Metadata of the chapter that will be visualized in SpringerLink

Book Title	Advances in Compositional Data Analysis	
Series Title		
Chapter Title	Finding the Centre: Compositional Asymmetry in High-Throughput Sequencing Datasets	
Copyright Year	2021	
Copyright HolderName	Springer Nature Switzerland AG	
Author	Family Name	**Wu**
	Particle	
	Given Name	**Jia R.**
	Prefix	
	Suffix	
	Role	
	Division	Department of Computer Science
	Organization	University of Waterloo
	Address	Waterloo, Canada
	Email	jr2wu@uwaterloo.ca
Author	Family Name	**Macklaim**
	Particle	
	Given Name	**Jean M.**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Department of Biochemistry
	Address	London, ON, Canada
	Email	jean.macklaim@gmail.com
Author	Family Name	**Genge**
	Particle	
	Given Name	**Briana L.**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Department of Biochemistry
	Address	London, ON, Canada
	Email	bgenge3@gmail.com
Corresponding Author	Family Name	**Gloor**
	Particle	
	Given Name	**Gregory B.**
	Prefix	
	Suffix	

	Role	
	Division	
	Organization	Department of Biochemistry
	Address	London, ON, Canada
	Email	ggloor@uwo.ca

Abstract	High-throughput sequencing datasets comprise millions of reads of genomic data and can be modelled as count compositions. These data are used for transcription profiles, microbial diversity, or relative cellular abundance in culture. The data are sparse and high dimensional. Moreover, they are often unbalanced, i.e. there is often systematic variation between groups due to presence or absence of features, and this variation is important to the biological interpretation of the data. The imbalance causes samples in the comparison groups to exhibit varying centres contributing to false positive and false negative identifications. Here, we extend the centred log-ratio transformation method used for the comparison of differential relative abundance between two groups in a Bayesian compositional context. We demonstrate the pathology in modelled and real unbalanced experimental designs to show how this causes both false negative and false positive inference. We examined four approaches to identify denominator features, and tested them with different proportions of modelled asymmetry; two were relatively robust, and recommended. We recommend the 'LVHA' transformation for asymmetric transcriptome datasets, and the 'IQLR' method for all other datasets when using the `ALDEx2` tool available on Bioconductor.

Finding the Centre: Compositional Asymmetry in High-Throughput Sequencing Datasets

Jia R. Wu, Jean M. Macklaim, Briana L. Genge, and Gregory B. Gloor

Abstract High-throughput sequencing datasets comprise millions of reads of genomic data and can be modelled as count compositions. These data are used for transcription profiles, microbial diversity, or relative cellular abundance in culture. The data are sparse and high dimensional. Moreover, they are often unbalanced, i.e. there is often systematic variation between groups due to presence or absence of features, and this variation is important to the biological interpretation of the data. The imbalance causes samples in the comparison groups to exhibit varying centres contributing to false positive and false negative identifications. Here, we extend the centred log-ratio transformation method used for the comparison of differential relative abundance between two groups in a Bayesian compositional context. We demonstrate the pathology in modelled and real unbalanced experimental designs to show how this causes both false negative and false positive inference. We examined four approaches to identify denominator features, and tested them with different proportions of modelled asymmetry; two were relatively robust, and recommended. We recommend the 'LVHA' transformation for asymmetric transcriptome datasets, and the 'IQLR' method for all other datasets when using the `ALDEx2` tool available on Bioconductor.

J. R. Wu
Department of Computer Science, University of Waterloo, Waterloo, Canada
e-mail: jr2wu@uwaterloo.ca

J. M. Macklaim · B. L. Genge · G. B. Gloor (✉)
Department of Biochemistry, London, ON, Canada
e-mail: ggloor@uwo.ca

J. M. Macklaim
e-mail: jean.macklaim@gmail.com

B. L. Genge
e-mail: bgenge3@gmail.com

P. Filzmoser et al. (eds.), *Advances in Compositional Data Analysis*,
https://doi.org/10.1007/978-3-030-71175-7_17

1 Background

AQ1 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 reference sequences for annotation of each feature that generates 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.

The instruments used for HTS have an upper limit on the total number of reads delivered, for example, the same library sequenced on an Illumina MiSeq or an Illumina NextSeq will deliver approximately 20 million or 400 million reads. HTS instruments thus deliver a fixed-sum random sample of the sequences that are in the library, which itself is a random sample of what was in the environment.

Data from HTS instruments are usually analysed by count-based methods which use a negative binomial or zero-inflated Gaussian model, and assume the features are independent and identically distributed for statistical tests, and that the majority of features are invariant [?,?]. In addition, the variance in HTS data is usually larger than the mean (overdispersed), and negative binomial and similar models are generally adequate when the underlying assumptions are approximated.

Many HTS datasets do not fit the usual statistical models. For example, in meta-transcriptomic datasets, the relative abundance of both organisms and of the genes expressed by the organisms can both change independently, while in metagenomic datasets there is no a priori reason for feature counts to follow any particular statistical model. The assumption that the majority of features are invariant 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 et al. 2015; Hummelen et al. 2010; Gajer et al. 2012). In the case of multi-organism RNA-seq (meta-transcriptomics), organisms resident in one condition may have a different expression profile and abundance than those resident in a second condition (Macklaim et al. 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 et al. 2014; Zhao et al. 2013; Gierliński et al. 2015). These differences are represented by either zeros or low-count features that occur systematically in only one group. Meta-transcriptomic datasets are thus some of the most difficult experiments to analyse (Macklaim et al. 2013; Fernandes et al. 2013) since the datasets do not fit multiple assumptions of existing tools. We have shown previously AQ2 that existing count-based tools fail to give sensible answers in meta-transcriptome and other datasets that exhibit asymmetry (Fernandes et al. 2013, 2014; Gloor et al. 2016b).

1.1 A Ratio-Based Approach to Analysing HTS

Another approach to analyse data taking into account the fixed capacity of the instrument and the random sampling of the library (and the environment) is to model the data as a multivariate probability distribution and to examine the result with the principles of compositional data analysis (Aitchison 1986; Fernandes et al. 2013). This method is instantiated in the `ALDEx2` Bioconductor package. Conceptually, the approach is similar to sequencing many different technical replicates of each of the underlying biological replicates, treating the resulting data as ratios of the features and reporting the mean value of the technical replicates. This approach has been found to be a generally useful tool that works for meta-transcriptomic datasets (Macklaim et al. 2013) and translates to many different experimental designs (Fernandes et al. 2014; McMurrough et al. 2014; Gloor et al. 2017; Macklaim and Gloor 2018; Almeida et al. 2019).

The basis for the approach used by `ALDEx2` is the centred log-ratio (CLR) transformation proposed by Aitchison (1986) and defined below. However, this approach requires that the denominator used to calculate the CLR be comparable across all samples. The ratio-based approach is similar to relative quantitative PCR (qPCR), a method in common use in molecular biology that measures relative abundance of molecules in a mixture (Thellin et al. 1999; Vandesompele et al. 2002) and is often used as the gold standard to validate HTS results. In this type of qPCR, the feature of unknown abundance is determined relative to the abundance of a feature of (presumed) known abundance. This standard can be a housekeeping gene or can be a DNA molecule of known amount, a so-called spike-in molecule, added to the mixture. It is well known that the relative abundance measure will change when a different spike-in species is used as the denominator, leading to the use of multiple (presumed invariant) species in some cases. Thus, one shortfall of any ratio-based approach is to determine the appropriate denominator.

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 miscentring 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. The asymmetry will also affect the scale invariance of the data, since a value of 0 is not scaled when multiplied by a constant. Note that it is also entirely possible for the dataset *as a whole* to be centred, but for the particular comparison of interest to not be centred. This could arise because of a systematic experimental bias that is unknown to the investigator.

For convenience, and because the datasets are generally very complex, the analyses and discussion here are drawn from RNA-seq, or transcriptome, experiments where the data are exploring the relative abundance of features that are gene transcripts found in cells in an environment. However, the examples, 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; McMurrough et al. 2014; Gloor et al. 2017).

1.2 Choosing the Denominator

The basic principle of compositional data analysis as developed by Aitchison is to convert the data into log-ratios between the features for each sample (Aitchison 1986). Formally, Aitchison defined a composition as a vector $\boldsymbol{x}$ of positive values $x_1 \ldots x_D$ whose features sum to an arbitrary constrained constant α. Absolute values of features in a composition are uninformative, and so the only information available in compositional data are the relative magnitudes, or the ratios between the pairs of components. For example, the only knowledge available is that the gene 1:gene 2 ratio is 5, but the absolute abundance of either is unavailable. This is the case in HTS since the count observed for a feature contains no information regarding the absolute number of molecules in either the sequencing library or the environment, although the magnitude of the count contains information on the precision of the estimate.

Data collected from high-throughput sequencing are count compositions (Fernandes et al. 2014; Gloor et al. 2016b). Count-based tools do not address the compositional nature of HTS data (Gloor et al. 2016b; Fernandes et al. 2014) and assume that the features are sufficiently independent when there are enough of them, or when they fulfill certain statistical properties [?]. Much effort is placed on 'normalizing' the data to have a consistent read depth instead of treating the data as compositional [?,?].

However, since all read count totals from a machine are arbitrary we should treat the data as an equivalence class where the composition contained in vector $\mathbf{x}$ can be scaled into an identical composition $\boldsymbol{y}$ by multiplication of a constant α (Barceló-Vidal et al. 2001). Thus, in the ideal case, we can discuss any composition as being a probability vector scaled by α without loss of precision (Fernandes et al. 2013).

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

$$\mathbf{x}_{clr} = log\left(\frac{x_i}{G(\mathbf{x})}\right)_{i=1\ldots D}, \tag{1}$$

where $G(\mathbf{x}) =$ the geometric mean of the D features of $\mathbf{x}$. When interpreting CLR-transformed values, care must be taken to ensure that the analyst understands that any changes observed are always relative to the denominator of the CLR in each sample.

AQ3 The CLR, and any other ratio-based method is, at least in theory, scale invariant because if the parts of $\boldsymbol{x}$ are counts with $\alpha = N$ reads, then

$$\mathbf{x}_{i,clr} = log\left(\frac{\alpha x_i}{G(\alpha\mathbf{x})}\right) = log\left(\frac{x_i}{G(\mathbf{x})}\right). \tag{2}$$

The important caveat that limits this ideal situation when dealing with high-throughput sequencing data is that the total read count, α, for each sample should be similar. The CLR is the default denominator used by `ALDEx2`. In practical terms, `ALDEx2` handles the problem of varying read depth by constructing a posterior probability distribution of the original count matrix (Fernandes et al. 2013). In this approach, features in low-count samples have broader probability distributions than features in high-count samples. Thus, the precision of the estimates varies when samples have different read depths.

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

$$\mathbf{x}_{alr} = log(\frac{x_i}{x_D})_{i=1\ldots D-1} \tag{3}$$

where, following from above, $\boldsymbol{x}_{alr}$ is the composition transformed by ALR. When using the ALR the denominator is the D^{th} feature of $\boldsymbol{x}$, which by convention is the feature chosen to be constant.

In the ALR, the log-ratio is determined by selecting one presumed invariant feature as the denominator, and so the ALR is similar to the relative qPCR approach in common use in molecular biology. The ALR and CLR can be viewed as the two limits of a continuum of incomplete knowledge about the proper internal standard, or basis, by which relative abundance should be judged. The ALR uses one presumed constant feature as the basis, while the CLR presumes that majority of features are not changed, leading to the use of the geometric mean of all features as the basis. Clearly, neither of these are satisfactory. We can, however, choose to use combinations of other features as the basis for comparison.

1.3 HTS Data Are Sparse

It is common for HTS data to be sparse, that is, for a given sample to contain features with counts of 0. One limitation of the log-ratio approach is that ratios have no meaning if the denominator contains a 0 value. The sparsity of a sample is affected by the total number of reads obtained for each sample, and by the number of features that the total reads are apportioned between. Each sample in a transcriptome contains between thousands and tens of thousands of features each of which may have a potential dynamic range of over four orders of magnitude. In many cases, a transcriptome dataset will be composed of several groups, where the expression of a feature (gene) 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 in a given environment are constrained by an absolute upper bound (Scott et al. 2010). In eukaryotic cells, the upper bound can be changed with surprisingly

small genetic changes (Lovén et al. 2012), and changes in the upper bound are known confounder of transcriptome experiments. In the case of a meta-transcriptome, a population of cells does not necessarily exhibit total gene expression with an upper bound, since the cells themselves can change in both absolute and relative abundance in a mixture.

The assumption being made when using the CLR transformation to identify features that differ between groups is that most features are either invariant or, at worst, vary at random when comparing the two groups. However, marked asymmetry can break this assumption, and this is the problem being addressed by identifying and using alternate denominators.

1.4 Alternative Denominators

The starting point for the alternative denominators is the n samples by D features matrix of counts. The data are CLR transformed using Eq. 1 sample-wise with a uniform prior of 0.5 to give a new matrix, and the objective is to identify a subset of the features that best represent those features that are least likely to exhibit systematic change in the dataset.

The *IQLR* (interquartile log-ratio) transformation uses as the denominator for the log-ratio those features with 'typical' variance across all samples. The variance is calculated for each feature group-wise, and features with variance between the first and third quartile in each group are extracted. Thus, the vector of features in group A, $\mathbf{f_A}$, is chosen from the variance vector of group $\mathbf{A}$ as follows: $\mathbf{var(A)_{Q1}} > \mathbf{f_A} < \mathbf{var(A)_{Q3}}$. The interquartile-variance feature *IQVF* subset of the D features is the intersect of the group-wise typical variance features, $\mathbf{IQVF} = \mathbf{f_A} \cap \mathbf{f_B}$ if there are two groups. The geometric mean of the *IQVF* set is used in the denominator to calculate a log-ratio for each sample vector $\mathbf{x_i}$ in matrix S:

$$\mathbf{x}_{\mathbf{i},IQLR} = \mathbf{log}\left(\frac{\mathbf{x_{i,j=1\ldots D}}}{\mathbf{G}(IQVF)}\right) \tag{4}$$

The non-zero log-ratio, *NZLR*, transformation uses as the denominator for the log-ratio those features that are non-zero in each group. The log-ratio is thus calculated using a different set of features as the denominator for each group. The indices of non-zero features in each group are retained $\mathbf{f_A} = \mathbf{min}(\mathrm{S}_{\mathbf{1\ldots i_A}}, \mathbf{1}\ldots\mathbf{D}) > \mathbf{0}; \mathbf{f_B} = \mathbf{min}(\mathrm{S}_{\mathbf{(i_B+1)\ldots n}}, \mathbf{1}\ldots\mathbf{D}) > \mathbf{0}$ and used group-wise.

Now the features in group A have their log-ratio computed using the non-zero features in the group as the denominator, and likewise for group B. Thus, the transformed vectors in group A are

$$\mathbf{x}_{\mathbf{iA},NZLR} = \mathbf{log}\left(\frac{\mathbf{x_{1\ldots i_A,j=1\ldots D}}}{\mathbf{G(f_A)}}\right). \tag{5}$$

Similarly, the vectors in group B are transformed using as the denominator the $G(\mathbf{f_B})$ of the set of non-zero features in group B.

The *LVHA* (low-variance high-abundance log-ratio) transformation uses as the denominator the intersect of those features that have low variance and high relative abundance in each group. These are chosen as the features that exhibit variance below the first quartile and relative abundance above the third quartile on a per-group basis. That is, the vector of features for group A is generated as follows: $\mathbf{f_A} = (\mathbf{A} < \mathbf{var(A)_{Q1}}) \cap \mathbf{A} > \mathbf{rab(A)_{Q3}}$, and so on for other groups. The final set is $LVHA = \mathbf{f_A} \cap \mathbf{f_B}$ if there are two groups.

The geometric mean of the *LVHA* set is used in the denominator to calculate a log-ratio for each sample vector $\mathbf{x_i}$ in matrix S:

$$\mathbf{x}_{\mathbf{i},LVHA} = \mathbf{log}\left(\frac{\mathbf{x_{i,j=1\ldots D}}}{\mathbf{G}(LVHA)}\right) \tag{6}$$

The log median transformation uses as the denominator the median of the clr-transformed values rather than the geometric mean.

$$\mathbf{x}_{\mathbf{i},\mathbf{MED}} = \mathbf{log}(\mathbf{x_{i,j=1\ldots D}}) - \mathrm{MED}(\mathbf{log}(\mathbf{x_i})) \tag{7}$$

2 Results

Throughout, we use two plots to summarize the location of the features in multivariate datasets. The Bland–Altman (BA) plot (Altman and Bland 1983) plots the mean log-ratio abundance on the x-axis and the difference between groups on the y-axis. The BA plot is efficient at showing the relationship between (relative, mean log-ratio) abundance and difference, but contains little information on the per-feature dispersion in the data. The effect size plot (Gloor et al. 2016a) complements the BA plot by showing the relationship between a measure of dispersion (on the x-axis) and the difference between groups (on the y-axis). All plots are in log units calculated with a base of two for convenience. The ratio between the within- and between-group differences is a proxy for the effect size statistic calculated by `ALDEx2`. Difference and dispersion are calculated using methods that are indifferent to distributional assumptions and are defined in the Implementation section. Figure 1 shows that incorrect estimates of the location of the data can be achieved with seemingly minor variation within simulated data, resulting in a clear asymmetry in the data. The goal is to identify a basis that best represents each sample so features can be accurately compared even when the data contains an asymmetry.

2.1 Simulated Data

RNA-Seq data was simulated for benchmarking purposes. Assemblies from *Saccharomyces cerevisiae* uid128 and a complete reference genome of *S. cerevisiae* were drawn from GenBank. The R package `polyester v1.10.0` (Frazee et al. 2016) was used to simulate an RNA-Seq experiment with 2 groups of 10 replicates with 20x average sequencing coverage across the simulation experiment. For the base dataset, 40 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` (Langmead and Salzberg 2012) to align the simulated reads to the *S. cerevisiae* reference genome. Labelling each group as A and B is arbitrary and hence the first 10 samples belong to condition A, and the final 10 samples belong to condition B. There are a total of 6349 features in these simulated data, but only the first 1000 genes by order were chosen for the majority of the figures.

An additional 51 datasets derived from the base dataset were generated in order to benchmark how well the generic log-ratio (*LR) transformations centre the data in the resultant set of datasets with sparsity ranging from 0 to 51% sparse in one condition relative to the other.

2.2 Four Alternative Methods

The CLR is used for these analyses rather than the more formally correct, and recently more popular, isometric log-ratio or balance-based approaches [?,?,?] primarily because of time and space complexity or because of additional requirements. For example, the `selbal` R package [?] fails to complete an analysis of the largest of the datasets used here in more than 24 hours of computation time using 10 concurrent threads on an Intel i9 class processor. Additionally, tools such as `phylofactor` [?] are tree-based methods that assume phylogeny is an important determinant of the outcome.

In its initial implementation, `ALDEx2` computes the CLR, a per-sample geometric mean using all features as the baseline for feature comparisons. The 'Symmetric dataset' panel in Fig. 1 is an 'effect plot' (Gloor et al. 2016a) and we can see 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. The remainder of the features have very small difference and a very large dispersion, and correspondingly have an effect size much less than 1, that is, the results are as expected.

The inset histogram shows that the distribution of difference values between groups A and B is symmetric and has a location of 0. However, the introduction of small amounts of asymmetry strongly affects the results. The asymmetric 2% dataset is the base dataset modified by choosing 20 random features from Group A

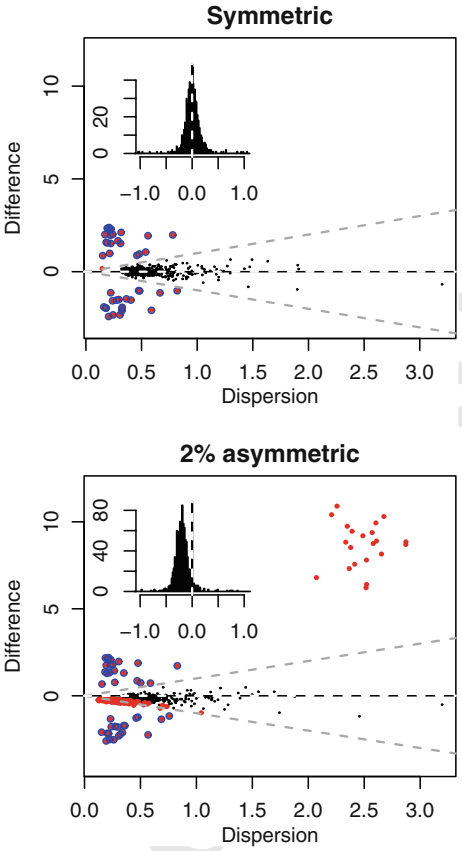

Fig. 1 Effect plots of simulated asymmetric data illustrate the problem. The effect plots show the difference between two conditions in simulated RNA-seq data with 1000 genes where 40 genes are modelled to have true difference between groups of an effect size of 1 or more. Each point is a feature (gene) and 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 the differences between groups as calculated by `ALDEx2`, and the vertical line shows a difference of 0. These x-axis of the histograms are truncated to show only differences near the midpoint

and setting their value to 0. It is apparent from the bottom panel of Fig. 1 that even this low level of simulated asymmetry breaks the assumption that most features are invariant, and the location of modal no-difference between groups is no longer at the origin. This small amount of asymmetry shifts the geometric mean of the data such that the two groups are calculated using different denominators, causing bias. It is unlikely that the problem will be as easy to diagnose in real data as in simulated data.

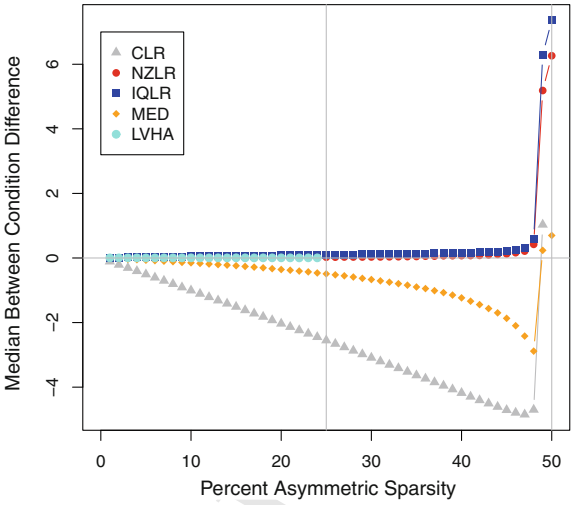

Fig. 2 The behaviour of each transform for datasets with varying asymmetry. Each point represents the median between condition difference for a given transformation in a dataset with a specified sparsity. Points closer to the location y=0 are favourable and indicate proper centring of the dataset. Abbreviations for each method for determining the denominator are the same as in the text

The major determinant of the centre of a sample when using the CLR is the denominator, or basis, used to compute the CLR, thus one obvious approach to address 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 examined four different approaches to identify the features to include in the denominator, and tested them with different proportions of modelled sparse asymmetry.

The transformation in Eq. 4 is termed the IQLR transformation, and relies on finding those features with 'typical' variance in the dataset. The assumption being made here is that features with typical variance are those with stochastic variance in the dataset, but that do not differ reproducibly between groups. The results of this method are shown in Fig. 2, and we see that the IQLR transformation restores the centre of the dataset to the origin even when over 40% of the modelled features are asymmetric.

The second approach uses as the denominator the set of non-zero features in each group calculated as in Eq. 5. Thus, in this case, the geometric mean of group A and group B are based on different, but potentially overlapping, sets of features, and this approach is called the non-zero log-ratio (NZLR). As shown in Fig. 2, the NZLR method also restores the centre of the data to the origin with similar efficiency in our test dataset as does the IQLR method. Note that if the asymmetry is not driven by sparsity, the NZLR method will obviously fail. This method is not recommended unless the investigator knows from prior inspection that asymmetric sparsity occurs.

The third approach was to identify the intersect between groups of those features that have variance which is in the bottom quartile and a relative abundance in the top quartile in each group. This is referred to as the low-variance high-abundance log-ratio *LVHA* method and is calculated as given in Eq. 6. In essence, the LVHA approach is an attempt to identify those features that would be similar to those chosen

by an experimentalist using qPCR, being relatively abundant with low variance in all groups in the dataset. The results of this method are shown in Fig. 2. We can see that the LVHA approach perfectly centres the data up to 25% asymmetry in our test dataset.

The fourth approach replaces the geometric mean in Eq. 1 with the median of the vector as given in Eq. 7 since this should be a robust estimate of the midpoint of the data. We can see in Fig. 2 that the median is better than the geometric mean at low proportions of asymmetry, but performs worse than the previous approaches in almost every case.

2.3 *Example of a Meta-RNA-seq Dataset*

We finally introduce the example of a real meta-transcriptome dataset collected to determine the differences in gene expression of the vaginal microbial community in the healthy (H) and bacterial vaginosis (BV) states (Macklaim et al. 2013; Deng et al. 2018). The vaginal community can be dominated either by a few members of the *Lactobacillus* genus in the H state, or by a mixed group of anaerobic bacterial genera in the BV state (Ravel et al. 2011). In either state the members of the bacterial consortium from the other state are either very rare or absent. The data presented in Fig. 3 show effect size plots for the CLR and three different denominators for a count table derived from data deposited at the European Nucleotide Archive metagenomics site under the accession number PRJEB31833 using the methods in Macklaim and Gloor (2018) and annotated using the SEED subsystems (Overbeek et al. 2014). Compatible count tables for other meta-transcriptomic or metagenomic datasets may be downloaded from the EBI metagenomics website (Mitchell et al. 2018).

The values in Fig. 3:denom=all were computed using the CLR. We can see that there is a large asymmetry in distribution, the most striking of which are the functions in BV located below the midline on the y-axis. This asymmetry is driven by the greater complexity of the BV microbial community, and the generally larger and more complex genomes in the set of bacteria found in BV (Macklaim et al. 2013). The asymmetry is composed of both presence–absence (sparsity) and large differences between groups. This can be seen with the sparsity overlay colour, where functions that contain one or more zeros are coloured in a dark grey.

Note that there appears to be two clusters of functions that are just above the y-axis midline. The ones at 4,2 are composed of sparse functions expressed at low levels in BV but that are absent or very rare in H. We also see a second group composed of functions found in common between the H and BV group that is expressed at very high relative levels, centred around 1,1.5 on the plot. These are generally housekeeping functions that are central and required by all living organisms, and we can see that the two sets of housekeeping functions that are highlighted cluster here: translational protein functions in blue (rib) and glycolysis (glycolic, core sugar metabolism)-related functions in magenta. Many of the functions in these groups are often used as internal standards for comparison by qPCR as it is assumed that their expression

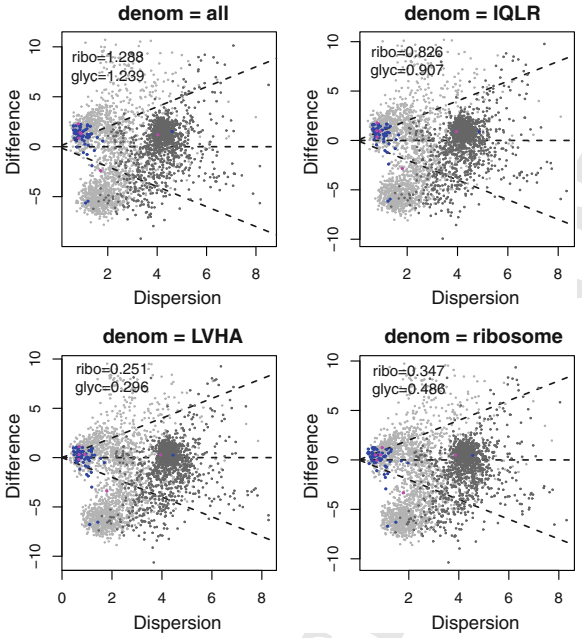

Fig. 3 Sparsity and asymmetry in a meta-transcriptome dataset of vaginal samples. The panel shows effect plots with log base 2 scaling of the data using the log-ratio values. Each point is a SEED-level 4 functional annotation category derived from the raw reads as in Macklaim and Gloor (2018) and they are coloured in light grey if no sample contains a 0 value (non-sparse), dark grey if at least one sample contains a 0 count, blue if annotated as a protein translation function or magenta if annotated as a glycolytic function. The median displacement of the functional subsets from 0 is given in the top left of the plot. The user-defined denominator was composed of functions annotated as ribosomal small or large subunit functions

is invariant (Scott et al. 2010), and we are attempting to minimize the displacement of these functions from a difference location of 0. The median offset of the two sets of housekeeping functions on the y-axis is given, and we can see that they are both observed to be substantially above the expected location of 0 when the CLR alone is used to determine differential relative abundance.

The other three panels in Fig. 3 show the result of applying three adjustment approaches to the asymmetric meta-RNA-seq dataset. The LVHA, IQLR, and user-defined adjustment (ribosome) methods all centred the data better than did the CLR. In the case of the user-defined functions, we used the set of translational protein functions in blue as the denominator in the denom=ribosomal panel. We can see that the bulk of the expected invariant groups are closer to the y-axis midline with two adjustments, and the offset of the translational and glycolytic functions is reduced substantially when compared to the CLR result. The LVHA adjustment brings the housekeeping functions very close to the centre of the dataset. Centring on the geo-

metric mean of the low variance but high (relative) abundance functions provides a substantial improvement over all other methods in this dataset.

3 Discussion

Biological data derived from high-throughput sequencing is rarely ideal and exhibits many pathologies. It is currently difficult to examine complex communities using RNA-seq because of a mismatch between the assumptions of the tools and the characteristics of the data. In particular, such data can be derived from asymmetric environments, where sets of genes, operational taxonomic units, or organisms can be present or abundant in one condition and absent or rare in another. Asymmetries also often exist in metagenomic, and in vitro selection (SELEX) datasets, and the approaches described here help to recentre the data (Almeida et al. 2019; McMurrough et al. 2018; Wolfs et al. 2016). Alternatively, an asymmetry in the data can arise because of a systematic failure in experimental design, for example, through improper blocking or the presence of outlier samples. In any of these instances, the presence of an asymmetry may not be obvious.

In a biological context, it is entirely reasonable that asymmetry could occur because counts rather than simple sparsity. 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 expression to very high expression. In the context of 16S rRNA gene sequencing study, it is possible for samples to be dominated not only by one very abundant organism but also to contain many other taxa at low abundance. Thus, we would have a count asymmetry that is not necessarily based on sparsity, and for this reason we do not recommend the NZLR approach unless the others fail. Furthermore, sparsity is strongly affected by read depth, the same samples derived from a sequencing dataset from an Illumina NextSeq run delivering a total count of 400M reads will be substantially less sparse than those derived from an Illumina MiSeq run delivering a total count of 25M reads; however, any underlying asymmetry will be preserved.

We demonstrated that even a small number of asymmetric features can change the location of the dataset, leading to both false positive and false negative differences being identified. When the asymmetry is moderate, the IQLR correction is most appropriate. This correction makes the assumption that those features with variance between the first and third quartile of variance are a suitable proxy for the expected 'typical' variance of the data. This approach can tolerate up to at least 40% asymmetry in the data when the geometric mean of these features are used as the denominator for a log-ratio normalization. In fact, we recommend that the IQLR be used as the default when performing differential relative abundance analysis, since this normalization makes no strong assumption about the data and appears to never perform worse than the CLR normalization until failure of the approach occurs. In this case, `ALDEx2` will report an error.

More extreme asymmetry, as found in the vaginal transcriptome dataset, forces the investigator to make strong assumptions about the underlying data and use the LVHA correction, or to explicitly choose a set of housekeeping genes. The assumptions made here are thus similar to the assumptions made when performing qPCR: that there are one or more invariant features in the data, and that these will typically be relatively abundant housekeeping functions. We found that these features could usually be identified as having low variance but high relative abundance and can be used as exemplars of 'invariant' features.

4 Conclusions

AQ5 We tested different methods to properly centre the data, and found that the IQLR- and LVHA-specified centring approaches were the most general purpose and thus recommended for use. All are implemented in the `ALDEx2` R package (Fernandes et al. 2014) and can be used in conjunction with the `propr` R package for compositional association (Quinn et al. 2017b). We observe that not all datasets have feature sets that are compatible with the LVHA approach; in these cases, the investigator can make an even stronger assumption and choose one or more features that prior knowledge suggests would be appropriate. In any case, it must be remembered that the results of any analysis must be interpreted as *abundance relative to the chosen invariant part of the dataset*, and not as changes in absolute abundance.

5 Abbreviations

ALR: additive log-ratio transformation
BA plot: Bland–Altman plot
BV: bacterial vaginosis state
CLR: centred log-ratio transformation
H: healthy (non-BV) state
HTS: high-throughput sequencing
IQLR: the interquartile log-ratio method
LVHA: the low-variance, high-abundance method
M: million
NZLR: the non-zero log-ratio method
OTU: operational taxonomic unit
qPCR: quantitative PCR

6 Declarations

7 Availability and Requirements

The methods are included in the `ALDEx2` R package available at Bioconductor (https://www.bioconductor.org/packages/release/bioc/html/ALDEx2.html).

7.1 Ethics Approval and Consent to Participate

Not applicable

7.2 Consent for Publication

Not applicable

7.3 Availability of Data and Material

All data used in this publication are publicly available. The raw data for the yeast genome was drawn from the reference *Saccharomyces cerevisiae* genome at: https://www.ncbi.nlm.nih.gov/genome/?term=txid4932. The raw data for Fig. 3 was obtained from European Nucleotide Archive metagenomics site under the accession number PRJEB31833 using the methods in Macklaim and Gloor (2018). All R scripts for figure generation and analysis are located on a public GitHub repository at: https://github.com/JRWu/Log-Ratio-Publication.

7.4 Competing Interests

The authors declare that they have no competing interests.

7.5 Funding

This work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada grant RGPIN-2015-03878 to GBG.

8 Author's Contributions

Designed the experiments GBG, JMM, JRW, BLG. Developed and applied corrections BLG, JRW, GBG, JMM. Wrote the manuscript, GBG. All authors have read and approved the manuscript.

Acknowledgements GBG first met Dr. Palwlowsky-Glahn in 2014 where she was an instructor and organizer of the CoDa summer school. At that time, our group was struggling with interpreting metatranscriptomic datasets, and found that compositional approaches were promising. GBG thanks Vera for her everlasting patience in answering any and all questions, and for making the entry into the world of CoDa a fruitful and rewarding experience.

References

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

A. Almeida, A.L. Mitchell, M. Boland, S.C. Forster, G.B. Gloor, A. Tarkowska, T.D. Lawley, R.D. Finn, A new genomic blueprint of the human gut microbiota. Nature **568**, 499 (2019). https://doi.org/10.1038/s41586-019-0965-1

D.G. Altman, J.M. Bland, Measurement in medicine: the analysis of method comparison studies. J. R. Stat. Soc. Ser. D (Stat.) **32**(3), 307–317 (1983), http://www.jstor.org/stable/2987937

C. Barceló-Vidal, J.A. Martin-Fernàndez, V. Pawlowsky-Glahn, Mathematical foundations of compositional data analysis, in *Proceedings of IAMG*, vol. 1 (Springer, 2001), pp. 1–20

G. Bian, G.B. Gloor, A. Gong, C. Jia, W. Zhang, J. Hu, H. Zhang, Y. Zhang, Z. Zhou, J. Zhang, J.P. Burton, G. Reid, Y. Xiao, Q. Zeng, K. Yang, J. Li, The gut microbiota of healthy aged Chinese is similar to that of the healthy young. mSphere **2**(5), e00327–17 (2017). https://doi.org/10.1128/mSphere.00327-17

Z.L. Deng, C. Gottschick, S. Bhuju, C. Masur, C. Abels, I. Wagner-Döbler, Metatranscriptome analysis of the vaginal microbiota reveals potential mechanisms for protection against metronidazole in bacterial vaginosis. mSphere **3**(3) (2018). https://doi.org/10.1128/mSphereDirect.00262-18

M.A. Dillies, A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloë, C. Le Gall, B. Schaëffer, S. Le Crom, M. Guedj, F. Jaffrézic, French StatOmique consortium. a comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief. Bioinform. **14**(6), 671–683 (2013). https://doi.org/10.1093/bib/bbs046

A.D. Fernandes, J.M. Macklaim, T.G. Linn, G. Reid, G.B. Gloor, ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-seq. PLoS One **8**(7), e67019 (2013). https://doi.org/10.1371/journal.pone.0067019

A.D. Fernandes, J.N. Reid, J.M. Macklaim, T.A. McMurrough, D.R. Edgell, G.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**, 15.1–15.13 (2014). https://doi.org/10.1186/2049-2618-2-15

A.C. Frazee, A.E. Jaffe, R. Kirchner, J.T. Leek, Polyester: simulate RNA-seq reads. R package version 1.10.0 (2016)

P. Gajer, 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, J. Ravel, Temporal dynamics of the human vaginal microbiota. Sci. Transl. Med. **4**(132), 132ra52 (2012). https://doi.org/10.1126/scitranslmed.3003605

M. Gierliński, C. Cole, P. Schofield, N.J. Schurch, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. Simpson, T. Owen-Hughes, M. Blaxter, G.J. Barton, Statistical models for RNA-seq data derived

509 from a two-condition 48-replicate experiment. Bioinformatics **31**(22), 3625–3630 (2015). https://
510 doi.org/10.1093/bioinformatics/btv425
511 G.B. Gloor, J.M. Macklaim, A.D. Fernandes, Displaying variation in large datasets: plotting a visual
512 summary of effect sizes. J. Comput. Graph. Stat. **25**(3C), 971–979 (2016a). https://doi.org/10.
513 1080/10618600.2015.1131161
514 G.B. Gloor, J.M. Macklaim, M. Vu, A.D. Fernandes, Compositional uncertainty should not be
515 ignored in high-throughput sequencing data analysis. Austrian J. Stat. **45**, 73–87 (2016b). https://
516 doi.org/10.17713/ajs.v45i4.122
517 G.B. Gloor, J.M. Macklaim, V. Pawlowsky-Glahn, J.J. Egozcue, Microbiome datasets are compo-
518 sitional: and this is not optional. Front. Microbiol. **8**, 2224 (2017). https://doi.org/10.3389/fmicb.
519 2017.02224
520 R. Hummelen, A.D. Fernandes, J.M. Macklaim, R.J. Dickson, J. Changalucha, G.B. Gloor, G. Reid,
521 Deep sequencing of the vaginal microbiota of women with HIV. PLoS One **5**(8), e12078 (2010).
522 https://doi.org/10.1371/journal.pone.0012078
523 K.S. Lang, T.J. Johnson, Transcriptome modulations due to A/C2 plasmid acquisition. Plasmid **80**,
524 83–89 (2015). https://doi.org/10.1016/j.plasmid.2015.05.005
525 B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat. Methods **9**(4), 357–
526 359 (2012)
527 M.I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq
528 data with DESeq2. Genome Biol. **15**(12), 550.1–550.21 (2014). https://doi.org/10.1186/s13059-
529 014-0550-8
530 D.R. Lovell, X.Y. Chua, A. McGrath, Counts: an outstanding challenge for log-ratio analysis of
531 compositional data in the molecular biosciences. NAR Genomics Bioinform. **2**(2), lqaa040 (2020)
532 J. Lovén, D.A. Orlando, A.A. Sigova, C.Y. Lin, P.B. Rahl, C.B. Burge, D.L. Levens, T.I. Lee, R.A.
533 Young, Revisiting global gene expression analysis. Cell **151**(3), 476–482 (2012). https://doi.org/
534 10.1016/j.cell.2012.10.012
535 J.M. Macklaim, G.B. Gloor, From RNA-seq to biological inference: using compositional data
536 analysis in meta-transcriptomics. Methods Mol. Biol. **1849**, 193–213 (2018). https://doi.org/10.
537 1007/978-1-4939-8728-3_13
538 J.M. Macklaim, A.D. Fernandes, J.M. Di Bella, J.A. Hammond, G. Reid, G.B. Gloor, Comparative
539 meta-RNA-seq of the vaginal microbiota and differential expression by Lactobacillus iners in
540 health and dysbiosis. Microbiome **1**(1), 12 (2013). https://doi.org/10.1186/2049-2618-1-12
541 J.M. Macklaim, J.C. Clemente, R. Knight, G.B. Gloor, G. Reid, Changes in vaginal microbiota
542 following antimicrobial and probiotic therapy. Microb. Ecol. Health Dis. **26**, 27799 (2015)
543 C. Martino, J.T. Morton, C.A. Marotz, L.R. Thompson, A. Tripathi, R. Knight, K. Zengler, A novel
544 sparse compositional technique reveals microbial perturbations. mSystems **4**(1) (2019). https://
545 doi.org/10.1128/mSystems.00016-19
546 T.A. McMurrough, R.J. Dickson, S.M.F. Thibert, G.B. Gloor, D.R. Edgell, Control of catalytic
547 efficiency by a coevolving network of catalytic and noncatalytic residues. Proc. Natl. Acad. Sci.
548 USA **111**(23), E2376–83 (2014). https://doi.org/10.1073/pnas.1322352111
549 T.A. McMurrough, C.M. Brown, K. Zhang, G. Hausner, M.S. Junop, G.B. Gloor, D.R. Edgell,
550 Active site residue identity regulates cleavage preference of LAGLIDADG homing endonucle-
551 ases. Nucleic Acids Res. **46**(22), 11990–12007 (2018). https://doi.org/10.1093/nar/gky976
552 A.L. Mitchell, M. Scheremetjew, H. Denise, S. Potter, A. Tarkowska, M. Qureshi, G.A. Salazar, S.
553 Pesseat, M.A. Boland, F.M.I. Hunter, P. Ten Hoopen, B. Alako, C. Amid, D.J. Wilkinson, T.P.
554 Curtis, G. Cochrane, R.D. Finn, EBI metagenomics in 2017: enriching the analysis of microbial
555 communities, from sequence reads to assemblies. Nucleic Acids Res. **46**(D1), D726–D735 (2018).
556 https://doi.org/10.1093/nar/gkx967
557 R. Overbeek, R. Olson, G.D. Pusch, G.J. Olsen, J.J. Davis, T. Disz, R.A. Edwards, S. Gerdes,
558 B. Parrello, M. Shukla, V. Vonstein, A.R. Wattam, F. Xia, R. Stevens, The seed and the rapid
559 annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res.
560 **42**(Database issue), D206–14 (2014). https://doi.org/10.1093/nar/gkt1226

J. Palarea-Albaladejo, J.A. Martín-Fernández, zCompositions — R package for multivariate imputation of left-censored data under a compositional approach. Chemom. Intell. Lab. Syst. **143**, 85–96 (2015). https://doi.org/10.1016/j.chemolab.2015.02.019, http://www.sciencedirect.com/science/article/pii/S0169743915000490

J. Peng, B. Hao, L. Liu, S. Wang, B. Ma, Y. Yang, F. Xie, Y. Li, RNA-seq and microarrays analyses reveal global differential transcriptomes of Mesorhizobium huakuii 7653R between bacteroids and free-living cells. PLoS One **9**(4), e93626 (2014). https://doi.org/10.1371/journal.pone.0093626

T.P. Quinn, I. Erb, M.F. Richardson, T.M. Crowley, Understanding sequencing data as compositions: an outlook and review. bioRxiv (2017a). https://www.biorxiv.org/content/early/2017/10/19/206425

T.P. Quinn, M.F. Richardson, D. Lovell, T.M. Crowley, propr: an R-package for identifying proportionally abundant features using compositional data analysis. Sci. Rep. **7**(1), 16252 (2017b). https://doi.org/10.1038/s41598-017-16520-0

J. Ravel, P. Gajer, Z. Abdo, G.M. Schneider, S.S.K. Koenig, S.L. McCulle, S. Karlebach, R. Gorle, J. Russell, C.O. Tacket, R.M. Brotman, C.C. Davis, K. Ault, L. Peralta, L.J. Forney, Vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. USA **108**, 4680–4687 (2011). https://doi.org/10.1073/pnas.100611107

M.D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. **11**(3), R25.1–R25.9 (2010). https://doi.org/10.1186/gb-2010-11-3-r25

M.D. Robinson, D.J. McCarthy, G.K. Smyth, edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**(1), 139–140 (2010). https://doi.org/10.1093/bioinformatics/btp616

M. Scott, C.W. Gunderson, E.M. Mateescu, Z. Zhang, T. Hwa, Interdependence of cell growth and gene expression: origins and consequences. Science **330**(6007), 1099–1102 (2010). https://doi.org/10.1126/science.1192588

O. Thellin, W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, E. Heinen, Housekeeping genes as internal standards: use and limits. J. Biotechnol. **75**(2–3), 291–295 (1999)

J. Vandesompele, K. De Preter, F. Pattyn, B. Poppe, N. Van Roy, A. De Paepe, F. Speleman, Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. Genome Biol. **3**(7), RESEARCH0034 (2002)

J.M. Wolfs, T.A. Hamilton, J.T. Lant, M. Laforet, J. Zhang, L.M. Salemi, G.B. Gloor, C. Schild-Poulter, D.R. Edgell, Biasing genome-editing events toward precise length deletions with an RNA-guided TevCas9 dual nuclease. Proc. Natl. Acad. Sci. USA (2016). https://doi.org/10.1073/pnas.1616343114

H. Zhao, 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, J. Xu, Global transcriptional and phenotypic analyses of Escherichia coli O157:H7 strain Xuzhou21 and its pO157_Sal cured mutant. PLoS One **8**(5), e65466 (2013). https://doi.org/10.1371/journal.pone.0065466

Author Queries

Chapter 17

Query Refs.	Details Required	Author's response
AQ1	Please note that '?' have been appeared throughout this chapter. Kindly check and confirm.	
AQ2	Please confirm if the section headings identified are correct.	OK
AQ3	Please check and approve the edit made in the sentence "The CLR, and any other..." and amend if necessary.	OK
AQ4	Please check and approve the edit made in the sentence "In the context of 16S rRNA gene..." and amend if necessary.	OK
AQ5	References 'Bian et al. (2017)', 'Dillies et al. (2013)', 'Love et al. (2014)', 'Lovell et al. (2020)', 'Martino et al. (2019)', 'Palarea-Albaladejo and Martín-Fernández (2015)', 'Quinn et al. (2017a)', 'Robinson and Oshlack (2010)' and 'Robinson et al. (2010)' are given in list but not cited in text. Please cite in text or delete from list.	Please remove all from the list

4. Anders, S., McCarthy, D.J., Chen, Y., Okoniewski, M., Smyth, G.K., Huber, W., Robinson, M.D.: Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. Nat Protoc 8(9), 1765–86 (2013). DOI 10.1038/nprot.2013.099

5. Auer, P.L., Doerge, R.W.: Statistical design and analysis of RNA sequencing data. Genetics 185(2), 405–16 (2010). DOI 10.1534/genetics.110.114983

24. McMurdie,P.J.,Holmes,S.:Waste not,want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 10(4), e1003531 (2014). DOI 10.1371/journal.pcbi.1003531

32. Rivera-Pinto, J., Egozcue, J.J., Pawlowsky-Glahn, V., Paredes, R., Noguera-Julian, M., Calle, M.L.: Balances: a new perspective for microbiome analysis. mSystems 3(4) (2018). DOI 10.1128/mSystems.00053-18

34. Silverman, J.D., Washburne, A.D., Mukherjee, S., David, L.A.: A phylogenetic transform enhances analysis of compositional microbiota data. Elife 6, 21887 (2017). DOI 10.7554/eLife.21887

35. Sun,J.,Nishiyama,T.,Shimizu,K.,Kadota,K.:TCC: an R package for comparing tag count data with robust normalization strategies. BMC Bioinformatics 14, 219.1–219.13 (2013). DOI 10.1186/1471-2105-14-219

38. Washburne,A.D.,Silverman,J.D.,Leff,J.W.,Bennett,D.J.,Darcy,J.L.,Mukherjee,S.,Fierer, N., David, L.A.: Phylogenetic factorization of compositional data yields lineage-level associ-ations in microbiome datasets. PeerJ 5, e2969 (2017). DOI 10.7717/peerj.2969

39. Weiss, S., Zu, Z.X., Peddada, S., Amir, A., Bittinger, K.,, Gonzalez, A., Lozupone, C., Zaneveld, J.R., Vazquez-Baeza, Y., Birmingham, A.,, Hyde, E.R., Knight, R.: Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome 5(1), 27 (2017). DOI 10.1186/s40168-017-0237-y

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

Instruction to printer	*Textual mark*	*Marginal mark*
Leave unchanged	··· under matter to remain	New matter followed by
Insert in text the matter indicated in the margin		or (x2)
Delete	through single character, rule or underline or through all characters to be deleted	or (x2)
Substitute character or substitute part of one or more word(s)	through letter or through characters	new character or new characters
Change to italics	under matter to be changed	
Change to capitals	under matter to be changed	
Change to small capitals	under matter to be changed	
Change to bold type	under matter to be changed	
Change to bold italic	under matter to be changed	
Change to lower case	Encircle matter to be changed	
Change italic to upright type	(As above)	
Change bold to non-bold type	(As above)	
Insert 'superior' character	through character or where required	or under character e.g. or
Insert 'inferior' character	(As above)	over character e.g.
Insert full stop	(As above)	
Insert comma	(As above)	
Insert single quotation marks	(As above)	or and/or or
Insert double quotation marks	(As above)	or and/or or
Insert hyphen	(As above)	
Start new paragraph		
No new paragraph		
Transpose		
Close up	linking characters	
Insert or substitute space between characters or words	through character or where required	
Reduce space between characters or words	between characters or words affected	