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.

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.

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. 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.

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

**How the analysis works**: Fortunately, there is a toolkit for compositional analysis that can be adapted for microbiome analysis. These tools were developed originally for multivariate compositional data in the fields of geology and economics. In addition, one of the workshop leaders (Gloor) sdeveloped 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.

Toolkit and code as appendix