Compositional analysis: a valid approach to analyze microbiome

high throughput sequencing data

Gregory B. Gloor (1,2)\*, Gregor Reid (2, 3, 4) 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. Ag Can

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

**Abstract**

A workshop held at the annual meeting of the Canadian Society of Microbiologists highlighted bioinformatics methods to analyse microbiome datasets. A summary of the material presented, methods of analysis and the importance of careful dataset analyses are presented herein. Given the growing number of studies uncovering the critical role of clusters of microbes in health and disease, and the need to understand alterations to their composition and function following intervention with fecal transplant, probiotics, diet and pharmaceutical agents, a clear understanding of bioinformatics is essential.

**Introduction**

The discipline of microbiology has had a resurgence with human microbiome studies that have shown a major link between health and disease (Fremont et al. 2013; Lourenço et al. 2014; Urbaniak et al. 2014). High throughput sequencing methodologies have made this possible, along with breakthroughs in culture techniques. The former has used approaches such as 16S rRNA gene sequencing, metagenomics, transcriptomics and meta-transcriptomics, leading to the generation of vast amounts of data that must be simplified and analyzed (Di Bella et al. 2013). Indeed, 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. Such has been the rapidity of these developments that some studies appear to have been published using methods that are at worst flawed and at best not sufficiently well utilized. The result can be papers that journals and the lay media publicize as major advances or breakthroughs, when in some cases the data are far from sufficient for such claims (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 teaching how to analyse large microbiome datasets, and increase the accuracy and dependability of the results. A summary of the workshop is hereby presented for the benefit of others interested in this area. The goals of the microbiome analysis are generally to determine (i) the complexity of an individual sample using one or several alpha diversity metrics; (ii) if the samples form subgroups that are congruent with the experimental design or some other metadata; and (iii) which taxa are driving the differences observed between groups. Participants in the workshop were encouraged to bring their own datasets and to learn how to explore and analyze their data.

**Starting out**

Such is the nature of this field that ‘beginners’ require some basic training in genetics and computational biology, with at least an understanding of the R statistical application (R Core Team) and DNA/RNA sequencing. All participants used their own laptop and some had their own datasets already acquired from 454, Illumina or other sequence runs. An assumption was made that a given sample has been appropriately collected, handled and processed. This is actually not something to be taken for granted, and much remains to be done to optimize sample collection, reduce the risk of contamination, and minimize degradation. Likewise, the primer selection can play a large role in what results emerge (Walker et al. 2015).

**Brief explanation of the data**

Data for microbiome analysis is collected by the following general workflow, and essentially every choice made affects the results (Walker et al. 2015). Once the sample (swab, stool, saliva, urine or whatever) is collected, the DNA is isolated and a small amount 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 (Schellenburg et al. 2009). Then, a random sample of the amplified product is used to make a library, and it is common to multiplex many samples in the library. A small aliquot of the library is processed on the high throughput-sequencing instrument.

It should be recognized 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, 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 (Aitchison 1986).

**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 1896). 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, and others. 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 (Friedman and Alm, 2013, Lovell et al. 2015). In addition, we would make incorrect univariate inferences as well, concluding that the blue part increased and the red parts decreased (Filmoser et al. 2009, Fernandes et al. 2013,2014). The result is that published analyses could be wrong since the tools in common use 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, however a frequent criticism is that the 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 data represents the counts per taxa 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 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). With this paradigm we 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 (Palarea-Albaladejo and Martin-Fernandez 2015)

**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 (Fernandes et al. 2013). This tool is available from Bioconductor (Gentleman et al. 2004). The workshop familiarized the participants with the problem of compositional data, and walked them 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.Rnw**, and the associated workshop.pdf. The .Rnw document contains both the code and annotations regarding what is happening behind the scenes, and the .pdf document contains the code and the resulting figures. We illustrate this by using a dataset from a clinical trial that examined the effect of treating women suffering from bacterial vaginosis with antibiotics and placebo or antibiotics plus a probiotic supplement (Macklaim et.al, 2015). For this example, we extracted only the ‘before’ (samples BXXX) and ‘after’ (AXXX) treatment samples. The samples 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 OTUs were aggregated to genus level using QIIME (Kuczynski et al. 2012), except for *Lactobacillus iners* and *Lactobacillus 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 an experiment. However, the weakness is that it is descriptive and exploratory, not quantitative. However, 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 one of the desirable 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%, 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.

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

Third, the longest link is between *L. iners* and *Megasphaera*. This means 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. 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.

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 any linear relationships containing three or more links.

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.

Some samples (A312\_bv, B312\_bv, A282\_n), were 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, and 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.

**Clustering to making conclusions from this dataset**

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 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, for example See the standard workflow in QIIME (Kuczynski et al. 2012). These metrics are much more sensitive to the makeup of the community 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.

The results of unsupervised clustering of the dataset are shown in Figure 3. It is important that the user 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 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. 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 a final step, 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 sequencing experiments, and uses the assumptions of compositional data analysis to normalize the data for sequencing effort.

The Bayesian estimate is used to model the sampling variation, and to estimate appropriate values for 0s 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 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 can 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* 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 |

**Use of graphics**

These data can also be examined and illustrated graphically as shown in Figure 4. The left panel 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 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. 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 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.

Examining the Hsiao Dataset using compositional approaches:

We include in the workshop.Rnw document a second worked example based on the data of Hsiao et al. (2013). We found that there is little evidence of difference between groups when analyzed using this approach. In addition, we found that the univariate differences reported in the original paper were not supported by a subsequent analysis with ALDEx2 which corrects for multiple hypothesis testing. In fact, 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. While it is tempting to be critical of this paper, others in the literature 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 inappropriate use of corrections for multiple hypothesis testing.

**Importance to future human microbiome studies**

With a body made up of bacteria, we need to understand why it evolved like this, how and when programing between microbes and human cells happens, and how and when we can influence it by either probiotic supplementation or nutrient or antimicrobial means of disrupting a microbiota profile. Indeed, 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 will require carefully designed studies with high quality clinical documentation, and samples that are processed using some of the methods described herein. As the toolkit 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 better 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 should consult the compositional data literature, but in particular three books: the original by Aitchison (1986) and a comprehensive book 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 a book outlining how to use the compositions R package by Van den Boogaart,and Tolosana-Delgado (2013) is particularly helpful, For others wishing to understand bioinformatics and data analysis of sequencing data in general terms, hopefully this paper is helpful, and encouragement will be given for members of their research group to enroll in specialized courses. Products, such as MinION, will emerge over the next few years to allow sequencing to be done more easily and data interpreted. The temptation may be to rely on 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 avoid making claims that are later disproven, but more importantly bring badly needed breakthroughs in prevention, treatment and cure of disease.

**Summary**

Using a 16S rRNA gene sequencing dataset, we have illustrated that these are best treated as ratios because the total number of reads is uninformative, and the resulting values are best interpreted as fold-changes. 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 centred log-ratio transformation 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 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 approaches (ref 14,16,18–20).

**Figure Legends**

**Figure 1:** The difference between counting and proportions. The left panel shows two samples composed of 49 invariant OTUs in red, and 1 OTU that changes in count 10-fold (blue). The top part of the panel represents this as a bar plot, and the bottom as a scatter plot. This is the type of data that most current analysis tools in the microbiome filed expect. The middle panel shows the same samples after they have been sequenced and so constrained to have a constant sum. With such a constraint, there 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 blue OTU still appears to be more abundant, but the red OTUs appear to have become less abundant! It is obvious that we would draw incorrect inferences regarding abundance changes in these data, yet this is the approach taken by existing tools. The third panel shows that much of this distortion can be removed using the centred log-ratio transformation. 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 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.

**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 the only distance that is robust to perturbations and sub-compositions of the data3. 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. 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 biplot of the Hsiao dataset. Here we can see that the control and experimental samples are intermingled and that the proportion of variance explained in the first component is not as compared to the other components is not as obvious as in the biplot in Figure 2.

**Figure 6:** An effect plot for the Hsiao dataset. Here we can see that all of the OTUs in this dataset are much more variable within a group than the size of the difference between groups. Thus, no OTU is significantly different between groups.

**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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24336217) Nature. 505(7484):559-63. doi: 10.1038/nature12820. Epub 2013 Dec 11.PMID:24336217.

Di Bella, J.M., Bao, Y., Gloor, G.B., Burton, J.P., Reid, G. 2013. [High throughput sequencing methods and analysis for microbiome research.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24029734) J Microbiol Methods. 2013 Dec;95(3):401-14. doi: 10.1016/j.mimet.2013.08.011. PMID: 24029734.

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24910773) Microbiome. 2:15. doi: 10.1186/2049-2618-2-15. eCollection 2014. PMID:24910773.

Filzmoser, P., Hron, K., Reimann, C. 2009. Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci total Environ. 407:6100-8. DOI 10.1016/j.scitotenv.2009.08.008

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

Frémont, M., Coomans, D., Massart, S., De Meirleir, K.. [High-throughput 16S rRNA gene sequencing reveals alterations of intestinal microbiota in myalgic encephalomyelitis/chronic fatigue syndrome patients.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/23791918) Anaerobe. 2013 Aug;22:50-6. doi: 10.1016/j.anaerobe.2013.06.002. Epub 2013 Jun 19. PMID:23791918.

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.

Gloor, G.B., Macklaim, J.M., Fernandes, A.F. 2016. Displaying variation in large datasets: a visual summary of effect sizes. Journal of Computational and Graphical Statistics (in press)

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24315484) Cell. 155(7):1451-63. doi: 10.1016/j.cell.2013.11.024. Epub 2013 Dec 5. PMID: 24315484

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

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/25244414) Eur J Gastroenterol Hepatol. 26(12):1353-9. doi: 10.1097/MEG.0000000000000214. PMID: 25244414.

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.

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/25139407) J Clin Periodontol. 41(11):1027-36. doi: 10.1111/jcpe.12302. Epub 2014 Sep 22.PMID:25139407

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. DOI 10.1371/journal.pcbi.1004075

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).

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

Macklaim, J.M., Clemente, J.C., Knight, R., Gloor, G.B., Reid, G. 2015. [Changes in vaginal microbiota following antimicrobial and probiotic therapy.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/26282697) Microb Ecol Health Dis. 26:27799. doi: 10.3402/mehd.v26.27799. eCollection 2015. PMID: 26282697.

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

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

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/.

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24788220) Inflamm Bowel Dis. 20(6):978-86. doi: 10.1097/MIB.0000000000000036. PMID:24788220.

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 **Doi** 10.1128/AEM.01640-08

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24939885) MBio. 5(3):e00893-14. doi: 10.1128/mBio.00893-14.PMID:24939885

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.](http://www-ncbi-nlm-nih-gov.proxy1.lib.uwo.ca/pubmed/24610844) Appl Environ Microbiol. 80(10):3007-14. doi: 10.1128/AEM.00242-14. Epub 2014 Mar 7. PMID: 24610844.

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

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 Doi 10.1186/s40168-015-0087-4