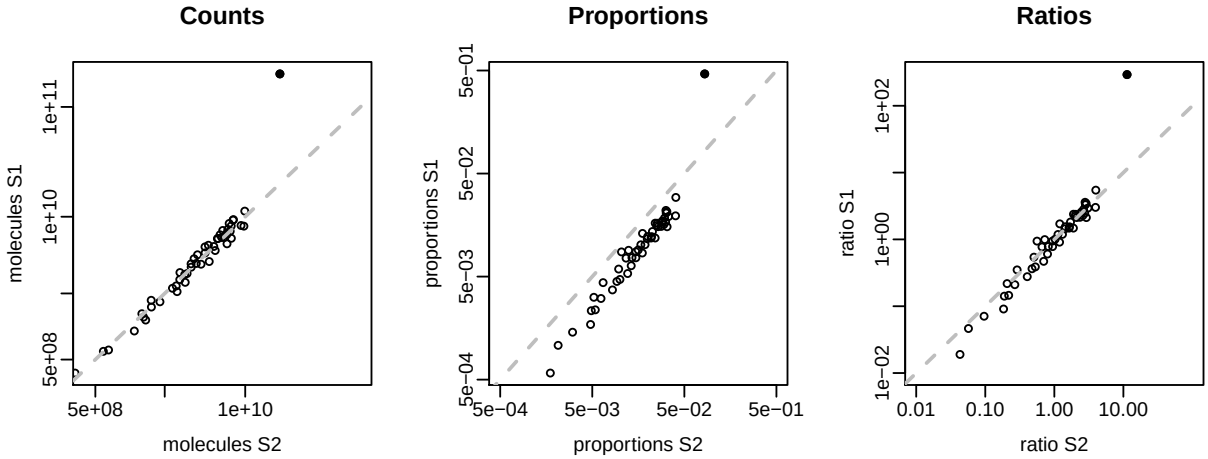


Figure 1: The difference between counts, proportions and ratios. Two samples of 50 molecules with 49 of those molecules differing in count only by random noise, and one molecule differing in count by a 10 fold increase, were randomly generated and plotted in the 'Counts' panel. The molecule that is different between the two samples is coloured in black, and the remainder fall on or near the dashed line of equivalence. The molecules in each sample were divided by the total sum in each sample, and converted to relative abundances. The result is plotted in the 'Proportions' panel, and it is obvious that the shape of the data has changed significantly. Taking the ratio of each molecule in a sample to the geometric mean abundance of each sample shows that the data are essentially reconstituted to their original shape, albeit with a different scale.

3 Compositional data analysis: more formal statement.

A dataset is defined as compositional if it contains D multiple parts, where each part is non-negative, and the sum of the parts is known (Aitchison 1986, pg 25). A composition containing D parts where the sum is 1 can be formally stated as: $C_D = \{(x_1, x_2, x_3, \dots x_D); x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, \dots x_D \geq 0; \sum_{x=1}^{D} = 1\}$. The sum of the parts is usually set to 1 or 100, but can take any value; i.e., any composition can be scaled to any arbitrary sum such as a ppm. It is important to know that the values of the parts of compositional datasets are constrained because of the constant sum. The constant sum constraint causes the parts to have a negative correlation bias since an increase in the value of one part must be offset by a decrease in value of one or more other parts. Thus any correlation-based analysis is invalid in these datasets, as originally noted by Pearson[1]. In addition, compositional datasets have the property that they are described by $D-1$ observations if the sum of the parts is known[2]. In other words, if we know that all parts sum to 1, then the last part can be known by subtracting the sum of all other parts from 1, i.e., $x_D = 1 - \sum_{x=1}^{D-1}$. Graphically, this means that compositions inhabit a space called the Aitchison simplex that contains 1 fewer dimensions than the number of parts. The distances between parts on the Aitchison simplex are not linear, especially at the boundaries. This is important because all common statistical tests assume a that differences between parts are linear (or additive). Thus, while standard tests will produce output, the output will be misleading because distances on the simplex are non-linear and bounded[3].

3.0.1 Sub-compositions:

Compositional data also exhibit the unusual property that the examination of a sub-composition of these data will provide different answers for those taxa in common in the full and sub-composition[2]. This is problematic because 16S rRNA gene sequencing experimental designs are *always* sub-compositions. Inspection of papers in the literature provide many examples. For example, it is common practice to discard rare OTU species prior to analysis and to re-normalize by dividing the counts for the remaining OTUs by the new sample sum. It is also common to use only one or a few taxonomic groupings to determine differences between experimental conditions. In the case of RNA-seq only the mRNA or miRNA is sequenced. All of these practices expose the investigator to the problem of non-coherence between sub-compositions.

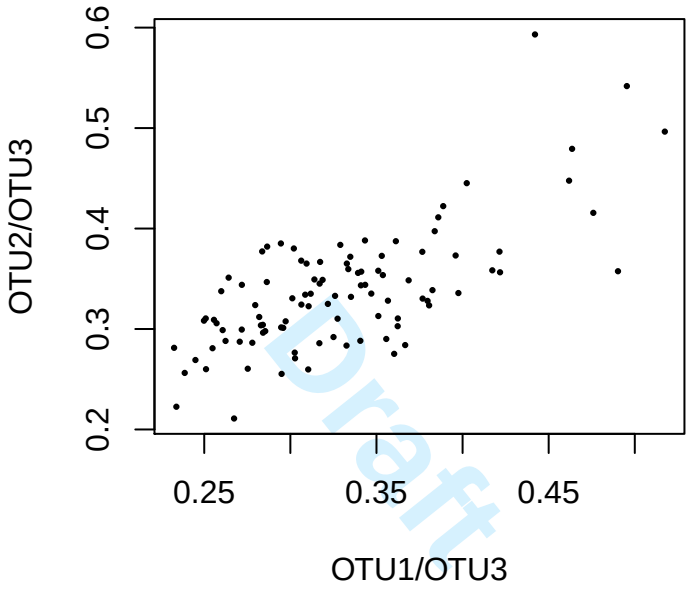


Figure 2: Spurious correlation in compositional data. Two random vectors drawn from a Normal distribution, were divided by a third vector also drawn at random from a Normal distribution. The two vectors have nothing in common, they should exhibit no correlation, and yet they exhibit a correlation coefficent of > 0.65 when divided by the third vector. See the introductory section of the Supplementary Information of Lovell[?] for a more complete description of this phenomenon.

3.0.2 Spurious correlations:

Finally, it is important to know that compositional data has the additional problem of spurious correlation[1], and in fact this was the first troubling issue identified with compositional data. This phenomenon is best illustrated with the following example from Lovell et. al[?] , where they show how simply dividing two sets of random numbers (say abundances of OTU1 and OTU2), by a third set of random numbers (say abundances of OTU3) results in a strong correlation. Note that this phenomenon depends only on there being a common denominator.

Practically speaking this means that *every microbial correlation network that has ever been published*

is suspect unless it was determined using SPARCC[4] or SPIEC-EASI[?] or the ϕ metric[5]. Lovell is in the process of producing an R package for the compositionally appropriate examination of correlations (personal communication).

Atichison[2], Pawlsky-Glahn[6], and Egozcue[7], have done much work to develop rigorous approaches to analyze compositional data[8]. The essential step is to reduce the data to ratios between the D parts as outlined above. This step moves the data from the Aitchison simplex and to the more familiar Euclidian space where the distances between parts are linear. However, the investigator must keep in mind that the distances are between ratios, not between counts. Several transformations are in common use, but the one most applicable to HTS data is the centred log-ratio transformation or clr, where the data in each sample is transformed by taking the logarithm of the the ratio between the count value for each part and the geometric mean count: i.e., for D features in sample X:

$$clr[x_1, x_2, x_3, \ldots x_D] = [log_2(x_1/gX), log_2(x_2/gX), log_2(x_3/gX) \ldots log_2(x_D/gX)],$$

where gX is the geometric mean of the features in sample X. This is the transformation described above.

4 So how can I analyze compositional data?

Fortunately, the analysis of compositional datasets has a well-developed methodology[? ?], much of which was worked out in the geological sciences. The following steps, and example code, is a step by step guide to examining a compositional 16S rRNA gene sequencing dataset in a more formally correct manner. This approach assumes that there is nothing really special about high-throughput sequencing data from the point of view of the analysis. The user should realize however that compositional data analysis is still an area of active research and the types of datasets typically found in high-throughput biology are particularly problematic because they are high-dimensional datasets that contain many 0 values.

4.1 An introduction to the compositional biplot

The compositional biplot is the essential workhorse tool for compositional analysis. Properly made and interpreted it summarizes all the essential results of your experiment. However, the weakness of this approach is that it is descriptive and exploratory, not quantitative unless at least 90% of the variance is explained on the first two principle components. Quantitative tools can be applied later to support the conclusions derived from the biplot.

We will illustrate this by examining a dataset from a clinical trial that examined the effect of treating women diagnosed as having bacterial vaginosis with either antibiotics, or antibiotics plus a probiotic supplement (Macklaim et.al, in press). For this example, I have extracted only the before and after treatment samples from the BV probiotic trial. Samples that are before treatment are identified as BXXX, where XXX is the sample identifier, and after treatment as AXXX. Samples are further identified as to their Nugent status, a rough indicator of whether the sample was from a women with BV or not: these are identified in the sample labels as '_bv' or '_n', some samples were indeterminate and are labeled as '_i'. In addition, for this analysis, individual OTUs have been aggregated to genus level using QIIME, except for *L. iners* and *L. crispatus* which remain as separate species in the tables.

We will use a dataset composed of the taxa that are more abundant than 0.1% in all samples. The following is a step-by-step guide in R with annotated code:

Then read the data in, adjust names, and convert to the centred log-ratio.

```
# load the data and the colours
d.pro.0 <- read.table("bbv_probiotic_samples.txt", header=T, row.names=1)

# remove awkward values from the names
rn <- gsub("_",".", rownames(d.pro.0))
rownames(d.pro.0) <- rn

# the first two rows and three columns of the data looks like this:
d.pro.0[1:2,1:3]
```

```
##                                B208_bv A208_n B210_bv
## Actinobacteria:Actinomyces           1     11       8
## Actinobacteria:Arcanobacterium       1      0       2
```

```
# a correspondence table of taxa and colours
col.tax <- read.table("bbv_colours.txt", header=T, row.names=1, comment.char="")

# again, change awkward characters in the row names
rownames(col.tax) <- gsub("_",".", rownames(col.tax))

# replace 0 values with the count zero multiplicative method and output counts
#
# this function expects the samples to be in rows and OTUs to be in columns
# so the dataset is turned sideways on input, and then back again on output
# you need to know which orientation your data needs to be in for each tool

d.pro <- t(cmultRepl(t(d.pro.0), method="CZM", output="counts"))
```

```
## No. corrected values:  42
```

```
# convert to proportions by sample (columns) using the apply function
d.pro.prop <- apply(d.pro, 2, function(x){x/sum(x)})

#####
# Make a dataset where the taxon is more abundant than 0.1% in all samples

# remove all taxa that are less than 0.1\% abundant in any sample
d.pro.abund.unordered <- d.pro[apply(d.pro.prop, 1, min) > 0.001,]

# add in the names again and sort by abundance
d.names <- rownames(d.pro.abund.unordered)[
    order(apply(d.pro.abund.unordered, 1, sum), decreasing=T) ]

# make a standard list of colours for plotting
colours <- as.character(col.tax[d.names,])
```

```
# get the taxa in the reduced dataset by name
d.pro.abund_unordered <- d.pro.abund.unordered[d.names,]

# order the taxa by their diagnosis bv, n or i
d.pro.abund <- data.frame(d.pro.abund_unordered[,grep("_bv", colnames(d.pro.abund_unordered))]
    d.pro.abund_unordered[,grep("_n", colnames(d.pro.abund_unordered))],
    d.pro.abund_unordered[,grep("_i", colnames(d.pro.abund_unordered))])

# make our compositional dataset
d.clr.abund <- t(apply(d.pro.abund, 2, function(x){log(x) - mean(log(x))}))

# more name plumbing!
colnames(d.clr.abund) <- gsub("\\w+:", "", colnames(d.clr.abund))
```

The first key function here is the 0 replacement function `cmultRepl` which has many options[9]. The bottom line is that the replacement of 0 values in these datasets is an area of ongoing research, and so there is no general way to treat 0 values in these datasets. The reader is encouraged to try different 0 replacement values and strategies and observe how it affects the conclusions.

The second key function is the `apply` function. This converts the data into centered log-ratios from the 0 replaced count dataset.

Compositional biplots show both the amount of variance of both samples and variables (taxa, shown with rays)[10]. Essentially, these plots are a projection (or shadow) of the multidimensional dataset onto two dimensions. They are oriented so that the maximum axes of variation are on components 1 and 2. If essentially all the variation is explained by the first two principle components, then the following rules can be used to examine the data:

1. the rays in this plot show the amount of variance exhibited by each taxon relative to the centre of the dataset where longer rays mean more variation across all samples.
2. the location of the sample name shows how variable it is relative to other samples.
3. samples that are highly variable, and that are in the same direction as a long ray for a taxon will contain that taxon in high abundance . The inverse is also true.
4. taxa where the tips of the rays are co-incident and of the same length indicate that the ratio between those two taxa are nearly identical across all samples up to the limit of the projection of the data. As we will see, the projections can be misleading if almost all the variation is not explained on the first two principle components.
5. taxa where the angles between the rays are orthogonal are uncorrelated.
6. taxa where the tips of the rays are very distant from each other, regardless of whether the link between the tips passes through the origin, indicated highly variable ratios across the samples.
7. three or more taxa lying on a common link will be positively or negatively correlated.
8. the angle between links contains information about correlations between pairs, or groups of ratios, more formally, the cosine of the angle is proportional the correlation between the pairs of ratios.

Let's generate a biplot of those taxa that are more abundant than 0.1% in any sample. First we make the singular value decomposition of the clr-transformed data. This makes a series of projections of the data that explain progressively less of the data on each component.

```
# Singlular value decompositon method of making a PCA (base R)
pcx.abund <- prcomp(d.clr.abund)

# getting info to color the samples
conds <- data.frame(c(rep(1,length(grep("_bv", rownames(d.clr.abund)))),
    rep(2, length(grep("_n", rownames(d.clr.abund)))),
    rep(3, length(grep("_i", rownames(d.clr.abund)))) ))
colnames(conds) <- "cond"

palette=palette(c(rgb(1,0,0,0.6), rgb(0,0,1,0.6), rgb(0,1,1,0.6)))
```

Now we can plot the data using the coloredBiplot function from the `compositions` R package. We are placing the amount of variance explained as part of the axes labels. The axes are unit scaled and values have little intrinsic meaning.

The covariance biplot made from the abundance filtered dataset in Figure 3 is somewhat informative. The first two components explain 0.61 of the variance in the data, and we can observe some structure both in the taxa and in the samples. The left side of the plot contains all of the organisms commonly observed in BV, and the right side of the plot contains only members of the genus *Lactobacillus*; indicating a clear split in the makeup of the samples. When we focus on the location of the samples, the majority of the before treatment samples are on the left hand side of the plot, and the majority of the after treatment samples are on the right hand side.

Note that biplots can use scale = 0 to generate a form biplot - you scale and interpret by the samples, or scale = 1 to generate a covariance biplot - you scale and interpret by the taxa

Applying the 8 rules of interpretation given above (and using only the genus name for brevity), we can see that:

1. *L. iners* has the longest ray, and among these taxa is exhibits the most variation relative to all taxa across samples, *Gardnerella* has the shortest ray and so is the least variable relative to all other taxa.
2. The sample A238_i is the sample that is least similar to any other sample because it is furthest from the centre (top left corner). It likely contains a substantial fraction of *Sneathia* and little *Megasphaera* and *BVAB2*.
3. The sample A238_i will contain a substantial proportion of *Sneathia* and a very small amount of *Lactobacillus* because the rays for these taxa are parallel to a ray that would connect this sample to the origin. The converse will be true for sample A314_n.
4. The two closest tips are for *Megasphaera* and *BVAB2*, thus the ratio between these two taxa should be relatively constant across samples, up to the limit of the projection. Thus, each will be abundant when the other is abundant and *vice versa*.
5. The abundance of *Sneathia* and the taxa *Gardnerella*, *Megasphaera*, *BVAB2* will be uncorrelated because these rays are approximately orthogonal.
6. The link between *L. iners* and *Sneathia* (and many others is very long), indicating that the

```
layout(matrix(c(1,2),1,2, byrow=T), widths=c(6,2), heights=c(8,3))
par(mgp=c(2,0.5,0))
# make a covariance biplot of the data with compositions function
coloredBiplot(pcx.abund, col="black", cex=c(0.6, 0.6), xlabs.col=conds$cond,
    arrow.len=0.05,
    xlab=paste("PC1 ", round (sum(pcx.abund$sdev[1]^2)/mvar(d.clr.abund),3), sep=""),
    ylab=paste("PC2 ", round (sum(pcx.abund$sdev[2]^2)/mvar(d.clr.abund),3), sep=""),
    expand=0.8,var.axes=T, scale=1, main="Biplot")

barplot(pcx.abund$sdev^2/mvar(d.clr.abund),  ylab="variance explained", xlab="Component", main=
```

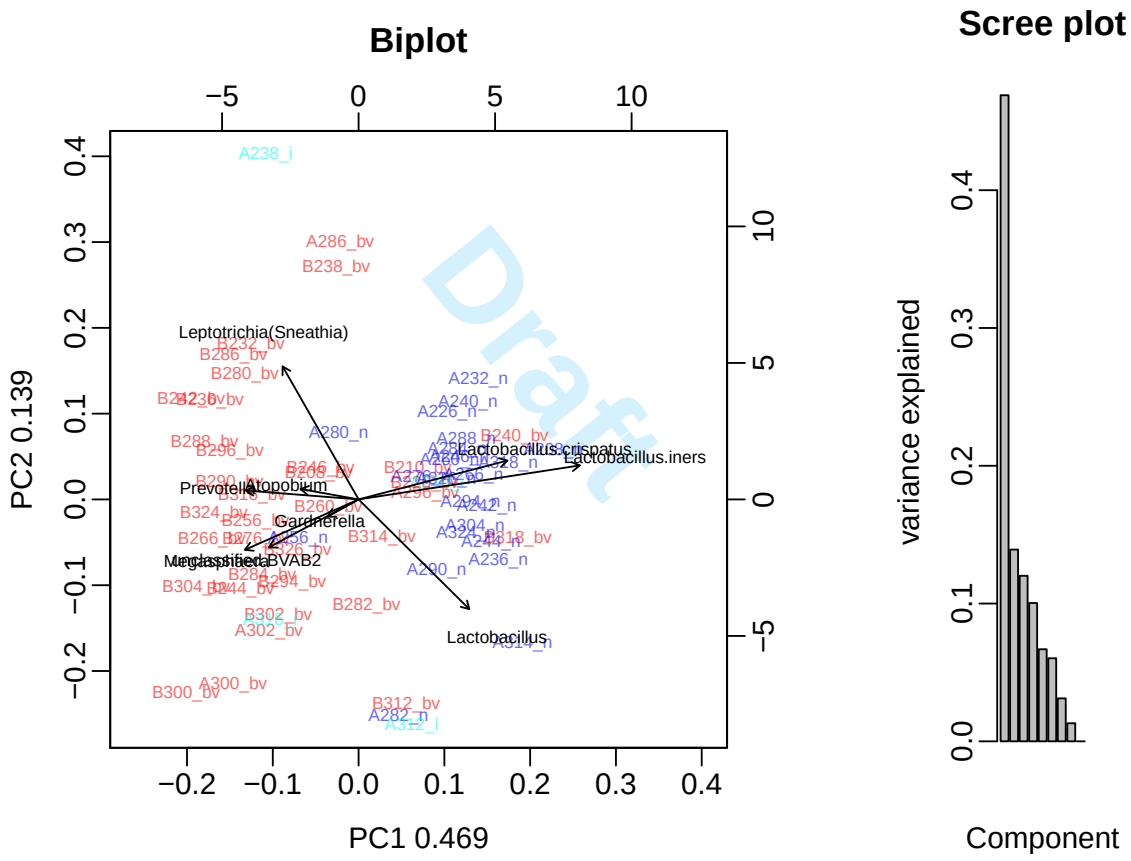


Figure 3: The left figure shows a covariance biplot of the abundance-filtered dataset, the right figure shows a scree plot of the same data. This exploratory analysis is much encouraging because the amount of variance explained is rather substantial with 0.469 of the variance being explained by component 1, and 0.139 being explained by component 2. The scree plot also shows that the majority of the variability is on component 1. We can interpret this biplot with some confidence.

ratios between these taxa are extremely variable. That is, when *L. iners* is abundant, there is little information about the abundance of *Sneathia*.

7. The link between *Prevotella* and *L. crispatus* passes directly through *Atopobium*. This indicates that these three taxa are linearly related. In this case, it is clear when *L. crispatus* increases, the other two will decrease.

8. The link between *BVAB2* and *Sneathia* and the link in the previous item intersect at approximately 90 degrees. Thus the ratios of the last two taxa will be uncorrelated with the ratios of the previous three taxa.

This biplot suggests some structure in the BV samples that is related to the abundance of *Sneathia*. The evidence for this is that the abundance of this genus is not correlated with the abundance of the others that are commonly found in BV. This 'pulls' several samples towards the top right corner. Investigation of other datasets would be required to test this observation.

4.1.1 Cluster analysis

The result of the biplot suggested that there were two 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* sp., and those lacking them. We can explore this by performing a cluster analysis. In the traditional microbiome analysis methods, clustering is based on the weighted or unweighted unifrac distances or on the Bray-Curtis dissimilarity metric. These metrics are much more sensitive to the makeup of the community than is the Aitchison distance used in compositional data analysis. Thus, here we will use the Aitchison distance metric which fulfills the criteria required for compositional data. In particular, by using a compositional approach, it is appropriate to examine a defined sub-composition of the data (i.e., we can make fairly robust conclusions even if not all taxa are included).

```
# generate the distance matrix
dd <- dist(d.clr.abund, method="euclidian")

# cluster the data
hc <- hclust(dd, method="ward.D2")

# now re-order the data to plot the barplot in the same order
d.order <- d.pro.abund[,hc$order]
d.order.acomp <- acomp(t(d.order))
```

The results of unsupervised clustering of the dataset is shown in Figure 4. Here we can use Euclidian distance because the Aitchison transformed data are linearly related and in placed in the familiar space. However, the user must remember that all distances are calculated from the ratios between taxa, and not on the taxa abundances themselves! For this figure we are using 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.

The cluster analysis shows the split between two types of samples rather clearly. Samples containing an abundance of *Lactobacillus* sp. are grouped together on the right, and samples with an abundance of other taxa are grouped together on the left.

```
layout(matrix(c(1,3,2,3),2,2, byrow=T), widths=c(6,2), height=c(4,4))
par(mar=c(2,1,1,1)+0.1)

# plot the dendrogram
plot(hc, cex=0.6)
# plot the barplot below
barplot(d.order.acomp, legend.text=F, col=colours, axisnames=F, border=NA, xpd=T)
par(mar=c(0,1,1,1)+0.1)
# and the legend
plot(1,2, pch = 1, lty = 1, ylim=c(-20,20), type = "n", axes = FALSE, ann = FALSE)
legend(x="center", legend=d.names, col=colours, lwd=5, cex=.6, border=NULL)
```

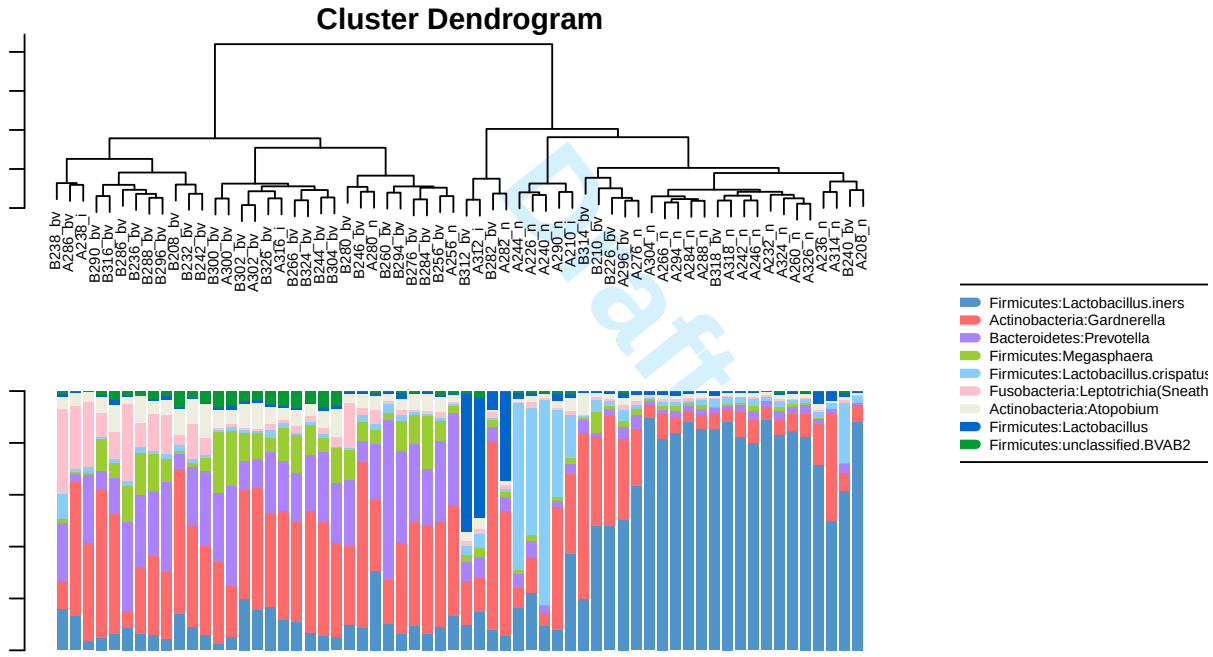


Figure 4: Unsupervised clustering of the reduced dataset. The top figure shows a dendrogram of relatedness generated by unsupervised clustering of the Aitchison distances, which is the only distance that is robust to perturbations and sub-compositions of the data[3]. The bottom figure shows a stacked bar plot of the samples in the same order. The legend indicating the colour scheme for the taxa is on the right side. Note that this confirms that *Lactobacillus* and *Sneathia* rich samples are outliers for the BV and H groups.

The results of the cluster analysis can help clarify the compositional biplot. For example, the four samples in the middle lower part of the biplot in Figure 3 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, The cluster plot and associated barplot show that they contain 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 two results it would be appropriate to exclude these four samples from further analysis because of their atypical makeup.

As indicated from the biplot, the abundance of *Gardnerella* sp. is not a good discriminator between the two groups because it may be abundant or rare in either group.

4.2 Univariate differences between groups

We will now conduct a univariate comparison between the B and A groups, for simplicity, we will keep the four outlier samples, but the reader is encouraged to remove them and see how the results change. For this, we will use the ALDEx2 tool, that incorporates a Bayesian estimate of the posterior probability of taxon abundance into a compositional framework. Here is the code:

```
# generate the dataset by making a data frame of
d.B <- colnames(d.pro.0)[grep("B", colnames(d.pro.0))] # Before samples
d.A <- colnames(d.pro.0)[grep("A", colnames(d.pro.0))] # After samples
d.aldex <- data.frame(d.pro.0[,d.B], d.pro.0[,d.A]) # make a data frame

# make the vector of set membership in the same order as
conds.aldex <- c(rep("Be", 31), rep("Af", 31))

# generate 128 Dirichlet Monte-Carlo replicates
x <- aldex.clr(d.aldex, mc.samples=128, verbose=FALSE)

## [1] "operating in serial mode"

# calculate p values for each replicate and report the mean
x.t <- aldex.ttest(x, conds.aldex)
# calculate mean effect sizes
x.e <- aldex.effect(x, conds.aldex, verbose=FALSE)

## [1] "operating in serial mode"

# save it all in a data frame
x.all <- data.frame(x.e,x.t)
```

The ALDEx2 tool estimates the distribution of taxon abundance by random sampling instances of the from a Dirichlet distribution that are consistent with the observed data. In more formal terms, we are generating a posterior distribution of the dataset using the observed dataset as the prior. This takes the original input data, and generates a distribution of posterior probabilities of observing each taxon. This distribution is transformed by the centred log-ratio transformation, and is used to conduct univariate statistical tests. These tests return a distribution of P and Benjamini-Hochberg adjusted P values, and the tool reports the mean of these distributions. In this way, we account for the large variation in these datasets, and identify only those taxa whose difference between the groups is robust to sampling variation.

We need to supply the table of counts and a list that outlines which group each sample belongs to. Following that, we generate the distribution of posterior probabilities using the `aldex.clr` function, then conduct the statistical tests and determine effect sizes using the `aldex.ttest` and `aldex.effect`. Finally, we can plot the results using `aldex.plot`.

The output table contained in `x.all` contains much information regarding your dataset, and is used to generate the output plot: see the documentation for ALDEx2 for a complete description of each entry in the table. The most important data for the purposes of comparison are those given in Table 2. All information in this table, except P value information, is on a log2 scale. Here we have the difference between groups (diff.btw), the maximum difference within groups (diff.win or variance), the effect size calculated as $\frac{diff.btw}{diff.win}$, the overlap between the Bayesian distributions of group A and B (overlap), and finally the raw expected P value from a Wilcoxon non-parametric test (wi.ep), and the expected Benjamini-Hochberg (wi.eBH) adjusted value.

When interpreting these results you should remember that you are actually examining ratios between values, rather than abundances. So abundances are determined as the ratio of the abundance of a taxon to all taxa in the sample. The user should also remember that all values reported are the mean values over the number of Dirichlet instances as given by the `mc.samples` variable in the `aldex.clr` function.

```
sig <- which(x.all$wi.eBH <= 0.05)
# make the table
xtable(
  x.all[sig,c(4:7, 10,11)], caption="Table of significant taxa", digits=3,
       label="sig.table", align=c("l",rep("r",6) )
)
```

	diff.btw	diff.win	effect	overlap	wi.ep	wi.eBH
Actinobacteria:Atopobium	0.857	1.521	0.519	0.295	0.007	0.038
Bacteroidetes:Prevotella	1.355	1.787	0.718	0.218	0.000	0.002
Firmicutes:Lactobacillus.crispatus	-1.101	1.790	-0.513	0.246	0.000	0.004
Firmicutes:Lactobacillus.iners	-2.262	2.680	-0.817	0.197	0.000	0.001
Firmicutes:Streptococcus	-1.121	2.315	-0.356	0.301	0.008	0.040
Firmicutes:Dialister	0.909	1.365	0.625	0.253	0.001	0.009
Firmicutes:Megasphaera	1.520	2.354	0.619	0.266	0.002	0.015

Table 1: Table of significant taxa

In the examples given in Table 2, we filtered to print only those taxa where the expected BH values was less than 0.05, meaning that the expected likelihood of a false positive identification *per taxon* is less than that threshold. Using *L. iners* as an example, we can see that the absolute difference between groups can be up to -2.26, implying that the absolute fold change in the ratio between *L. iners* and all other taxa between groups for this organism is on average $2^{2.26} = 4.8$ fold across samples. However, note that the difference within is even larger, giving an effect size of -0.82. Thus, we can see that the difference between groups is less than the variability within a group, a result that is typical for microbiome studies.

We can examine these data graphically as shown in Figure 5. The left panel of this figure shows a plot of the within to between condition differences[11], with the red dots representing those that have a BH adjusted P value of 0.05 or less. Taxa that that are more abundant than the mean in the BV

samples have positive y values, and those that are more abundant than the mean in the N samples have negative y values. We refer to these as 'effect size' plots, and they summarize the data in an intuitive way. The grey lines represent the line of equivalence for the within and between group values. Black dots are taxa that are less abundant than the mean taxon abundance: here it is clear that the abundance of these taxa, in general, are difficult to estimate with any precision.

The middle plot in Figure 5 shows a plot of the effect size vs. the BH adjusted P value, and we can see the strong correspondence between these two measures. In general, we prefer to use an effect size cutoff because this is more robust than are P values (eg. [11]? . The right plot in this figure shows a familiar volcano plot for reference.

4.3 Correlation with ϕ

Correlation is very problematic in these datasets. In fact, any correlation reported in compositional data must be treated as suspect if it was performed with *any of the standard correlation tools* such as Pearson's or Spearman's or Kendall's measures. The code to calculate the expected value of ϕ across Dirichlet Monte-Carlo instances is below. Note that the cutoff value of ϕ is relatively large, illustrating that the data are extremely variable.

```
x.p <- aldex.clr(d.pro.abund, mc.samples=128)

## [1] "operating in serial mode"

min.sma.df <- aldex.phi(x.p)
min.sma.df$row <- gsub(".+:", "", min.sma.df$row)
min.sma.df$col <- gsub(".+:", "", min.sma.df$col)
phi.cutoff <- 0.5
min.sma.lo.phi <- subset(min.sma.df, phi < phi.cutoff)

## generate a graphical object
g <- graph.data.frame(min.sma.lo.phi, directed=FALSE)
## # get the clusters from the graph object
g.clust <- clusters(g)

### # data frame containing the names and group memberships of each cluster
g.df <- data.frame(Systematic.name=V(g)$name, cluster=g.clust$membership,
   cluster.size=g.clust$csize[g.clust$membership])

# get clusters of a given size here all clusters are captured
big <- g.df[which(g.df$cluster.size >= 1),]
colnames(big) <- colnames(g.df)
```

Now we can plot the compositionally associated taxa.

```
layout(matrix(c(1,2,3,1,2,3),2,3, byrow=T), widths=c(5,2,2), height=c(4,4))
par(mar=c(5,4,4,1)+0.1)
aldex.plot(x.all, test="wilcox", cutoff=0.05, all.cex=0.8, called.cex=1)
plot(x.all$effect, x.all$wi.eBH, log="y", pch=19, main="Effect",
    cex=0.5, xlab="Effect size", ylab="Expected Benjamini-Hochberg P")
    abline(h=0.05, lty=2)
plot(x.all$diff.btw, x.all$wi.eBH, log="y", pch=19, main="Volcano",
    cex=0.5, xlab="Difference", ylab="Expected Benjamini-Hochberg P")
    abline(h=0.05, lty=2)
```

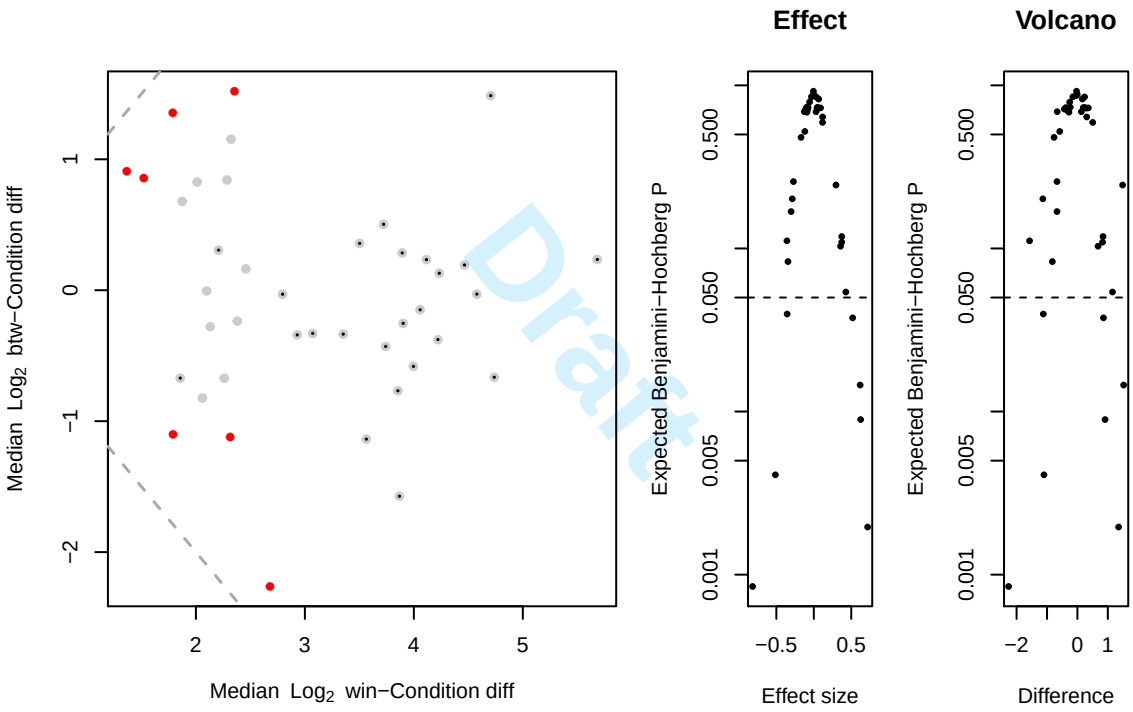


Figure 5: Examination of univariate differences between groups. The left plot shows a plot of the maximum variance within the B or A group vs. the difference between groups. Red points indicate those that have a mean Benjamini-Hochberg adjusted P-value of 0.05 or less using P values calculated with the 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.

```
plot(g)
```

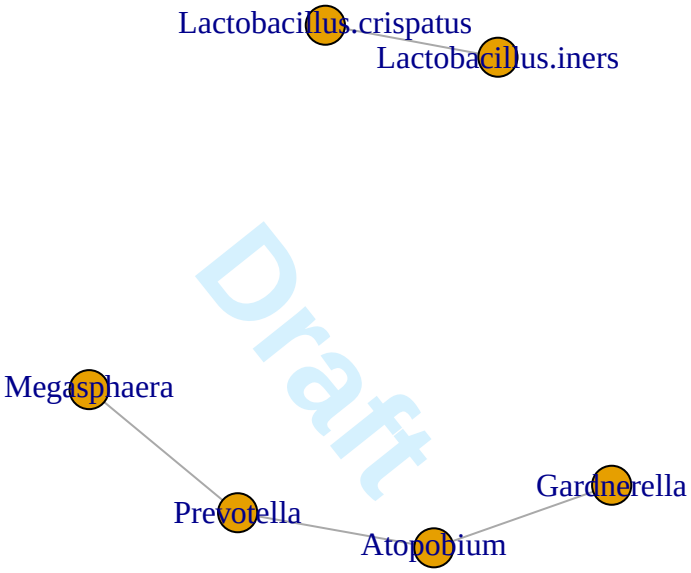


Figure 6: Correlation structure within the dataset. All pairs of taxa where $\phi <= 0.5$ are shown as edges connecting nodes. Note that the four taxa associated with BV in the biplot form one natural compositional-associated group, and the two *Lactobacillus* species associated with health form the other.

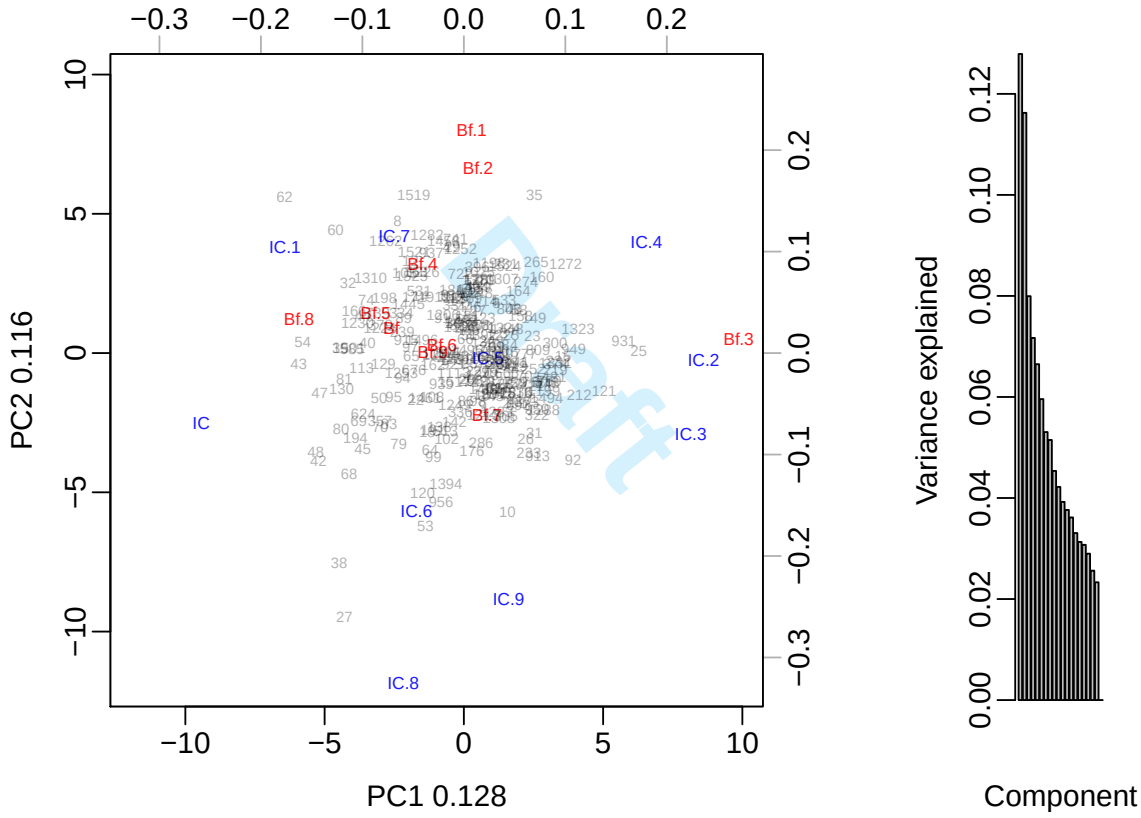
5 Examining the Hsiao et al. dataset

Hsiao et al. (2013)[12] conducted a study that examined the effect of *Bacillus fragilis* treatment in a mouse model of autism and concluded that there was a difference in the gut microbiota between *b. fragile* treated and control groups. This data is publicly available, and here we will analyze these two groups to determine if the conclusions are supported when we incorporate sampling variation, and use compositional analysis methods. The original dataset contained 1474 taxa, of which only 703 were found to occur with a count of 5 or more in the 10 samples under consideration. We will use this reduced dataset for analysis because it contains all the taxa identified as significantly different in the original analysis. Samples are labeled Bf if treated, and IC if control.

This section contains fewer annotations to the code, since we are recapitulating the same analysis with a different dataset.

```
# read the table
d <- read.table("hsiao5.txt", header=T, row.names=1)
tax.d <- read.table("tax.txt", row.names=1, header=T, sep="\t")

colnames(d) <- gsub("PolyIC...", "", colnames(d))
colnames(d) <- gsub("Poly", "", colnames(d))

# apply the same cutoff as before to simplify the data, since rare OTUs do not change the bipl
cutoff = 1-0.50

d.0 <- data.frame(d[which(apply(d, 1, function(x){length(which(x == 0))/length(x)}) < cutoff),]
tax.0 <- tax.d[which(apply(d, 1, function(x){length(which(x == 0))/length(x)}) < cutoff),1]

# replace 0 values with imputed value
d.bf <-cmultRepl(t(d.0),  label=0)
## No. corrected values:  107
# convert to clr transformed data
bi <- t(apply(d.bf, 1, function(x){log(x) - mean(log(x))}) )

# here we have to do singular value decomposition and make a PCA object as before
# arrows will represent samples, not taxa
# however because we care about the links between taxa, and the arrows don't contain any real
pcx <- prcomp(bi)

# getting info to color the samples
conds.bi <- data.frame(c(rep(1,length(grep("IC", rownames(bi)))),
    rep(2, length(grep("Bf", rownames(bi)))) ))
colnames(conds.bi) <- "cond"

palette=palette(c(rgb(1,0,0,0.9), rgb(0,0,1,0.9), rgb(0,1,1,0.6)))

# scale = 0 is a form biplot - you scale and interpret by the arrows
# scale = 1 is a covariance biplot - you scale and interpret by the samples
```

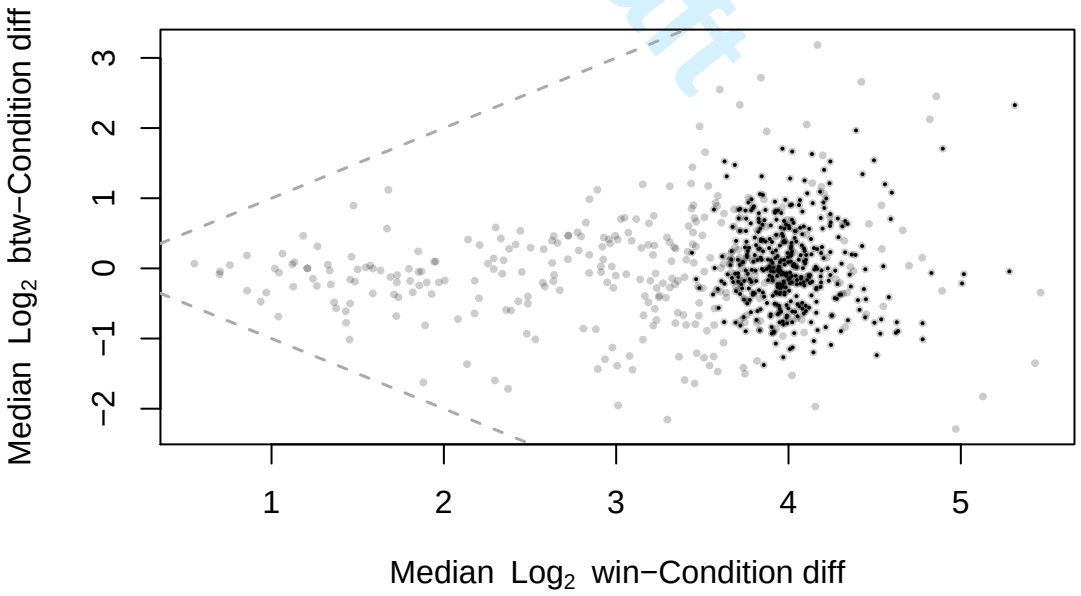
```
layout(matrix(c(1,2),1,2, byrow=T), widths=c(6,2), height=c(6,4))
par(mgp=c(2,0.5,0))
# here we use a form biplot to scale by 1
coloredBiplot(pcx, col=rgb(0,0,0,0.3), cex=c(0.6, 0.5),
    arrow.len=0.05, xlabs.col=conds.bi$cond, expand=0.8,var.axes=F, scale=0,
    xlab=paste("PC1 ", round (sum(pcx$sdev[1]^2)/mvar(bi),3), sep=""),
    ylab=paste("PC2 ", round (sum(pcx$sdev[2]^2)/mvar(bi),3), sep="")
)

barplot(pcx$sdev^2/mvar(bi), ylab="Variance explained", xlab="Component") # scree plot
```



Here we see some problems. First, while the scree plot shows that the first two components contain more information than the remainder of the components, it is not nearly as explanatory as for the BV dataset above. In fact, the first three components explain only 0.128, 0.116, and 0.08 of the proportion of variability of the data, a very low amount given the small size of the dataset, suggesting that the data are not particularly informative. It is clear that the BF and IC samples are intermixed, and do not separate, although the Bf samples may have lower overall dispersion.

We can now examine if there are any univariate differences between the two sample groups. Intuitively, this should be unlikely given the lack of multivariate separation.

```
conds <- c(rep("Bf", 10), rep("C", 10))

# generate technical replicates and perform the clr transformation
# here we will use all the OTUs since each OTU is treated independently of the others
x <- aldex.clr(d, mc.samples=128, verbose=FALSE)

## [1] "operating in serial mode"

# conduct the statistical tests and calculate FDR corrected values
# data are medians of all Dir instances for each OTU
x.t <- aldex.ttest(x, conds)

# calculate the effect sizes for plotting
x.e <- aldex.effect(x, conds, verbose=FALSE)

## [1] "operating in serial mode"

# merge into one data frame for plotting and examination
x.all <- data.frame(x.t, x.e)

# explore the dataset
aldex.plot(x.all)
```

We can address this issue by examining the data in a univariate way using ALDEx2 that will determine if there are differences in abundance of rare taxa. The plot shows that there are no significant hits (i.e., no red dots), and that the dataset displays extremely high variability, likely because of low OTU counts. We can examine the OTUs that pass a P value cutoff of 0.1 (to

be generous), and display their summary data in a table as before. Here we see that there were some 'significant' OTU's, but that none of these reach significance when the P values are adjusted for multiple test corrections (wi.eBH). Inspection of the original paper indicates that a multiple hypothesis test correction *was not* done on the reported P values. Thus, both the methods in the paper and ALDEx2 identified a small number of 'significant' OTUs, but the multiple test correction indicates that these are almost certainly false positive identifications. In the end, we should conclude that the treated and untreated samples were not different by either the multivariate or univariate criteria.

```
sig <- which(x.all$wi.ep <= 0.1)
# make the table
xtable(
  x.all[sig,c(4:7, 10,11)], caption="Table of  taxa with $P<0.1$", digits=3,
       label="sig.table", align=c("l",rep("r",6) )
)
```

	wi.eBH	rab.all	rab.win.Bf	rab.win.C	effect	overlap
1288	0.752	3.477	2.994	3.970	0.577	0.236
53	0.739	1.422	0.294	3.124	0.629	0.217
533	0.756	3.328	3.796	2.736	-0.609	0.231
145	0.728	0.108	-1.248	1.949	0.688	0.217
64	0.770	1.285	0.287	2.593	0.490	0.250
4	0.803	2.813	3.275	1.953	-0.559	0.231
602	0.680	2.299	3.085	1.367	-0.608	0.198
26	0.792	4.868	4.383	5.164	0.591	0.245
1262	0.758	1.499	2.359	0.255	-0.598	0.220
956	0.709	0.813	-0.422	2.257	0.640	0.202
36	0.712	0.480	1.534	-0.615	-0.606	0.218
1248	0.793	5.770	6.134	5.456	-0.566	0.248
8	0.717	5.499	6.106	4.228	-0.839	0.195
1519	0.750	3.873	4.664	2.911	-0.666	0.208

Table 2: Table of taxa with $P < 0.1$

References

1) Karl 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*, 60:489–498, 1897.

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

3) JA Martín-Fernández, C Barceló-Vidal, V Pawlowsky-Glahn, A Buccianti, G Nardi, and R Potenza. Measures of difference for compositional data and hierarchical clustering methods. In *Proceedings of IAMG*, volume 98, pages 526–531, 1998.

4) Jonathan Friedman and Eric J Alm. Inferring correlation networks from genomic survey data. *PLoS Comput Biol*, 8(9):e1002687, 2012.

5) David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler.

Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol*, 11(3):e1004075, Mar 2015.

6) V. Pawlowsky-Glahn and J. J. Egozcue. Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications*, 264(1):1–10, 2006.

7) JJ Egozcue and V. Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7):795–828, 2005.

8) Vera Pawlowsky-Glahn and Antonella Buccianti. *Compositional data analysis: Theory and applications*. John Wiley & Sons, 2011.

9) Javier Palarea-Albaladejo and Josep Antoni Martín-Fernández. zcompositions — r package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems*, 143(0):85 – 96, 2015.

10) John Aitchison and Michael Greenacre. Biplots of compositional data. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 51(4):375–392, 2002.

11) Gregory B. Gloor, Jean M. Macklaim, and Andrew D. Fernandes. Displaying variation in large datasets: a visual summary of effect sizes. *Journal of Computational and Graphical Statistics*, accepted, 2015.

12) Elaine Y Hsiao, Sara W McBride, Sophia Hsien, Gil Sharon, Embriette R Hyde, Tyler McCue, Julian A Codelli, Janet Chow, Sarah E Reisman, Joseph F Petrosino, Paul H Patterson, and Sarkis K Mazmanian. Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell*, 155(7):1451–63, Dec 2013.