Why many high throughput sequencing experiments are irreproducible: an example from 16S rRNA gene sequencing.

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

June 16, 2015

Contents

1 Abstract **1**

2 Introduction **2**

3 Results and Discussion **3**
3.1 Sequencing technologies randomly sample a pool of input DNA. 3
3.2 Data are multivariate and, as a minimum, must be corrected for multiple hypothesis tests. 5
3.3 Sequencing can change the shape of the data: 5
3.4 Commonly used transformations are misleading 7
3.5 A count of 0 does not mean that you expect 0! 8
3.6 The uncertainty is relative 11
3.7 Scaling the data 12
3.8 The power of compositional data analysis: more formal statement. 12
3.8.1 Sub-compositions: 13
3.8.2 Spurious correlations: 13

4 Conclusions **15**

1 Abstract

Fundamentally, many high throughput sequencing approaches generate similar data: samples contain many different features, the count of reads-per-feature is tabulated for each sample, these counts are normalized, and finally the samples in each group are compared in some way. The standard statistical tools used to analyze RNA-seq, ChIP-seq, 16S rRNA gene sequencing, metagenomics, etc. are fundamentally different for each approach despite the underlying similarity in the data because the methods were developed in isolation for each experimental design, and often do not take into account the multivariate nature of the data. Here we show, using an example from 16S rRNA gene sequencing, that the approaches taken towards the analysis of these datasets suffers from several common pitfalls and that widely used tools that treat the data as point estimates of the counts can lead to very misleading inferences.

2 Introduction

The human microbiome project has initiated the large-scale culture-independent analysis of microbial communities. However, many studies fail to replicate earlier studies even when the same technologies and strategies are used. For example, a multitude of studies have examined the link between autism and the human gut microbiota. These have variously implicated x,y,...,and z microbe as being linked to the condition. In a recent high-profile paper[1], *Bacillus fragilis* was suggested to restore some taxonomic groups to the gut microbiome of a mouse autism model. However, examination of the dataset shows that the conclusion was likely due to chance alone. While the autism dataset serves as a facile example, the literature are replete with other examples.

All 16S rRNA gene sequencing datasets share a common origin. A microbial population is sampled, and DNA is isolated. One or more rRNA gene variable regions are PCR amplified using primers specific for the flanking constant regions. An aliquot of the resulting mixture of DNA fragments are then on a particular platform, generating hundreds of thousands to hundreds of millions of sequences that represent random samples of the PCR-amplified mixture.

It is important to remember that the number of sequences obtained after sequencing contains no information about the *number* of sequences in the PCR amplified pool, nor does it contain information about the *number* of molecules in the original DNA sample that was amplified. Instead the only information available is the *relative* proportion of individual sequences in the PCR amplified mixture, which is assumed to approximate the proportion of sequences in the input DNA sample. Furthermore, the number of sequences obtained for a particular sample is determined entirely by the capacity of the sequencing platform. Thus, the outputs of the sequencing event are reads per operational taxonomic unit (OTU) per sample. The values across samples are often normalized by subsampling (rarefaction) or by converting to proportions or percentages; these latter values are widely spoken of in the literature as 'relative abundances'. Subsampling is frequently used to estimate the associated sampling error. Some groups have begun advocating the use of normalization methods prevalent in the RNA-seq field[2], but still treat the data as point estimates of the true abundance. There are three main data analysis issues that must be acknowledged.

First, the nature of these data are misunderstood. As outlined above, the number of counts observed per OTU are determined entirely by the capacity of the instrument and provide no information about the number of molecules in the input sample. Recall that both bacterial growth, and PCR are doubling processes and not linear processes, and so would be better modelled as $\log_2$ differences.

Understanding that we are dealing with fold-change data is an explicit acknowledgement that the data do not map to normal Euclidian space where differences are linear. Commonly used statistical tests expect linear differences between values and so are compromised to some degree, often catastrophically[3,4]. Therefore, the often-used approaches of converting the OTU count values to proportions or percentages and conducting statistical tests on those values, or of using data reduction strategies such as Principle Component Analysis on the values are inappropriate because the differences between values are not linear. An alternative approach is to convert the OTU counts to ratios[3,5–7] which makes the differences between values linear, and so allows the use of common statistical tests.

Second, high throughput sequencing (HTS) data represent samples of an unknown underlying large number of molecules. Thus, there is a large and unappreciated error of estimation that is problematic when dealing with these data[8]. The high error of estimation often results in false positive

identification of differences, in fact, the statistical result can often be explained entirely by sampling variation. This error is not captured by rarefaction or even acknowledged by other normalization methods and should be estimated and accounted for when deciding what is a significant difference.

Third, 16S rRNA gene sequencing surveys, and similar experiments, contain many variables in each sample. Thus, any analysis that attempts to characterize the individual differences between groups must correct for the many hypotheses that are being tested. This step is often ignored, even in work published in very high profile journals subject to rigorous peer review.

The purpose of these notes is to show why HTS data for 16S rRNA gene sequencing, and any similar experiments such as RNA-seq, should be treated as ratio data and that it is possible to do so simply. We show that this approach accurately recapitulates the shape of the data for both constrained and unconstrained datasets. We show that converting the data to ratios can accurately model the very high variability at the low count margins, and that rarefaction under-estimates this variability. We use an example from the literature to show how ignoring these factors leads to improper conclusions.

3 Results and Discussion

3.1 Sequencing technologies randomly sample a pool of input DNA.

When generating a high throughput sequencing datseta it is important to understand the source of the data. In the case of 16S rRNA gene sequencing surveys, DNA is isolated and subsequently amplified using primers specific for constant regions that flank one or more 16S rRNA gene variable regions. A portion of the amplified DNA is then taken and used to generate a library, and only a portion of the library is sequenced. This workflow can be thought of as three random samples from an urn containing a random assortment of molecules. The first random sample is the sample from the environment itself that is used to make the DNA: what is swabbed, and how much is collected relative to the environment. Depending on the environment, this can be a large and complete sample or a small, incomplete and unrepresentative sample. The second random sample is the DNA molecules that are input into the initial PCR reaction. Here, the number of molecules available depends on the initial concentration of the DNA and the mean genome size of the organisms composing the sample. The third random sample comprise the DNA molecules from the library that are actually sequenced on the machine.

As an example, if the mean genome size is 4 Mb, then 1 ng of DNA would provide approximately 1×10^6 amplification templates if we assume 3-4 rRNA loci per genome. These templates would be amplified up to 2^{25} fold by PCR amplification, but only between 1×10^6 and 2×10^8 sequences are generated on the machine. These are apportioned across dozens or hundreds of samples giving typically 10000 - 100000 reads per sample.

In the case of RNA-seq, the first random sample is the sample from the environment, the second is the mRNA placed into the cDNA reaction, the third random sample is the fraction of the cDNA that is used for the library.

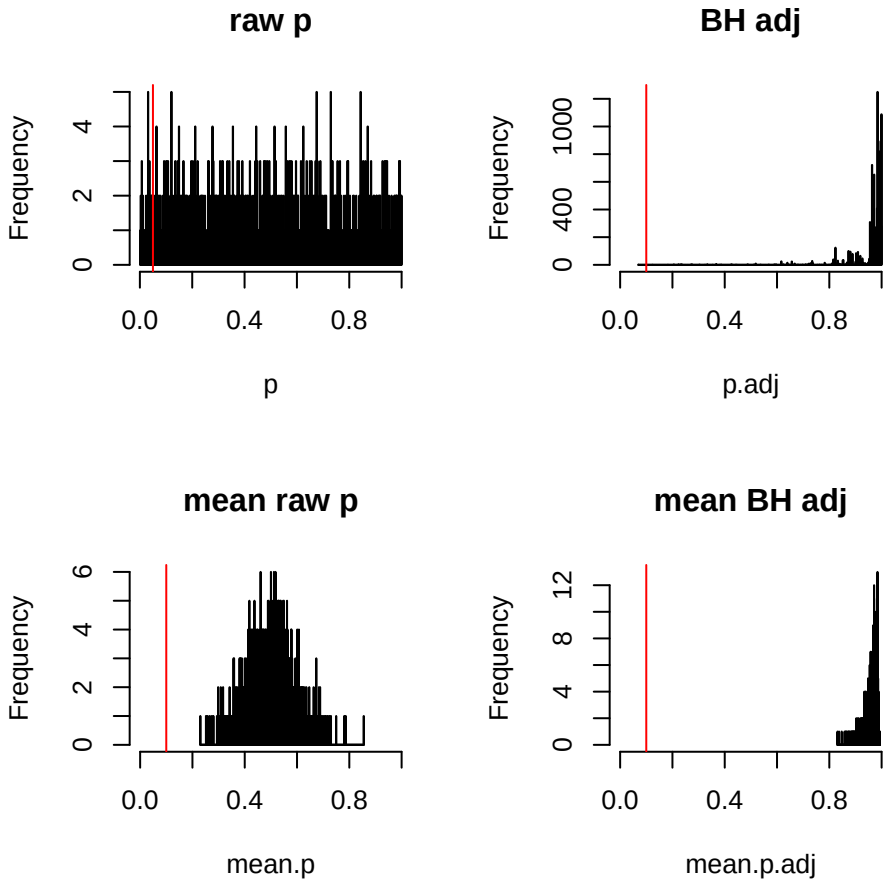

Figure 1: P values represent the "probability of finding the observed sample results, or "more extreme" results, when the null hypothesis is actually true". In other words, a p-value is the probability of seeing a difference between groups when that difference is actually due to chance. The upper left panel in the figure shows the distribution of p values for a simulated experiment where there are 500 samples in group A and 500 samples in group B and 1000 variables in each sample. The upper right panel shows the Benjamini-Hochberg corrected p-values for the same experiment. The bottom left panel shows the mean p and mean BH corrected value of 10 replicates of the upper panels.

3.2 Data are multivariate and, as a minimum, must be corrected for multiple hypothesis tests.

A p value is the likelihood of observing the result, or one more extreme, by chance alone. The commonly used cutoff of 0.05 means that we are almost certain to observe many false positive results when the samples contain many variables. Figure 1 shows an example. Here, I have chosen 1000 random numbers for each of 1000 samples. Then, I split the samples, arbitrarily, into two groups. Finally, I conducted an unpaired Welch's t-test on each of the 1000 variables in the two groups of 500 samples. The upper left panel shows a histogram of the results: p<= 0.05 for approximately 50 of the 1000 variables. These are shown as the p values to the left of the red line. You should be very wary of anyone who shows multivariate data, and then makes conclusions from raw p values.

One commonly used approach is to correct your p values for multiple hypothesis tests using one of the many corrections. The Benjamin-Hochberg procedure is widely used, and corrects the raw p values such that the likelihood of observing a given corrected value is the adjusted value. This is called a False Discovery Rate correction. For example, if the p value is 0.001, and the BH corrected p value is 0.1, then the likelihood that the difference is observed by chance is 10%. Thus, if you have 100 variables that have a BH value less than 0.1, you expect that 10 of them will be false positives - you just don't know which ones. The alternative is a Family Wide Error Rate correction, the most famous of which is the Bonferroni correction. In this case the value reported is the likelihood that any of the values reported are wrong. So if 100 values are reported using a FWER cutoff of 0.1, then the likelihood that any of them are wrong is 10%.

Another approach is to determine if the p values are stable to sampling variation. This approach is shown in the lower two panels of Figure 1. The bottom left shows the mean p values for 10 random replicates of the data in the top left panel. Here we see that the mean p value is approximately 0.5 because the expected p value for a randomly chosen comparison is 0.5. Note that the mean BH adjusted p values approach 1.

3.3 Sequencing can change the shape of the data:

It is assumed that the output from a high-throughput sequencing experiment represents in some way the underlying abundance of the input DNA molecules. The input counts panels on the left side of Figure 2 shows two idealized experiments. The top left shows the case where the total count of all nucleic acid species in the input is constrained, the bottom left illustrates the case where the total count is unconstrained. These are modelled as a time series, but any process would produce the same results.

Constrained datasets occur if the increase or decrease in any component is exactly compensated by the increase or decrease of one or more others. Here the total count remains constant across all experimental conditions. Examples of constrained datasets would include allele frequencies at a locus where the total has to be 1, and the RNA-seq where the induction of genes occurs in a steady-state cell culture. In this case, any process, such as sequencing that generates a proportion simply recapitulates the data with sampling error. The unspoken assumption in most high throughput experimental designs is that this assumption is true— *it is not!*

An unconstrained dataset results if the total count is free to vary. Examples of unconstrained datasets would include ChIP-Seq, RNA-seq where we are examining two different conditions or cell populations, metagenomics, etc. Importantly, 16S rRNA gene sequencing analyses are almost

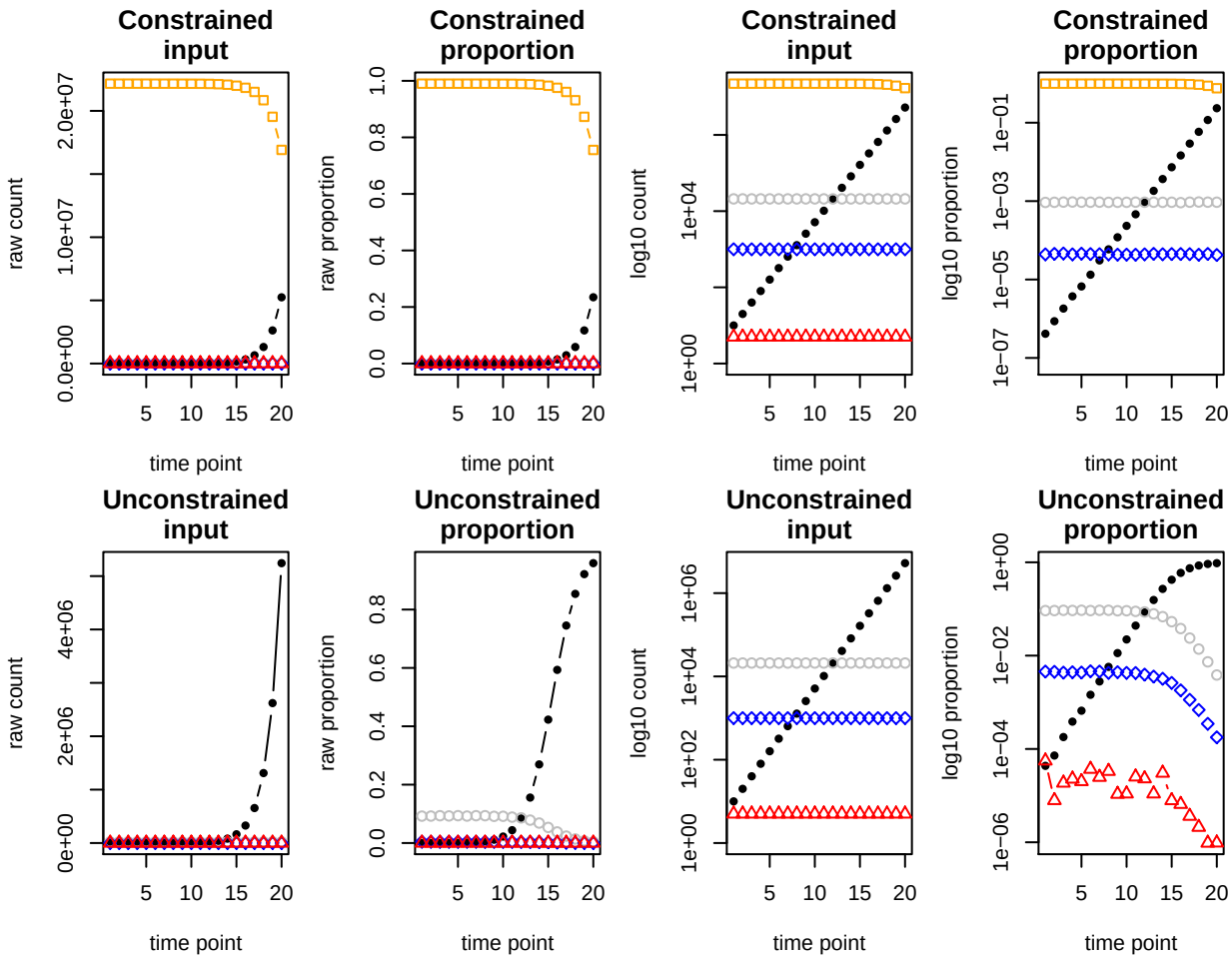

Figure 2: High-throughput sequencing affects the shape of the data differently on constrained and unconstrained data. The two left panels show the absolute number of reads in the input tube for 20 steps where the green and black OTUs are changing abundance by 2-fold each step. The gray, blue and red OTUs are held at a constant number in each step in both cases. The second column shows the output in proportions (or ppm, or FPKM) after random sampling to a constant sum, as occurs on the sequencer. The orange OTU in the constrained data set is much more abundant than any other, and is changing to maintain a constant number of input molecules. Samples in the two right columns are the same values plotted on a log scale on the Y-axis for convenience. Note how the constrained data is the same before and after sequencing while the unconstrained data is severely distorted.

always free to vary; that is, the total bacterial load is rarely constant in an environment. Thus, the unconstrained data type would be the predominant type of data that would be expected.

The relative abundance panels on the right side of Figure 2 shows the result of random sampling with a defined maximum value in these two types of datasets. This random sampling reflects the data that results from high throughput sequencing where the total number of reads is constrained by the instrument capacity. The data is represented as a proportion, but scales to parts per million or parts per billion without changing the shape. Here we see that the shape of the data after sequencing is very similar to the input data in the case of constrained, but is very different in the case of non-constrained data. In the unconstrained dataset, observe how the blue and red features appear to be constant over the first 10 time points, but then appear to decrease in abundance at later time points. Conversely, the black feature appears to increase linearly at early time points, but appears to become constant at late time points. Obviously, we would misinterpret what is happening if we compared early and late timepoints in the unconstrained dataset. It is also worth noting how the act of random sampling makes the proportional abundance of the rare OTU species uncertain in both the constrained and unconstrained data, but has little effect on the relative apparent effect on the relative abundance of OTUs with high counts.

3.4 Commonly used transformations are misleading

Current practice is to examine the datasets using 'relative abundance' values, that is, the proportional abundance of the OTUs either before or after normalization for read depth. This approach is equivalent to examining the input unconstrained data of the type seen in Figure 2 in the relative abundance sample space in the bottom right panel of the figure. This approach will obviously lead to incorrect assumptions in at least some cases. For example, depending upon the steps chosen to compare, the blue OTU, that has constant counts in the input, will be seen to either increase or decrease in abundance. Conversely, the green OTU, that is always decreasing in abundance will be seen to be constant if comparing samples 1-8.

The ecological literature offers many different transformations for such data, often as a way of making the data appear 'more normal'. Figure 3 shows the results of five such transformations.

- The frequency transform divides the each OTU value by the largest OTU count, and then divides the resulting values by the number of OTUs in the sample that had non-zero counts.
- The Hellinger transformation that takes the square root of the relative abundance (proportion) value.
- The range transform standardizes the values to have a range from 0 to 1. OTUs with 0 counts are set to 0.
- The standardize transform standardizes the values for each sample to have a mean of 0 and a variance of 1. I
- The log transform divides each OTU count in a sample by the minimum non-zero count value, then takes the logarithm of the resulting value and adds 1. Counts of 0 are assigned a value of 0 to avoid taking the logarithm of 0.
- The centred log-ratio transformation divides the OTU values by the geometric mean OTU abundance, and then takes the logarithm.

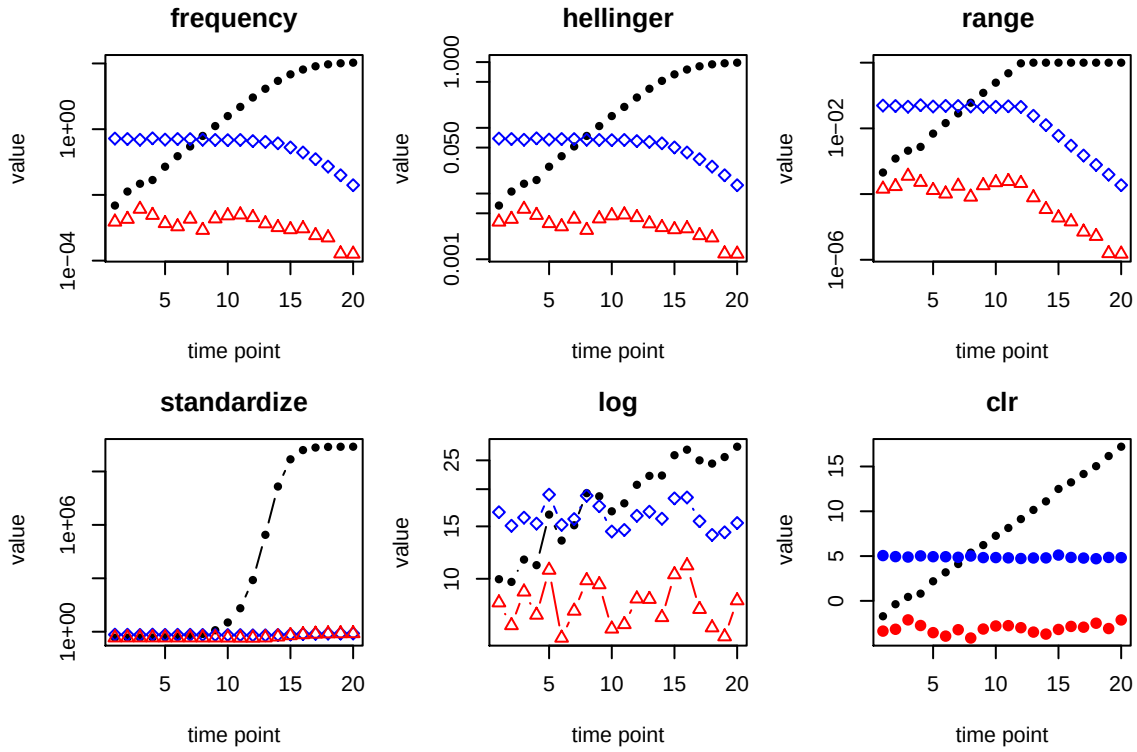

Figure 3: The effect of ecological transformations on unconstrained high throughput sequencing datasets. Data generated as in Figure 2 were transformed with five different approaches implemented in the vegan ecological analysis package, and with the entered log-ratio approach suggested by Aitchison.

It is obvious that the first four transformations result in data that badly mis-represents the shape of the actual input data. The log transformation, however results in the shape of the output data approximating the shape of the input data, except that the uncertainty of each data point is large. The ratio transform his transformation accurately recapitulates the shape of the original input data, and more accurately represents the uncertainty of each data point.

3.5 A count of 0 does not mean that you expect 0!

A common misconception to normalizing by the geometric mean is that the geometric mean is not defined if any of the values in a sample are 0. While this is true, values of 0 for an OTU can arise because the OTU sequence could not occur in the experiment, or because the OTU could exist in one group but not the other, or because the OTU was very rare in one or more samples making its selection from the library subject to chance. In the first case, the OTU would not be found in any sample, and that OTU could simply be deleted from the dataset without effect. In the second case, the OTU would be represented in one group but not the other. In the third case, where an OTU has at least one count in at least one sample, a value of 0 could arise in other samples because of sequencing depth. In the latter two cases, it is possible that the OTU could have been detected if more reads per sample were obtained or if more replicates of the library were sequenced.

Current practice in 16S rRNA gene sequencing studies is to assume that an observed value of 0

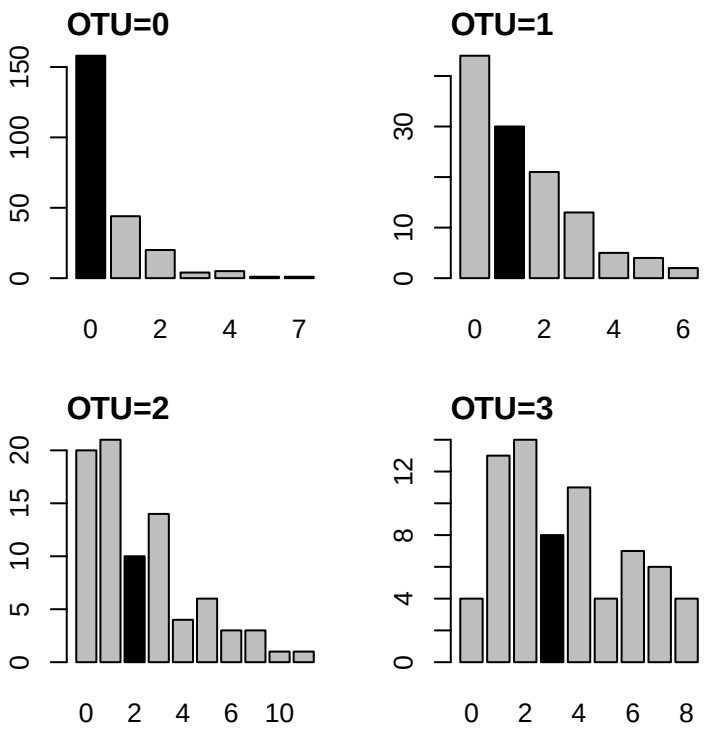

Figure 4: The distribution of counts in a replicate OTU when the first OTU has counts between 0 and 3. The same library of sixteen different 16S rRNA gene amplification samples were sequenced on two different Illumina HiSeq runs, and the count of OTUs that had values of 0 to 3 in one replicate, shown by the black bar, were tabulated for the other replicate, shown by the grey bar. Sequencing depth for each replicate was within 10% of the the other replicate.

in a sample represents the actual value. In RNA-seq it is common to remove all genes where the total sum across all samples is small (usually with a mean of 2 or less and no more than about 10 counts in any sample). In either case, the assumption is that variables with very low counts are irrelevant.

This assumption was tested by sequencing the same library from 16 different samples on two individual Illumina runs, and then determining the OTU count in one run if the OTU had a count of 0, 1, 2, or 3 in the other run. Figure 4 shows the result, and it can be seen that the count observed for an OTU in one replicate is often very different from the count observed for the other replicate. Similar observations hold for RNA-seq. It is clear that the absolute number of counts observed varies between replicates and as expected the underlying distributions approximate what would be expected for random sample of the input library. The uncertainty in ascertaining the true values for OTUs with low counts is the reason that the log transformation in Figure 3 injects undesired variability into the data.

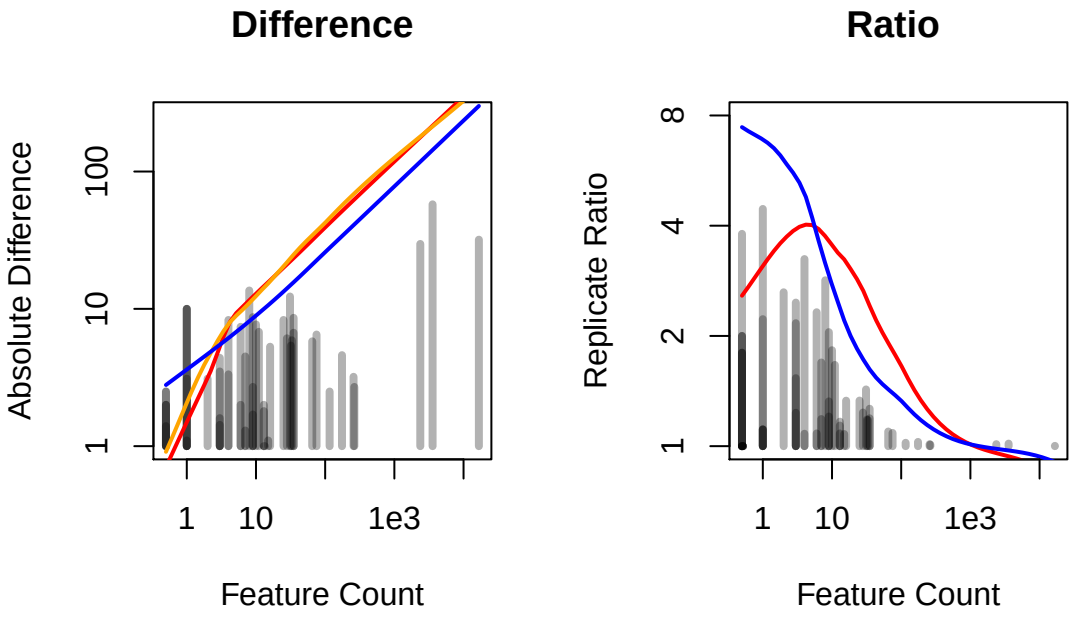

Figure 5: Examining technical replicate variability of 16S rRNA gene sequencing experiments as linear differences and as ratios. The same input DNA library for fourteen samples was sequenced on two different HiSeq lanes to estimate the technical variability. The difference between the counts for each OTU in replicate A and B were plotted in two ways. First, as differences between counts in replicates A and B, and second, as the ratio between the counts observed in replicates A and B. The grey bars show the actual difference or ratio observed plotted vs. the minimum value observed for each OTU in the replicates. The estimated variability in the data was determined in three ways: 1), by rarefying the data with the Jacknife or 2) with the Bootstrap approach in QIIME, or 3) by drawing instances from the Dirichlet distribution. The 99th percentile of the difference between these random instances and the actual data found by these three approaches is shown as the red, orange or blue line.

3.6 The uncertainty is relative

That the true count of an OTU cannot be determined from a single sequencing run, suggests that we might be better to estimate the range of values that an OTU can assume, and to determine how these estimates change the results of the analysis performed. case, we do not know the true underlying value for the OTU count and must estimate it. There are two approaches used to estimate the uncertainty of 16S rRNA gene sequencing experiments.

Subsampling, or rarefaction, is performed to normalize all samples in 16S rRNA gene sequencing to a common sequencing depth, and to estimate the variability in the data for downstream procedures such as α- and β-diversity analyses. An alternative to subsampling is the draw instances of the data from the Dirichlet distribution that treats each variable as a multinomial Poisson instance with the constraint that the values sum to 1. We have shown previously that the instances of the Dirichlet distribution accurately reflect the underlying technical variation in RNA-seq datasets (see Figure 1 in[8]). Note that when sampling occurs from an Poisson process that the error expected is relatively large for low counts and relatively low for high counts because the expected value and the variance of the data are equal. Thus, the expected error is 100% for a count of 1, and 10% for a count of 100 and 1% for a count of 10000, etc.

One caveat with Dirichlet subsampling is that it is a Baysian approach to estimate the distribution of frequencies for each OTU. Thus, we must include our prior belief of the actual abundance of each OTU before sampling. This is often thought to inject investigator bias into the analysis. The bias will be most extreme at the margins of the estimation because a value close to 0 indicates a prior belief that the OTU should never have been detected, while a value close to 1 indicates a prior belief that the OTU should have been detected with certainty. Thus the least biased value in general should be 0.5[9]. In practice we observe that the choice of prior has little effect on the outcome provided it is not close to 0 or 1.

We examined how well both procedures modelled the actual underlying variation in OTU counts from the replicate samples used in Figure 4. Monte-Carlo instances of the data drawn from the Dirichlet distribution were generated and multiplied by the number of total reads for the sample. Subsampling was performed with QIIME[10] using the Jacknife procedure (i.e., subsampling from an original distribution without replacement[11]), and with the Bootstrapping procedure[11] (i.e., subsampling with replacement) using a sampling depth of 10000 using the multiple_rarefactions.py script in QIIME v1.8.0. This reduced the read depth for the samples by a factor between 2 and 4.5 fold, which is well within the subsampling parameters of recent experiments. The difference between the read counts observed in the subsampled, Dirichret, and actual data was calculated as for the replicate difference.

Figure 5 shows an example plot of the technical variability that occurs at the OTU level for the same samples sequenced to approximately the same sequencing depth on two separate Illumina lanes. Zero counts, if they occurred in one replicate were replaced with 0.5[8,12]. The actual technical variability is shown as the bars, and is plotted as the absolute difference between replicate 1 and replicate 2. The density of the bars illustrate the number of observations at that co-ordinate, with darker bars representing more observations. The red and orange lines in Figure 5 show the loess line of best fit for the 99^{th} percentile of the difference between the actual and the Jackknifed and bootstrapped rarefied data, and the blue line shows the 99th percentile for the Dir instances.

The left plot shows that when treating the data as absolute counts, the difference between the actual replicates is greatest when the read counts are the highest, and that OTUs with very small

counts tend to have small differences. Plots such as this form the basis for using the Negative Binomial to estimate variability in RNA-seq datasets of this type, and are one of the motivations to use RNA-Seq based statistical procedures for 16S rRNA sequencing datasets[2].

Note that the 99th percentile of the variability observed for all three sampling methods underestimates the variability in the data when the counts are close to 0. Interestingly, the Jacknife and bootstrap estimation methods become non-linear in this region and severely under-estimate the variability seen for OTUs with less than 10 counts. In contrast the Dirichlet instances become slightly more variable when the feature counts are low. The reasons for the under-counting of technical variance by subsampling are obvious upon reflection: subsampling is constrained by the actual observations being sampled and so can only result in estimating *fewer* reads per feature, never more.

The right plot shows the same data plotted as the ratio between the counts observed for replicates A and B. When plotted in this way, it is clear that the rare OTUs, for which we have the least accurate estimate of their underlying abundance, show the largest differences in ratio abundance. Conversely, the most abundant OTUs show the smallest amount of variability. In this plot the range of variability estimated by Dirichlet sampling brackets the observed variability, and the subsampling approaches again under-estimate the variability of OTUs with very low numbers of counts. The Loess line of best fit appears to under-represent the variation when feature counts are high, but this is an artefact of the line fitting procedure.

3.7 Scaling the data

It is obvious that the data for each sample must be placed on a common scale, and while the logarithm of the counts reduces the difference between extreme values, the data is skewed because the minimum is the logarithm of the value assigned to 0 (in the examples above, this is $log2(0.5) = -1$) and the maximum value is proportional to count value of the most abundant OTU in the sample. So simply taking the logarithm while commonly used for proportional data, does not scale the data properly[3]. The solution devised by Aitchison, and validated and extended others[7] is to centre the logarithmic transformation on the geometric mean as was done for the ratio transform used in Figure 3. This is referred to as the centred log-ratio transformation. This transformation has several effects. First, the data is on a common scale, where the values can be interpreted as the ratio between the count for each OTU and the geometric mean count for all OTUs in the sample. Second, the differences between values represent the ratios between values in that sample; i.e., a difference of 1 represents a two-fold abundance difference if the base of the logarithm is 2. Third, converting the count values to ratios preserves the 1:1 correspondence between the original values. Fourth, the relationships between values is not affected if particular OTUs are included or excluded from analysis. Finally, the ratio values are now linearly related and so can now be used in standard statistical analyses. Atichison[3], Pawlsky-Glahn[6], and Egozcue[13], have done much work to develop rigorous approaches to analyze such data types[7], and we have developed an R package that can be useful to determine differential abundances of OTUs in these datasets[8,14].

3.8 The power of 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[15]. In addition, compositional datasets have the property that they are described by $D-1$ observations if the sum of the parts is known[3]. 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 2). 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.

3.8.1 Sub-compositions:

Compositional data also exhibit the unusual property that the examination of a sub-composition of these data will provide different answers than will be obtained with the full dataset[3]. 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.

3.8.2 Spurious correlations:

Finally, it is important to know that compositional data has the additional problem of spurious correlation[15], 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[16], 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[17], a tool that accounts for this spurious correlation.

Atichison[3], Pawlsky-Glahn[6], and Egozcue[13], 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

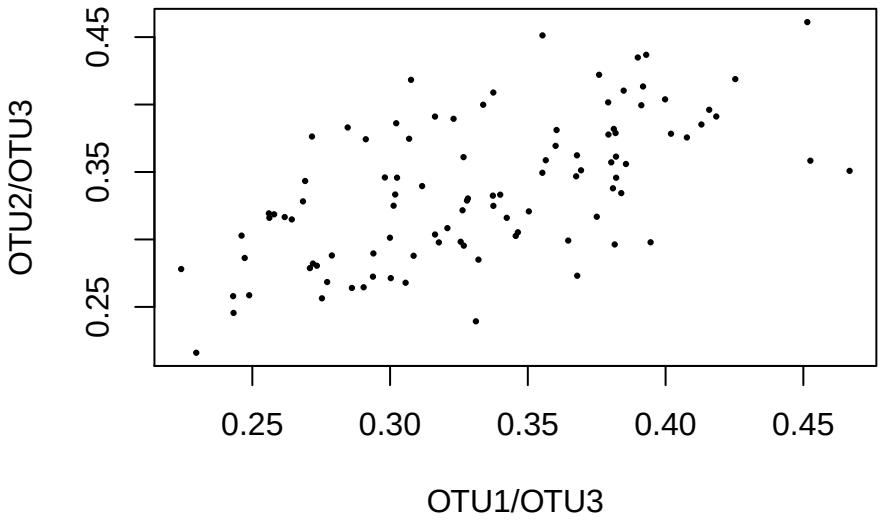

Figure 6: Spurious correlation in compositional data. Two random vectors drawn from a Normal distribution, were divided by a third vector also drawn at random from a Normal distribution. The two vectors have nothing in common, they should exhibit no correlation, and yes they exhibit a correlation coefficent of > 0.65 when divided by the third vector. See the introductory section of the Supplementary Information of Lovell [16] for a more complete lay description of this phenomenon.

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.

4 Conclusions

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 uninformative, 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 significance in a real dataset can substantially inflate the observed P value because of random partitioning of low count values across datasets.

References

1) 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.

2) Paul J McMurdie and Susan Holmes. Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput Biol*, 10(4):e1003531, Apr 2014.

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

4) K. Gerald van den Boogaart and R. Tolosana-Delgado. "compositions": A unified R package to analyze compositional data. *Computers & Geosciences*, 34(4):320 – 338, 2008.

5) J. Aitchison and J. J. Egozcue. Compositional data analysis: Where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850, 2005.

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

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

8) A. D. Fernandes, J. M. Macklaim, T.G Linn, G. Reid, and G. B. Gloor. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. *PLoS ONE*, 8(7):e67019, July 2013.

9) E. T Jaynes and G. Larry Bretthorst. *Probability theory: the logic of science*. Cambridge University Press, Cambridge, UK, 2003.

10) J Gregory Caporaso, Justin Kuczynski, Jesse Stombaugh, Kyle Bittinger, Frederic D Bushman, Elizabeth K Costello, Noah Fierer, Antonio Gonzalez Peña, Julia K Goodrich, Jeffrey I Gordon, Gavin A Huttley, Scott T Kelley, Dan Knights, Jeremy E Koenig, Ruth E Ley, Catherine A Lozupone, Daniel McDonald, Brian D Muegge, Meg Pirrung, Jens Reeder, Joel R Sevinsky, Peter J Turnbaugh, William A Walters, Jeremy Widmann, Tanya Yatsunenko, Jesse Zaneveld,

and Rob Knight. Qiime allows analysis of high-throughput community sequencing data. *Nat Methods*, 7(5):335–6, May 2010.

11) B. Efron. Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68(3):589, 1981.

12) Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.

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

14) Andrew D Fernandes, Jennifer Ns Reid, Jean M Macklaim, Thomas A McMurrough, David R Edgell, and Gregory B Gloor. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome*, 2:15, 2014.

15) 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, 1896.

16) David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol*, 11(3):e1004075, Mar 2015.

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