Understanding sequencing data as compositions: an outlook and review

Thomas P. Quinn[1,*], Ionas Erb[2,3], Mark F. Richardson[1,4], and Tamsyn M. Crowley[1,5]

[1]Bioinformatics Core Research Group, Deakin University, Geelong, 3220, Australia
[2]Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Barcelona, Spain
[3]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[4]Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, Geelong, 3220, Australia
[5]Poultry Hub Australia, University of New England, Armidale, New South Wales, 2351, Australia
* *contacttomquinn@gmail.com*

Abstract

Motivation: Although seldom acknowledged explicitly, count data generated by sequencing platforms exist as compositions for which the abundance of each component (e.g., gene or transcript) is only coherently interpretable relative to other components within that sample. This property arises from the assay technology itself, whereby the number of counts recorded for each sample is constrained by an arbitrary total sum (i.e., library size). Consequently, sequencing data, as compositional data, exist in a non-Euclidean space that renders invalid many conventional analyses, including distance measures, correlation coefficients, and multivariate statistical models.

Results: The purpose of this review is to summarize the principles of compositional data analysis (CoDA), provide evidence for why sequencing data are compositional, discuss compositionally valid methods available for analyzing sequencing data, and highlight future directions with regard to this field of study.

1 From raw sequences to counts

Automated Sanger sequencing served as the primary sequencing tool for decades, ushering in significant accomplishments including the sequencing of the entire human genome ([50]). Since the mid-2000s, however, attention has shifted away this "first-generation technology" toward new technologies collectively know as next-generation sequencing (NGS) ([50]). A number of NGS products exist, each differing in the sample preparation required and chemistry used ([50]). Although each product tends toward a different application, they all work by determining the base order (i.e., sequence) from a population of nucleotides, such that it becomes possible to estimate the abundances of unique sequences ([50]). However, these sequence abundances are not absolute abundances because the total number of sequences measured by NGS technology (i.e., the library size) ultimately depends on the chemistry of the assay, not the input material.

Depending on the input material, NGS has many uses. These include (1) variant discovery, (2) genome assembly, (3) transcriptome assembly, (4) epigenetic and chromatin profiling (e.g., ChIP-seq, methyl-seq, and DNase-seq), (5) meta-genomic species classification or gene discovery, and (6) transcript abundance quantification ([50]). The application of NGS to catalog transcript abundance is better known as RNA-Seq ([50]) and can be used to estimate the portional presence of transcript isoforms, gene archetypes, or other. RNA-Seq works by taking a population of (total or fractionated) RNA, converting them to a library of cDNA fragments, optionally amplifying the fragments, and then sequencing those fragments in a "high-throughput manner" ([73]). When sequencing smaller RNA (e.g., microRNA), an additional size selection step is used to ensure a uniform size of the RNA product ([36]).

The result of RNA-Seq is a virtual "library" of many short sequence fragments that are converted to a numeric data set through alignment (most often to a previously established reference genome or transcriptome) and quantification ([33]). The alignment and quantification steps summarize the raw sequence data (i.e., reads) as a "count matrix", a table containing the estimated number of times a sequence successfully aligns to a given reference annotation. The "count matrix" therefore provides a numeric distillation of the raw sequence reads collected by the assay; as such, it constitutes the data routinely used in statistical modeling, including differential expression analysis ([33]). Two factors complicate alignment and quantification. First, assembled references (e.g., genomes or transcriptomes) are only just references: sequences measured from biological samples will have an expected amount of variation, either systematic or random, when compared with the reference. This variation necessitates that the alignment procedure accommodates (at least optionally) a certain amount

of mismatch ([16]). Meanwhile, some reads (notably short reads) can ambiguously map to multiple reference sites, an undesired outcome that is amplified by mismatch tolerance ([16]). Many alignment and quantification methods exist (e.g., TopHat ([68]), STAR ([18]), Salmon ([51]), and others) and are reviewed elsewhere (e.g., [29]; [27]; [42]; [21]; [34]; [9]; [72]; [8]).

The "count matrix" (or equivalent) produced by alignment and quantification is routinely analyzed using statistical hypothesis testing (e.g., generalized linear models) or data science techniques (e.g., clustering or classification). Most commonly, data are studied using differential expression analysis, a constellation of methods that seek to identify which unique sequence fragments (if any) differ in abundance across the experimental condition(s). Like alignment and quantification, many differential expression methods exist (e.g., Cufflinks ([69]), limma ([55]), edgeR ([56]), DESeq ([7]), and others) and are reviewed elsewhere (e.g., [17]; [54]; [63]; [26]; [35]; [59]; [61]; [65]; [49]). However, it is important to note that conclusions drawn from RNA-Seq data appear to have a certain "robustness" to the choice in the alignment and quantification method, such that the choice in the differential expression method impacts the final result most ([75]).

The focus of this review is not to elaborate on the niceties of alignment, quantification, or differential expression, but rather to discuss the relative (i.e., compositional) nature of sequencing count data and the implications this has on many analyses (including differential expression analysis). In this review, we show how sequencing count data measure abundances as portions, rendering many conventional methods invalid. We then discuss methods available for dealing with portional data. Finally, we conclude by discussing challenges specific to these analyses and by considering advancements to this field of study. Although we emphasize RNA-Seq data throughout this paper, the principles discussed here apply to any NGS abundance data set.

2 Counts as parts of a whole

2.1 Image brightness as portions

As an analogy, let us imagine that we instructed two photographers to take a series of black and white photographs using a digital camera. We can represent the captured images as a set of N-dimensional vectors where each element (i.e., pixel) records the amount of light that hit a corresponding part of the film sensor. Considering this data set, let us ask a pointed experimental question: which photographer captured their photographs in brighter light? Better yet, for which pixels, on average, did Photographer A capture brighter light than Photographer B?

On first glance, this appears straight-forward. However, we want to know about the amount of light present when the photograph was taken, not the amount of light recorded by the film sensor. Although related, many factors influence the light measured at a given pixel. These include, for example, exposure time, aperture diameter, and the sensitivity of the film sensor. Changing any one of these parameters will change the image. Of course, such a change in the image does not mean a change in the reality.

At each pixel, we could then define two variables: luminance, the amount of light present at the moment of the photograph, and brightness, the amount of light perceived by the film sensor. Intuitively, we can understand brightness (the observed value, o), as a function, f, of luminance (the actual value, a):

$$\mathbf{o} = f(\mathbf{a}) \tag{1}$$

Even if we do not know the function, f, that relates these two measures, we see here that the total brightness recorded (i.e., $\sum \mathbf{o}$) is an artifact of the conditions under which the luminance is measured. Yet, if we can assume that the film sensor responds proportionally to light and does not clip (an unrealistic and idealized assumption), then the portional brightness would equal the portional luminance:

$$\frac{\mathbf{o}}{\sum \mathbf{o}} = \frac{\mathbf{a}}{\sum \mathbf{a}} \tag{2}$$

In this scenario, we can understand each element of $\mathbf{o}$ as a portion of the whole. As such, the brightness of a single pixel is only meaningful when interpreted relative to the total brightness (or to the brightness of the other pixels). Importantly, it follows that the ratio of any two parts of brightness will equal the ratio of any two parts of luminance.

2.2 Sequence abundance as portions

RNA-Seq data, through alignment and quantification, measure transcript abundance as counts. However, like the brightness of a digitalized image, the amount of RNA estimated for each transcript depends on some factors other than the amount of RNA molecules present in the assayed cell. Like a photograph, it is possible to change the observed magnitude while keeping the actual input the same. As such, RNA-Seq count data are not actually counts *per se*, but rather portions of a whole.

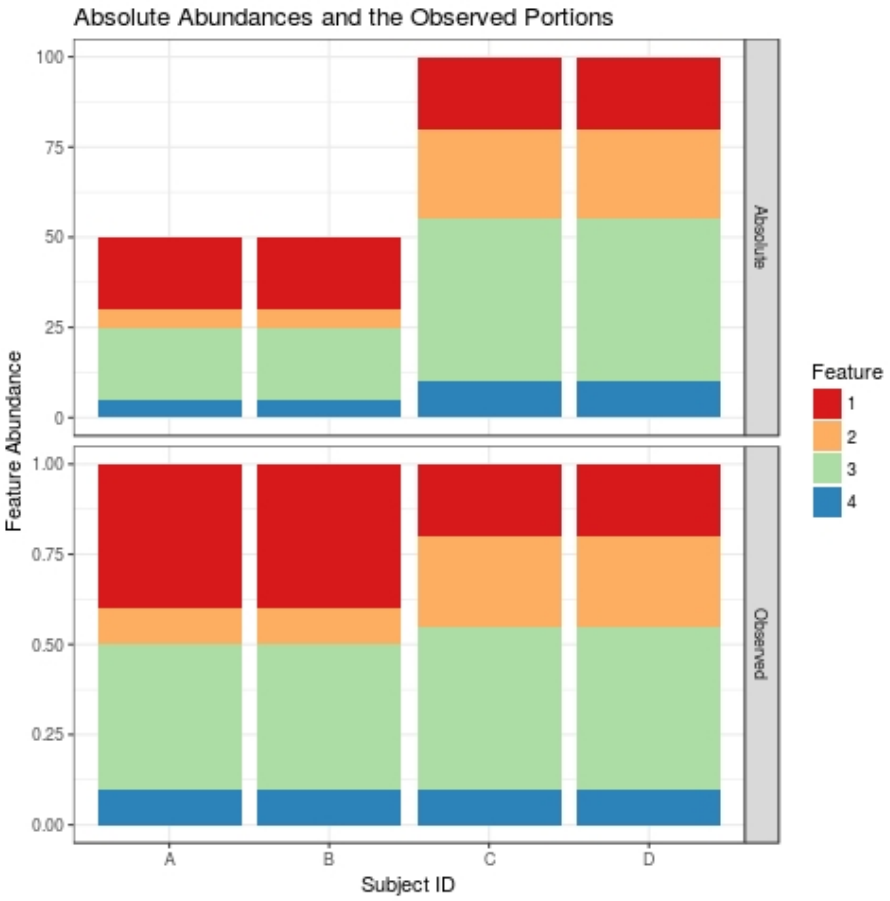

Figure 1: The top panel shows the absolute abundances of 4 features measured across 4 samples. The bottom panel shows their relative (i.e., portional) abundances for those same samples. Although Feature 4 is less abundant in samples A and B, it appears equally abundant across all samples when viewed as portions. Although Feature 1 is equally abundant across all samples, it appears more abundant in samples A and B when viewed as portions.

In fact, this is a property of all NGS abundance data: the abundances for each sample are constrained by an arbitrary total sum (i.e., the library size) ([63]). Since the library size is arbitrary, the individual values of the observed counts are irrelevant. However, the relative abundances of the observed counts still carry meaning. We can understand this by considering how, for a given sample, $\mathbf{o}$, the library size (i.e., $\sum \mathbf{o}$) cancels for a ratio of any two transcripts, i and j:

$$\frac{o_i}{o_j} = \frac{o_i / \sum \mathbf{o}}{o_j / \sum \mathbf{o}} \tag{3}$$

Analogous to how the relationship between luminance and brightness is unique to each photograph, the relationship between the actual abundances and the observed abundances is unique to each sample. Each independent sample, whether derived from a human subject or a cell line, may have undergone systematic or random differences in processing at any stage of RNA extraction, library preparation, or sequencing, causing between-sample biases ([63]). As such, library sizes typically differ between samples, making direct comparisons impossible ([63]). However, because the counts are portions of a whole, the interpretation is complicated even when library sizes are constant. For example, a large increase (or large decrease) in only a few transcripts will necessarily lead to a decrease (or increase) in all other measured counts ([63]). Figure 1 provides an abstracted visualization of how this might happen.

3 Counts as compositional data

3.1 The definition of compositional data

Compositional data measure each sample as a *composition*, a vector of non-zero positive values (i.e., components) carrying relative information ([2]). Compositional data have two unique properties. First, the total sum of all

component values (i.e., the library size) is an artifact of the sampling procedure ([71]). Second, the difference between component values is only meaningful proportionally (e.g., the difference between 100 and 200 counts carries the same information as the difference between 1000 and 2000 counts ([71]).

Examples of compositional data include anything measured as a percent or proportion. It also includes other data that are incidentally constrained to an arbitrary sum. NGS abundance data have compositional properties, but differ slightly from the formally defined compositional data in that they contain integer values only. However, except for possibly at near-zero values, we can treat so-called *count compositional data* as compositional data ([43]; [53]). Note that it is not a requirement for the arbitrary sum to represent complete unity: many data sets (including possibly NGS abundance data) lack information about potential components and hence exist as incomplete compositions ([1]).

3.2 The consequences of compositional data

Compositional data do not exist in real Euclidean space, but rather in a sub-space known as the simplex ([2]). Yet, many commonly used metrics implicitly assume otherwise; such metrics are invalid for relative data. This includes distance measures, correlation coefficients, and multivariate statistical models ([12]). For compositional data, the distance between any two variables is erratically sensitive to the presence or absence of other components ([4]). Meanwhile, correlation reveals spurious (i.e., falsely positive) associations between unrelated variables ([52]). In addition, multivariate statistics yield erroneous results because representing variables as portions of the whole makes them mutually-dependent, multivariate objects (i.e., increasing the abundance of one decreases the portional abundance of the others) ([12]). All of this applies to NGS abundance data too ([43]).

In the life sciences, count data are usually modeled using the Poisson distribution or negative binomial distribution ([11]). For NGS abundance data, the negative binomial model is preferred because it accommodates situations in which the variance is much larger than the mean, a common feature of biological replicates in RNA-Seq studies ([63]). These models are necessary because analyzing non-normalized and non-transformed count data as if they were normally distributed would imply that it is possible to sample negative and non-integer values, contradicting the assumptions behind many statistical hypotheses ([15]) (although it is possible to extend Gaussian analysis to counts by use of precision weights ([39])). Moreover, NGS abundance data are compositional counts, not counts, meaning that the measured variables (i.e., components) are not univariate objects ([13]).

3.3 Normalization to effective library size

Although the negative binomial distribution is still used to model NGS abundance data ([63]), doing so necessitates (at the very least) an additional normalization step ([63]). The simplest normalization would involve rescaling counts by the library size (i.e., the total number of mapped reads from a sample) ([63]), but this does not transform compositional counts into absolute counts. Instead, analysts most often use other, more elaborate normalization methods that (generally speaking) adjust the individual counts of each sample based on the counts of a reference (or pseudo-reference) sample ([17]). The sum of these rescaled counts is called the effective library size.

Effective library size normalization for RNA-seq data was first proposed in an attempt to address the relative (i.e., closed) nature of the data through a method known as the trimmed mean of M (TMM) ([57]). This normalization works by inferring an ideal (i.e., unchanged) reference from a subset of transcripts based on the assumption that the majority of transcripts remain unchanged across conditions. Here, the reference was chosen to be a trimmed mean ([57]), although others have proposed using the median over the transcripts as the reference ([7]). The TMM normalizes data to an effective library size based on the principle that if counts are evaluated relative to (i.e., divided by) an unchanged reference, the original scale of the data is recovered. In the language of compositional data analysis, this approach is described as an attempt to "open" the closed data, and is often criticized on the basis that "there is no magic powder that can be sprinkled on closed data to make them open" ([3]). Yet, if the data were open originally (and only incidentally closed by the sequencing procedure), this point of view is perhaps extreme. On the other hand, if the cells themselves produce closed data by default (e.g., due to their limited capacity for mRNA production ([60])), any attempt to open the data might prove futile.

Given the difficulties in identifying a truly unchanged reference (and in interpreting it correctly in the case that closed data is being produced by the cells themselves) avoiding normalization altogether would seem desirable. After all, the choice of normalization method impacts the final results of an analysis. For example, the number and identity of genes reported as differentially expressed change with the normalization method ([41]), as do false discovery rates ([40]). This also holds true for compositional metabolomic data ([58]). Moreover, at least some normalization methods are sensitive to the removal of lowly abundant counts ([41]), as well as to data asymmetry ([63]).

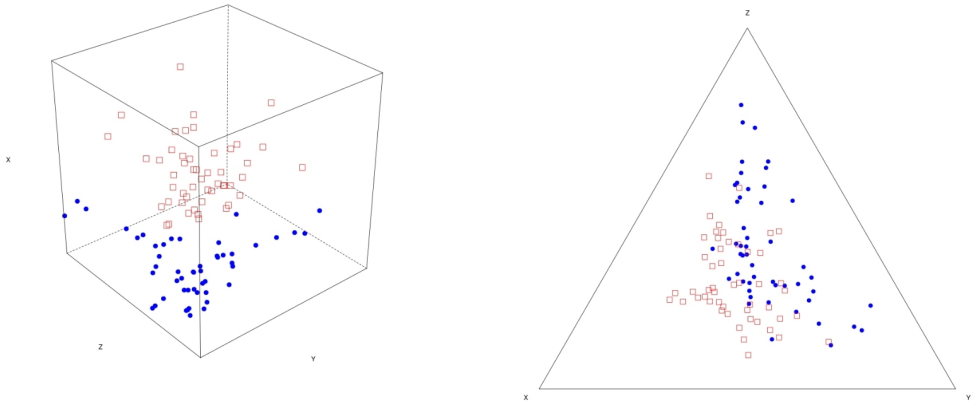

Figure 2: The left panel shows a 3D scatter of 100 samples, belonging to one of two groups, measuring the absolute abundances of 3 features. The right panel shows a ternary diagram of those same samples with the 3 features measuring relative (i.e., portional) abundances. Although the difference between the two groups is apparent in absolute space, the boundary between them becomes unclear in relative space. Note that for these relative data, it is not possible to reclaim a clear separation of the groups through transformation owing to the limited number of features available and the magnitude of the noise.

4 Principles of compositional data analysis

4.1 Approaches to compositional data

In lieu of normalization, many compositional data analyses begin with a transformation. Although compositional data exist in the simplex, Aitchison first documented that these data could get mapped into real space by use of the log-ratio transformation ([2]). By transforming data into real space, measurements like Euclidean distance become meaningful ([4]). However, it is also possible to analyze compositional data without log-ratio transformations. One approach involves performing calculations on the components themselves (called the "staying-in-the-simplex" approach) ([47]). Another involves performing calculations on ratios of the components (called the "pragmatic" approach) ([32]). Nevertheless, many compositional data analyses still begin with a log-ratio transformation.

Unlike normalizations, log-ratio transformations do not claim to open the data. Instead, the interpretation of the transformed data (and some of their results) depend on the reference used. In contrast, normalizations assume that an unchanged reference is available to recover the data (i.e., up to a proportionality constant) as they existed prior to closure by sequencing. Yet, while log-ratio transformations are conceptually distinct from normalizations, they are sometimes interpreted as if they were normalizations themselves ([25]). Although this contradicts compositional data analysis principles, conceiving of transformations as normalizations is helpful in understanding their use in some RNA-Seq analyses. Such log-ratio "normalizations", like conventional normalizations, aim to recast compositional data in absolute terms, allowing for a straight-forward univariate interpretation of the data. Like effective library size normalization, this is done through use of an ideal reference.

4.2 The log-ratio transformation

First, let us consider a small relative data set with only 3 features measured across 100 samples. These samples belong to one of two groups. One of the features, "X", can differentiate these groups perfectly. The other features, "Y" and "Z", constitute noise. We can turn an absolute data set into a compositional data set by dividing each element of the sample vector by the total sum. Figure 2 shows how the relationship between the samples (represented as points) changes when made compositional. Although the two groups appear clearly linearly separable in absolute space, the boundaries between groups become unclear in relative space. Meanwhile, the distances between samples become arbitrary.

When analyzing compositional data, it is sometimes possible to reclaim the discriminatory potential of relative data through transformation. For example, by setting all or some of the features relative to (i.e., divided by) a reference feature, one might discover that the resultant ratios can separate the groups ([66]). In

fact, any separation revealed by such ratios can be analyzed by standard statistical techniques ([66]). This illustrates the concept behind the additive log-ratio (alr) transformation, achieved by taking the logarithm of each measurement within a composition (i.e., each sample vector containing relative measurements) as divided by a reference feature (i.e., x_D) ([2]):

$$\text{alr}(\mathbf{x}) = \left[\ln \frac{x_i}{x_D}; ...; \ln \frac{x_{D-1}}{x_D}\right] \tag{4}$$

Instead of a specific reference feature, one could use an abstracted reference. In the case of the centered log-ratio (clr) transformation, the geometric mean of the composition (i.e., sample vector) is used in place of x_D ([2]). We use the notation $g(\mathbf{x})$ to indicate the geometric mean of the sample vector, $\mathbf{x}$. Note that because these transformations apply to each sample vector independently, the presence of an outlier sample does not alter the transformation of the other samples:

$$\text{clr}(\mathbf{x}) = \left[\ln \frac{x_i}{g(\mathbf{x})}; ...; \ln \frac{x_D}{g(\mathbf{x})}\right] \tag{5}$$

Likewise, other transformations exist that use the geometric mean of a feature subset as the reference. For example, the ALDEx2 package introduces the inter-quartile log-ratio (iqlr) transformation, which includes only features that fall within the inter-quartile range of total variance in the geometric mean calculation ([24]; [25]). Another, more complex, transformation, called the isometric log-ratio (ilr) transformation ([20]), also exists and is used in geological studies ([15]) and at least one analysis of RNA-Seq data ([67]). The ilr transforms the data with respect to an orthonormal coordinate system that is constructed from sequential binary partitions of features ([13]). Its default application to standard problems has been criticized by Aitchison on the basis that it lacks interpretability ([5]). Applications where the basis construction follows a microbiome phylogeny seem an interesting possibility ([74]).

4.3 The log-ratio "normalization"

In some instances, the log-ratio transformation is technically equivalent to a normalization. For example, let us consider the case where we know about our data the identity of a feature with a fixed abundance in absolute space across all samples. We could then use a log-ratio procedure to "sacrifice" this feature in order to "back-calculate" the absolute abundances. This is akin to using the alr transformation as a kind of normalization. However, because a single unchanged reference is rarely available or knowable (although synthetic RNA spike-ins may represent one way forward ([37])), we could try to approximate an unchanged reference from the data. For this, one might use the geometric mean of a feature subset, thereby using a clr (or iqlr) transformation as if it were a normalization.

Although log-ratio "normalizations" differ from log-ratio transformations only in the interpretation of their results, transformations alone are still useful even when they do not normalize the data. This is because they provide a way to move from the simplex into real space ([4]), rendering Euclidean distances meaningful. Importantly, clr- and ilr-transformed data impart four key properties to analyses: *scale invariance* (i.e., multiplying a composition by a constant k will not change the results), perturbation invariance (i.e., converting a composition between equivalent units will not change the results), *permutation invariance* (i.e., changing the order of the components within a composition will not change the results), and *sub-compositional dominance* (i.e., using a subset of a complete composition carries less information than using the whole) ([13]). Yet, the interpretation of transformation-based analyses remains complicated because the analyst must consider their results with respect to the chosen reference, or otherwise translate the results back into compositional terms.

4.4 Measures of distance

Euclidean distances do not make sense for compositional data ([4]). In contrast, the Aitchison distance does, providing a measure of distance between two d-dimensional compositions, $\mathbf{x}$ and $\mathbf{X}$ ([4]):

$$d(\mathbf{x}, \mathbf{X}) = \sqrt{\sum_{i=1}^{d} \left[\ln \frac{x_i}{g(\mathbf{x})} - \ln \frac{X_i}{g(\mathbf{X})}\right]^2} \tag{6}$$

Although the Aitchison distance is simply the Euclidean distance between clr-transformed compositions, this distance (unlike Euclidean distance) has scale invariance, perturbation invariance, permutation invariance, and sub-compositional dominance. Few other distance measures satisfy all four of these properties, including

none of the metrics routinely used in hierarchical clustering ([45]) (a routine part of RNA-Seq analysis). The property of sub-compositional dominance is especially important: even if the log-ratio transformation does not normalize the data, the addition of more sequence data will never make two samples appear *less* distant. This follows logically: as the amount of information available grows, the distance between samples should not shrink.

4.5 Measures of association

Like the Aitchison distance, there also exists a compositionally valid measure of association: the log-ratio variance (VLR) measures the agreement between two components (**a** and **b**) across two or more compositions. Specifically, it computes the variance of the logarithm of one component as divided by a second component. As such, a D-component data set contains D^2 associations (albeit with symmetry). Unlike Aitchison distance, however, the VLR does not require a log-ratio transformation whatsoever; in fact, if using log-ratio transformed data, the reference denominators would cancel out. Note that, while distances occur between compositions (i.e., between samples), associations occur between components (i.e., between transcripts).

$$\mathrm{VLR}(\mathbf{a}, \mathbf{b}) = \mathrm{var}\left[\ln \frac{a_i}{b_i}; ...; \ln \frac{a_k}{b_k}\right] \tag{7}$$

We can gain an intuition of the VLR by considering its formula. Recall that the relationship between components is one of relative importance: for the feature pair $[\mathbf{a}, \mathbf{b}]$, the coordinates $[2, 4]$ and $[4, 8]$ have equivalent meaning. Therefore, it follows that the features **a** and **b** are associated if $\frac{\mathbf{a}}{\mathbf{b}}$ remains constant across all samples. Hence, we measure the variance of the (log-)ratios, such that VLR ranges from $[0, \mathrm{inf}]$ where 0 indicates a perfect association. Unfortunately, VLR lacks an intuitive scale, making non-zero values difficult to interpret ([43]).

Importantly, the VLR is *sub-compositionally coherent*: the removal of a third feature **c** would have no bearing on the variance of the (log-)ratio $\frac{\mathbf{a}}{\mathbf{b}}$. Yet, the VLR suffers from a key limitation: it is unscaled with respect to the variances of the log components ([43]). In other words, the magnitude of VLR depends partially on the variances of its constituent parts (i.e., var(**a**) and var(**b**)). This makes it difficult to compare VLR across pairs (e.g., comparing $\frac{\mathbf{a}}{\mathbf{b}}$ with $\frac{\mathbf{b}}{\mathbf{c}}$) ([43]). Still, unlike correlation, the VLR does not produce spurious results for compositional data, and in fact, provides the same result for both relative data and the absolute counter-part, all without requiring normalization or transformation.

4.6 Principal Component Analysis

Just as there are problems regarding between-sample distances and between-feature correlations, it follows that Principal Component Analysis (PCA) should not get applied directly to compositional data. Instead, analysts could apply PCA to clr-transformed data (resulting in an additional centering of the rows after log-transformation) ([6]). However, analysts must take care when interpreting the resultant PCA: covariances and correlations between features now exist with respect to the geometric mean reference. As such, when plotting features as arrows in the new coordinate space, the angles between them (i.e., the correlations) will usually change when subsets of the data are analyzed. However, the distances between feature pairs (i.e., the links between the arrow heads) remain invariable with respect to sub-compositions: these correspond to their log-ratio variance ([6]). Meanwhile, the usual PCA plot (with samples as points in a new coordinate space) projects the distances between samples using the Aitchison distance (which has the desired property of sub-compositional dominance). In combining these into a joint visualization of features and samples, the resultant log-ratio biplot (i.e., the "relative variation biplot") reveals associations between samples and features, and can also be used to infer power law relationships between features in an exploratory analysis ([6]). Such biplots are reminiscent of the visualizations obtained by Correspondence Analysis (CA). In fact, CA can indeed be used to approximate relative variation biplots provided the data are raised to a (small) power ([30]), the optimal size of which can be obtained by analyzing sub-compositional incoherence ([31]). Using CA with power transformation has the advantage that zeros in the data are handled naturally by the technique.

5 Compositional methods for sequence data

5.1 Methods for differential abundance

The ALDEx2 package, available for the R programming language, uses compositional data analysis principles to measure differential expression between two or more groups ([24]; [25]). Unlike conventional approaches to differential expression, ALDEx2 uses log-ratio transformation instead of effective library size normalization. The algorithm has five main parts. First, ALDEx2 uses the input data to create randomized instances based

on the compositionally valid Dirichlet distribution ([24]; [25]). This renders the data free of zeros. Second, each of these so-called Monte Carlo (MC) instances undergoes log-ratio transformation, most usually clr or iqlr transformation ([24]; [25]). Third, conventional statistical tests (i.e., Welch's t and Wilcoxon tests for two groups; glm and Kruskal-Wallis for two or more groups) get applied to each MC instance to generate p-values (p) and Benjamini-Hochberg adjusted p-values (BH) for each transcript ([24]; [25]). Fourth, these p-values get averaged across all MC instances to yield expected p-values ([24]; [25]). Fifth, one considers any transcript with an expected BH $< \alpha$ as statistically significant ([24]; [25]).

Although popular among meta-genomics researchers for analyzing the differential abundance of operational taxonomic units (OTUs) (e.g., [48]; [70]), the ALDEx2 package has not received wide-spread adoption in the analysis of RNA-Seq data. In part, this may have to do with our observation that ALDEx2 requires a large number of samples. This requirement may stem from its use of non-parametric testing, as suggested by the reduced power of other non-parametric differential expression methods ([61]; [75]), for example NOISeq ([64]). However, competing software packages like limma ([62]) and edgeR ([56]) also benefit from moderated t-tests that "share information between genes" to reduce per-transcript variance estimates and increase statistical power.

Still, even in the setting of large sample sizes, ALDEx2 has one major limitation: its usefulness depends largely on interpreting the log-ratio transformation as a normalization. If the log-ratio transformation does not sufficiently approximate an unchanged reference, the statistical tests will yield results that are hard to interpret. Another tool developed for analyzing the differential abundance of OTUs suffers from a similar limitation: ANCOM ([44]) uses presumed invariant features to guide the log-ratio transformation. The tendency to interpret differential abundance results as if they were derived from log-ratio "normalizations" highlights the importance of pursuing numeric and experimental techniques that can establish an unchanged reference. It also highlights the benefit of seeking novel methods that do not require using log-ratio transformations as a kind of normalization.

5.2 Methods for association

The SparCC package, available for the R programming language, replaces Pearson's correlation coefficient with an estimation of correlation based on its relationship to the VLR (and other terms) ([28]). The algorithm works by iteratively calculating a "basis correlation" under the assumption that the majority of pairs do not correlate (i.e., a sparse network) ([28]). Another algorithm, SPIEC-EASI, makes the same assumption that the underlying network is sparse, but bases its method on the inverse covariance matrix of clr-transformed data ([38]).

The propr package ([53]), available for the R programming language, implements proportionality as introduced in ([43]) and expounded in ([22]). Proportionality provides an alternative measure of association that is valid for relative data. One could think of proportionality as a modification to the VLR that uses information about the variability of individual features (gained by a log-ratio transformation) to give the VLR scale. It can be defined for the i-th and j-th features (e.g., transcripts) of a log-ratio transformed data matrix, $\tilde{\mathbf{a}}_i$ and $\tilde{\mathbf{a}}_j$, and thus also depends on the reference used for transformation. Unlike SparCC and SPIEC-EASI, proportionality does not assume an underlying sparse network.

At least three measures of proportionality exist. The first, ϕ, ranges from $[0, \inf]$ with 0 indicating perfect proportionality ([43]):

$$\phi(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) = \frac{\mathrm{var}(\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_j)}{\mathrm{var}(\tilde{\mathbf{a}}_i)}. \tag{8}$$

Its definition adjusts the VLR (in the numerator) by the variance of one of the log-ratio transformed features in that pair (in the denominator). The use of only one feature variance in the adjustment makes ϕ asymmetric (i.e., $\phi(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) \neq \phi(\tilde{\mathbf{a}}_j, \tilde{\mathbf{a}}_i)$).

The second, ϕ_s, also ranges from $[0, \inf]$ with 0 indicating perfect proportionality, but has a natural symmetry ([53]). Its definition adjusts the VLR by the variance of the log-product of the two features:

$$\phi_s(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) = \frac{\mathrm{var}(\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_j)}{\mathrm{var}(\tilde{\mathbf{a}}_i + \tilde{\mathbf{a}}_j)}. \tag{9}$$

The third, ρ_p, like correlation, takes on values from $[-1, 1]$, where a value of 1 indicates perfect proportionality ([22]). Its definition adjusts the VLR by the sum of the variances of the log-ratio transformed features in that pair (as subtracted from the value 1). Thus, ρ_p is symmetric.

$$\rho_p(\tilde{\mathbf{a}}_i, \tilde{\mathbf{a}}_j) = 1 - \frac{\mathrm{var}(\tilde{\mathbf{a}}_i - \tilde{\mathbf{a}}_j)}{\mathrm{var}(\tilde{\mathbf{a}}_i) + \mathrm{var}(\tilde{\mathbf{a}}_i)}. \tag{10}$$

Note that ρ_p and ϕ_s are monotonic functions of one another (i.e., you can compute ρ_p directly from ϕ_s and *vice versa*) (e.g., see ([22]) where ϕ_s is called $\tilde{\beta}^2$). Unlike Pearson's correlation coefficient, proportionality coefficients tend not to produce spurious results ([53]). Instead, proportionality serves as a robust measure of association when analyzing relative data ([43]). Although proportionality gives VLR scale, it is limited in that its interpretation still depends partly on using transformation as a kind of normalization (i.e., for the calculation of individual feature variances) ([22]). Still, its interpretability, along with its observed resilience to spurious results, makes it a good choice for inferring co-expression from RNA-Seq data ([43]) or co-abundance from meta-genomics data ([10]).

6 Challenges to compositional analyses

6.1 Challenges unique to count compositions

Compositional data analysis, because it relies on log-transformations, does not work when the data contain zeros. Yet, count compositional data are notably prone to zeros, those of which could signify either that a component is absent from a sample or otherwise only present at a quantity below the detection limit ([14]). For NGS abundance data, the difference between a zero and a one might be stochastic. How best to handle zeros remains a topic of ongoing research. However, it is common to replace zeros with a number less than the detection limit ([14]). Other replacement strategies would include adding a fixed value to all components, replacing zeros with the value one, or omitting zero-laden components altogether. A more principled (yet computationally expensive) way of replacing zeros is the Dirichlet sampling procedure implemented in ALDEx2 (as described above). Note that the simple addition of a pseudo-count to all components does not preserve the ratios between them, which can be amended by modifying the non-zero components in a multiplicative way ([46]).

Moreover, while count compositional data carry relative information, they differ from true compositional data in that they contain integer values only. Restricting the data to integer space can introduce problems with an analysis because the sampling variation becomes more noticeable as the measurements approach zero ([53]). In other words, the difference between 1 and 2 counts is not exactly the same as the difference between 1,000 and 2,0000 counts ([53]). While it is not mathematically necessary to remove low counts, analysts should proceed carefully in their presence.

6.2 Challenges unique to sequencing data

In the second section, we discussed how between-sample biases render NGS abundances incomparable between samples, thus necessitating normalization or transformation. However, we did not address two important sources of within-sample biases for sequencing data. The first is read length bias, in which more reads map to longer transcripts ([63]). The second is GC content bias, in which more reads map to high GC regions ([19]). Such biases distort the ratios between features and are thus relevant to compositional analysis as well. Yet, because within-sample biases are usually assumed to have the same proportional impact across all samples, they are usually ignored ([63]). For the same reason, one might also ignore these biases when interpreting NGS abundance data as compositions (as long as we are only interested in between-sample effects). However, if a sample were to contain, for example, a polymorphic or epigenetic change which alters the size or GC content of a transcript, the compositional nature of sequencing data could cause a skew in the observed abundances for all other transcripts (for reasons suggested by Figure 1). More work is needed to understand the extent to which within-sample biases impacts compositional data analysis in practice.

6.3 Limitations of transformation-based analysis

Formal transformation-based approaches often suffer from a lack of interpretability or otherwise get interpreted erroneously. For example, when using the centered log-ratio (clr) transformation, one may be tempted to interpret the transformed data as if they referred to single features (e.g., transcripts); however, the transformed data actually refer to the ratios of the transcripts to their geometric mean. As such, an analyst must interpret results with regard to their dependence on this mean. Moreover, because the geometric mean can change with the removal of features, the transformed data are incoherent with respect to sub-compositions.

When log-ratio transformations are used for scaled measures of association (i.e., proportionality), the resulting covariations depend on the implicitly chosen reference. Therefore, they will not give the same results for absolute and relative data (unless both data were transformed). The formal relationship of results when applying ρ_p with and without transformation is investigated elsewhere ([22]). Although lacking a natural scale, the log-ratio variance (VLR) has an advantage in that it provides identical results for both absolute and relative data, without requiring normalization or transformation.

6.4 The merits of ratio-based analysis

Aitchison's preferred summary of the covariance structure of a compositional data set was a matrix containing the log-ratio variances for all feature pairs (i.e., the variation matrix) ([2]). Although this matrix formally contains a lot of redundant information, an analyst who is familiar with the features might still find this kind of representation useful. Recently, the focus on ratios has been called the "pragmatic" approach to compositional data analysis ([32]), and offers some benefits. For one, transformation (i.e., the restriction to ratios with the same denominator) is not needed. Instead, the ratios can be dealt with directly as if they were unconstrained (i.e., absolute) data ([66]). Moreover, ratios may carry a clear meaning to the analyst interpreting them. Recently, Greenacre proposed a formal procedure to select a non-redundant subset of feature pairs that contains the entire variability of the data ([32]).

Such ratio-based analyses are also applicable to NGS abundance data. For example, Erb et al. proposed a method to identify the differential expression of gene ratios, a technique comprising part of what is termed differential proportionality analysis ([23]). When comparing gene ratios across two groups, this method selects ratios in which only a small portion of the total log-ratio variance (i.e., VLR) is explained by the sum of the within-group log-ratio variances ([23]). These selected gene ratios tend to show differences in the group means of those ratios, analogous to how genes selected by differential expression analysis show differences between their means ([23]). Reinforcing the analogy further, Erb et al. have shown how it is possible to use the limma package to apply an empirical Bayes model with underlying count-based precision weights ([62]; [39]) to gene ratios, thus quantifying "second order" expression effects while still avoiding normalization ([23]).

In addition to measuring differences in the means of gene ratios between groups, ratio-based methods (such as those used in differential proportionality analysis) can also help identify differences in the coordination of gene pairs. Such "differential coordination analysis" would otherwise depend on correlation ([76]), and therefore fall susceptible to spurious results. Instead, we can harness the advantages of the VLR to define a sub-compositionally coherent measure that tests for changes in the magnitude (i.e., slope of association) or strength (i.e., coefficient of association) of co-regulated gene pairs. Moreover, ratio-based analyses could work as normalization-free feature selection methods for data science applications (such as clustering and classification). Such techniques would especially suit large data sets aggregated from multiple sequencing centers, platforms, or modalities, where heterogeneity and batch effects are not easily normalized.

7 Summary

All NGS abundance data are compositional because sequencers sample only a portion of the total input material. However, RNA-Seq data might have compositional properties regardless owing to constraints on the cellular capacity for mRNA production. Whatever the reason, compositional data cannot undergo conventional analysis directly, at least without prior normalization or transformation. Otherwise, measures of differential expression, correlation, distance, and principal components become unreliable.

In the analysis of RNA-Seq data, effective library size normalization is used to recast the data in absolute terms prior to analysis. However, successful normalization requires meeting certain (often untestable) assumptions. Alternatively, log-ratio transformations provide a way to interrogate the data using familiar methods, but analysts must interpret their results with respect to the chosen reference. Sometimes, log-ratio transformations can be used to normalize the data, but this requires an approximation of an unchanged reference. Instead, shifting focus to the analysis of ratios yields methods that avoid normalization and transformation entirely. These ratio-based methods may represent an important future direction in the compositional analysis of relative NGS abundance data.

References

[1] J. Aitchison. The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):139–177, 1982.

[2] J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London, UK, UK, 1986.

[3] J Aitchison. A concise guide to compositional data analysis. *2nd Compositional Data Analysis Workshop; Girona, Italy*, 2003.

[4] J. Aitchison, C. Barceló-Vidal, J. A. Martín-Fernández, and V. Pawlowsky-Glahn. Logratio Analysis and Compositional Distance. *Mathematical Geology*, 32(3):271–275, April 2000.

[5] John Aitchison. The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies. *Proceedings of CoDaWork'08, The 3rd Compositional Data Analysis Workshop; Girona, Spain*, 2008.

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

[7] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[8] Giacomo Baruzzo, Katharina E. Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A. FitzGerald, and Gregory R. Grant. Simulation-based comprehensive benchmarking of RNA-seq aligners. *Nature Methods*, 14(2):135–139, February 2017.

[9] Ashlee M. Benjamin, Marshall Nichols, Thomas W. Burke, Geoffrey S. Ginsburg, and Joseph E. Lucas. Comparing reference-based RNA-Seq mapping methods for non-human primate data. *BMC Genomics*, 15:570, July 2014.

[10] Gaorui Bian, Gregory B. Gloor, Aihua Gong, Changsheng Jia, Wei Zhang, Jun Hu, Hong Zhang, Yumei Zhang, Zhenqing Zhou, Jiangao Zhang, Jeremy P. Burton, Gregor Reid, Yongliang Xiao, Qiang Zeng, Kaiping Yang, and Jiangang Li. The Gut Microbiota of Healthy Aged Chinese Is Similar to That of the Healthy Young. *mSphere*, 2(5):e00327–17, October 2017.

[11] C. I. Bliss and R. A. Fisher. Fitting the Negative Binomial Distribution to Biological Data. *Biometrics*, 9(2):176–200, 1953.

[12] K. Gerald van den Boogaart and Raimon Tolosana-Delgado. Descriptive Analysis of Compositional Data. In *Analyzing Compositional Data with R*, Use R!, pages 73–93. Springer, Berlin, Heidelberg, 2013. DOI: 10.1007/978-3-642-36809-7_4.

[13] K. Gerald van den Boogaart and Raimon Tolosana-Delgado. Fundamental Concepts of Compositional Data Analysis. In *Analyzing Compositional Data with R*, Use R!, pages 13–50. Springer Berlin Heidelberg, 2013. DOI: 10.1007/978-3-642-36809-7_2.

[14] K. Gerald van den Boogaart and Raimon Tolosana-Delgado. Zeroes, Missings, and Outliers. In *Analyzing Compositional Data with R*, Use R!, pages 209–253. Springer, Berlin, Heidelberg, 2013. DOI: 10.1007/978-3-642-36809-7_7.

[15] Antonella Buccianti. Is compositional data analysis a way to see beyond the illusion? *Computers & Geosciences*, 50:165–173, January 2013.

[16] Ana Conesa, Pedro Madrigal, Sonia Tarazona, David Gomez-Cabrero, Alejandra Cervera, Andrew McPherson, Michał Wojciech Szcześniak, Daniel J. Gaffney, Laura L. Elo, Xuegong Zhang, and Ali Mortazavi. A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17:13, 2016.

[17] Marie-Agnès Dillies, Andrea Rau, Julie Aubert, Christelle Hennequet-Antier, Marine Jeanmougin, Nicolas Servant, Céline Keime, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom, Mickaël Guedj, and Florence Jaffrézic. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*, 14(6):671–683, November 2013.

[18] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[19] Juliane C. Dohm, Claudio Lottaz, Tatiana Borodina, and Heinz Himmelbauer. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 36(16):e105, September 2008.

[20] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. Isometric Logratio Transformations for Compositional Data Analysis. *Mathematical Geology*, 35(3):279–300, April 2003.

[21] Pär G. Engström, Tamara Steijger, Botond Sipos, Gregory R. Grant, André Kahles, The RGASP Consortium, Gunnar Rätsch, Nick Goldman, Tim J. Hubbard, Jennifer Harrow, Roderic Guigó, and Paul Bertone. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods*, 10(12):1185–1191, December 2013.

[22] Ionas Erb and Cedric Notredame. How should we measure proportionality on relative gene expression data? *Theory in Biosciences*, January 2016.

[23] Ionas Erb, Thomas Quinn, David Lovell, and Cedric Notredame. Differential Proportionality - A Normalization-Free Approach To Differential Gene Expression. *Proceedings of CoDaWork 2017, The 7th Compositional Data Analysis Workshop; available under bioRxiv*, page 134536, May 2017.

[24] Andrew D. Fernandes, Jean M. Macklaim, Thomas G. Linn, Gregor Reid, and Gregory B. Gloor. ANOVA-Like Differential Expression (ALDEx) Analysis for Mixed Population RNA-Seq. *PLOS ONE*, 8(7):e67019, July 2013.

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

[26] Nuno A. Fonseca, John Marioni, and Alvis Brazma. RNA-Seq gene profiling–a systematic empirical comparison. *PloS One*, 9(9):e107026, 2014.

[27] Nuno A. Fonseca, Johan Rung, Alvis Brazma, and John C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, December 2012.

[28] Jonathan Friedman and Eric J. Alm. Inferring correlation networks from genomic survey data. *PLoS computational biology*, 8(9):e1002687, 2012.

[29] Gregory R. Grant, Michael H. Farkas, Angel D. Pizarro, Nicholas F. Lahens, Jonathan Schug, Brian P. Brunk, Christian J. Stoeckert, John B. Hogenesch, and Eric A. Pierce. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, September 2011.

[30] Michael Greenacre. Power transformations in correspondence analysis. *Computational Statistics & Data Analysis*, 53(8):3107–3116, June 2009.

[31] Michael Greenacre. Measuring Subcompositional Incoherence. *Mathematical Geosciences*, 43(6):681–693, August 2011.

[32] Michael Greenacre. Towards a pragmatic approach to compositional data analysis. Technical Report 1554, Department of Economics and Business, Universitat Pompeu Fabra, January 2017.

[33] Malachi Griffith, Jason R. Walker, Nicholas C. Spies, Benjamin J. Ainscough, and Obi L. Griffith. Informatics for RNA Sequencing: A Web Resource for Analysis on the Cloud. *PLoS computational biology*, 11(8):e1004393, August 2015.

[34] Ayat Hatem, Doruk Bozdağ, Amanda E. Toland, and Ümit V. Çatalyürek. Benchmarking short sequence mapping tools. *BMC Bioinformatics*, 14:184, June 2013.

[35] Katharina E. Hayer, Angel Pizarro, Nicholas F. Lahens, John B. Hogenesch, and Gregory R. Grant. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics (Oxford, England)*, 31(24):3938–3945, December 2015.

[36] Steven R. Head, H. Kiyomi Komori, Sarah A. LaMere, Thomas Whisenant, Filip Van Nieuwerburgh, Daniel R. Salomon, and Phillip Ordoukhanian. Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, 56(2):61–passim, February 2014.

[37] Lichun Jiang, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. Synthetic spike-in standards for RNA-seq experiments. *Genome Research*, 21(9):1543–1551, September 2011.

[38] Zachary D. Kurtz, Christian L. Müller, Emily R. Miraldi, Dan R. Littman, Martin J. Blaser, and Richard A. Bonneau. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. *PLOS Computational Biology*, 11(5):e1004226, May 2015.

[39] Charity W. Law, Yunshun Chen, Wei Shi, and Gordon K. Smyth. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, 15:R29, January 2014.

[40] Jun-Hao Li, Shun Liu, Ling-Ling Zheng, Jie Wu, Wen-Ju Sun, Ze-Lin Wang, Hui Zhou, Liang-Hu Qu, and Jian-Hua Yang. Discovery of protein–lncRNA interactions by integrating large-scale CLIP-Seq and RNA-Seq datasets. *Bioinformatics and Computational Biology*, 2:88, 2015.

[41] Yanzhu Lin, Kseniya Golovnina, Zhen-Xia Chen, Hang Noh Lee, Yazmin L. Serrano Negron, Hina Sultana, Brian Oliver, and Susan T. Harbison. Comparison of normalization and differential expression analyses using RNA-Seq data from 726 individual Drosophila melanogaster. *BMC Genomics*, 17, January 2016.

[42] Robert Lindner and Caroline C. Friedel. A Comprehensive Evaluation of Alignment Algorithms in the Context of RNA-Seq. *PLOS ONE*, 7(12):e52403, December 2012.

[43] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: A Valid Alternative to Correlation for Relative Data. *PLoS Computational Biology*, 11(3), March 2015.

[44] Siddhartha Mandal, Will Van Treuren, Richard A. White, Merete Eggesbø, Rob Knight, and Shyamal D. Peddada. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microbial Ecology in Health and Disease*, 26:27663, 2015.

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

[46] JA Martín-Fernández and S Thió-Henestrosa. Rounded zeros: some practical aspects for compositional data. *Geological Society, London, Special Publications*, 264(1):191–201, 2006.

[47] Glòria Mateu-Figueras, Vera Pawlowsky-Glahn, and Juan José Egozcue. The Principle of Working on Coordinates. In Vera Pawlowsky-Glahn and Antonella Buccianti, editors, *Compositional Data Analysis*, pages 29–42. John Wiley & Sons, Ltd, 2011. DOI: 10.1002/9781119976462.ch3.

[48] Amy McMillan, Stephen Rulisa, Mark Sumarah, Jean M. Macklaim, Justin Renaud, Jordan E. Bisanz, Gregory B. Gloor, and Gregor Reid. A multi-platform metabolomics approach identifies highly specific biomarkers of bacterial diversity in the vagina of pregnant and non-pregnant women. *Scientific Reports*, 5:14174, September 2015.

[49] Gabriela A Merino, Ana Conesa, and Elmer A Fernandez. A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. *bioRxiv*, 2017.

[50] Michael L. Metzker. Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46, January 2010.

[51] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nature methods*, 14(4):417, 2017.

[52] Karl Pearson. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity, and Panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318, 1896.

[53] Thomas Quinn, Mark F. Richardson, David Lovell, and Tamsyn Crowley. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. *bioRxiv*, page 104935, February 2017.

[54] Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E. Mason, Nicholas D. Socci, and Doron Betel. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biology*, 14(9):R95, 2013.

[55] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, April 2015.

[56] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, January 2010.

[57] Mark D. Robinson and Alicia Oshlack. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11:R25, 2010.

[58] Edoardo Saccenti. Correlation Patterns in Experimental Data Are Affected by Normalization Procedures: Consequences for Data Analysis and Network Inference. *Journal of Proteome Research*, November 2016.

[59] Nicholas J. Schurch, Pietá Schofield, Marek Gierliński, Christian Cole, Alexander Sherstnev, Vijender Singh, Nicola Wrobel, Karim Gharbi, Gordon G. Simpson, Tom Owen-Hughes, Mark Blaxter, and Geoffrey J. Barton. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA (New York, N.Y.)*, 22(6):839–851, June 2016.

[60] Matthew Scott, Carl W Gunderson, Eduard M Mateescu, Zhongge Zhang, and Terence Hwa. Interdependence of cell growth and gene expression: origins and consequences. *Science*, 330(6007):1099–1102, 2010.

[61] Fatemeh Seyednasrollah, Asta Laiho, and Laura L. Elo. Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, 16(1):59–70, January 2015.

[62] Gordon K. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:Article3, 2004.

[63] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14:91, 2013.

[64] Sonia Tarazona, Pedro Furió-Tarí, David Turrà, Antonio Di Pietro, María José Nueda, Alberto Ferrer, and Ana Conesa. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. *Nucleic Acids Research*, 43(21):e140–e140, December 2015.

[65] Mingxiang Teng, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, Christopher E. Mason, Sara Olson, Dmitri Pervouchine, Cricket A. Sloan, Xintao Wei, Lijun Zhan, and Rafael A. Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17:74, March 2016.

[66] C. W. Thomas and J. Aitchison. Log-ratios and geochemical discrimination of Scottish Dalradian limestones: a case study. *Geological Society, London, Special Publications*, 264(1):25–41, January 2006.

[67] Hande Topa and Antti Honkela. Analysis of differential splicing suggests different modes of short-term splicing regulation. *Bioinformatics*, 32(12):i147–i155, June 2016.

[68] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[69] Cole Trapnell, Brian A. Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J. van Baren, Steven L. Salzberg, Barbara J. Wold, and Lior Pachter. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, 28(5):511–515, May 2010.

[70] Camilla Urbaniak, Michelle Angelini, Gregory B. Gloor, and Gregor Reid. Human milk microbiota profiles in relation to birthing method, gestation and infant gender. *Microbiome*, 4:1, 2016.

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

[72] W. A. Wang, C. T. Wu, T. P. Lu, M. H. Tsai, L. C. Lai, and E. Y. Chuang. Comparisons and performance evaluations of RNA-seq alignment tools. In *2014 International Conference on Electrical Engineering and Computer Science (ICEECS)*, pages 215–218, October 2014.

[73] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, January 2009.

[74] Alex D. Washburne, Justin D. Silverman, Jonathan W. Leff, Dominic J. Bennett, John L. Darcy, Sayan Mukherjee, Noah Fierer, and Lawrence A. David. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ*, 5, February 2017.

[75] Claire R. Williams, Alyssa Baccarella, Jay Z. Parrish, and Charles C. Kim. Empirical assessment of analysis workflows for differential expression analysis of human samples using RNA-Seq. *BMC Bioinformatics*, 18, January 2017.

[76] Tianwei Yu and Yun Bai. Capturing changes in gene expression dynamics by gene set differential coordination analysis. *Genomics*, 98(6):469–477, December 2011.