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:	30-Mar-2016
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 as solid black) to increase 10-fold between samples 1 and 2, while the abundance of the other 49 parts (in open circles) 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 black part still appears to become more abundant, although it is less than a 10-fold change. However, the 49 other parts appear to become less abundant. This property leads to the *negative correlation bias* observed in compositional data, and renders invalid any type of correlation or covariance based analysis such as correlation networks, 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).

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 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.

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 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 currently 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 significantly different between groups after accounting for the random sampling that occurs when the DNA is extracted and loaded onto the sequencing instrument. In either case, both ANCOM and ALDEx2 explicitly acknowledge the multivariate compositional nature of the data, and control for false positive identifications much better than do the usual approaches.

The third objective is to determine if there are taxa in the 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 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. 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 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 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 (Gloor 2016). 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 first worked example is a vaginal microbiome dataset. This 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 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 are 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,

386 and uses the assumptions of compositional data analysis to normalize the data for the
387 differing number of reads in each sample (Fernandes et al. 2013, Lovell et al. 2015).
388 When interpreting these results, it is important to remember that we are actually
389 examining ratios between values, rather than abundances. Thus, we are examining the
390 change in abundance of a taxon *relative to all others* in the dataset. The user should
391 also remember that all values reported are the means or medians over the number of
392 Dirichlet instances as given by the mc.samples variable in the aldex.clr function and
393 explained more fully in the supplementary material and the original papers (Fernandes
394 et al. 2013, 2014).
395 In the examples given in Table 1, we filtered to show only those taxa where the
396 expected Bejamini-Hochberg (1995) adjusted P value was less than 0.05, meaning that
397 the expected likelihood of a false positive identification per taxon is less than 5%, with
398 the actual value per taxon given in the wi.eBH column. Using *L. iners*, we note that the
399 absolute difference between groups can be up to −2.25. Thus, the absolute fold change
400 in the ratio between *L. iners* and all other taxa between groups for this organism is on
401 average 4.76 fold ($1/2^{-2.25}$): being more abundant in the A samples than in the B
402 samples. However, the difference within the groups (roughly equivalent to the standard
403 deviation) is even larger, giving an effect size of −0.79. Thus, the difference between
404 groups is less than the variability within a group, a result that is typical for microbiome
405 studies.
406 These quantitative results are largely congruent with the biplot, which showed
407 that the taxa represented here were the ones that best explained the variation between
408 groups, and that the *Leptotrichia* (*Sneathia*) and *Lactobacillus* taxa were not contributing

409 to the separation of the two large groups and so would not be expected to be
410 significantly different, despite being highly variable.

411 The left panel of Figure 4 shows a plot of the within (diff.win) to between (diff.btw)
412 condition differences, with the large black dots representing those that have a BH
413 adjusted P value of 0.05 or less. Taxa that are more abundant than the mean in the B
414 samples have positive y values, and those that are more abundant than the mean in the
415 A samples have negative y values. These are referred to as ‘effect size’ plots, and they
416 summarize the data in an intuitive way (Gloor et al. 2015). The grey lines represent the
417 line of equivalence for the within and between group values. Small black dots represent
418 taxa that are less abundant than the mean taxon abundance: here it is clear that the
419 abundance of rare taxa, are generally difficult to estimate with any precision.

420 The middle plot in Figure 4 shows a plot of the effect size vs. the BH adjusted P
421 value, with a strong correspondence between these two measures. In general, an effect
422 size cutoff is preferred because it is more robust than P values. The right plot in this
423 figure shows a volcano plot for reference.

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

430 In the context of microbiome datasets, the phi metric (Lovell et al. 2015) seeks to
431 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 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 with *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 agrees, it is generally not advised

454 to conduct a univariate analysis since it is likely that only false positive results would be
455 obtained (Hubert and Wainer 2012).

456 However, these authors went on to identify a number of univariate differences in
457 taxon abundance between groups using the LEfSe and Metastats tools that are
458 standard in the field (White et al. 2009, Segata et al. 2012), but that do not assume the
459 data are multivariate compositions. When examining univariate differences with the
460 ALDEx2 tool, we found that none of the univariate differences reported in the original
461 paper were supported by subsequent analysis. In particular, the authors indicated that
462 the largest differences between groups were found for six taxa labeled as 53, 145, 638,
463 836, 837, and 956 in Figure 4 of the paper. The reason for this discrepancy is that
464 inspection of the original paper reveals that raw, and not Benjamini-Hochberg adjusted
465 P values were reported. Thus it is likely that the majority, if not all, of the taxa different
466 between the control and treatment groups are false positive identifications. This result is
467 congruent with the multivariate results found in both the original paper, and by the
468 compositional biplot. Finally, in support of this assertion, we observe that all of these
469 predicted differences become insignificant following a multiple test correction using
470 either the P values reported in the paper, or P values calculated using the ALDEx2
471 software.

472 While we have been critical of the microbiome analysis methods used in this
473 paper, we must acknowledge that other published papers exhibit many of the same
474 flaws: namely an over-reliance on tools that do not treat the data as compositions, the
475 identification of extremely rare taxa as the most ‘significantly different’ taxa between
476 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 where the data are ratios between the OTU count in a sample and geometric mean for that 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 here and in the supplementary material, a book outlining how to use the compositions R package by Van den Boogaart and Tolosana-Delgado (2013) is particularly helpful, although none of the examples are drawn from the biological literature. For others wishing to understand bioinformatics and data analysis of sequencing data in general terms, hopefully this paper will prove helpful, and encourage people to enroll in specialized courses. The temptation may be to rely on proprietary third party systems, even at a cost, but the ‘devil is in the details’ and for

thoroughness we recommend developing the highest level of skill possible, especially to continue to create new analytical tools.

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

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). Work in the lab of GG is also supported by an NSERC Discovery Grant. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aitchison, J. 1986. The statistical analysis of compositional data, Chapman and Hall, London England. ISBN 1-930665-78-4

Aitchison, J and Greenacre, M. 2002. Biplots of compositional data. J. Royal Stat. Soc: Series C. 51:375-92

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.

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.

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.

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.

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.

557 Filzmoser, P., Hron, K., Reimann, C. 2009. Univariate statistical analysis of
558 environmental (compositional) data: problems and possibilities. Sci tTtal Environ.
559 407:6100-8.
560 Frémont, M., Coomans, D., Massart, S., De Meirleir, K.. 2013. High-throughput 16S
561 rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic
562 encephalomyelitis/chronic fatigue syndrome patients. Anaerobe. 22:50-6.
563 Friedman, J., Alm, E. J. 2012. Inferring correlation networks from genomic survey data.
564 PLoS Comput. Biol. 8(9): e1002687
565 Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis,
566 B. et al. 2004. Bioconductor: open software development for computational biology
567 and bioinformatics. Gen. Biol. 5 (10): R80.
568 Gloor, G.B., Macklaim, J.M., Fernandes, A.F. 2016. Displaying variation in large
569 datasets: a visual summary of effect sizes. J. Comput. Graph. Stat. (in press),
570 **DOI:**10.1080/10618600.2015.1131161.
571 Gloor, G.B., Macklaim, J.M., Vu, M, Fernandes, A.F. 2016. Compositional uncertainty
572 should not be ignored in high-throughput sequencing data analysis. Austrian Journal of
573 Statistics (in press).
574 Gloor, G.B. 2016. Compositional data analysis for high throughput sequencing: an
575 example from 16S rRNA gene sequencing. Supplementary Information at:
576 http://github.com/ggloor/CJM_Supplement. DOI:10.5281/zenodo.49579.
577 Hsiao, E. Y., McBride, S.W., Hsien, S., Sharon, G., Hyde, E.R., McCue, T., Codelli, J.A.,
578 Chow, J., Reisman, S.E., Petrosino, J.F., Patterson, P.H., Mazmanian,S.K. 2013.

579 Microbiota modulate behavioral and physiological abnormalities associated with
580 neurodevelopmental disorders. Cell. 155(7):1451-63
581 Hubert, L., Wainer, H. 2012. A statistical guide for the ethically perplexed. CRC Press,
582 London, UK.
583 Kuczynski, J., Stombaugh, J., Walters, W. A., González, A., Caporaso, J. G., Knight, R.
584 2012. Using QIIME to analyze 16S rRNA gene sequences from microbial communities.
585 Curr. Prot. Microbiol. 1E-5.
586 Kurtz, Zachary D and Müller, Christian L and Miraldi, Emily R and Littman, Dan R and
587 Blaser, Martin J and Bonneau, Richard A 2015. Sparse and compositionally robust
588 inference of microbial ecological networks. PLoS Comp. Bio. 11:e1004226
589 Kwak, D.S., Jun, D.W., Seo, J.G., Chung, W.S., Park, S.E., Lee, K.N., Khalid-Saeed,
590 W., Lee, H.L., Lee, O.Y., Yoon, B.C., Choi, H.S. 2014. Short-term probiotic therapy
591 alleviates small intestinal bacterial overgrowth, but does not improve intestinal
592 permeability in chronic liver disease. Eur J Gastroenterol Hepatol. 26(12):1353-9.
593 Lourenço, T.G., Heller, D., Silva-Boghossian, C.M., Cotton, S.L., Paster, B.J., Colombo,
594 A.P. 2014. Microbial signature profiles of periodontally healthy and diseased patients. J
595 Clin Periodontol. 41(11):1027-36.
596 Lovell, D., Pawlowsky-Glahn, V., Egozcue, J. J. Marguerat, S., Bähler, J. 2015.
597 Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol
598 11:e1004075.
599 Lozopone, C., Knight, R. 2005. Unifrac: a new phylogenetic method for comparing
600 microbial communities. Applied Env. Micro. 71:8228-8235.

601 Macklaim, J.M., Clemente, J.C., Knight, R., Gloor, G.B., Reid, G. 2015. Changes in
602 vaginal microbiota following antimicrobial and probiotic therapy. Microb Ecol Health Dis.
603 26:27799.
604 Mandal, S., Van Treuren, W., White, RA., and Eggesbø, M., Knight, R., Peddada, S. D.
605 2015. Analysis of composition of microbiomes: a novel method for studying microbial
606 composition. Microl. Ecol. Health Dis. 26:27663.
607 Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. 1998. Measures of
608 difference for compositional data and hierarchical clustering methods. In A. Buccianti,
609 G. Nardi, & R. Potenza (Eds.), Proc. IAMG (Vol. 98, pp. 526-531).
610 Palarea-Albaladejo J., Antoni Martín-Fernández, J. 2015. zCompositions --- R package
611 for multivariate imputation of left-censored data under a compositional approach.
612 Chemometrics and Intelligent Laboratory Systems. 143:85-96
613 Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, R. 2015. Modeling and
614 Analysis of Compositional Data. John Wiley & Sons. Springer. 258 pg, London, UK.
615 Pearson, K. 1896. Mathematical contributions to the theory of evolution. -- on a form of
616 spurious correlation which may arise when indices are used in the measurement of
617 organs. Proc. Royal Soc. Lond. 60:489-498
618 R Core Team 2015. R: A language and environment for statistical computing. R
619 Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
620 Rajca, S., Grondin, V., Louis, E., Vernier-Massouille, G., Grimaud, J.C., Bouhnik, Y.,
621 Laharie, D., Dupas, J.L., Pillant, H., Picon, L., Veyrac, M., Flamant, M., Savoye, G.,
622 Jian, R., Devos, M., Paintaud, G., Piver, E., Allez, M., Mary, J.Y., Sokol, H., Colombel,
623 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.

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

Schellenberg, J., Links, M. G., Hill, J. E., Dumonceaux, T. J., Peters, G. A., Tyler, S., 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.

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

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.

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

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.

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

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

White, J.R., Nagarajan, N., Pop, M. 2009. Statistical methods for detecting differentially abundant features in clinical metagenomic samples. PLoS Comput. Biol. 5:e1000352

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 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. Samples are colored in red if diagnosed as BV, blue if healthy, and purple if

intermediate. The scree plot also shows that the majority of the variability is on component 1. We can interpret this biplot with some confidence, although it is likely that any associations will be found to have large variation.

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. Large black points indicate those that have a mean Benjamini-Hochberg adjusted P-value of 0.05 or less using P values calculated with the Wilcoxon rank test. The middle plot shows a plot of the effect size vs. the adjusted P value. In general, effect size measures are more robust than are P values and are preferred. For a large sample size such as this one, an effect size of 0.5 or greater will likely correspond to biological relevance. The right plot shows a volcano plot where the difference between groups is plotted vs the adjusted P value.

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.

702

703 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

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

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

706 effect: effect size of the difference, median of diff.btw/diff.win

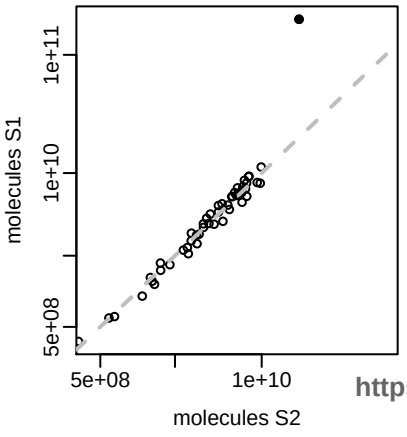
707 overlap: confusion in assigning an observation to H or BV group. Smaller is better

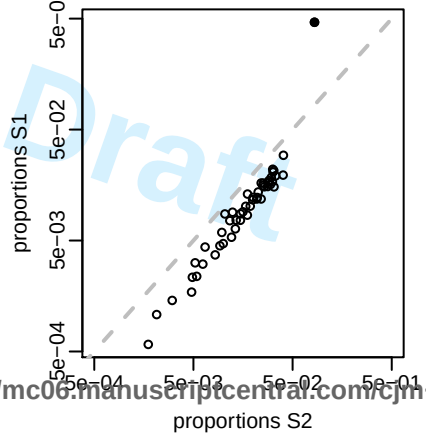
708 wi.ep: expected value of the Wilcoxon Rank Test P-value

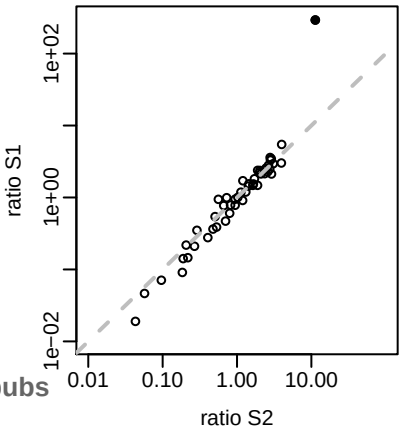
709 wi.eBH: expected value of the Benjamini-Hochberg corrected P-value

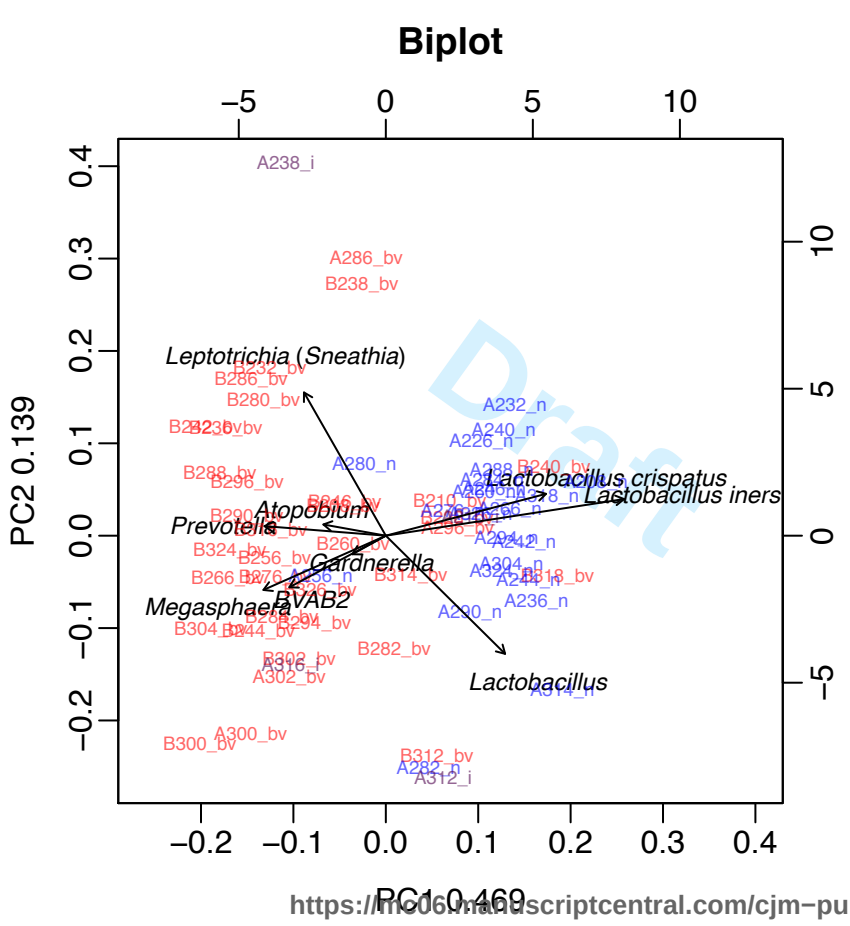
710

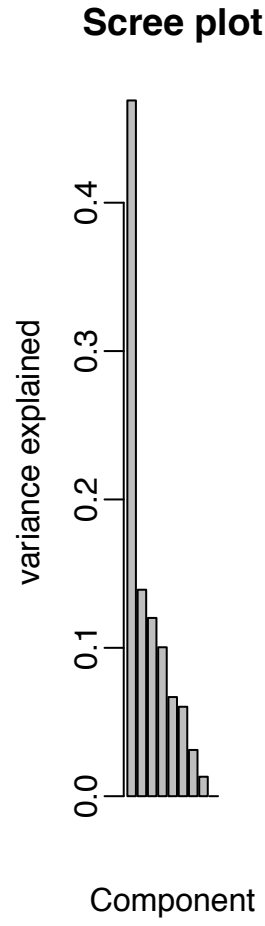
711

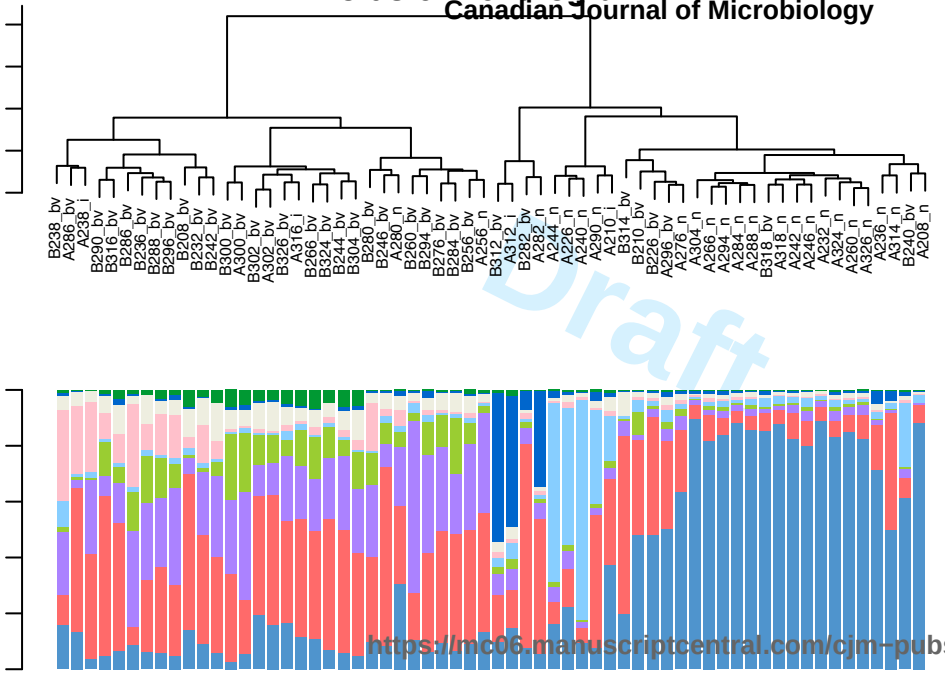


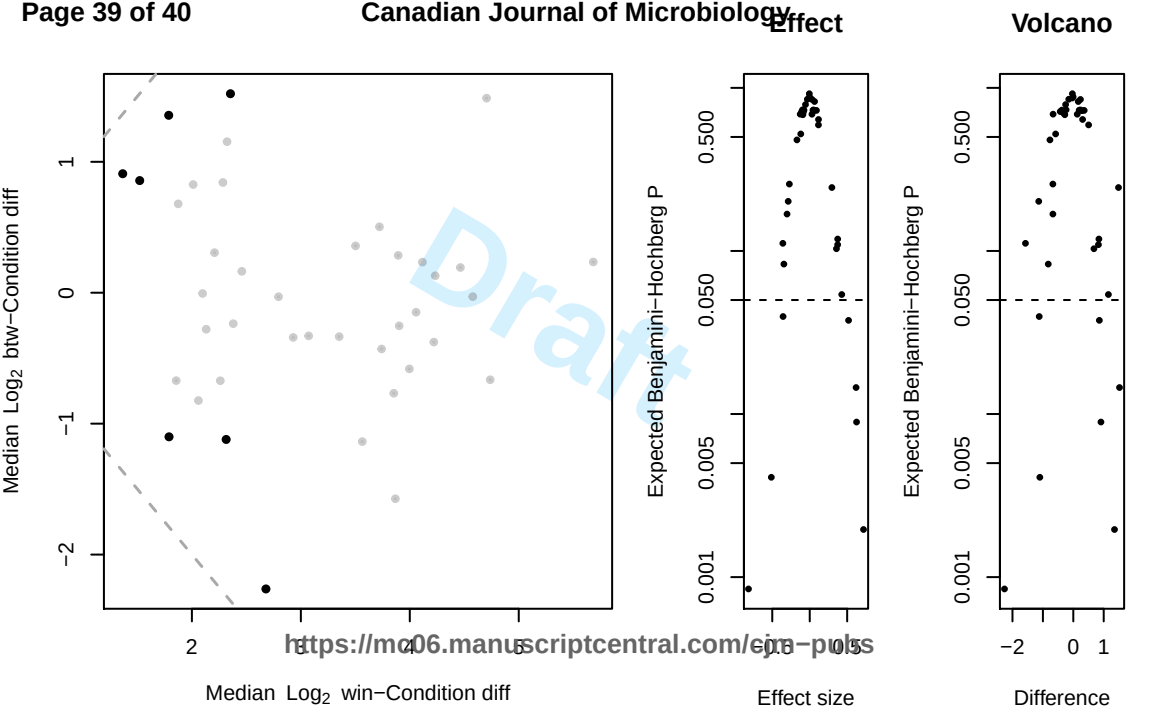


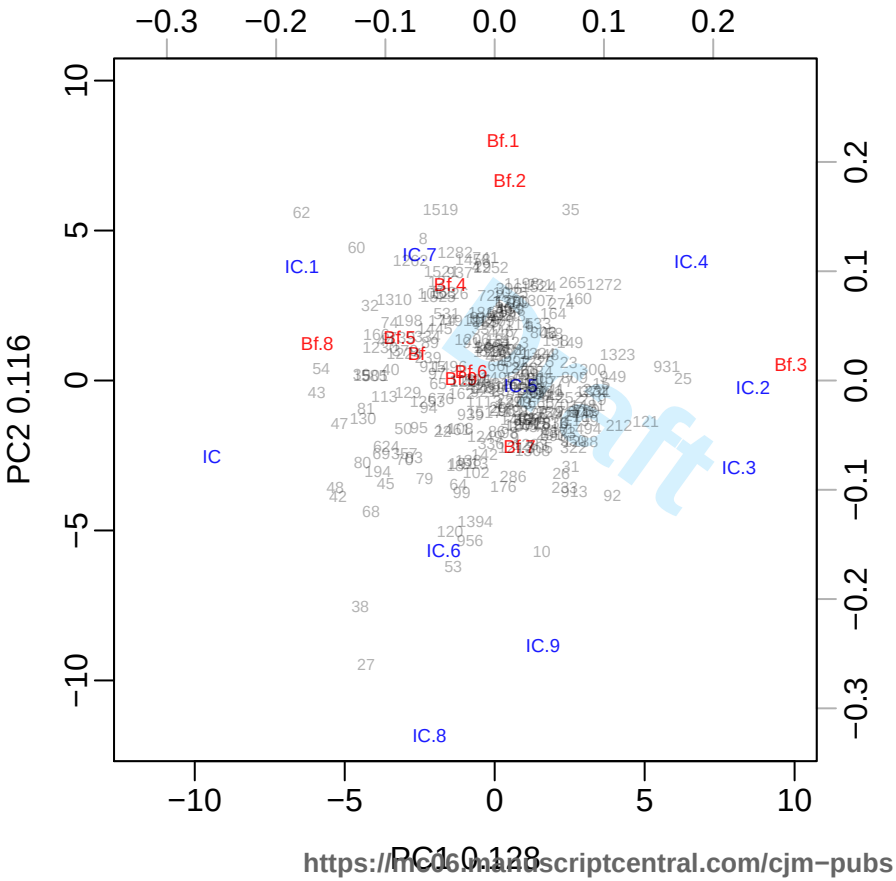


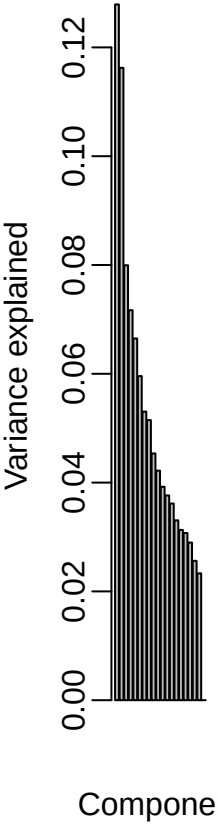

Biplot
PC2 0.139
PC1 0.469
Leptotrichia (Sneathia)
Atopobium
Prevotella
Gardnerella
BVAB2
Megasphaera
Lactobacillus crispatus
Lactobacillus iners
Lactobacillus
Draft


Scree plot
variance explained
Component






−0.3
−0.2
−0.1
0.0
0.1
0.2
10
5
0
−5
−10
PC2 0.116
0.2
0.1
0.0
−0.1
−0.2
−0.3
Bf.1
Bf.2
Bf.3
Bf.4
Bf.5
Bf.6
Bf.7
Bf.8
Bf
IC.1
IC.2
IC.3
IC.4
IC.5
IC.6
IC.7
IC.8
IC.9
IC
−10
−5
0
5
10
PC1 0.128


Variance explained
0.12
0.10
0.08
0.06
0.04
0.02
0.00
Component