Instructions for authors - CoDaWork 2017

Jia R. Wu[1], **Jean M. Macklaim**[1], **and Gregory B. Gloor**[1,2]
[1]Dep't of Biochemistry, U. Western Ontario, London, Canada, N6A 5C1
[2]Dep't of Applied Mathematics, U. Western Ontario, London, Canada, N6A 5C1
gbgloor@gmail.com

Abstract

High throughput sequencing is a technology that allows for the generation of millions of reads of genomic data regarding a study of interest, and data from high throughput sequencing platforms are usually count compositions. Subsequent analysis of this data can yield information on transcription profiles, microbial diversity, or even cellular abundance in culture. These data have many pathologies: because of the high cost of acquisition the data are usually sparse, and often contain far fewer observations than variables. However, an under-appreciated pathology of these data are their often unbalanced nature: i.e, there is often be systematic variation between groups simply due to presence or absence of features, and this variation is important to the biological interpretation of the data. A simple example would be comparing transcriptomes of yeast cells with and without a gene knockout. This causes samples in the comparison groups to exhibit widely varying centres. Despite the compositional nature of sequencing data, the majority tools inappropriately model the the underlying features in sequencing data as linearly independent counts. This work extends a previously described log-ratio transformation method that allows for variable comparisons between samples in a compositional context. We demonstrate the pathology in several modelled and real unbalanced experimental designs that have a unidirectional direction of change to show how this dramatically causes both false negative and false positive inference in both traditional and compositional approaches. We then introduce several measures drawn from the RNA-seq and robust CoDa analysis fields to demonstrate how the pathologies can be addressed. An extreme example is presented where only the use of a predefined basis is appropriate. The transformations are implemented as an extension to a general compositional data analysis tool known as ALDEx2 or ANOVA Like Differential Expression.

Key words: ALDEx2, Bayesian estimation, sparse data, high throughput sequencing, robust estimation, qPCR

1 Introduction

High throughput sequencing (HTS) technology is used to generate information regarding the relative abundance of features. In these designs, DNA or RNA is isolated, a library is made from a sample of the nucleic acid, and a random sample of the library is sequenced on an instrument. The output is a set of short sequence tags, called reads, which are mapped to example sequences for each feature to generate a table of read counts per feature for every sample. Traditionally, samples comprise a set of features whose identity depends on the experimental design. For example, features are genes in the case of RNA-seq or metagenomic sequencing, or are operational taxonomic units (OTUs) when the objective is identifying microbial diversity.

These data are often analyzed by count based methods, such as negative binomial or zero-inflated Gaussian models, that assume the features are independent and identically distributed for statistical tests (Auer and Doerge 2010; Anders et al. 2013). However, the capacity of the instrument used for HTS imposes an arbitrary upper limit on the total number of reads observed. Thus, data collected from high throughput sequencing are compositions, and so counts per feature are not independent when collected in this way. Traditional tools do not address the compositional nature of HTS data (Fernandes et al. 2014; Gloor et al. 2016) and assume that the features are sufficiently independent when there are enough of them, or when they fulfill certain statistical properties (Weiss et al. 2016), although much effort is placed on 'normalizing' the data to have a consistent read depth (Sun et al. 2013; McMurdie and Holmes 2014).

Formally, Aitchison (1986) defined a composition as a vector $\boldsymbol{x}$ of positive values $x1 \ldots xD$ whose components sum to an arbitrary constrained constant c. The constant sum is usually set to 1, leading to these vectors being of unit length. Absolute values of components in a composition are uninformative, and the only information provided in compositional data are the relative magnitudes of the ratios between the pairs of components, and Aitchison demonstrated that compositional data can be efficiently analyzed by log-ratios between the parts, since these data carry only relative information between components (Aitchison 1986).

One way of satisfying the need to examine the ratios between parts is to use the centred-log-ratio (CLR) transformation proposed by Aitchison, defined as:

$$\begin{aligned} clr_x &= log\big(\frac{x_i}{g(x)}\big)_{i=1,\ldots,D} \\ \text{where } clr_x &= \text{A composition transformed by CLR} \\ x_i &= \text{A feature of the non-transformed composition (x)} \\ D &= \text{The number of features of x} \\ g(x) &= \text{Geometric mean of D features of x} \end{aligned} \tag{1}$$

Since all arbitrary sums are the same this led to the concept of a composition as an equivalence class where composition $\mathbf{x}$ can be scaled into an identical composition $\boldsymbol{y}$ by multiplication of a constant α (Barceló-Vidal et al. 2001). Thus, we can discuss any composition as being a proportion scaled by α without loss of precision. The CLR, and indeed any ratio-based method is scale-invariant because if the parts of $\boldsymbol{x}$ are counts with $\alpha = N$ reads, then:

$$clr_x = log\big(\frac{Nx_i}{g(Nx)}\big) = log\big(\frac{x_i}{g(x)}\big). \tag{2}$$

Aitchison (1986) also defined the ALR, the additive log-ratio as:

$$\begin{aligned} alr_x &= log(\frac{x_i}{x_D})_{i=1,\ldots,D-1} \\ \text{where } alr_x &= \text{A composition transformed by CLR} \\ x_i &= \text{A feature of the non-transformed composition (x)} \\ D &= \text{The number of features of x} \\ x_D &= \text{The D}^{\text{th}} \text{ feature of x} \end{aligned} \tag{3}$$

In the ALR, the log-ratio is determined by using one, presumed invariant, feature as the denominator, and by convention this is written as the last feature. The ALR is surprisingly similar to the qPCR approach in common use in molecular biology to measure relative abundance of molecules in a mixture. Here, the species of unknown abundance is determined relative to the abundance of a species of known abundance, which can be a housekeeping gene or can be a DNA molecule of known amount added to the mixture. It is well known that the relative abundance measure will change when a different DNA species is used as the denominator, leading to the use of multiple species in some cases. In some ways, the ALR and CLR can be viewed as the two limits of a continuum of incomplete knowledge about the proper internal standard by which relative abundance should be judged. The ALR uses one presumed constant feature as the internal standard; while the CLR presumes that the majority of features are not changed, leading to the geometric mean of all features as the internal standard. We can however, choose to use combinations of other features as presumed invariant features.

For convenience, the analyses and discussion here are drawn from RNA-seq, or transcriptome, experiments where the data are exploring the relative abundance of gene species that are found in cells in an environment. However, the results and conclusions apply without restriction to metagenomic sequencing, microbial diversity sampling (by 16S rRNA gene sequencing) or to in-vitro selection experiments (Fernandes et al. 2014).

1.0.1 Sparsity and asymmetry in HTS data

It is common for HTS data to be sparse, that is, for a given sample to contain features with one or more counts of 0. Furthermore, the sparsity of the samples is affected by the total number of reads obtained for each sample. It is common for samples in a transcriptome to contain thousands of features each of which may have a dynamic range of over 4 logs. In many cases a transcriptome dataset will be composed of several groups, where the expression of a gene (feature) is so low that it is below the detection limit in one group, and very high in another group. The expression of genes in biological systems is linked, and some genes control the expression of other genes, either by increasing or decreasing their relative abundance. Furthermore, the cell has a built-in control system whereby gene expression itself appears to be a composition, that is, the expression levels of all genes in a cell are constrained by an absolute upper bound (Scott et al. 2010). Note however, that this does not mean that a population of cells will have total gene expression with an upper bound, since the cells themselves can change in both absolute and relative abundance in a mixture.

The potential for a change in cell number and the potential for expression linkage of genes in biological systems, coupled with the inability to collect a large enough number of sequence reads, can lead to experiments with an apparent or a real asymmetry in relative abundance of many genes or features. Such an asymmetry will result in a mis-centering of the data when conducting differential abundance analyses, largely, but not exclusively because of the effect on the geometric mean upon which the CLR depends.

2 Statement of the Problem

The assumption being made when using the CLR transformation to identify features that vary between groups is that most features are either invariant or varying at random. This assumption is broken if there is any sort of systematic variation between groups. For example, when comparing microbial diversity between sampling sites or conditions, organisms present in one sub-site or condition may be absent from another (Macklaim, Clemente, Knight, Gloor, and Reid 2015; Hummelen, Fernandes, Macklaim, Dickson, Changalucha, Gloor, and Reid 2010; Gajer, Brotman, Bai, Sakamoto, Schütte, Zhong, Koenig, Fu, Ma, Zhou, Abdo, Forney, and Ravel 2012). In the case of multi-organism RNA-seq (meta-rna-sea), organisms that reside in one condition may have a different expression profile and abundance than those that reside in a second condition (Macklaim, Fernandes, Di Bella, Hammond, Reid, and Gloor 2013). In the case of a single-organism RNA-seq, samples from one condition may contain more genes than samples from another condition (Lang and Johnson 2015; Peng, Hao, Liu, Wang, Ma, Yang, Xie, and Li 2014; Zhao, Chen, Xiong, Xu, Lan, Wang, Yao, Bai, Liu, Meng, Zhang, Sun, Zhao, Bai, Cheng, Chen, Ye, and Xu 2013). These differences are represented by either zeroes or low count features that occur in only one group.

It will be demonstrated that exaggerated differences can be achieved with seemingly little variation within simulated data. Additionally, zeroes are problematic as they cannot be represented in logarithmic space. Therefore, prior to the log ratio transformation, zero count features are discarded if present in exclusive or conditions, or adjusted with an uninformative prior. The goal is to identify a non-deterministic geometric mean that best represents each sample so features can be accurately compared.

2.1 Simulated Data

RNA-Seq data was simulated for benchmarking purposes. Assemblies from *Saccharomyces cerevisiae* uid 128 and a complete reference genome of *S. cerevisiae* were drawn from GenBank. The R package `polyester v1.10.0` was used to simulate an RNA-Seq experiment with 2 groups of 10 replicates with 20x sequencing coverage across the simulation experiment. For the base dataset, forty genes were chosen at random to have 2-5 fold expression difference, and these were apportioned equally between the two groups. These 40 features serve as an internal control of true positives for each dataset as their fold changes are explicit and should always be displayed as differentially expressed. We used bowtie2 to align the simulated reads to the *S. cerevisiae* reference genome. Labeling each group as A and B is arbitrary and hence samples s_1, s_2, ..., s_10 belong to condition A, and samples s_11, s_12, ..., s_20 belong to condition B. There are a total of 6349 features in these simulated data, and only the first 1000 genes were chosen for the majority of the figures. On average, samples in the unmodified base dataset have a simulated read depth of approximately 870000 reads.

An additional 98 datasets derived from the base dataset are generated in order to benchmark how well the IQR (robust?) log-ratio transformation supports the assumption that most features are invariant and unchanging. As the original dataset has approximately 6000 nonzero features, 60 features are incrementally removed from the samples of condition A in each simulated dataset for a resultant set of datasets with sparsity ranging from 0% to 98% sparse in condition A.

2.2 An illustration

In its current implementation, ALDEx2 computes a per-sample geometric mean for the features and declares this as the baseline for feature comparisons. Figure 1:symmetric is an effect plot demonstrating that the 40 internal control features are found to be both statistically significant and to have an effect size greater than 1 between the two groups, and the remainder of the features have very small difference, and correspondingly have an effect size much less than 1. The inset histogram shows the distribution of difference values between groups A and B, and it is clear that

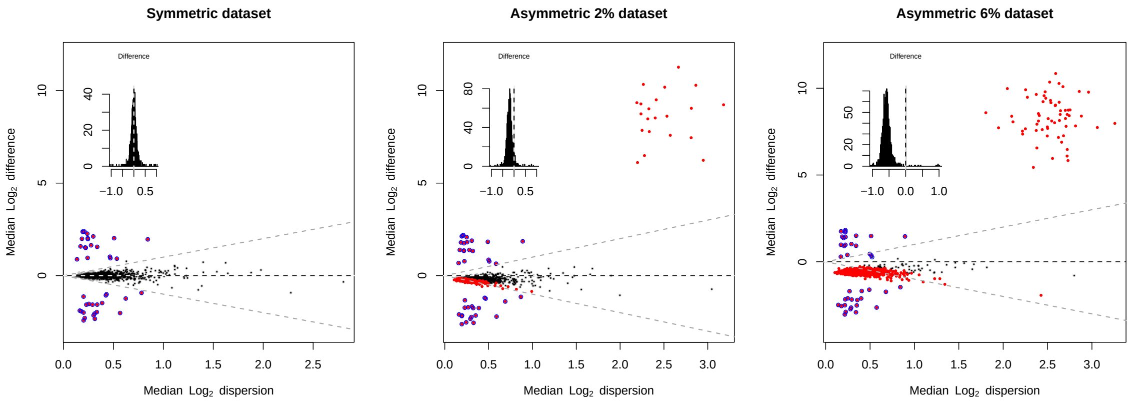

Figure 1: Effect plots of simulated asymmetric data that illustrates the problem. The effect plots show the difference between two conditions simulated RNA-seq data with 1000 genes where 40 are modelled to have true difference between groups. Each point is a feature (gene), they are coloured in black if not different between groups, red if identified as being statistically significantly different between groups, and red with a blue circle if they are one of the 40 genes modelled to be 'true positives'. The red points in the top right quadrant are the genes modelled to be asymmetrically variable between groups. These are also true positive features, but are not part of the initial modelled true positives. The inset histograms show the distribution of differences between groups as calculated by ALDEx2, and the vertical line shows a difference of 0.

it is symmetric and has a location of 0. However, the introduction of small amounts of asymmetry strongly affect the results. The asymmetric 2% dataset is the base dataset modified by setting the count value to 0 for 20 features chosen at random from Group A, and likewise the asymmetric 6% dataset has 60 features from group A set to 0. It is apparent from the two right panels of Figure **??** that this low level of simulated asymmetry breaks the assumption that most features are invariant, and the location of the difference between groups is no longer at the origin. Supplementary Figure 1 shows compositional biplots of the same data, and here it is obvious that the centre of the data is not at the origin. Thus, the small amount of asymmetry is shifting the geometric mean of the data, causing bias. Thus, if there are a large proportion of features in a sample that do not follow the central tendency of the data, the geometric mean can be unreliable as a baseline. It is unlikely that the problem will be as easy to diagnose in real data as in simulated data.

As can be seen in Equation 1, the major determinant of the centre of a sample is the denominator, or basis, used to compute the CLR. Thus, one obvious approach to solving the problem is to compute the geometric mean of a subset of features that are more representative of the central tendency of the data, and to use this value as the denominator in the equation. We next examined three different approaches to identifying the features to include in the denominator.

The first approach was to identify those features that have variance which is most typical across all the samples. This was done by calculating the variance of each feature after CLR transformation of the data, then identify tin those features with a variance between the first and third quartiles of the datase: this is referred to as the *iqlr* set of features. Thus, equation 1 becomes modified to be:

$$\begin{aligned} IQLR_x &= log\Big(\frac{x_i}{g(iqlr)}\Big)_{i=1,\ldots,D} \\ \text{where } IQLR_x &= \text{The transformed composition} \\ x_i &= \text{A feature of the non-transformed composition (x)} \\ D &= \text{The number of features of x} \\ g(iqlr) &= \text{Geometric mean of the iqlr features of x} \end{aligned} \tag{4}$$

This transformation in Equation 4 is termed the IQLR, transformation, since the denominator for Equation 1 is the features from the inter-quartile range of the variance. The results of this method are shown in Figure 2:IQLR on the Asymmetric 2% dataset. The IQLR transformation restores the centre of the dataset to the origin, and the proper set of features is identified as being both significant, and having an effect size greater than 1.

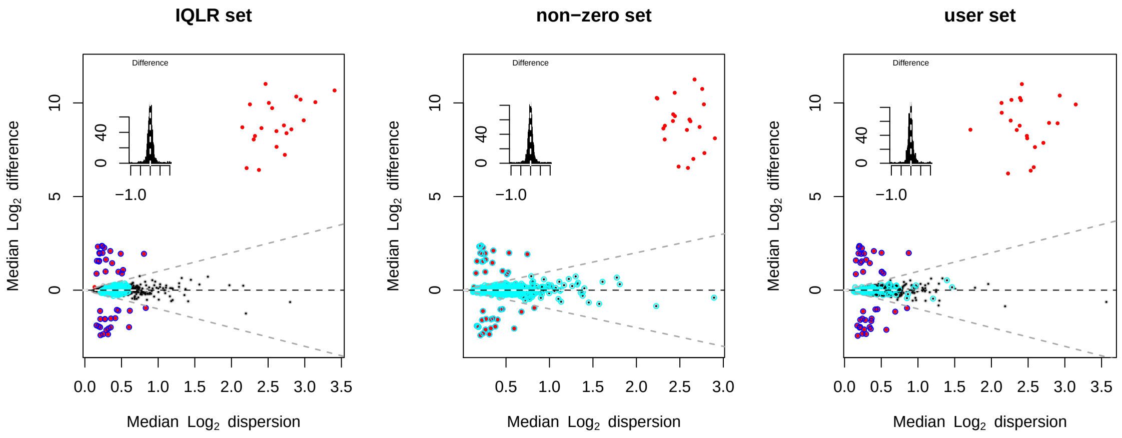

Figure 2: Effect plots of simulated asymmetric data with transformations, based on the CLR that can result in a more accurate centring of the data.

The second approach uses as the denominator the set of non-zero features in each group. Thus, in this case the geometric mean of group A and group B are based on different, but potentially overlapping, sets of features. As shown in Figure 2:non-zero, this method also restores the centre of the data to the origin and identifies the proper set of features as differential.

The final approach uses as the denominator a set of user-defined features. Thus, the user could choose to use one feature, in which case the approach will be the same as the ALR, or all features, in which case the result would be the same as the CLR, or a subset chosen based upon prior information. In the case of RNA-seq, this could include the set of genes involved in translation as these have been shown to be relatively stable across multiple conditions (Scott et al. 2010), and can be presumed to represent a set of genes that are representative of the overall growth state of the cell.

2.3 Limitations of the approaches

We next explored the limitations of these approaches in two ways. First, we examined how the sparsity affected the ability of the approaches to properly centre the data when dealing with asymmetric sparse data. Figure **??**:All shows that the centre of the CLR-transformed data deviates from 0 when the data have even very small amounts of asymmetric sparsity. However, both the IQLR and Zero approaches are able to properly centre the data when large amounts of asymmetric sparsity are present. The breakdown point for the IQLR method is 25% sparsity in this dataset, and is approximately 45% sparsity for the Zero basis approach. Both, are obviously better choices that is the CLR when asymmetric sparsity is present.

Next, rather than modelling sparsity, we modelled low-count asymmetry by changing the asymmetry to a defined count of 1: note that any defined count will behave similarly. In a biological context is entirely reasonable that a low-count asymmetry could occur rather than a sparsity asymmetry. For example, the default gene expression condition for many genes is low-level expression, and the inclusion of a transcriptional activator could increase expression of many genes from very low

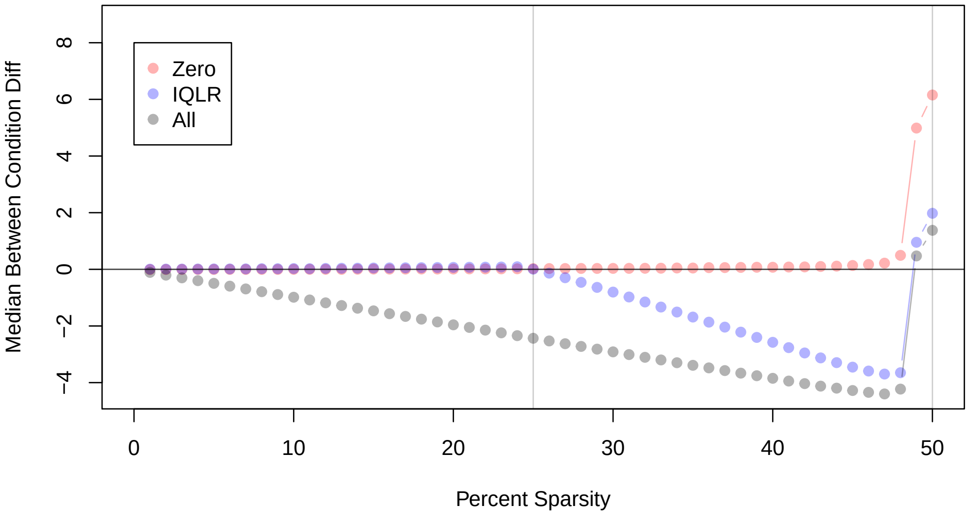

Figure 3: A figure depicting the behaviour of the transformations for datasets with varying sparsity. Each point represents the median between condition difference for a given transformation given a dataset with a specified sparsity. Points closer to the location y=0 are favourable as it implies that the transformations support the assumption that the majority of features are invariant and unchanging. The original ALDEx2 transformation fails as soon as sparsity is introduced. Inter Quantile Log Ratio transformation is effective on datasets with up to 25% sparsity from zeroes or extreme count features. Zero-removal transformation is effective on datasets with up to 50% sparsity exclusively from zeroes. Zero removal is an extended special case of IQR centering as it only affects datasets whose sparsity is due to zeroes.

expression to very high expression. Thus, we would have a low-count asymmetry where expression of many genes changes from very low to very high.

Remember that the final electronic paper –without limit in the number of pages– will be due **before May 4, 2017**. It will be sent in PDF format. It will be published in the CoDaWork'17 CD (with ISBN) and will be available in the sessions of the workshop. **Only papers of participants registered before May 4, 2017 will be included in the proceedings CD**.

3 First level headings

The workshop on Compositional Data is intended as a forum for discussion of important points related to the statistical treatment and modelling of compositional data, as well as their applications and interpretation. The goal of such discussions is to get some insight into the most appealing future lines of research in the field.

3.1 Second level headings

In order to meet this general but clear goal, we intend to bring together a significant number of specialists, users and interested people to collect critical contributions and start a stimulating brainstorming.

3.1.1 Third level headings

The Introductory course to statistical analysis of compositional data will work from a variety of practical compositonal problems. Different case studies will be presented and analyzed using CoDaPack, a freeware software based on EXCEL. This software is oriented to users coming from the applied sciences. No extensive background in using computer packages is required.

4 Citations, figures and references

4.1 Citations in text

Citations within the text should include the author's last name and year: "The air conditioner data (Proschan, 1963) ...", or when the author is used as a noun in the sentence: "Proschan (1963) presented a data set ...".

In text, captions, and table headings, list all authors if two or fewer, and just the first author followed by "and others" for more. Examples:

(Jones and Johnson, 1986; Emmanuel and others, 1989)

or

Emmanuel and others (1989) showed that ..., whereas Jones and Johnson (1986) found that ...

When giving a quote or referring to a specific fact or formula in a book or from an article of more than 8 pages, the citation should include the page number. Examples:

(Chayes, 1956, p. 55) or (Matheron, 1975, p. 229).

Page numbers should not be given in the text when referring to the work as a whole. As with figures, you do not need to direct the reader to "see" a citation to the literature. Be sure your references are accurate and formatted correctly.

4.2 Figures

All figures should be inserted within the text exactly as they should appear when printed. All figures must be centered. Figure number and caption always appear below the figure. Leave 2 line spaces between the figure and the caption.

Figure 4: This is a figure caption.

When you refer to an illustration, capitalize and spell out the word "Figure" if not in parenthesis, as in "Figure 2 shows that the distribution of permeability is skewed ..."; or abbreviate if in parenthesis, as in "The distribution of permeability is skewed (Fig. 2) ...".

If you have multiple parts in a figure, then label them with capital letters A,B,C, etc. Refer to them in the text as Figure 2A, or (Fig. 2A). In captions, follow this example:

Figure 5: Density functions: (A) Permeability; (B) Porosity.

4.3 Tables

All tables must be centered, neat, clean, and legible. Table number and title always appear above the table (see the example below). Use one line space before the table title, one line space after the table title, and one line space after the table.

Table 1: This is an example of a table.

Income	$42.94
Expenses	$26.12
Rest	$16.82

The word “Table” should be capitalized, and not abbreviated even in parentheses.

4.4 Equations

The word “Equation” should be capitalized and spelled out in the text, as in “It follows from Equation (3) that ...” but capitalized and abbreviated in parenthesis, as in “It follows [Eq. (3)] that ...”. If you use any other word to refer to an equation, such as “expression” or “relationship”, do not capitalize.

Acknowledgements and appendices

Use non-numbered first level headings for the acknowledgements. They should follow text, and precede the list of references. Appendices follow references, and should be headed “**Appendix** A” etc. if more than one.

References

REFERENCES

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

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth, W. Huber, and M. D. Robinson (2013, Sep). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat Protoc 8*(9), 1765–86.

Auer, P. L. and R. W. Doerge (2010, Jun). Statistical design and analysis of RNA sequencing data. *Genetics 185*(2), 405–16.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (2001). Mathematical foundations of compositional data analysis. In *Proceedings of IAMG*, Volume 1, pp. 1–20.

Fernandes, A. D., J. N. Reid, J. M. Macklaim, T. A. McMurrough, D. R. Edgell, and G. B. Gloor (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*, 15.1–15.13.

Gajer, P., R. M. Brotman, G. Bai, J. Sakamoto, U. M. E. Schütte, X. Zhong, S. S. K. Koenig, L. Fu, Z. S. Ma, X. Zhou, Z. Abdo, L. J. Forney, and J. Ravel (2012, May). Temporal dynamics of the human vaginal microbiota. *Sci Transl Med 4*(132), 132ra52.

Gloor, G. B., J. M. Macklaim, M. Vu, and A. D. Fernandes (2016, September). Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian Journal of Statistics 45*, 73–87.

Hummelen, R., A. D. Fernandes, J. M. Macklaim, R. J. Dickson, J. Changalucha, G. B. Gloor, and G. Reid (2010). Deep sequencing of the vaginal microbiota of women with HIV. *PLoS One 5*(8), e12078.

Lang, K. S. and T. J. Johnson (2015, Jul). Transcriptome modulations due to a/c2 plasmid acquisition. *Plasmid 80*, 83–9.

Macklaim, J. M., J. C. Clemente, R. Knight, G. B. Gloor, and G. Reid (2015). Changes in vaginal microbiota following antimicrobial and probiotic therapy. *Microb Ecol Health Dis 26*, 27799.

Macklaim, M. J., D. A. Fernandes, M. J. Di Bella, J.-A. Hammond, G. Reid, and G. B. Gloor (2013). Comparative meta-RNA-seq of the vaginal microbiota and differential expression by *Lactobacillus iners* in health and dysbiosis. *Microbiome 1*, 15.

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

Peng, J., B. Hao, L. Liu, S. Wang, B. Ma, Y. Yang, F. Xie, and Y. Li (2014). Rna-seq and microarrays analyses reveal global differential transcriptomes of mesorhizobium huakuii 7653r between bacteroids and free-living cells. *PLoS One 9*(4), e93626.

Scott, M., C. W. Gunderson, E. M. Mateescu, Z. Zhang, and T. Hwa (2010, Nov). Interdependence of cell growth and gene expression: origins and consequences. *Science 330*(6007), 1099–102.

Sun, J., T. Nishiyama, K. Shimizu, and K. Kadota (2013). TCC: an R package for comparing tag count data with robust normalization strategies. *BMC Bioinformatics 14*, 219.1–219.13.

Weiss, S., W. Van Treuren, C. Lozupone, K. Faust, J. Friedman, Y. Deng, L. C. Xia, Z. Z. Xu, L. Ursell, E. J. Alm, A. Birmingham, J. A. Cram, J. A. Fuhrman, J. Raes, F. Sun, J. Zhou, and R. Knight (2016, Jul). Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. *ISME J 10*(7), 1669–81.

Zhao, H., C. Chen, Y. Xiong, X. Xu, R. Lan, H. Wang, X. Yao, X. Bai, X. Liu, Q. Meng, X. Zhang, H. Sun, A. Zhao, X. Bai, Y. Cheng, Q. Chen, C. Ye, and J. Xu (2013). Global transcriptional and phenotypic analyses of Escherichia coli O157:H7 strain Xuzhou21 and its pO157_Sal cured mutant. *PLoS One 8*(5), e65466.