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

Microbiome datasets are large and complex. They qualify as high-dimensional big data and this brings a number of methodological and analytical problems. For example, each sample may have tens to hundreds of thousands of sequence reads associated with it, and the entire dataset across all samples can easily have hundreds of millions to billions of reads. The challenge is to distill these data into a form that can be analyzed and interpreted. The goals when analyzing any microbiome dataset are very common, and there are a number of all-in-one packages that facilitate the traditional analyses, with the most commonly used being QIIME and mothur (refs).

The first goal is to determine the complexity of each individual sample using one or several alpha diversity metrics. These approaches are descriptive, but provide insights into overall composition, the number of taxa, and whether the samples are composed of an equal mix of many taxa, or are dominated by a small number of taxa (Di Bella paper).

The second goal is to determine if the samples form subgroups that are congruent with the experimental design or some other metadata. This analysis depends on a difference or distance metric. The most commonly used metric in ecology is the Bray-Curtis difference (ref), but in microbiome analysis the most common metric is most often the tree based metrics, the weighted or unweighted unifrac distances (ref). These three measures are widely used, but are sensitive to the number of taxa in the sample, the sequencing depth (ref). More recently, the weighted unifrac distance has been found to be sensitive to the topology of the tree suggesting that this metric may be more sensitive to the initial conditions than believed. Once a distance metric is chosen, it can be used to determine distances between samples, and exploratory methods such as PCoA can be used to visualize differences, or discriminatory methods such as PerMANOVA, ANOSIM or MiRKAT can be used to quantitate those differences (refs).

The third goal is to determine which taxa are driving the differences observed between groups. Here there are several methods with the most common being Metastats and LEFSe (refs). The fourth and final goal is to determine correlation between taxa in the dataset, and it is common to use Spearman’s correlations, the ReBoot or SparCC for this (refs).

However, whatever the analysis approach chosen, it is often found that results from one group contradict the results from others. For example, out of five primary research articles examining the microbiome in autism spectrum disorder, there was essentially no overlap in the taxa described (refs). Thus one priority of the field is to determine why such contradictions occur so that they can be resolved.

**Brief explanation of the data:** Data for microbiome analysis is collected by the following general workflow, and essentially every choice made affects the results (ref). First a random sample of the environment is conducted: this can be a swab or other method of sampling. Second, DNA is isolated from the sample. Third, a small amount of the DNA isolated is amplified using the PCR with primers specific to one or more variable regions of the 16S rRNA gene, or another conserved gene such as the *cpn60* gene. Fourth, a random sample of the amplified product is used to make a library, and it is common to multiplex many samples in the library. Finally, a small aliquot of the library is sequenced on a high throughput-sequencing instrument.

It is clear from this description that we are sequencing a *random sample* of the DNA in the environment. Thus, we need to ensure that any analysis takes this random component into account. Perhaps less obvious is that there number of reads obtained for a sample bears no relationship to the number of molecules of DNA isolated from the environment. This is because the number of reads obtained for a sample is determined by the capacity of the instrument: 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 if microbiome datasets. However, this constant arbitrary sum constraint generates significant unanticipated problems for analysis as described below, and is the main reason to use a compositional data approach.

**What is compositional data?** Compositional data is a term used to describe datasets where the parts in each sample have an arbitrary or non-informative sum (Aitchison 1986) such as high throughput sequencing data. These data have long been known to be problematic when analyzed using standard approaches (Pearson 1897). The essential problem is illustrated in Figure 1 where we can see the difference between counting molecules in part A, and the same data converted to relative abundances or proportions in part B. Note that in part B when the blue part increases in the sample, the 49 red parts *appear* to become less abundant, when in fact they remain at the same count in the actual sample. 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, etc. Thus the current tools used to examine the second and fourth analysis goals give results that may be inconsistent, difficult to interpret and in many cases completely wrong (Lovell 2015). In addition, we would make incorrect univariate inferences as well, concluding that the blue part increased and the red parts decreased. Thus, many published analyses could be wrong since the tools in common use for the third analysis goal do not take this property into account.

The essential feature 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 have a vector of numbers *X* that contains *D* parts:

X = [x1,x2, … xD],

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

Xclr = [log[x1/gX], log[x2/ gX], … log[xD/ gX],

where *gX* 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), and as shown in Figure 1C, reconstitutes the shape of the data so that univariate analyses are also more likely to be valid.

This approach would be ideal if microbiome datasets did not contain any 0 values, because a geometric mean cannot be computed if any of the values in the vector are 0. However, it is here that we have to remember that our date represents the counts per taxa through the process of *random sampling* (Fernandes 2014). Thus, some 0 values could arise simply by random chance, while others arise because of true absence in the environment. Fortunately, the compositional analysis approach allows us to dispose of taxa with very 0 counts in all or most samples, and to assign an estimate the likelihood of the 0 being a sampling artifact to the remainder, the mechanism to do this is presented below (Martin-Fernandez ????).

**How the analysis works**: Fortunately, there is a toolkit for compositional analysis that can be adapted for microbiome analysis (van den Boogaart 2013). These tools were developed originally for multivariate compositional data in the fields of geology and economics. In addition, one of the workshop leaders (Gloor) developed a tool called ALDEx2 for univariate analyses of microbiome and transcriptome data. The workshop familiarized the participants with the problem of compositional data, and walked the student through a compositional analysis using the R statistical programming language.

**A worked example:** The figures that follow can be reproduced by downloading and running the file **workshop.R**, which contains both the code and annotations regarding what is happening behind the scenes. 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 antibiotics, or antibiotics plus a probiotic supplement (Macklaim et.al, 2015). 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.

The **compositional biplot** is the essential workhorse tool for compositional analysis. These are principle component, or singular value decomposition plots of the data that seek to display the major axes of variance (or change) in the dataset (Aitchison and Greenacre 2002). Properly made and interpreted they summarize all the essential results of your experiment. However, the weakness of this approach is that it is descriptive and exploratory, not quantitative. Quantitative tools can be applied later to support the conclusions derived from the biplot.

For simplicity, we will filter the dataset to include only those taxa that are at least 0.1% abundant in any sample, although one of the properties of compositional data analysis is that subsets of the dataset will give 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 covariance biplot for this dataset along with the associated scree plot. The point values (labeled in red, blue or cyan) illustrate the variance between the samples, and the taxa values (represented by the black rays) illustrate the variance between the taxa. There are many interpretation rules for biplots of compositional data, but these boil down to remembering that we are examining the *ratios* between taxa, and that we can only examine links between the tips of the rays, or links between samples. Keeping this in mind, we can see the following:

First, the proportion of variance explained in the first component is very good, being 47%, and then falling to 13% on component 2, and decreasing rapidly thereafter. This indicates that the major difference between our samples can be captured in essentially one direction along component 1.

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

Third, the longest link is between *L. iners* and *Megasphaera*. Thus we can conclude that these two taxa 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 is between *Megasphaera* and *BVAB2*. Thus, we can conclude that the ratio of these two taxa is relatively constant across all samples. That is, their ratio abundance will be highly correlated.

Fifth, 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.  Likewise, this property can be extended to linear relationships containing three or more parts.

Sixth, the link between *L. iners* and *Megasphaera*, and the link between *Leptotrichia* 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 the abundance relationship between the latter two taxa.

Finally, we can extend this methodology to the samples as well. We observe that some samples (A312\_bv, B312\_bv, A282\_n), are tightly clustered, indicating that they contain similar sets of taxa at similar ratio abundances. We would expect that these samples contain an abundance of *Lactobacillus* and be depleted in *Leptotrichia*. 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.

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 (Martin Fernandez ????). Thus, here we will use 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.

The results of unsupervised clustering of the dataset are shown in Figure 3. Here we can use Euclidian distance because the Aitchison transformed data are linearly related. 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, although Ward.D and Ward.D2 are usually the most appropriate.

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.

The results of the cluster analysis can help explain and clarify the compositional biplot. For example, the four samples in the middle lower part of the biplot in Figure 8 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.

Finally, we will conduct a univariate comparison between the B and A groups. For simplicity of coding, 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 taxon abundance into a compositional framework, with the results shown in Table 1 and Figure 4. The Bayesian estimate is used to model the sampling variation that is expected in these datasets. See the documentation and the Fernandes, 2014 paper for a full description.

When interpreting these results it is important to remember that we are actually examining ratios between values, rather than abundances. 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.

In the examples given in Table 1, we filtered to show only those taxa where the expected BH 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* as an example, we can see 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, note that the difference within the groups (roughly equivalent to the standard deviation) is even larger, giving an effect size of −0.79. 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.

Note that 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* and *Lactobacillus* taxa were uncorrelated with the proportional abundances of these 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 |

We can also examine these data graphically as shown in Figure 4. The left panel of this figure shows a plot of the within to between condition differences, 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 B samples have positive y values, and those that are more abundant than the mean in the A 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 10 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. The right plot in this figure shows a volcano plot for reference.

Summary:

We have shown that 16S rRNA gene sequencing datasets, and others of the same type including RNA-Seq datasets, are logically best treated as ratios because the total number of reads is uninfor- mative, and the resulting values are best interpreted as fold-changes. We showed that treating the data as ratios where the denominator is the geometric mean for a sample accurately recapitulates the shape and the error profile of the input data. We used with Dirichlet Monte-Carlo replicates coupled with the centred log-ratio transformation to show that point-estimates of statistical signifi- cance in a real dataset can substantially inflate the observed P value because of random partitioning of low count values across datasets. Finally, we provide a worked example of how to examine a published 16S rRNA gene sequencing dataset and have given guidance in how to interpret the results in both a multivariate and univariate way.

In essence, we argue 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 approach. 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 approaches14,14,16,18–20 . Interested readers should consult the compositional data literature, but in particular three books are useful: the original by Aitchison in 19863, a book outlining how to use the compositions R package21, and finally a comprehensive book that outlines the essential geometric problem of compositional data as it is understood at present22.

Toolkit and code as appendix