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

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

December 8, 2015

Contents

1 What is this?

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

2 Compositional data analysis: more formal statement.

A dataset is defined as compositional if it contains D multiple parts, where each part is non-negative, and the sum of the parts is known (Aitchison 1986, pg 25). A composition containing D parts where the sum is 1 can be formally stated as: $C_D = \{(x_1, x_2, x_3, \ldots x_D); x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, \ldots x_D \geq 0; \sum_{x=1}^{D} = 1\}$. The sum of the parts is usually set to 1 or 100, but can take any value; i.e., any composition can be scaled to any arbitrary sum such as a ppm. It is important to know that the values of the parts of compositional datasets are constrained because of the constant sum.

The constant sum constraint causes the parts to have a negative correlation bias since an increase in the value of one part must be offset by a decrease in value of one or more other parts. Thus any correlation-based analysis is invalid in these datasets, as originally noted by Pearson[1]. In addition, compositional datasets have the property that they are described by $D-1$ observations if the sum of the parts is known[2]. In other words, if we know that all parts sum to 1, then the last part can be known by subtracting the sum of all other parts from 1, i.e., $x_D = 1 - \sum_{x=1}^{D-1}$. Graphically, this means that compositions inhabit a space called the Aitchison simplex that contains 1 fewer dimensions than the number of parts. The distances between parts on the Aitchison simplex are not linear, especially at the boundaries (see Figure **??**). This is important because all common statistical tests assume a that differences between parts are linear (or additive). Thus, while standard tests will produce output, the output will be misleading because distances on the simplex are non-linear and bounded[3].

2.0.1 Sub-compositions:

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

2.0.2 Spurious correlations:

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

Practically speaking this means that *every microbial correlation network that has ever been published is suspect* unless it was determined using SPARCC[4], a tool that at least partially accounts for this spurious correlation. Lovell is in the process of producing an R package for the compositionally appropriate examination of correlations (personal communication).

Atichison[2], Pawlsky-Glahn[5], and Egozcue[6], have done much work to develop rigorous approaches to analyze compositional data[7]. The essential step is to reduce the data to ratios between the D parts as outlined above. This step moves the data from the Aitchison simplex and to the more familiar Euclidian space where the distances between parts are linear. However, the investigator must keep in mind that the distances are between ratios, not between counts. Several transformations are in common use, but the one most applicable to HTS data is the centred log-ratio transformation or clr, where the data in each sample is transformed by taking the logarithm of the the ratio between the count value for each part and the geometric mean count: i.e., for D features in sample X,

$clr[x_1, x_2, x_3, \ldots x_D] = [log_2(x_1/gX), log_2(x_2/gX), log_2(x_3/gX) \ldots log_2(x_D/gX)]$ where gX is the geometric mean of the features in sample X. This is the transformation described above.

3 So how can I analyze compositional data?

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

3.1 An introduction to the compositional biplot

The compositional biplot is the essential workhorse tool for compositional analysis. Properly made and interpreted it summarizes all the essential results of your experiment. However, the weakness of this approach is that it is descriptive and exploratory, not quantitative. Quantitative tools can be applied later to support the conclusions derived from the biplot.

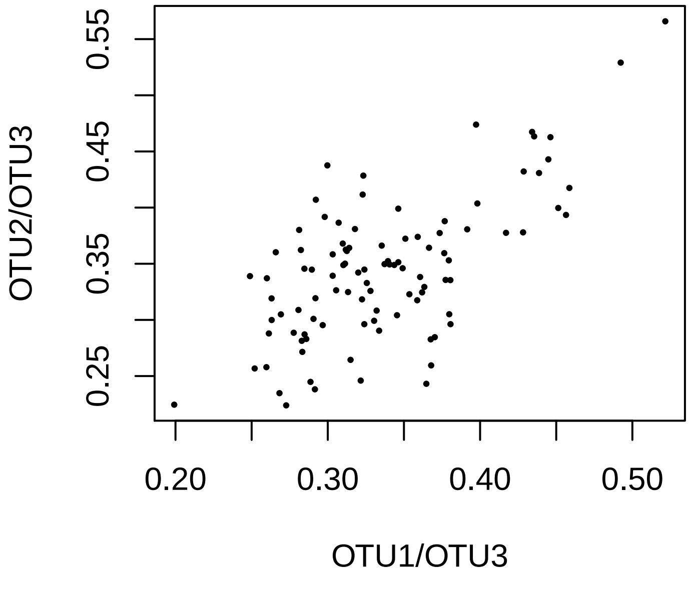


Figure 1: Spurious correlation in compositional data. Two random vectors drawn from a Normal distribution, were divided by a third vector also drawn at random from a Normal distribution. The two vectors have nothing in common, they should exhibit no correlation, and yes they exhibit a correlation coefficent of > 0.65 when divided by the third vector. See the introductory section of the Supplementary Information of Lovell[?] for a more complete description of this phenomenon.

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

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

```
# load the required R packages
require(compositions) # exploratory data analysis of compositional data
require(zCompositions) # used for 0 substitution
require(ALDEx2) # used for per-OTU comparisons
require(xtable) # used to generate tables from datasets

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

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

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

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

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

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

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

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

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

```
# convert to proportions by sample (columns) using the apply function
```

```
d.pro.prop <- apply(d.pro, 2, function(x){x/sum(x)})

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

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

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

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

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

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

# make our compositional dataset
d.acomp.abund <- acomp(t(d.pro.abund))

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

The first key function here is the 0 replacement function `cmultRepl` which has many options[8]. The bottom line is that the replacement of 0 values in these datasets is an area of ongoing research, and so there is no general way to treat 0 values in these datasets. The reader is encouraged to try different 0 replacement values and strategies and observe how it affects the conclusions.

The second key function is the `acomp` function. This makes an `R S3` class dataset from the 0 replaced count dataset. This is a dataset where the counts are treated as centre log-ratio values for all subsequent operations.

Compositional biplots show both the amount of variance of both samples and variables (taxa, shown with rays)[9]. If substantial variation is explained by the first two principle components, then the following rules can be used to examine the data:

1. the rays in this plot show the amount of variance exhibited by each taxon relative to the centre of the dataset where longer rays mean more variation across all samples.

2. the location of the sample name shows how variable it is relative to other samples.

3. samples that are highly variable, and that are in the same direction as a long ray for a taxon will contain that taxon in high abundance . The inverse is also true.

4. taxa where the tips of the rays are co-incident and of the same length indicate that the ratio between those two taxa are nearly identical across all samples.

5. taxa where the angles between the rays are orthogonal are uncorrelated.

6. taxa where the tips of the rays are very distant from each other, regardless of whether the link between the tips passes through the origin, indicated highly variable ratios across the samples.

7. three or more taxa lying on a common link will be positively or negatively correlated.

8. the angle between links contains information about correlations between pairs, or groups of ratios, more formally, the cosine of the angle is proportional the correlation between the pairs of ratios.

Let's generate a biplot of those taxa that are more abundant than 0.1% in any sample:

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

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

1. *L. iners* has the longest ray, and among these taxa is the most variable organism across samples, *Gardnerella* has the shortest ray and so is the least variable.

2. The sample A238_i is the sample that is least similar to any other sample because it is furthest from the centre (top left corner). It likely contains a substantial fraction of *Sneathia* and little *Megasphaera* and *BVAB2*.

3. The sample A238_i will contain a substantial proportion of *Sneathia* and a very small amount of *Lactobacillus* because the rays for these taxa are parallel to a ray that would connect this sample to the origin. The converse will be true for sample A314_n.

4. The two closest tips are for *Megasphaera* and *BVAB2*, thus the ratio between these two taxa is relatively constant across samples. Thus, each will be abundant when the other is abundant and *vice versa*.

5. The abundance of *Sneathia* and the taxa *Gardnerella*, *Megasphaera*, *BVAB2* will be uncorrelated because these rays are approximately orthogonal.

6. The link between *L. iners* and *Sneathia* (and many others is very long), indicating that the ratios between these taxa are extremely variable. That is, when *L. iners* is abundant, then others will be correspondingly rare.

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

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

```
pcx.abund <- princomp(d.acomp.abund)


conds <- data.frame(c(rep(1,length(grep("_bv", rownames(d.acomp.abund)))), rep(2, length(grep(

colnames(conds) <- "cond"

palette=palette(c(rgb(1,0,0,0.5), rgb(0,0,1,0.5), rgb(0,1,1,0.5)))

# scale = 0 is a form biplot - you scale and interpret by the arrows
# scale = 1 is a covariance biplot - you scale and interpret by the samples
layout(matrix(c(1,2),1,2, byrow=T), widths=c(6,2), height=c(6,4))

coloredBiplot(pcx.abund, col="black", cex=c(0.6, 0.7), xlabs.col=conds$cond, arrow.len=0.05, e
#biplot(pcx.abund, cex=0.7, col=c("black", "red"), arrow.len=0, scale=1, expand=0.8)

plot(pcx.abund, type="variance")
```

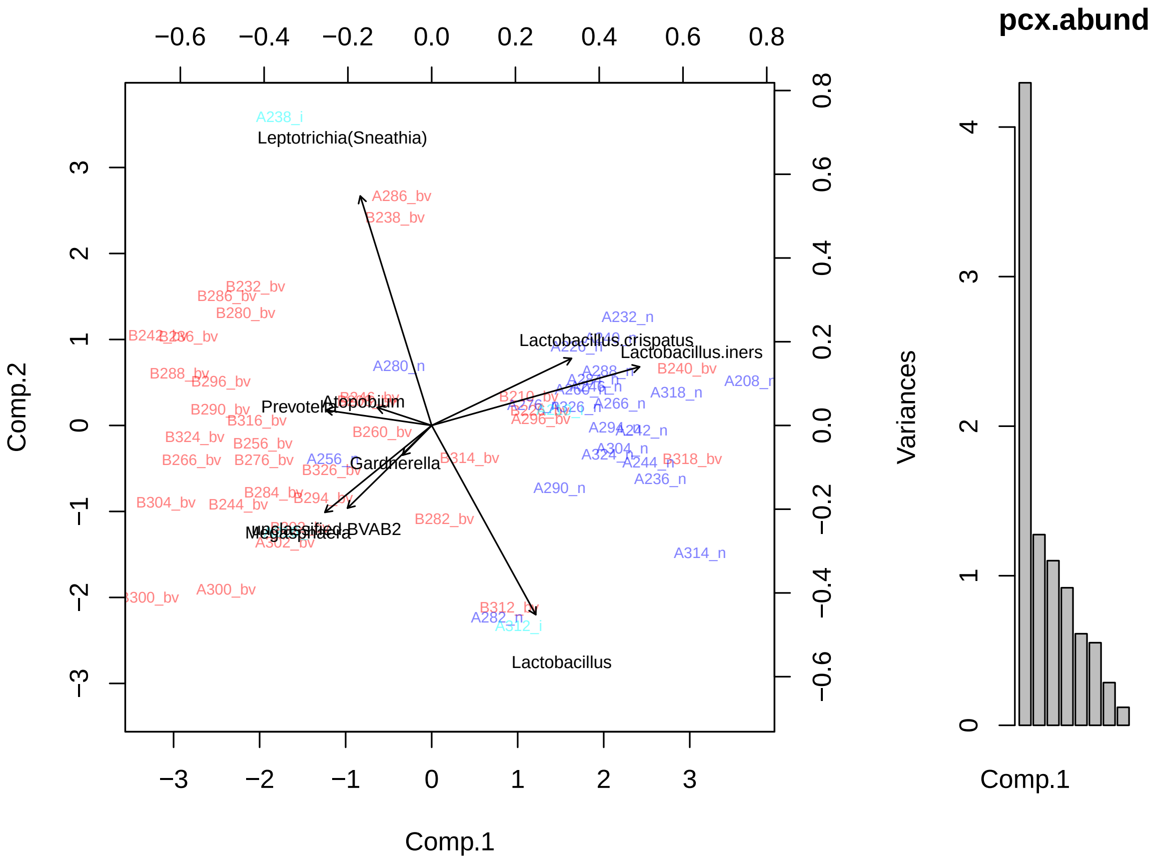


Figure 2: The left figure shows a covariance biplot of the abundance-filtered dataset, the right figure shows a scree plot of the same data. This exploratory analysis is much encouraging because the amount of variance explained is rather substantial with 0.469 of the variance being explained by component 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.

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

3.1.1 Cluster analysis

The result of the biplot suggested that there were two groups that could be defined with this set of data. With a few exceptions, there appears to be a fairly strong separation between the samples containing a majority of *Lactobacillus* sp., and those lacking them. We can explore this by performing a cluster analysis. In the traditional microbiome analysis methods, clustering is based on the weighted or unweighted unifrac distances or on the Bray-Curtis dissimilarity metric. These metrics are much more sensitive to the makeup of the community than is the Aitchison distance used in compositional data analysis. Thus, here we will use the Aitchison distance metric which fulfills the criteria required for compositional data. In particular, by using a compositional approach, it is appropriate to examine a defined sub-composition of the data.

The results of unsupervised clustering of the dataset is shown in Figure 3. Here we can use Euclidian distance because the Aitchison transformed data are linearly related and in placed in the familiar space. However, the user must remember that all distances are calculated from the ratios between taxa, and not on the taxa abundances themselves! For this figure we are using the ward.D2 method which clusters groups together by their squared distance from the geometric mean distance of the group. There are many other options, and the user should choose one that best represents the data.

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

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 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 marplot show that they contain substantially more of the *Lactobacillus* taxon, and somewhat more of the taxa normally found in BV than in the other N samples. Based on these two results it would be appropriate to exclude these four samples from further analysis because of their atypical makeup.

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

3.2 Univariate differences between groups

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

```r
# generate the dataset
d.B <- colnames(d.pro.0)[grep("B", colnames(d.pro.0))]
```

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

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

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

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

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

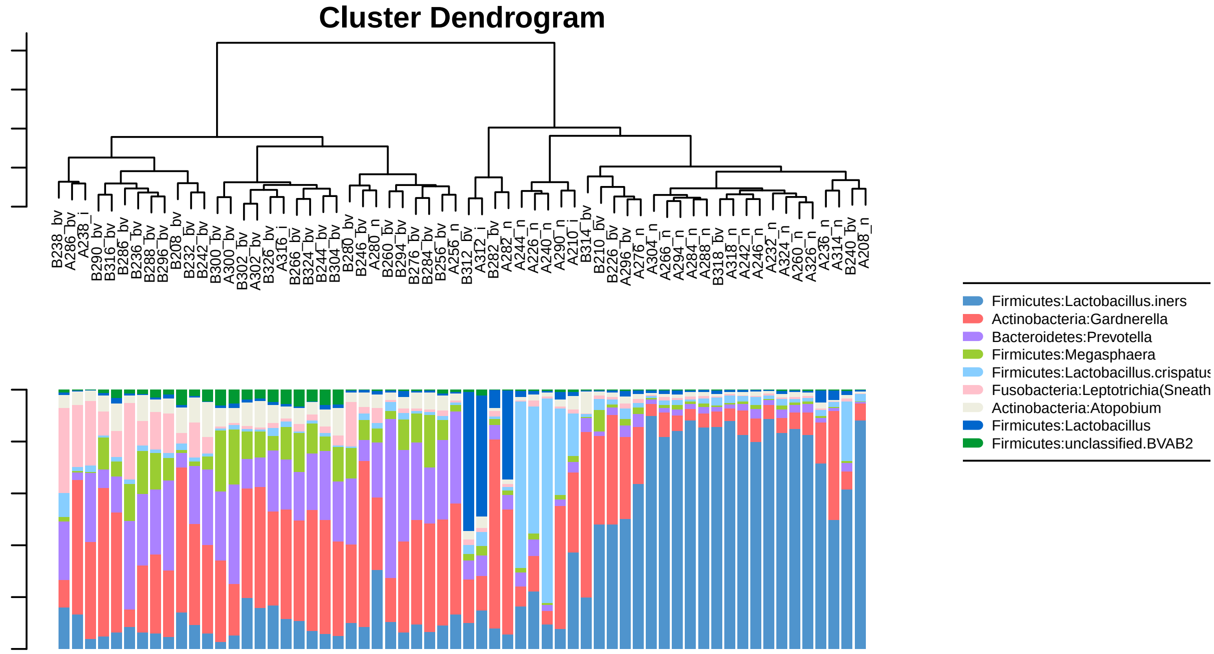


Figure 3: Unsupervised clustering of the reduced dataset. The top figure shows a dendrogram of relatedness generated by unsupervised clustering of the Aitchison distances, which is the only distance that is robust to perturbations and sub-compositions of the data[3]. The bottom figure shows a stacked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

```r
d.A <- colnames(d.pro.0)[grep("A", colnames(d.pro.0))]
d.aldex <- data.frame(d.pro.0[,d.B], d.pro.0[,d.A])

# make the list of set membership
conds <- c(rep("Be", 31), rep("Af", 31))

x <- aldex.clr(d.aldex, mc.samples=256, verbose=FALSE)
```

```
## [1] "operating in serial mode"
```

```r
x.t <- aldex.ttest(x, conds)
x.e <- aldex.effect(x, conds, verbose=FALSE)
```

```
## [1] "operating in serial mode"
```

```r
x.all <- data.frame(x.e,x.t)
```

The ALDEx2 tool estimates the distribution of taxon abundance by sampling from a Dirichlet distribution with the results outlined in Figure ??. This takes the original input data, and generates a distribution of posterior probabilities of observing each taxon. This distribution is transformed by the centred log-ratio transformation, and is used to conduct univariate statistical tests. These tests return a distribution of P and Benjamini-Hochberg adjusted P values, and the tool reports the mean of these distributions. In this way, we account for the large variation in these datasets, and identify only those taxa whose difference between the groups is robust to sampling variation.

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

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

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

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

In the examples given in Table 2, we filtered to print only those taxa where the expected BH values

	diff.btw	diff.win	effect	overlap	wi.ep	wi.eBH
Actinobacteria:Atopobium	0.884	1.518	0.535	0.298	0.007	0.035
Bacteroidetes:Prevotella	1.393	1.775	0.745	0.214	0.000	0.002
Firmicutes:Lactobacillus.crispatus	-1.054	1.780	-0.483	0.240	0.000	0.005
Firmicutes:Lactobacillus.iners	-2.276	2.705	-0.814	0.201	0.000	0.001
Firmicutes:Streptococcus	-1.126	2.378	-0.365	0.306	0.008	0.042
Firmicutes:Dialister	0.900	1.391	0.574	0.257	0.001	0.009
Firmicutes:Megasphaera	1.588	2.353	0.621	0.271	0.002	0.014

Table 1: Table of significant taxa

was less than 0.05, meaning that the expected likelihood of a false positive identification *per taxon* is less than that threshold. Using *L. iners* as an example, we can see that the absolute difference between groups can be up to -2.28, implying that the absolute fold change in the ratio between *L. iners* and all other taxa between groups for this organism is on average 4.84 fold across samples. However, note that the difference within is even larger, giving an effect size of -0.81. Thus, we can see that the difference between groups is less than the variability within a group, a result that is typical for microbiome studies.

We can examine these data graphically as shown in Figure 4. The left panel of this figure shows a plot of the within to between condition differences[10], with the red dots representing those that have a BH adjusted P value of 0.05 or less. Taxa that that are more abundant than the mean in the BV samples have positive y values, and those that are more abundant than the mean in the N samples have negative y values. We refer to these as 'effect size' plots, and they summarize the data in an intuitive way. The grey lines represent the line of equivalence for the within and between group values. Black dots are taxa that are less abundant than the mean taxon abundance: here it is clear that the abundance of these taxa, in general, are difficult to estimate with any precision.

The middle plot in Figure 4 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.

4 Examining the Hsiao et al. dataset

Hsiao et al. (2013)[11] conducted a study that examined the effect of *Bacillus fragile* treatment in a mouse model of autism and concluded that there was a difference in the gut microbiota between *b. fragile* treated and control groups. This data is publicly available, and here we will analyze these two groups to determine if the conclusions are supported when we incorporate sampling variation, and use compositional analysis methods. The original dataset contained 1474 taxa, of which only 703 were found to occur with a count of 5 or more in the 10 samples under consideration. We will use this reduced dataset for analysis because it contains all the taxa identified as significantly different in the original analysis. Samples are labeled Bf if treated, and IC if control.

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

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

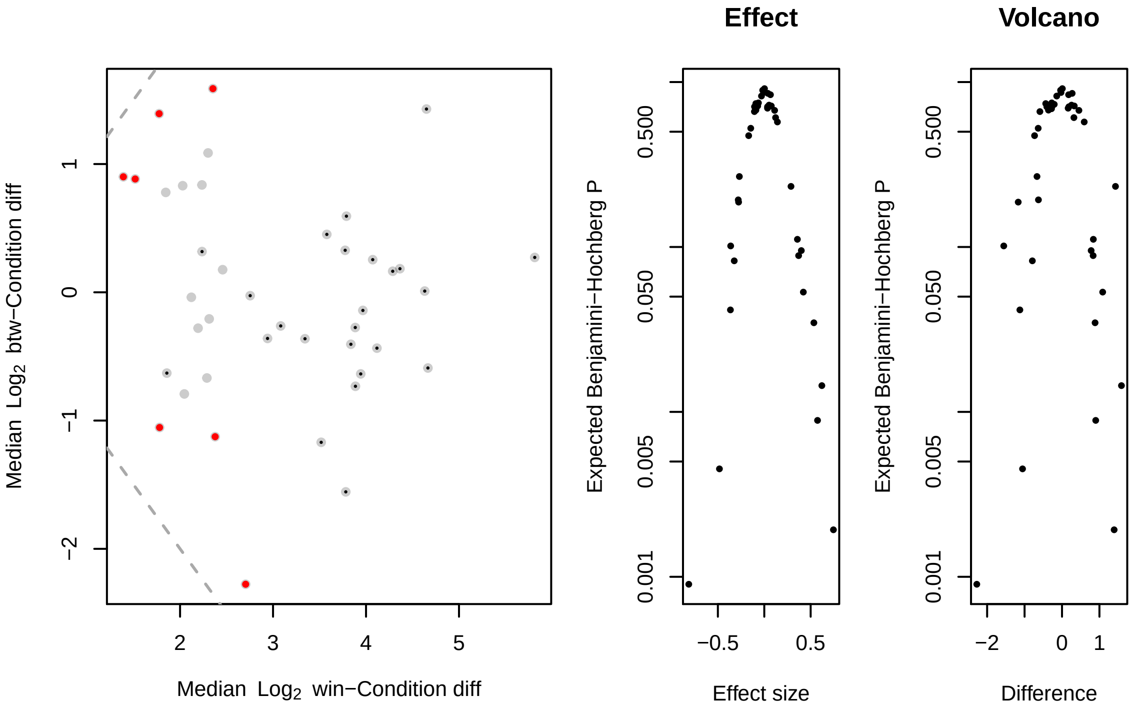


Figure 4: Examination of univariate differences between groups. The left plot shows a plot of the maximum variance within the B or A group vs. the difference between groups. Red points indicate those that have a mean Benjamini-Hochberg adjusted P-value of 0.05 or less using P values calculated with the Wilcoxon rank test. The middle plot shows a plot of the effect size vs. the adjusted P value. In general, effect size measures are more robust than are P values and are preferred. For a large sample size such as this one, an effect size of 0.5 or greater will likely correspond to biological relevance. The right plot shows a volcano plot where the difference between groups is plotted vs the adjusted P value.

```
## Warning in file(file, "rt"):  cannot open file 'tax.txt':  No such file or directory
```

```
## Error in file(file, "rt"):  cannot open the connection
```

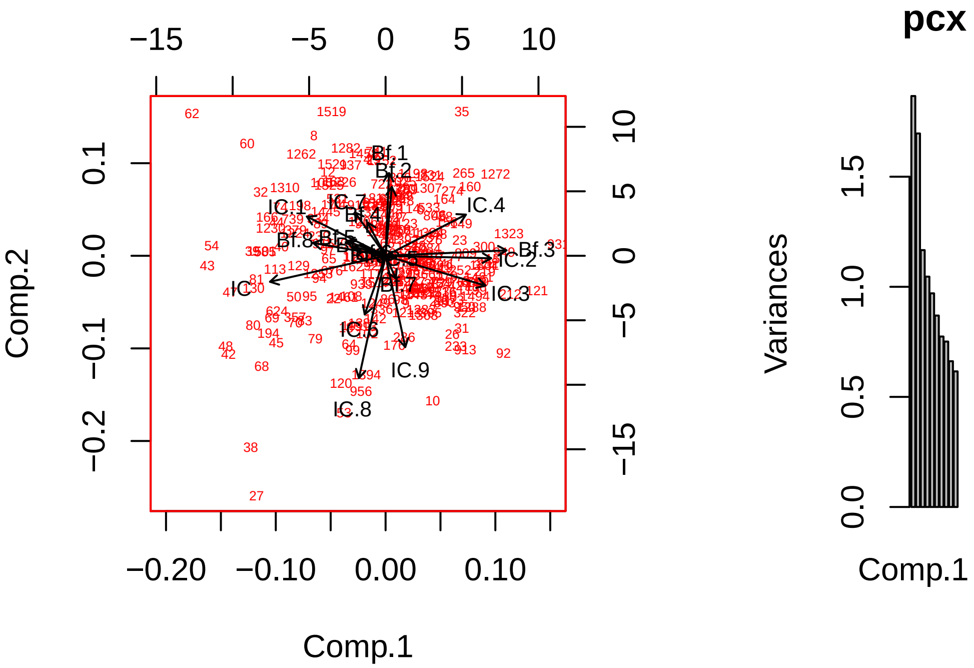
```
colnames(d) <- gsub("PolyIC...", "", colnames(d))
colnames(d) <- gsub("Poly", "", colnames(d))

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

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

```
## Error in eval(expr, envir, enclos):  object 'tax.d' not found
```

```
# replace 0 values with imputed value
d.bf <-cmultRepl(t(d.0),  label=0)
```

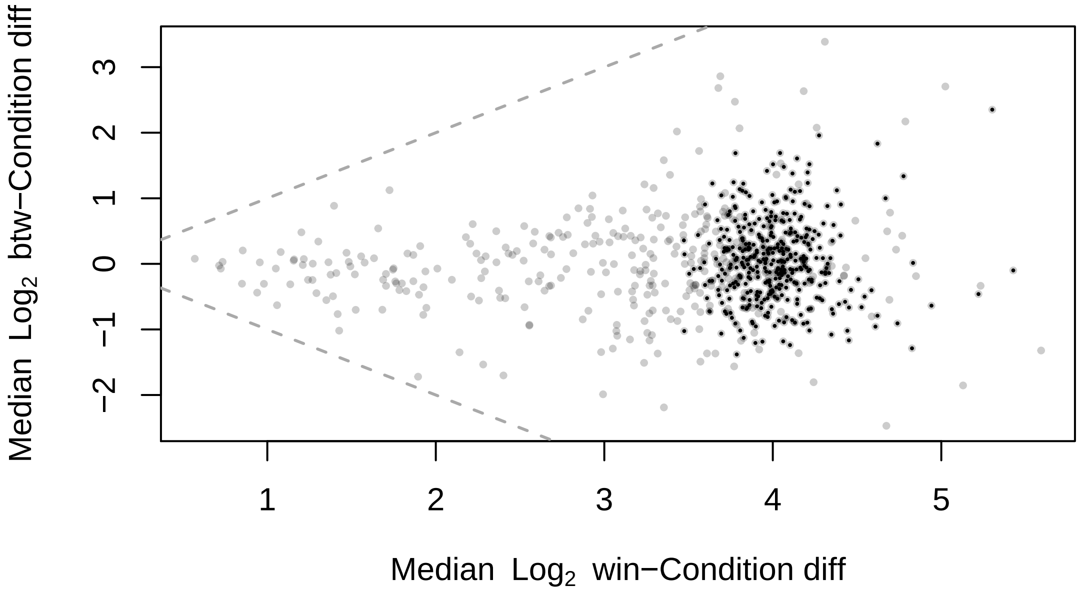
```
## No. corrected values:  107
```

```
# convert to an acomp object (clr transformed data)
bi <- acomp(d.bf)

# here we have to do the PCA on the rotated dataset because we only have a few samples
# arrows will represent samples, not taxa
# however because we care about the links between taxa, and the arrows don't contain any real
pcx <- princomp(t(bi))


# scale = 0 is a form biplot - you scale and interpret by the arrows
# scale = 1 is a covariance biplot - you scale and interpret by the samples
layout(matrix(c(1,2),1,2, byrow=T), widths=c(6,2), height=c(6,4))

# here we use a form biplot to scale by 1
coloredBiplot(pcx, col=c("red", "black"), cex=c(0.4, 0.7), arrow.len=0.05, expand=0.8,var.axes=

plot(pcx, type="variance")
```



Here we see some problems. First, we needed to calculate our PCA using the rotated dataset because we have fewer samples than we have OTUs. We should get the same explanatory power, but this is a first hint that we are not strictly 'legal' with respect to the assumptions of normal multivariate statistics. Second, while the scree plot shows that the first two components contain more information than the remainder of the components, it is not nearly as explanatory as for the BV dataset above. In fact, the first three components explain only 0.009, 0.008, and 0.006 of the proportion of variability of the data, a very low amount, suggesting that these data are not particularly informative. In the original paper, the authors found a difference with unweighed unfired, suggesting that rare taxa were important. In addition, it is clear that the BF and IC samples are intermixed, and do not separate. So perhaps, the compositional biplot is not displaying the contribution of the taxa that were found to be different in the original experiment.

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

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

## [1] "operating in serial mode"

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

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

## [1] "operating in serial mode"
## [1] "sanity check complete"
## [1] "rab.all  complete"
## [1] "rab.win  complete"
## [1] "rab of samples complete"
## [1] "within sample difference calculated"
## [1] "between group difference calculated"
## [1] "group summaries calculated"
## [1] "effect size calculated"
## [1] "summarizing output"

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

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



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

We can address this issue by examining the data in a univariate way using ALDEx2 that will determine if there are differences in abundance of rare taxa. The plot shows that there are no significant hits (i.e., no red dots), and that the dataset displays extremely high variability, likely

	wi.eBH	rab.all	rab.win.Bf	rab.win.C	effect	overlap
1288	0.755	3.475	2.950	3.998	0.583	0.230
53	0.722	1.412	0.260	3.022	0.665	0.220
533	0.755	3.303	3.808	2.743	-0.686	0.228
145	0.708	0.113	-1.282	2.029	0.654	0.203
64	0.773	1.302	0.278	2.730	0.538	0.231
4	0.805	2.802	3.281	1.966	-0.527	0.251
602	0.710	2.243	3.079	1.405	-0.593	0.197
26	0.799	4.876	4.411	5.187	0.586	0.242
78	0.762	0.537	-0.625	1.910	0.588	0.218
1262	0.752	1.464	2.360	0.185	-0.618	0.236
956	0.694	0.750	-0.576	2.270	0.700	0.194
36	0.740	0.480	1.487	-0.658	-0.575	0.216
1248	0.808	5.783	6.126	5.449	-0.515	0.242
8	0.732	5.483	6.102	4.210	-0.806	0.228
1519	0.765	3.860	4.661	3.018	-0.664	0.225

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

because of low OTU counts. We can examine the OTUs that pass a P value cutoff of 0.1 (to be generous), and display their summary data in a table as before. Here we see that there were some 'significant' OTU's, but that none of these reach significance when the P values are adjusted for multiple test corrections (wi.eBH). Inspection of the original paper indicates that a multiple hypothesis test correction *was not* done on the reported P values. Thus, both the methods in the paper and ALDEx2 identified a small number of 'significant' OTUs, but the multiple test correction indicates that these are almost certainly false positive identifications. In the end, we should conclude that the treated and untreated samples were not different by either the multivariate or univariate criteria.

References

1) Karl Pearson. Mathematical contributions to the theory of evolution. – on a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60:489–498, 1897.

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

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

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

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

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

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

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

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

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

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