Austrian Journal of Statistics
MMMMMM YYYY, Volume VV, 1–xx.
http://www.ajs.or.at/

Compositional uncertainty in high-throughput sequencing data analysis should not be ignored

Gregory B. Gloor
The University
of Western Ontario

Jean M. Macklaim
The University
of Western Ontario

Michael Vu
The University
of Western Ontario

Andrew D. Fernandes
YouKaryote Genomics
London, Canada

Abstract

The abstract of the article in English

Keywords: Bayesian estimation, centred log-ratio, transcriptome, metagenome, 16S rRNA gene sequencing, ALDEx2, R.

1. Introduction

High throughput sequencing studies, that generate as outputs thousands to billions of sequence tags, are becoming the norm in the life sciences. That these experiments generate compositional data can be understood with two statements. First, the total number of sequence tags obtained in an experiment are of no importance. Second, the sequence tags are binned into features where the difference between features is exponential and best explained by ratios. These features can represent genes as in 16S rRNA gene sequencing, transcriptomics and metagenomics or single-nucleotide variant abundances after differential growth experiments. The experimentalist in these experiments is interested in knowing which features, if any, are differentially abundant between two or more distinct groups. Furthermore, all experiments of this type explicitly or implicitly examine sub-compositions. Finally, each individual experimental design is analyzed using different sets of underlying assumptions that are derived from historical dogma, despite having the same underlying data structure. Fernandes 2014 demonstrated that tools developed for one experimental design (e.g. RNA-Seq) do not translate well to other experimental designs (e.g. 16S RNA gene sequencing).

These data are necessarily sparse, and complex. There are often hundreds or thousands of features, and the expense of these experiments prevents the collection of sufficient sequence tags to ensure that all features are covered by at least one sequence tag. Thus, the treatment of features with zero counts is a pervasive problem when treating these data as compositions (Lovell, Müller, Taylor, Zwart, Helliwell, Pawlowsky-Glahn, and Buccianti 2011). It is assumed that features with zero counts across all samples are removed because they are uninformative. For the remainder where one approach is to delete features where one or more samples have zero counts (Lovell *et al.* 2011; Lovell, Pawlowsky-Glahn, Egozcue, Marguerat, and Bähler 2014). This removes the problem of zero count features at the expense of po-

tentially excluding the most important features from consideration. Another approach, is to replace the zero counts with an expected value calculated in some way. Several approaches with differing underlying assumptions are in use, and Martín-Fernández 2014 suggested that a Bayesian-Laplace approach to be the most reasonable. Regardless of the method used to treat zero count features, these analyses use maximum-likelihood approaches to determine feature abundance prior to analyses.

We have found that the variation due to sampling alone (technical variation) in compositional datasets derived from high-throughput sequencing is large and inversely related to the number of reads mapping to a fragment (Fernandes, Macklaim, Linn, Reid, and Gloor 2013). Ignoring this technical variation can lead to false positive inferences regarding differential abundance if the data are not treated as compositions. We have found that a two-step procedure incorporating a Bayesian estimate of feature abundance along with analyses conducted after centred-log-ratio transformation markedly improves specificity with no loss of sensitivity, and that the increase in specificity derived almost entirely from the exclusion of low-count (including zero count) features (Fernandes *et al.* 2014).

Our paper explores how the analyses differ when the value of zero is assigned using different approaches with, and without Bayesian estimation of the technical variation. Our initial work showed that a uniform prior added to all values was able to encompass the estimated technical variation in a sparse dataset (Fernandes *et al.* 2013). However, we observed that this approach slightly overestimated technical variation of low count and zero count features, suggesting that this approach had less than optimum power.

We will compare uniform priors that replace 0, uniform priors added to all values, and the prior estimation methods from the zCompositions package (Palarea-Albaladejo and Martín-Fernández 2015) that produce non-uniform estimates of the actual zero value. We will examine a real differential growth experimental dataset for which an objective standard of truth is known. We argue that these results are generalizable across other datasets including RNA-seq datasets and 16S rRNA gene sequencing experiments.

2. Statement of the problem

High throughput sequencing is a technology that delivers thousands to millions of reads that correspond uniquely to genes or other features in a genome, or to bins that represent sequence variants. Figure 1A shows several different study designs that are common in the literature. Regardless of design a very large number of molecules, shown in the orange box in Figure 1A are randomly sampled to produce a library that is then sequenced. The sequencing instrument delivers a much smaller random sample of the actual input and the act of sequencing converts the data from unconstrained to constrained proportional data because the instrument delivers a fixed number of sequence reads. This hard upper bound means that all such analyses generate compositional data regardless of the actual study design. In general, these experiments aim to ask the question, "what gene or feature has a different abundance between groups A and B?"

Figure 1B shows how sequencing distorts the data. Many processes examined by high throughput sequencing can be thought of as linear compositional processes. Consider a mixture of many distinctive molecules in vector $x = [x_1, x_2, ...x_n]$ over time or space increments i. For each increment we can determine the abundance of each molecule using Equation 1:

$$x_i = x_0 \times 2^{(\lambda i)} \tag{1}$$

where λ is the incremental rate. If $\lambda = 0$ for all but one of the members of vector x and $\lambda = 1$ for one member, then one member will double in abundance at each increment and all remaining members will be unchanged. Figure 1B.1 shows such a thought experiment where the values are plotted as counts of molecules. Producing and sequencing a library generates a set of counts per gene that are scaled by the maximum number of reads delivered by the

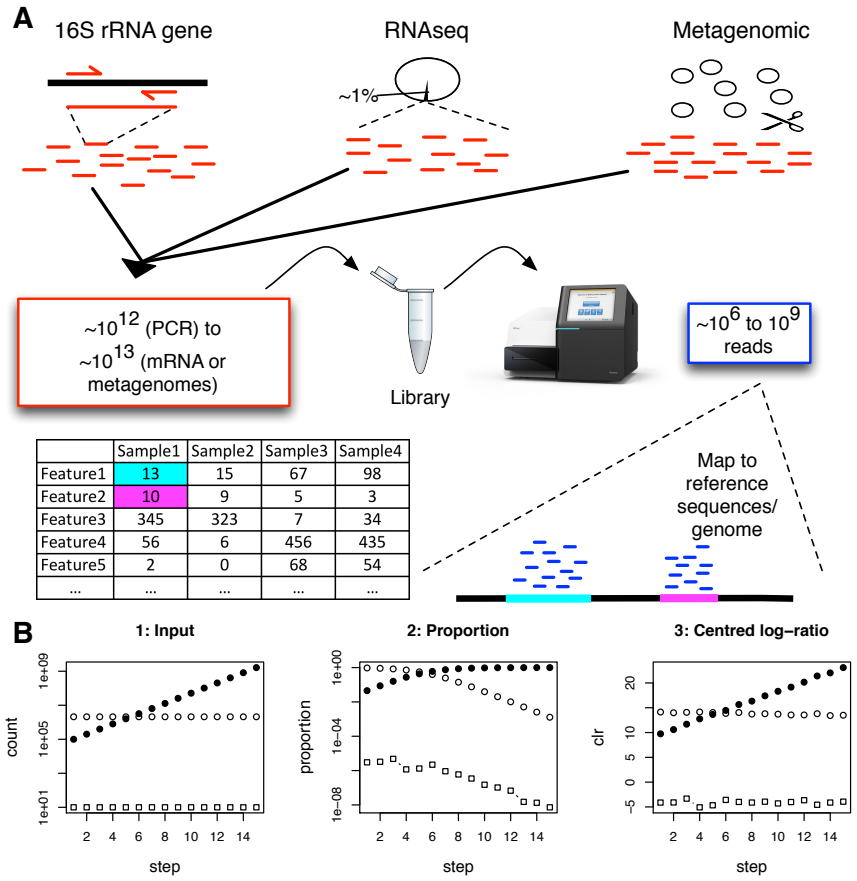

Figure 1: High-throughput sequencing affects the shape of the data. Panel A illustrates the workflow by which high throughput sequencing samples the DNA or RNA from an environment. There are many more molecules that are sampled than can be incorporated into the library, or that can be sequenced on the instrument. The capacity of the instrument itself determines the number of reads observed. The orange box shows the number of molecules in typical initial samples, and the blue box shows the maximum number of reads that are obtained from the instrument. These reads are assigned to features such as genes or operational taxonomic units or other bins, and a table of the reads per feature is output. Panel B illustrates how the data is distorted during the process. The input DNA or RNA usually has no fixed sum and is randomly sampled sequentially during the library preparation and sequencing steps. The output from the instrument is compositional because the instrument can deliver only a fixed upper limit of reads, regardless of the number of molecules in the input. Panel B.1 shows the number of reads in the input tube for 15 steps where the open square and circular features are held at a constant number and the black feature is increasing in abundance by 2-fold each step. Panel B.2 shows the output in proportions (or ppm) after random sampling to a constant sum, as occurs on the sequencer. Panel B.3 shows the shape of the data following centre log-ratio transformation. Note that panels B.1 and B.2 have the y axis on a logarithmic scale, and that the natural scale for centred log-ratio data is logarithmic.

machine. In other words, the counts for gene x_i are per-gene probabilities p_i and are formally equivalent to a random multivariate Poisson sample of the original group of DNA molecules. We model this process by sampling from the Dirichlet distribution according to Equation 2:

$$[p_1, p_2, ...p_n] \sim \text{Dirichlet}[x_1, x_2, ...x_n]. \tag{2}$$

A single Dirichlet instance generates a single Bayesian estimate of the underlying posterior probabilities for each feature, and multiple samples generate a full posterior distribution

(Holmes, Harris, and Quince 2012; La Rosa, Brooks, Deych, Boone, Edwards, Wang, Sodergren, Weinstock, and Shannon 2012; Fernandes *et al.* 2013). Figure and panel 1B.2 shows the posterior values for a single Dirichlet instance from the counts in Panel 1B.1. Here we can see that the constant sum constraint resulting from the finite read limit of the instrument severely distorts the underlying shape of the data. Figure 1B.3 demonstrates that applying the standard centred log-ratio transform of Aitchison 1986 to the vector of probabilities p in Panel 1B.1

$$clr(p) = [log_2\frac{p_1}{g(p)}, log_2\frac{p_2}{g(p)}, ...log_2\frac{p_n}{g(p)}] \tag{3}$$

reconstitutes the essential shape of the data, with the actual data points now showing some variability because of random sampling. In equation 3, $g(p)$ denotes the geometric mean of the vector p. This transformation is convenient because it reconstitutes the essential shape of the original data, and because there is a one to one mapping between the values in the original and in the transformed dataset. Furthermore, this transformation can easily interpreted for the experimentalist because it is simply a ratio between the abundance of a gene or feature in the sample and the average abundance of all genes or features in the sample. Of particular note is that $g(p)$ cannot be calculated when 0 values are present, and it is the influence of different means of estimating 0 values that are the primary focus of this report.

2.1. Data from high-throughput sequencing is highly variable

Data from high throughput sequencing experiments are often thought of as point estimates despite being random samples of the input molecules, and despite several experiments showing that sequencing the same DNA library will produce somewhat different count tables at the same sequencing depth (Marioni, Mason, Mane, Stephens, and Gilad 2008; Bottomly, Walter, Hunter, Darakjian, Kawane, Buck, Searles, Mooney, McWeeney, and Hitzemann 2011; Gierliński, Cole, Schofield, Schurch, Sherstnev, Singh, Wrobel, Gharbi, Simpson, Owen-Hughes, Blaxter, and Barton 2015). Figure 2 shows an example of this variability. Marioni et al. 2008 did an experiment where two aliquots of the same RNA-seq library were run in duplicate, and the resulting reads were mapped to the $>$ 20000 genes in the human genome. Replicate runs did not return exactly the same number of reads per gene: for example, when the genes in one replicate contained zero counts, the same genes in the other replicate often had non-0 reads. This imprecision extends across the range of per-gene counts as shown for a few replicate read values in Figure 2. This imprecision is proportionally larger for small count values, and smaller for large count values. For example, the range of counts observed in replicate B when genes in replicate A contain one count span the range of 0-14 in this example: a difference of over 10-fold. By comparison, when genes in replicate A contain 64 counts the corresponding genes in replicate B span counts from 38-91: a difference of less than 50%. See Figure 1 of Fernandes et al. 2013 for a demonstration that the proportional error does indeed span the entire range of expression values in this dataset.

The imprecision can be modelled by sampling instances from a Dirichlet distibution (Fernandes *et al.* 2013, 2014) as in Equation 2. Figure 2 shows quantile-quantile plots comparing the distribution of true technical variation to the distribution of the estimated technical variation obtained by drawing instances from the Dirichlet distribution. That is for a vector of counts x, $x_{Dir} = \text{Dirichlet}[x] \times \sum x$. Sampling multiple Dirichlet instances thus returns a distribution of the posterior probabilities of each feature in the vector x, and conserves probability. These plots show that the Dirichlet instances slightly over-estimate the tails of the distributions, although these conclusions need to be tempered by the lack of datapoint for technical replicates containing double-digit counts. We conclude that drawing Dirichlet instances is an acceptable method to model posterior probabilities in these datasets.

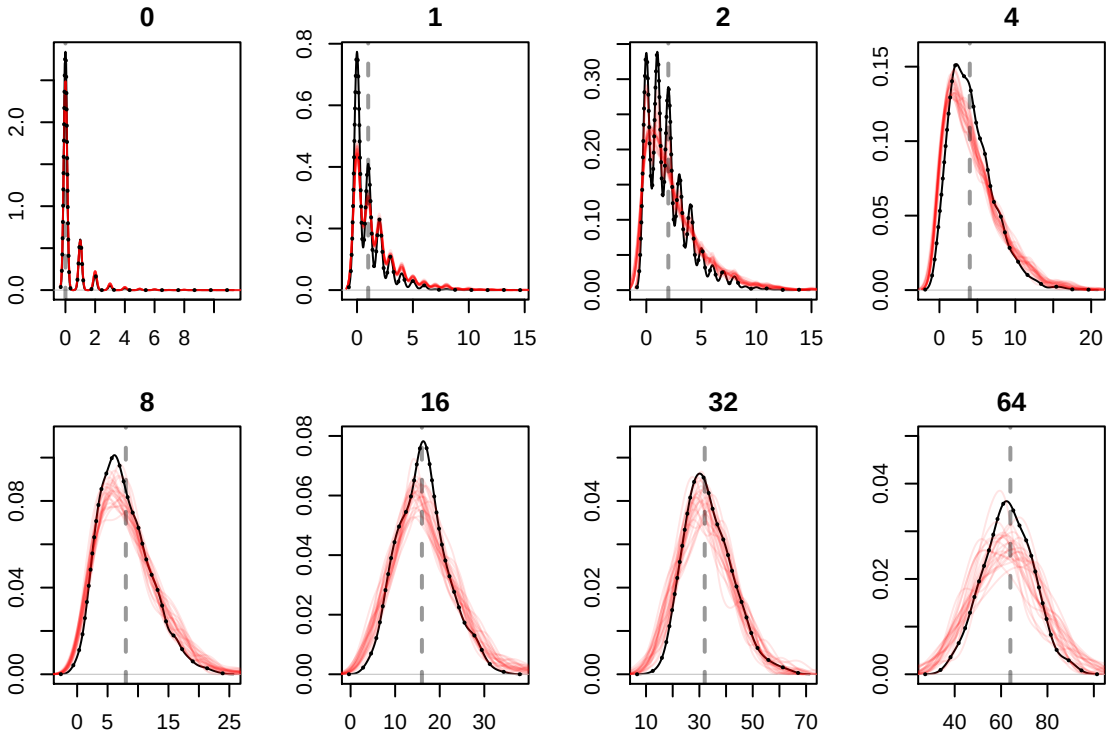

Figure 2: Technical replicate variation, and density plots of the estimation of that variation for an RNA-seq experiment. The black dotted lines show the density of the technical variation of features from one replicate of an RNA-seq dataset when compared to another as a function of the counts in the first replicate. The count value of the first replicate is give above each plot, and the location of this value is shown as the dotted grey vertical line. The red lines show density plots of the inferred technical generated through 25 random instances drawn from a Dirichlet distribution. Data are from the Marioni et al. 2008 dataset.

2.2. False positive results because of unaccounted variation

One problem when analyzing such data is that the available datasets – whether derived from 16S rRNA gene sequencing, transcriptomics, or other experiments – are exploratory and so generally lack a standard of truth. This makes it difficult to develop and test tools without modelling a dataset. While modelled datasets have some allure because the parameters can be closely controlled, we prefer to examine the behaviour of different approaches in real biological datasets because they often have unanticipated error and less predictable behaviour than modelled datasets.

McMurrough et al. 2014 generated a selective growth dataset, hereafter called the 'selex' dataset, for which a standard of truth for many variables is known, and that can be inferred for many others. This dataset compares the growth of a set of 1600 sequence variants in the I-LtrI endonuclease under two conditions. The first condition is a non-restrictive condition where the growth of all variants is unconstrained. The second condition is restrictive for growth, unless the I-LtrI endonuclease is active and can cleave and inactivate the gene encoding *Ccdb*, a DNA gyrase toxin. The gyrase toxin is dose-dependent so cleavage of a fraction of the plasmids containing the gene confers slower growth (Smith and Maxwell 2006), and under the conditions of the assay, the toxin would be bacteriostatic if no cleavage occurred. Thus in this experimental design the difference between inactive variants between the two conditions would be one of dilution alone, and *no variant should become less abundant during the experiment.* Variants that cleave the toxin gene would confer a growth advantage, and would become more abundant over the time of the assay. Furthermore, McMurrough et al. 2014 showed that the *in vitro* enzymatic activity of the endonuclease is strongly correlated with the output of the

selective growth experiment.

The abundance of each variant in the mixture can be modelled by Equation 1. At time zero if each variant is contained in vector $n_0 = [n_1, n_2, n_3 ... n_{1600}]$, over time increments, the change in abundance in the non-selected growth conditions can be modelled with $\lambda = 1$ and the variation in λ being small. The experimental conditions allowed for approximately 16 doublings, or time increments. Therefore at the last increment of the non-selected time series, we anticipate that the initial relationships between the abundances of each of the 1600 variants will be essentially unchanged. In contrast, the selected variants are under strongly differential selection. Here the most active variants will have $\lambda \approx 1$, that is, these variants grow at the same rate in the selected and unselected conditions. The least active variants will have $\lambda = 0$, that is, these variants will not change in actual abundance during selection, but will become relatively less abundant when compared to their active counterparts. Inactive variants are known to be by far the most prevalent in the samples. Intermediate positive values of λ are expected, and no negative values are expected. Finally, it is possible for individual samples to demonstrate differences in apparent λ under selection. This can occur if a variant is partially active, and cleaves different proportions of the toxin genes in a particular cell by chance. This event is heritable and so would allow cells carrying the same variant to grow at slightly different rates. Thus, the sample in which this occurred would have an apparent increase in λ for that variant in that sample.

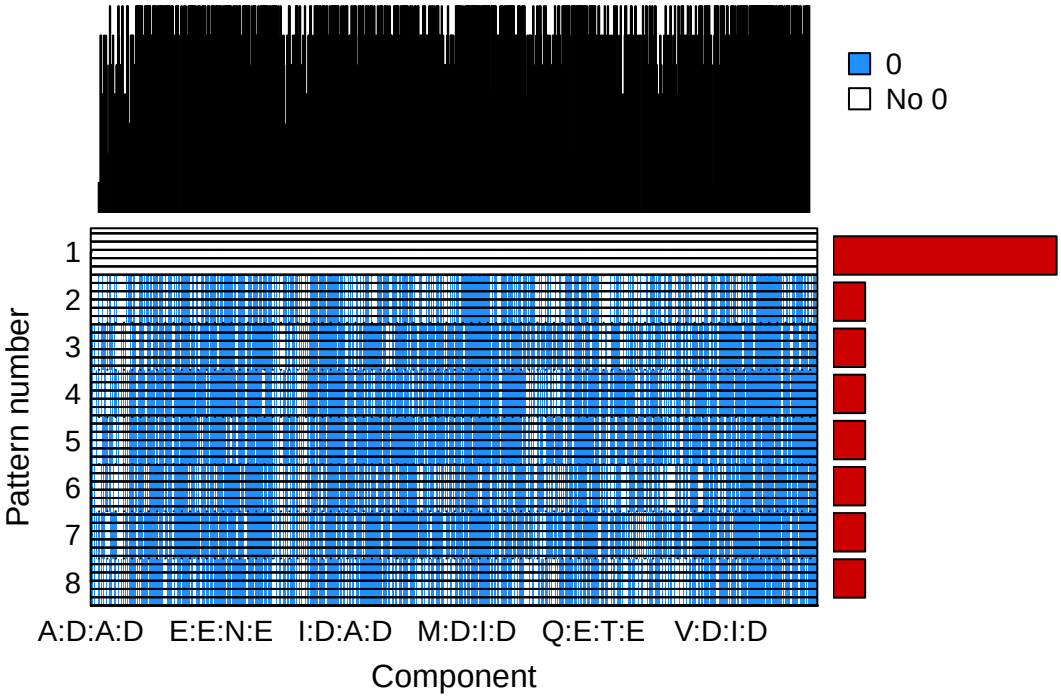

Figure 3: Characteristics of the McMurrough et al. 2014 dataset summarized by the zCompositions package. The top panel shows that all parts contain 0 values, and the bottom panel shows that there are 8 patterns in the data. Seven of the samples contain no 0 values, these are the control samples grown in the absence of selection. The samples derived from the selective growth display individual seven different patterns for 0 values likely due to random sampling.

The question we wish to address with this dataset is: can we identify from the growth experiment alone which variants are likely to be active? Active variants will have had a maximum of 16 cell doublings becoming much more abundant, inactive variants will stay at the same abundance and variants with partial activity will become only somewhat more abundant. In

addition, we wanted to know the effect on our inference of the different approaches to estimating the zero values. We first examined the dataset using a biplot to show the relationship between the samples and the variants.

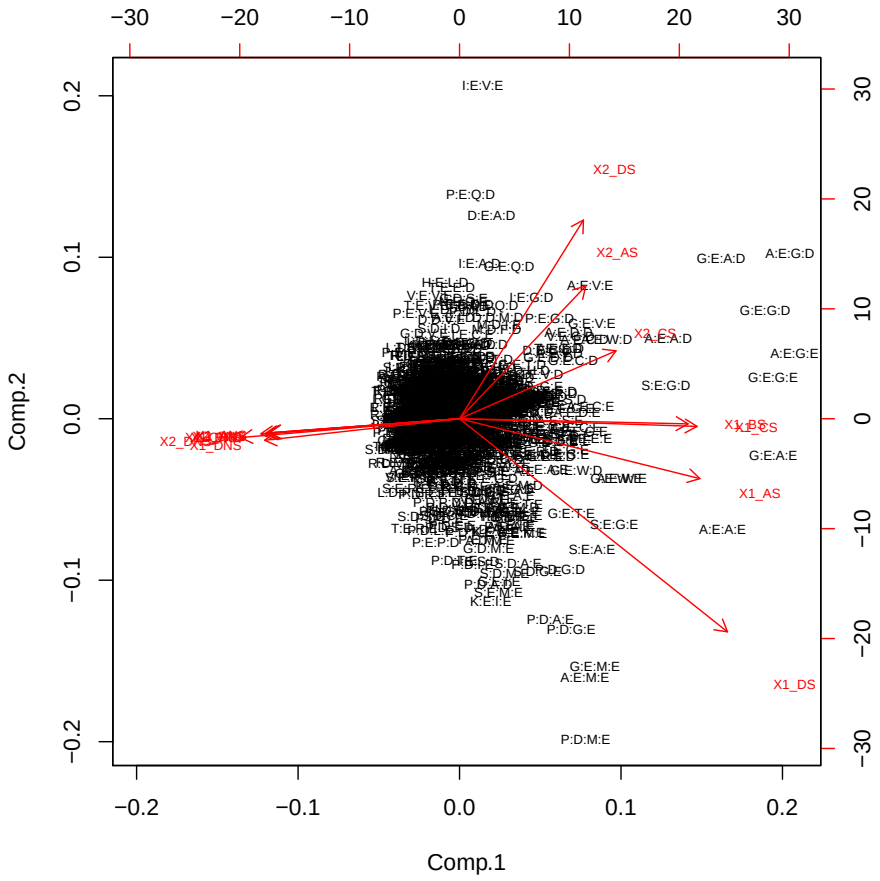

Figure 4: Biplot showing the relationship between samples and variables in the selex dataset. Zero values were adjusted using the count zero multiplicative approach using the zCompositions R package, and the biplot was generated using the compositions R package. Samples ending with 'NS' are the control non-selected growth samples and samples ending in 'S'are from the selected growth samples. The vast majority of the variables cluster around the centre of the dataset. The red arrows to the left show that the 7 non-selected replicates are very similar, and the selected replicates on the right exhibit some variability. The differences between samples are driven by variation in a small number of variables. In this dataset, component 1 explains 52.4% of the variability and component 2 explains 10.4%.

Figure 3 shows the density and distributions of 0 values in this dataset as summarized by the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015). We can see that this dataset will be very challenging to analyze because the control samples do not contain any 0 values, but most of the variants in the experimental samples contain 0 values in several samples. This high density of 0 values comes about because the number of sequence reads was insufficient, and not because we expected a 0 value in any of the variants. Thus, we must impute the most likely value of 0 in each sample before analysis.

A compositional biplot generated with the compositions R package (van den Boogaart and Tolosana-Delgado 2008) following zero replacement using the CZM approach from the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015) is shown in Figure 4. The first two components of this biplot explained 52.4% and 10.4% of the variance in the data, indicating that this is a good summary of the data. The selected and non-selected

samples separate clearly on the first component, and this separation is driven largely by the abundance of the variants on the right side, such as A:E:G:E, G:E:G:E, G:E:G:E, etc, which McMurrough et al. 2014 demonstrated to be the highly active variants. However, it is difficult to quantitate the magnitude of the abundance change of the variants from this analysis. This figure also shows that the non-selected samples, which cluster on the left side, are essentially redundant since the links between them are exceedingly short. The selected samples on the right side are much more diverse, and some of the links from the origin are nearly orthogonal. The differences between the samples is largely on component 2, and is driven by the abundance of a small number of variants that also exhibit variation from the bulk on component 2. Inspection of the data finds that this diversity is driven by only a few variables such as I:E:V:E, P:D:M:E, A:E:M:E etc, and that these variables separate the X1 and X2 sample sets: interestingly, these sets are from identical experiments performed with different batches of the same cell type. Examination of the underlying count table shows that these variants are indeed different in abundance between the X1 and X2 sets. For example the I:E:V:E variable has 17933 reads in sample X2_DS but has zero reads in sample X1_DS. This is an example of a single stochastic event that conferred a growth advantage to this variant in this sample. It is important to note that this large stochastic variation has important consequences when examining datasets because such variation is not unusual in real biological datasets.

Not accounting for sampling results in many false positive identifications

The selex dataset is unique because we have a validated truth for some of the features that differentiate the conditions (McMurrough *et al.* 2014). In this dataset we have unambiguously biochemically identified variants that are active and those that are not. Based on this prior information, we expect that approximately 60 variants would exhibit substantial activity in this assay, and so substantial deviation from this number would indicate many false positive results.

One approach that is widely used in the literature is to reduce the values from the count table to proportions or normalized counts through Maximum Likelihood approaches, then to conduct univariate statistical tests for each variant, and (sometimes) to correct for multiple testing (see for example the 16S rRNA analysis methods in Hsiao 2013). These approaches often treat zero values as actual zeros, making no adjustments. Applying this simple method using an unpaired Wilcoxon test, and applying the Benjamini-Hochberg correction (Benjamini and Hochberg 1995) to the resulting P values reveals that 1593 of 1600 variants are identified as having a differential abundance between the selected and non-selected conditions with an adjusted P value cutoff of 0.05. This is clearly at odds with the known biology of the underlying dataset.

A potentially more rigorous, yet still simple approach is to adjust the zero values in this dataset using one of the available methods that are implemented in the zCompositions R package (Palarea-Albaladejo and Martín-Fernández 2015) and then to treat the data as compositions by applying Equation 3 before performing univariate statistical tests. Recall from Figure 1 that this transformation recapitulates the essential shape of the data, and there is a one to one mapping of variant counts to centred log-ratio values. The range of values for zero replacement by different methods are given in the Prior column of Table 1. Interestingly, three of these zero correction methods returned values greater than 1 for some of the zero values, this likely was a result of the very large difference between the selected and non-selected count values in the two groups. In addition, we applied two other approaches to deal with zero values. The first, labeled uniform replacement, replaces all zero values with 0.5 but does not adjust other values in the dataset. This is akin to adding a pseudo count to zero values. The second labelled uniform prior, applies a uniform prior adjustment to *all values* in the dataset. For this we use the minimally informative Jeffrey's prior of 0.5. The Comp 1 and Comp 2 columns in Table 1 show the percentage variation explained by a clr compositional biplot using each of these zero adjustments in the first two principle components. Only the biplot that used the square-root Bayesian multiplicative method appears to result in a transformation that

explains substantially less of the variation in the dataset.

Table 1: Numbers of distinguishing features identified in the selective growth experiment observed with different approaches to assign prior expectations to zero count features.

Prior assignment	Prior	Comp 1	Comp 2	Point	Dir
Count zero	0.325452 - 0.325910	0.524	0.104	874	91
Geometric Bayesian	0.061279 - 4.890273	0.504	0.108	355	82
Square root	0.006854 - 3.102299	0.452	0.118	1008	DNR
Bayes-Laplace	0.030497 - 4.883747	0.480	0.108	435	133
Uniform replacement	0.5	0.556	0.098	958	74
Uniform Prior	0.5	0.528	0.102	868	84

The utility of these zero replacement approaches to detect univariate differences in this experiment was tested closing the vectors after prior assignment, applying the centred log-ratio transform to each sample and then subjecting the features to unpaired Wilcoxon tests. Again P values were adjusted using the Benjamini-Hochberg method and an adjusted P value of 0.05 was used as the threshold for significance. Table 1-Point shows the results of this approach. Here we see that all of the methods substantially improve upon the naive approach, with between one-quarter and two-thirds of the variables being identified as differential. In this dataset, the square root Bayesian multiplicative method provides the largest number of positive identifications, and the Geometric Bayesian multiplicative correction provides the smallest number of positive identifications, although no method is able to strongly distinguish the known small number of true positives from a much larger number of false positives.

Accounting for sampling reduces false positive identifications

One substantial shortcoming of these approaches is that the inherent technical variation in the dataset is not taken into account. It is becoming an accepted practice to account for the sampling using Dirichlet multinomial mixture models, where each sample is represented by a vector of probabilities, rather than point estimates (Holmes *et al.* 2012). For example, Ding 2014 recently used this approach to partition microbiomes into different community states in a robust manner. This approach thus generates a Bayesian posterior estimate of the probabilities associated with each count prior to analysis.

This approach was tested by generating 128 Dir instances of the selex dataset using Equation **??** with a uniform prior of 0.5, and then conducted per-variant Wilcoxon tests on each instance. The mean Benjamini-Hochberg adjusted P value for each variant was tabulated, and again the cutoff used was an adjusted P value of 0.5. Surprisingly, this approach, which takes into account the inferred technical variation, again resulted in 1593 of the 1600 variants as being differentially abundant between the selected and non-selected groups. This is more than the 868 variants detected when variation was not taken into account but the centred log-ratio transform was applied, and equivalent to the naive method accounting for neither variation nor the compositional nature of the data. Thus, simple averaging across inferred the technical variates is not sufficient to screen out false positive variants in this dataset.

Finally, we combined the Bayesian posterior estimated from 128 Dirichlet instances of the data and the centred log-ratio transformation of the posterior and used this as the input to significance tests. This method is implemented in the ALDEx2 R package for the analysis of high throughput sequencing datasets (Fernandes *et al.* 2013, 2014), and is available at Bioconductor.

As implemented, the ALDEx2 package uses the uniform zero replacement value of 0.5. One purpose of this investigation was to determine if using one of the more rigorous zero replacement models from the zCompositions package would increase our selectivity because, as shown

in Table 1, these adjustments output non-uniform estimates of the underlying value of zero based on abundances of the same feature in different samples.

We applied the same seven methods to adjust the value of zero in the selex dataset, and an overview of the results are shown in the Dir column of Table 1. We again used Wilcoxon tests on the two groups and corrected the resulting P values using the Benjamini-Hochberg approach. Significance was assumed if the mean adjusted P value across all 128 instances was less than 0.05. In this analysis the substituted values of zero in the adjusted datasets serve as prior estimates of the range of values that zero could assume in each of the Dirichlet instances. The square root Bayesian multiplicative approach was incompatible with generating Dirichlet instances because many of the prior values that replaced zero generated Dirichlet posterior estimates that were not distinguishable from zero. Modelling uniform priors indicated that this occurred when the prior for zero was less than approximately 0.05. The remaining six approaches were compatible with the approach, and resulted in substantially smaller numbers of variants being identified as significantly different between the selected and non-selected groups. In this analysis, the Geometric Bayesian multiplicative, uniform replacement and uniform prior approaches were approximately similar, the count zero multiplicative approaches was nearly as selective, and the Bayes-Laplace approach was least selective.

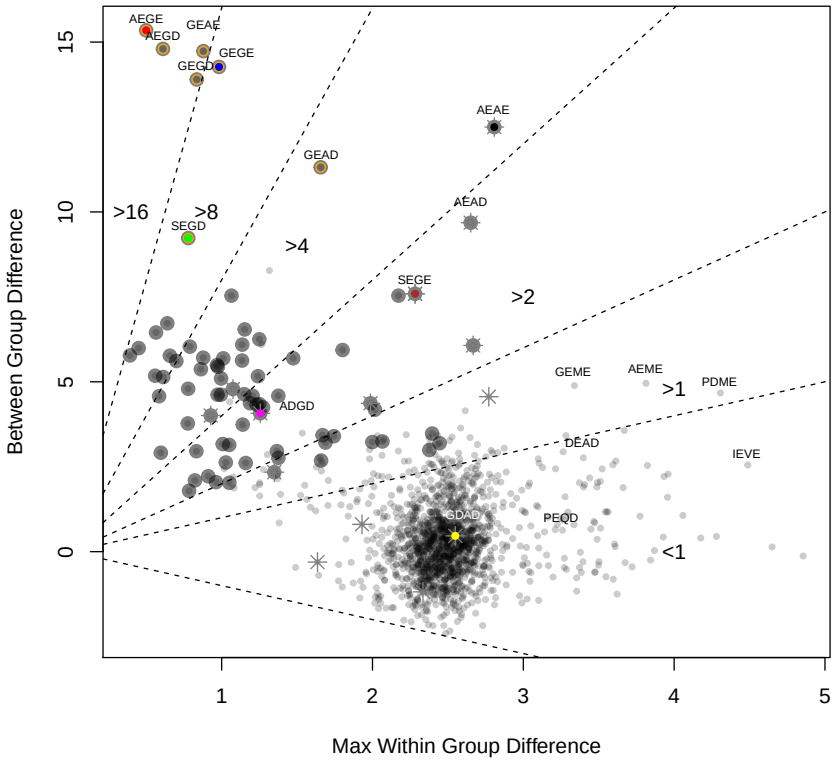

Figure 5: Variance-variance plot showing the median maximum centred log-ratio scaled difference within each group plotted vs. median between group difference for each variant. Dotted lines represent the approximate location of effect sizes, which is calculated as the median between to within group difference. Variants are coloured if their activity was validated *in vitro*, have a star if they failed to grow reliably in individual culture *in vitro*. Variants that exhibit a significant increase in abundance using the Wilcoxon test with a mean Benjamini-Hochberg adjusted P value of > 0.05 are shown as large grey dots. The analysis was done with a uniform prior of 0.5 applied to the dataset. Also shown are the six variants that were outliers on component 2 of the clr biplot in Figure 4

Figure 5 shows a variance-variance plot of the output from an analysis using the uniform prior replacement with a value of 0.5. Note that in this plot the vast majority of variants have an estimated between group difference of approximately zero, that only a small number have a positive between group difference, and no variants have a strong negative between group difference. This fits with the experimental design where variants could increase in abundance if the endonuclease was active, but not decrease in abundance if it was not. In this plot the variants with a mean Benjamini-Hochberg adjusted P value determined by an unpaired Wilcoxon test are indicated by the large grey dots. Variants that were tested for enzymatic activity *in vitro* are indicated by coloured central dots. Variants that had near wild type enzymatic activity *in vitro* are in the sector marked as > 8. There were four variants that had partial enzymatic activity *in vitro*. Many variants were tested for growth in pure culture. Variants AEAE, SEGE, ADGD and GDAD exhibited variable, partial growth under these conditions, with the GDAD variant exhibiting the weakest growth. Thus, there is a strong relationship between the observed results in this experiment, and the results observed *in vitro*.

There is remarkable concordance between the data viewed in this way, and the same data viewed as a point estimate in the compositional biplot. The biplot shows that the most distinguishing variants between the selected and non-selected groups, i.e., the variants that drive the separation on principle component 1, are those in the upper left quadrant of Figure 5. In addition, the variants that drive the separation on principle component 2, are those that exhibit the largest within-condition difference. For example, the GEME, AEME, PDME, IEVE, PEQD, and DEAD variants that were strongly separated on component 2 on the biplot, are among those with the largest within group difference on the variance-variance plot.

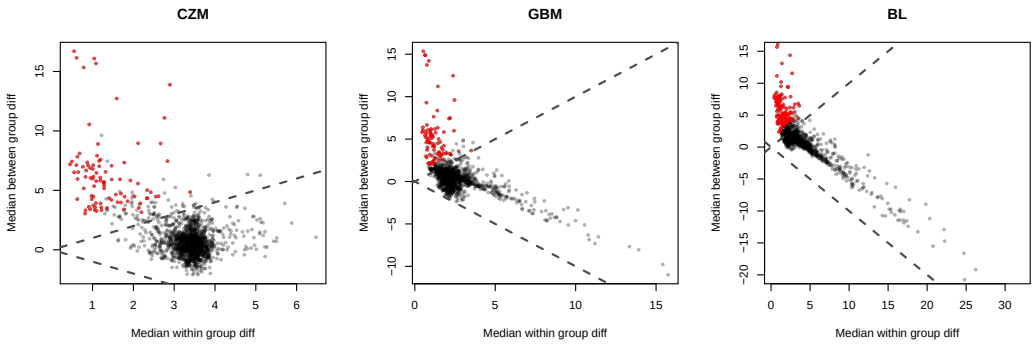

Figure 6: Variance-variance plots showing how the prior values for zero determined by the count zero multiplicative (CZM), geometric Bayesian multiplicative (GBM) and Bayes-Laplace (BL) methods alter the variation of the data. Red dots represent those that are called differential, black dots are not differential and the lines represent effect sizes of 1 and -1. The cutoff used was a Benjamini-Hochberg adjusted P value of 0.05 from an unpaired Wilcoxon test.

Finally, we examined the effect of the different zero replacement methods on the shape of the variance-variance plot to determine why these different approaches deliver slightly different results after Dirichlet sampling log-ratio transformation. As shown in Figure 6 all the prior estimation methods delivered similar differences between conditions for the true positive variants. These all exhibited an increase in abundance of about 2^{16} relative to their mean abundance in the unselected group. In particular, the CZM plot was remarkably similar to the plot that used a uniform prior of 0.5, with the major difference between the two approaches being a slight broadening of the within-group difference. This is perhaps not surprising since the prior values of zero using this approach are non-uniform in a narrow range of near 0.325. In contrast the non-uniform prior values for zero count variants from both the GBM and BL ranged over much larger values. The vast majority of values were between zero and one, but the GBM method had an average of 197.4 zero replacements that were greater than one,

and the BL had an average of 55 replacements that were greater than one. Examination of the variance-variance plots of these two approaches showed that between-group difference for many variants was not significant, but tended to be strongly negative. This result is incompatible with the known biology of the experiment, where no variant is expected to become less abundant than average in the selected dataset. Therefore, in this dataset, the geometric Bayesian multiplicative and the Bayes-Laplace substitution methods are distorting the underlying data. This distortion likely contributes to the greater number of variants identified as significantly different between the selected and non-selected groups.

3. Discussion

High throughput sequencing datasets are different from other types of datasets to which compositional approaches are often applied, but fall into the general class of 'count compositional' data. However, it is useful to remember that high throughput sequencing datasets result from random sampling of a large number of DNA fragments, and that the act of sequencing these DNA fragments on the instrument results in data that has the constant sum constraint. The estimation of the true abundance of genes or features with low counts exhibits a very large proportional error.

It is tempting to imagine that the large number of counts observed for a given sample, ranging from the thousands to billions, provides great precision in estimating the true values of the genes or OTUs (parts) being examined. This is an erroneous assumption because there are hundreds to thousands of parts in each sample, and many of the parts will be represented by zero reads in some samples. Thus, it is more useful to think of these data as *one instance of the data observed from a single random sample*. When thinking about the data in this way, the reads per part in each sample can be represented as prior values for a Bayesian estimation of their posteriors. The posterior distribution of the underlying abundance of each feature can be estimated by generating multiple instances of the data by sampling from a Dirichlet distribution.

As noted above, these datasets are necessarily very sparse, but in many cases the sparsity is informative. For example, in 16S rRNA gene sequencing it is difficult to argue that a particular taxonomic group would *never* be observed if we generated sufficient sequencing reads. As another example, gene expression is stochastic, and the number of transcripts for a given gene is observed not to be zero when large populations of cells are sampled, even if the gene in question is 'not expressed' (Munsky, Neuert, and van Oudenaarden 2012).

The centred log-ratio approach is intrinsically attractive in a biological context for two reasons. First, it can be intuitively explained to biologists as being similar to quantitative PCR, a familiar technique where the ratio between the gene of interest and a gene assumed to be at a constant level is determined. The centred log-ratio approach merely extends this analogy to the ratio between the gene of interest and all other genes in the system. Second, biologists understand that many of the processes that they study, cell growth, enzyme kinetics, etc, are exponential processes. Less well understood is that the underlying data is not 'set in stone' but actually represents a snapshot of what would have been observed had the experiment been done again.

A common criticism of using log-ratio approaches when analyzing such sparse data is the problem of zero observed counts. Structural zeros, those features that contain zero in every sample, are always excluded, and do not cause problems. However, count zeros that occur in one condition but not the other are problematic because log-ratio transformations cannot be performed when the underlying data contains one or more features with a zero value (Aitchison 1986). Much work has been put into this problem because of the prevalence of features with values of zero are common in many kinds of datasets. Several approaches have been developed to determine the best point estimate of the actual underlying value of zero in these datasets (Pawlowsky-Glahn, Egozcue, and Tolosana-Delgado 2015), and they are implemented in zCompositons R package (Palarea-Albaladejo and Martín-Fernández 2015).

Less work has been done modelling this in a Bayesian framework where the distribution of probable values for each variable are taken into account.

Here we have examined the effect of using various approaches to estimating the value of zero on both point estimates and Bayesian distributions derived from Dirichlet multinomial sampling. We have found that point estimates, whether modelled as proportions or centre log-ratio transformed values, cannot distinguish features that differ between conditions in a problematic dataset. We found that estimating the technical variation alone is also unsuitable. However, the combination of estimating technical variation and the centre log-ratio transformation provides a large increase in selectivity. We further observe that methods that generate priors in a narrow range give outputs that closely mimic a dataset derived from a differential growth experiment, and that methods that generate priors with broad ranges generate posterior distributions that are different from the known underlying distribution.

The selex dataset is an extreme example of the type of data that is analyzed by high throughput sequencing. It has a small number of features that exhibit a marked difference in abundance between conditions, and is very sparse. Other experimental designs will have much smaller difference in abundance of features. For example, in the case of RNA-seq it is more common to examine differential abundance of a small number of genes that are themselves relatively rare in the cell, and from carefully controlled experiments where the total number of input molecules is similar between conditions. This would be akin to comparing steps 1 and 2 in Figure 1B.1, where no gene or set of genes perturbs the system signficantly. In this simple case, any approach would likely give reasonable answers. However, comparing gene expression between cells from different tissues, or gene expression in RNA from environmental samples, would introduce extreme distortions in the underlying data and could give false positive and false negative results (Fernandes *et al.* 2013; Macklaim, Fernandes, Di Bella, Hammond, Reid, and Gloor 2013; Fernandes *et al.* 2014). In the case of 16S rRNA gene sequencing experiments, it is likely that many conditions would have wildly divergent underlying abundances because bacterial growth is an exponential process, and such samples are more difficult to analyze.

Acknowledgements

This work was supported by a grant to Greg Gloor by the National Science and Engineering Research Council of Canada.

Correspondence addresses of author(s) should be added at the end of the manuscript.

References

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

Benjamini Y, Hochberg Y (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 289–300.

Bottomly D, Walter NAR, Hunter JE, Darakjian P, Kawane S, Buck KJ, Searles RP, Mooney M, McWeeney SK, Hitzemann R (2011). "Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-seq and microarrays." *PLoS One*, **6**(3), e17820. doi:10.1371/journal.pone.0017820.

Ding T, Schloss PD (2014). "Dynamics and associations of microbial community types across the human body." *Nature*, **509**(7500), 357–60. doi:10.1038/nature13178.

Fernandes AD, Macklaim JM, Linn T, Reid G, Gloor GB (2013). "ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq." *PLoS ONE*, **8**(7), e67019.

Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB (2014). "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. doi:10.1186/2049-2618-2-15.

Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson G, Owen-Hughes T, Blaxter M, Barton GJ (2015). "Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment." *Bioinformatics*. doi:10.1093/bioinformatics/btv425.

Holmes I, Harris K, Quince C (2012). "Dirichlet multinomial mixtures: generative models for microbial metagenomics." *PLoS One*, **7**(2), e30126. doi:10.1371/journal.pone.0030126.

Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T, Codelli JA, Chow J, Reisman SE, Petrosino JF, Patterson PH, Mazmanian SK (2013). "Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders." *Cell*, **155**(7), 1451–63. doi:10.1016/j.cell.2013.11.024.

La Rosa PS, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD (2012). "Hypothesis testing and power calculations for taxonomic-based human microbiome data." *PLoS One*, **7**(12), e52078. doi:10.1371/journal.pone.0052078.

Lovell D, Müller W, Taylor J, Zwart A, Helliwell C, Pawlowsky-Glahn V, Buccianti A (2011). "Proportions, percentages, ppm: do the molecular biosciences treat compositional data right?" *Compositional Data Analysis: Theory and Applications*, pp. 193–207.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2014). "Proportionality: a valid alternative to correlation for relative data." *bioRxiv*, p. 008417.

Macklaim MJ, Fernandes DA, Di Bella MJ, Hammond JA, Reid G, Gloor GB (2013). "Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis." *Microbiome*, **1**, 15. doi:doi:10.1186/2049-2618-1-12.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008). "RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays." *Genome Res*, **18**(9), 1509–17. doi:10.1101/gr.079558.108.

Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J (2014). "Bayesian-multiplicative treatment of count zeros in compositional data sets." *Statistical Modelling*, **doi:10.1177/1471082X14535524**, 1:25.

McMurrough TA, Dickson RJ, Thibert SMF, Gloor GB, Edgell DR (2014). "Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues." *Proc Natl Acad Sci U S A*, **111**(23), E2376–83. doi:10.1073/pnas.1322352111.

Munsky B, Neuert G, van Oudenaarden A (2012). "Using gene expression noise to understand gene regulation." *Science*, **336**(6078), 183–7. doi:10.1126/science.1216379.

Palarea-Albaladejo J, Martín-Fernández JA (2015). "zCompositions — R package for multivariate imputation of left-censored data under a compositional approach." *Chemometrics and Intelligent Laboratory Systems*, **143**(0), 85 – 96. ISSN 0169-7439. doi:http://dx.doi.org/10.1016/j.chemolab.2015.02.019. URL http://www.sciencedirect.com/science/article/pii/S0169743915000490.

Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015). *Modeling and Analysis of Compositional Data*. John Wiley & Sons.

Smith AB, Maxwell A (2006). "A strand-passage conformation of DNA gyrase is required to allow the bacterial toxin, CcdB, to access its binding site." *Nucleic Acids Res*, **34**(17), 4667–76. doi:10.1093/nar/gkl636.

van den Boogaart KG, Tolosana-Delgado R (2008). ""compositions": A unified R package to analyze compositional data." *Computers & Geosciences*, **34**(4), 320 – 338. ISSN 0098-3004. doi:http://dx.doi.org/10.1016/j.cageo.2006.11.017. URL http://www.sciencedirect.com/science/article/pii/S009830040700101X.

Affiliation:

Gregory B. Gloor
Department of Biochemistry
The University of Western Ontario
London, Ontario, Canada
E-mail: ggloor@uwo.ca
URL: http://www.academicbiography.uwo.ca/profile.php?&n=ggloor