





## TECHNICAL NOTE

# A field guide for the compositional analysis of any-omics data

Thomas P. Quinn <sup>1,2,\*</sup>, Ionas Erb<sup>3</sup>, Greg Gloor <sup>4</sup>, Cedric Notredame <sup>3</sup>, Mark F. Richardson<sup>1,5,6,†</sup> and Tamsyn M. Crowley <sup>7,†</sup>

<sup>1</sup>Bioinformatics Core Research Group, Deakin University, 1 Gheringhap Street, Geelong Victoria 3220, Australia; <sup>2</sup>Centre for Molecular and Medical Research, Deakin University, 1 Gheringhap Street, Geelong Victoria 3220, Australia; <sup>3</sup>Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Dr Aiguader 88, Barcelona 08003, Spain; <sup>4</sup>Department of Biochemistry, University of Western Ontario, 1151 Richmond Street, London ON N6A 3K7, Canada; <sup>5</sup>Genomics Centre, School of Life and Environmental Sciences, Deakin University, 1 Gheringhap Street, Geelong Victoria 3220, Australia; <sup>6</sup>Centre for Integrative Ecology, School of Life and Environmental Sciences, Deakin University, 1 Gheringhap Street, Geelong Victoria 3220, Australia and <sup>7</sup>Poultry Hub Australia, University of New England, Elm Avenue, Armidale New South Wales 2351, Australia

\*Correspondence address. Thomas P. Quinn, Deakin University, 1 Gheringhap Street, Geelong Victoria 3220, Australia. E-mail: [contacttomquinn@gmail.com](mailto:contacttomquinn@gmail.com)  <http://orcid.org/0000-0003-0286-6329>

†Contributed equally.

## Abstract

**Background:** Next-generation sequencing (NGS) has made it possible to determine the sequence and relative abundance of all nucleotides in a biological or environmental sample. A cornerstone of NGS is the quantification of RNA or DNA presence as counts. However, these counts are not counts per se: their magnitude is determined arbitrarily by the sequencing depth, not by the input material. Consequently, counts must undergo normalization prior to use. Conventional normalization methods require a set of assumptions: they assume that the majority of features are unchanged and that all environments under study have the same carrying capacity for nucleotide synthesis. These assumptions are often untestable and may not hold when heterogeneous samples are compared. **Results:** Methods developed within the field of compositional data analysis offer a general solution that is assumption-free and valid for all data. Herein, we synthesize the extant literature to provide a concise guide on how to apply compositional data analysis to NGS count data. **Conclusions:** In highlighting the limitations of total library size, effective library size, and spike-in normalizations, we propose the log-ratio transformation as a general solution to answer the question, “Relative to some important activity of the cell, what is changing?”

## Introduction

The advent of next-generation sequencing (NGS) has allowed scientists to probe biological systems in unprecedented ways. For an ever-decreasing sum of money, it is possible to determine the sequence and relative abundance of all nucleotide fragments in a sample [1]. NGS works by sequencing a population of

DNA fragments, including reverse-transcribed RNA isolates. In addition to its general use for variant discovery and genome assembly, NGS is used to quantify relative abundances of (i) RNA species from tissue (RNA sequencing [RNA-Seq]) [1], (ii) organism diversity from the environment (metagenomics) [2], (iii) RNA species from the environment (meta-transcriptomics) [3], and (iv) regions of the genome targeted by a protein (chromatin im-

Received: 14 February 2019; Revised: 10 July 2019; Accepted: 12 August 2019

© The Author(s) 2019. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

munoprecipitation sequencing) [4], among others. Recently, improvements in the sequencing protocols have allowed for these measurements to be carried out at the single-cell level, with single-cell RNA-Seq being the most mature technology. Most applications share an analogous procedure whereby DNA or RNA are isolated from samples, optionally filtered by size or other property [5], converted to a complementary DNA (cDNA) library of nucleotide fragments, sequenced on a sequencer, and then mapped to a reference to quantify relative abundance. Because all data derive from the same assay, one might expect that they would undergo the same analysis. However, this is not true: rather, methods tailored for one mode of data do not generalize to another (e.g., RNA-Seq methods have inflated false discovery rates [FDRs] when applied to metagenomics data [6, 7]).

Fernandes et al. posited that the analysis of all NGS data can be conceptually unified by recognizing the compositional nature of these data [8]. By “compositional,” we mean that the abundance of any 1 nucleotide fragment is only interpretable relative to another. This property emerges from the sequencer itself; the sequencer, by design, can only sequence a fixed number of nucleotide fragments. Consequently, the final number of fragments sequenced is constrained to an arbitrary limit so that doubling the input material does not double the total number of counts. This constraint also means that an increase in the presence of any 1 nucleotide fragment necessarily decreases the observed abundance of all other transcripts [9], and applies to bulk and single-cell sequencing data alike. It is especially problematic when comparing cells that produce more total RNA than their comparator (e.g., high-c-Myc cells, which up-regulate 90% of all transcripts without commensurate down-regulation [10]). However, even if a sequencer could directly sequence every RNA molecule within a cell, the cells themselves are compositional because of the volume and energy constraints that limit RNA synthesis, as evidenced by the observation that smaller cells of a single type contain proportionally less total messenger RNA (mRNA) [11].

Compositional data only carry relative information. Consequently, they exist in a simplex space with 1 fewer dimensions than components. Analyzing relative data as if they were absolute can yield erroneous results for several common techniques [12–14] (also demonstrated in Supplementary Analysis S1). First, statistical models that assume independence between features are flawed because of the mutual dependency between components [15]. Second, distances between samples are misleading and erratically sensitive to the arbitrary inclusion or exclusion of components [16]. Third, components can appear definitively correlated even when they are statistically independent [17]. For these reasons, compositional data pose specific challenges to the differential expression, clustering, and correlation analyses routinely applied to NGS data, as well as other data that measure the relative abundance of small molecules (e.g., spectrometric peak data [18]). For compositional NGS data, each sample is called a “composition” and each nucleotide species is called a “component” [13, 14].

There are 3 general approaches to analyzing compositional data. First, the “normalization-dependent” approach seeks to normalize the data in order to reclaim absolute abundances. However, normalizations depend on assumptions that may not hold true outside of tightly controlled experiments. For example, popular RNA-Seq normalization methods assume that most transcripts have the same absolute abundance across samples [19, 20], an assumption that does not hold for the aforementioned high-c-Myc cells [10]. Second, the “transformation-dependent” approach transforms the data with regard to a ref-

erence to make statistical inferences relative to the chosen reference [12]. Third, the “transformation-independent” approach performs calculations directly on the components [21] or component ratios [22].

The latter 2 approaches constitute compositional data analysis (CoDA). Unlike normalization-based methods, CoDA methods will generalize to all data, relative or absolute. In this article, we describe a unified pipeline for the analysis of NGS count data, with all parts fully capable of modeling the uncertainty of counts with low abundance. First, we show how existing CoDA software tools can be used to draw compositionally valid and biologically meaningful conclusions. Second, we illustrate how these methods can accommodate complex study design, facilitate the analysis of horizontally integrated multi-omics data, and accommodate machine learning applications. Third, we show how compositionality can systematically bias results if ignored. Finally, we conclude with a discussion of key problems associated with spike-in normalization, and show how the CoDA framework applies specifically to single-cell sequencing data.

## Methods

### Overview of pipeline

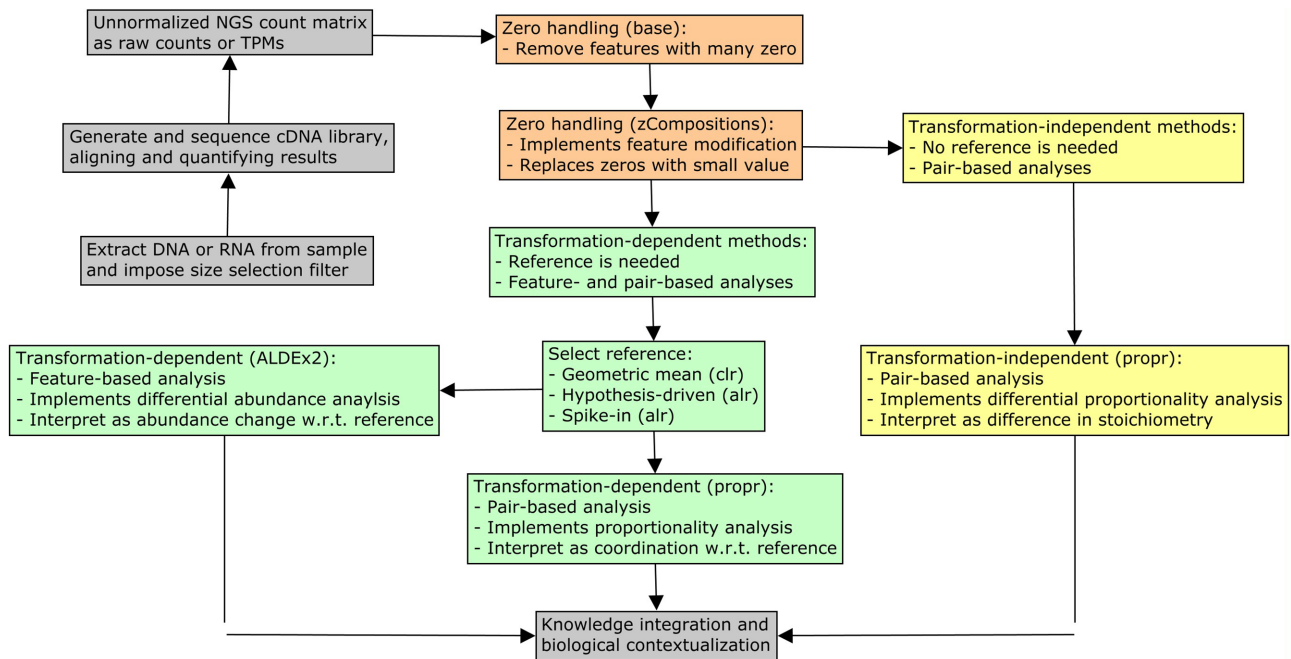
Our pipeline uses software tools made freely available for the R programming language. It begins with an unnormalized “count matrix” generated from the alignment and read-mapping of a sequence library. Details regarding quality control, assembly, alignment, and read-mapping are beyond the scope of this article and have been covered extensively elsewhere (e.g., [23, 24]). This count matrix records the number of times each feature (e.g., transcript or operational taxonomic unit [OTU]) appears in each sample. Most software returns measurements as integer counts, although some uses continuous values (e.g., Salmon quasi-counts [25]) or another proportional unit (e.g., transcripts per million [TPM] [26]). For many CoDA methods, units have no importance. However, small counts carry more uncertainty than large counts, and our pipeline can model this directly. Therefore, we recommend using unadjusted “raw counts.” TPM can also be used with CoDA methods but can bias the modeling of small counts if the library size differs greatly between samples. Otherwise, the data should not undergo further normalization or standardization and must never contain negative values. Fig. 1 provides a schematic of our unified NGS pipeline.

### Data acquisition

To demonstrate the utility of our pipeline, we use publicly available time course data of the RNA and protein expressed by mouse dendritic cells following lipopolysaccharide (LPS) exposure, a potent immunogenic stimulus. RNA-Seq and mass spectrometry (MS) data were acquired already preprocessed to measure the relative abundance of 3,147 genes in TPM-equivalent units [27]. The RNA-Seq and MS data had 28 overlapping samples, spanning 2 conditions with 7 time points and 2 replicates each.

```
# Read in the RNA-Seq data
rnaseq <- read.csv("rnaseq-x.csv", row.names=1)
rnaseq.annot <- read.csv("rnaseq-y.csv", row.names=1)

# Read in the Mass Spec HL data
masshl <- read.csv("masshl-x.csv", row.names=1)
masshl.annot <- read.csv("masshl-y.csv", row.names=1)
```



**Figure 1:** This figure illustrates how our unified NGS pipeline might sit within a larger workflow. Colored boxes indicate procedures that would apply to any relative data set. In orange, we describe the optional zero removal and modification steps presented in section “Zero handling.” In green, we describe the log-ratio transformation-dependent methods presented in section “Transformation-dependent analyses.” This includes the differential abundance analysis of individual features and the proportionality analysis of feature pairs. In yellow, we describe the transformation-independent methods presented in section “Transformation-independent analyses.” This includes the analysis of the differences in the log-ratio means of feature pairs. In gray, we describe other essential steps unique to the data type under study but not covered here. w.r.t.: with regard to.

```
# We will subset Mass Spec to include timepoints
# with a corresponding RNA-Seq measurement
# (used in “Vertical Data Integration”)
inRNAandMS <- masshl.annot$Time
masshl <- masshl[,inRNAandMS]
masshl.annot <- masshl.annot[inRNAandMS,]
```

## New analyses

In presenting this workflow, we perform a new analysis of the Jovanovic et al. [27] data in order to learn how mRNA transcript abundance and protein abundance change in response to LPS stimulation. This includes a relative differential abundance analysis, an analysis of gene-gene coordination, and an analysis of differential gene-gene coordination. In addition, we integrate the 2 data types with a differential proportionality analysis to evaluate how mRNA stoichiometry differs from protein stoichiometry in response to LPS treatment. Unlike the original analysis presented by Jovanovic et al. [27], we do not use transcripts per million (TPM) normalization. Rather, we argue that TPMs recast an already compositional data set as yet another compositional data set (just with a different denominator). In Supplementary Analysis S1, we show how TPMs introduce systematic errors. This is because when a reference is not explicitly chosen, an arbitrary reference is still implicitly present. We also include an appendix (Supplementary Analysis S2) that benchmarks how several zero-handling procedures affect proportionality and differential proportionality analysis.

## Software contributions

This workflow primarily uses 3 open source software packages, all of which are available for the R programming language. They include zCompositions [28], ALDEx2 [8, 29], and propr [30, 31]. The reader can download these software packages from Bioconductor and CRAN.

```
install.packages(“zCompositions”)
install.packages(“propr”)
install.packages(“BiocManager”)
# Read ‘::’ as ‘the install function from the
# BiocManager package”
BiocManager::install(“ALDEx2”)
library(zCompositions)
library(ALDEx2)
library(propr)
```

In preparing this workflow, we have made several contributions to the compositional data analysis software universe. First, we present the new `propr::aldex2propr` function that integrates the ALDEx2 and propr packages by calculating an average proportionality coefficient over ALDEx2-generated Monte Carlo instances. Second, we present the new `propr::updateCutoffs` function that permutes FDR across varying proportionality coefficient cut-offs. Third, we present the `propr::propd` function that implements the differential proportionality method described by Erb et al. [31], including an implementation of a zero-handling procedure based on the Box-Cox transform. These new contributions make a complete compositional data analysis workflow possible.

## Benchmark validation

Although one can devise a “normalizing” reference by invoking a set of assumptions, we prefer an alternative framework that does not require any normalization. We use this framework because it provides a more general solution to the analysis of -omics data. As such, our proposed workflow could be used to analyze bulk RNA-Seq, single-cell RNA-Seq, metagenomics, metabolomics, lipidomics, and other data.

Although the software tools presented here do not normalize the data, they can be benchmarked against conventional methods by invoking the assumption that the explicit reference performs a kind of “log-ratio normalization.” Under these conditions, ALDEx2 can identify differential abundance with high precision in RNA-Seq data [8, 32] and control false-positive rates in highly sparse 16S metagenomics count data [6]. Meanwhile, proportionality analysis has been shown to outperform all 15 competing measures of association in single-cell clustering and network inference tasks across 213 data sets [33]. Although differential proportionality analysis has not yet been benchmarked, it is formally related to an ANOVA, a foundational test in most biological research. As a statistical test for significance, it is valid wherever an ANOVA is valid. We also include an appendix (Supplementary Analysis S2) that benchmarks how several zero-handling procedures affect proportionality and differential proportionality analysis.

## Zero handling

### General strategies for zero handling

CoDA methods depend on logarithms that do not compute for zeros. Therefore, we must address zeros prior to, or during, the pipeline. Before handling zeros, the analyst must first consider the nature of the zeros. There exists 3 types of zeros: (i) “rounding,” also called “sampling,” where the feature exists in the sample below the detection limit; (ii) “count,” where the feature exists in the sample, but counting is not exhaustive enough to see it at least once; and (iii) “essential,” where the feature does not exist in the sample at all [34]. The approach to zero handling depends on the nature of the zeros [34]. For NGS data, a nucleotide fragment is either sequenced or not, and would not contain rounding zeros. Because there is no general methodology for dealing with essential zeros within a strict CoDA framework [34], we assume that any feature present in  $\geq 1$  sample could appear in another sample if sequenced with infinite depth, and thus treat all NGS zeros as “count zeros.” Others have also suggested that the essential zeros of NGS count data are sufficiently modeled as sampling zeros [35].

There are 2 general approaches to zero handling. In “feature removal,” components with zeros get excluded, yielding a subcomposition that can be analyzed by any CoDA method. Feature removal is usually appropriate when a feature contains many zeros and can always be justified for essential zeros. In “feature modification,” zeros get replaced with a non-zero value, with or without modification to non-zeros. Analysts may choose 1 or both zero-handling procedures but should always demonstrate that the removal or modification of zero-laden features does not change the overall interpretation of the results.

### Feature modification with zCompositions

For “count zeros,” Martin-Fernández et al. recommend replacing zeros by a Bayesian-multiplicative replacement strategy that preserves the ratios between the non-zero components [34], im-

plemented in the zCompositions package as the `cmultRepl` function [28]. Alternatively, one could use a multiplicative simple replacement strategy, whereby zeros get replaced with a fixed value  $< 1$  in a compositionally robust manner. Here, we use zCompositions to replace zeros.

```
# Standard functions expect rows as samples
# so we will transpose the matrix
rnaseq <- t(rnaseq)
masshl <- t(masshl)

# Now we can replace zeros with a small value
# the ‘p-counts’ option has the function return
# pseudo-counts instead of proportions
library(zCompositions)
rnaseq.no0 <- cmultRepl(rnaseq, output = ‘p-counts’).
```

One can interpret this “up-regulation” to mean that the gene increases its expression in response to LPS stimulation more than nuclear factor  $\kappa$ B (NF $\kappa$ B). All P-values correspond to the expectation of the Benjamini-Hochberg adjusted P-values computed from a Welch’s t-test over 128 simulated instances of the data. By choosing a reference that is relevant to the biological system under study, we can gain meaningful insights from the data without any need for normalization. In Table 1, between-group differences are the differences between the 2 conditions (defined for each Dirichlet instance), within-group differences are the maximum difference across Dirichlet instances (defined for each condition), and effect sizes are the ratio of the between-group differences to the maximum of within-group differences (defined for each Dirichlet instance). The columns “Effect size,” “Difference (between),” and “Difference (within)” report the median effect size, median between-group difference, and median within-group difference, respectively.

```
masshl.no0 <- cmultRepl(masshl, output = ‘p-counts’)
```

Many compositional software tools have their own built-in zero-handling procedures. Although zCompositions is not necessarily better than these built-in procedures, we recognize that removing zeros right away has a practical advantage: by using zCompositions in combination with a log-ratio transformation, analysts can apply most conventional analyses to their compositional data right away. Because zCompositions empowers readers to use methods beyond those presented here, we decided to include it as the first part of our field guide. However, we recommend that readers look at Supplementary Analysis S2, which benchmarks how several zero-handling procedures affect proportionality and differential proportionality analysis.

## Transformation-dependent analyses

### The log-ratio transformation

All components in a composition are mutually dependent features that cannot be understood in isolation. Therefore, any analysis of individual components is done with respect to a reference. This reference transforms each sample into an unbounded space where any statistical method can be used. The centered log-ratio (clr) transformation uses the geometric mean of the sample vector as the reference [36]. The additive log-ratio (alr) transformation uses a single component as the reference [36]. Other transformations use specialized references based on the geometric mean of a subset of components (collectively called multi-additive log-ratio [malr] transformations [32]). One

**Table 1.** The 47 genes selected as significantly up-regulated by ALDEx2 when using the NF $\kappa$ B subunits as a reference

Gene	Effect size	Difference (between)	Difference (within)	Expected Benjamini-Hochberg P-value
Il1b	4.7372	3.9576	0.6912	0.0000
Irg1	4.3462	3.8904	0.7888	0.0000
Il1a	3.5950	3.8242	0.9037	0.0000
Cd40	2.2887	5.3325	2.0422	0.0000
Ifih1	2.2056	2.8529	1.1157	0.0000
Isg15	1.9678	4.4490	1.8330	0.0000
Oasl1	1.9304	5.6562	2.1200	0.0000
Ifit1	1.8317	5.6101	2.0773	0.0000
Ptgs2	1.6923	4.0869	2.0606	0.0002
Gbp5;Gbp1	1.6523	2.4494	1.2349	0.0000
Rsad2	1.4933	6.2747	2.4692	0.0001
Marcks1	1.4886	1.0748	0.5740	0.0001
BC006779	1.4686	2.2184	1.2465	0.0001
Mndal	1.4163	2.1047	1.5182	0.0000
Parp14	1.3139	1.7655	0.9357	0.0002
Ifi205	1.2916	5.3159	3.4587	0.0026
Slc7a2	1.2883	1.3797	0.9920	0.0002
Ifit2	1.2292	5.4975	2.6744	0.0002
Clic4	1.2037	0.8486	0.5765	0.0003
Sp140	1.1612	1.0030	0.7385	0.0005
Cmpk2	1.1149	5.7323	2.1088	0.0003
Stat5a	1.0806	0.8666	0.6461	0.0017
Ifi47	1.0443	2.0495	1.5704	0.0030
Pyhin1	1.0152	1.9150	1.4752	0.0024
Ifit3	0.9978	4.7313	3.2116	0.0012
Ccl5	0.9962	2.0765	1.6671	0.0015
Acsl1	0.9937	1.0837	1.0073	0.0009
Il1rn	0.9811	0.6795	0.6366	0.0017
Irgm1	0.9755	1.7076	1.0634	0.0094
IIGP;ligp1	0.9588	3.5610	3.1760	0.0023
Rnf213;AK217856	0.9541	1.2867	1.0478	0.0041
Daxx	0.9118	1.1938	0.9013	0.0119
Flnb	0.8639	1.6654	1.8185	0.0122
Cd274	0.8299	0.6050	0.6354	0.0051
Trex1	0.8171	0.5647	0.6350	0.0090
Car13	0.7586	1.1455	1.2839	0.0140
Xaf1	0.7550	1.5118	1.4338	0.0214
Gbp3	0.7478	1.5118	1.4837	0.0128
Ehd1	0.7460	0.3648	0.4812	0.0078
Gm4902	0.7413	1.9614	1.7899	0.0151
Rasa4	0.7254	0.8805	0.9109	0.0478
Oas3	0.7089	1.5673	1.7756	0.0213
Serpinb2	0.7048	1.7770	2.1734	0.0272
Dhx58;D11lgp2	0.6947	1.4875	1.6956	0.0425
Gbp2	0.6597	1.5376	1.7339	0.0212
Saa3	0.6291	1.0259	1.5384	0.0187
Sbds	0.5522	0.3107	0.5363	0.0443

One can interpret this “up-regulation” to mean that the gene increases its expression in response to LPS stimulation more than NF $\kappa$ B. All P-values correspond to the expectation of the Benjamini-Hochberg adjusted P-values computed from a Welch’s t-test over 128 simulated instances of the data. By choosing a reference that is relevant to the biological system under study, we can gain meaningful insights from the data without any need for normalization. In this table, between-group differences are the differences between the 2 conditions (defined for each Dirichlet instance), within-group differences are the maximum difference across Dirichlet instances (defined for each condition), and effect sizes are the ratio of the between-group differences to the maximum of within-group differences (defined for each Dirichlet instance). The columns “Effect size,” “Difference (between),” and “Difference (within)” report the median effect size, median between-group difference, and median within-group difference, respectively.

malr transformation is the inter-quartile log-ratio (iqlr) transformation, which uses components in the interquartile range of variance [37]. Another, the robust centered log-ratio (rclr) transformation, only uses the non-zero components [38].

Importantly, transformations are not normalizations: while normalizations claim to recast the data in absolute terms, transformations do not. The results of a transformation-based analysis must be interpreted with respect to the chosen reference. Of

these, the clr transformation is most common:

$$\text{clr}(\mathbf{x}_j) = \left[ \ln \frac{x_{1,j}}{g(\mathbf{x}_j)}, \dots, \ln \frac{x_{D,j}}{g(\mathbf{x}_j)} \right], \quad (1)$$

where  $\mathbf{x}_j$  is the  $j$ th sample and  $g(\mathbf{x}_j)$  is its geometric mean. The other transformations replace  $g(\mathbf{x}_j)$  with a different reference.

The isometric log-ratio (ilr) transformation uses an orthonormal basis as the reference [39] and is preferred when a non-singular covariance matrix is needed [21]. When the basis is a branch of a dendrogram, the ilr offers an intuitive way to contrast 1 set of components against another set of components. These contrasts, called balances, have been used to analyze metagenomics data based on evolutionary trees [40, 41] but could be applied to any data if a similarly meaningful tree were available.

Each transformation implies its own reference(s). In most practical settings, the choice of transformation will depend on the preferred interpretation. An analysis of clr data will reveal how genes (or OTUs) behave relative to the per-sample average. An analysis of alr and malr data will reveal how genes (or OTUs) behave relative to 1 or more explicitly chosen internal references. An analysis of iqlr data will reveal how genes (or OTUs) behave relative to the per-sample interquartile (“robust”) average. In a compositional framework, none of these are normalizations: each new variable is a log-ratio of the original variable divided by the reference and therefore should get interpreted as a kind of within-sample log-fold difference. Although the difference between transformation and normalization may seem subtle, it can have a profound impact on the conclusions drawn from the analysis. Although the temptation will exist, one must never confuse the transformed data with absolute abundances.

## Differential abundance analysis with ALDEx2

Differential abundance (DA) analysis seeks to identify which features differ in abundance between experimental groups. The ALDEx2 package tests for DA in compositional data by performing univariate statistical analyses on log-ratio transformed data [8, 29]. It does so with a layer of complexity that controls for technical variation by finding the expectation of  $B$  simulated instances of the data, each sampled from the Dirichlet distribution. This procedure implicitly models the uncertainty of low counts while also handling zeros.

Importantly, ALDEx2 identifies DA with respect to the chosen reference. By default, this reference is the geometric mean of the composition. It is possible, if not likely, that the mean centers are not the ideal references; if so, differences in the transformed abundances would not reflect differences in the absolute abundances. On the other hand, if one could assume that the chosen reference did have fixed absolute abundance across all samples, then the log-ratio transformation can be benchmarked as a “log-ratio normalization” [14]. Under these conditions, ALDEx2 can identify DA with high precision in RNA-Seq data [8, 32], and control false-positive rates in highly sparse 16S metagenomics count data [6]. However, the “log-ratio normalization” interpretation implies a similar assumption implied by other DA tools: that the majority of transcript species remain unchanged [42]. Alternatively, one could select an arbitrary reference based on a biological hypothesis to identify “relative DA,” even if the reference does not have fixed abundance across samples. Fig. 2 shows how the chosen reference changes the interpretation of DA.

To run ALDEx2, the user must provide count data with integer values, a vector of group labels, and a reference. The reference could be “all” (for clr), “iqlr” (for iqlr), or 1 or more user-specified features (for alr or malr). Here, we use the geometric mean of 2 NF $\kappa$ B subunits as a hypothesis-based reference, chosen because LPS activates NF $\kappa$ B to control the transcription of other immune genes [43]. With this reference, up-regulation signifies that a gene’s expression increases beyond that of NF $\kappa$ B, allowing for a clear biological interpretation. Table 1 lists 47 genes up-regulated relative to NF $\kappa$ B.

```
# Let's use Nfkb sub-units as alr reference
ref <- grep("Nfkb", colnames(rnaseq))

# ALDEx2 expects:
# 'reads': integer counts with columns as samples
# 'conditions': the experimental outcome
# 'denom': the log-ratio transform reference
library(ALDEx2)
conditions <- factor(rnaseq.annot$Treatment,
  levels = c("MOCK", "LPS"))
tt <- aldex(reads = t(ceiling(rnaseq)),
  conditions = conditions,
  denom = ref)

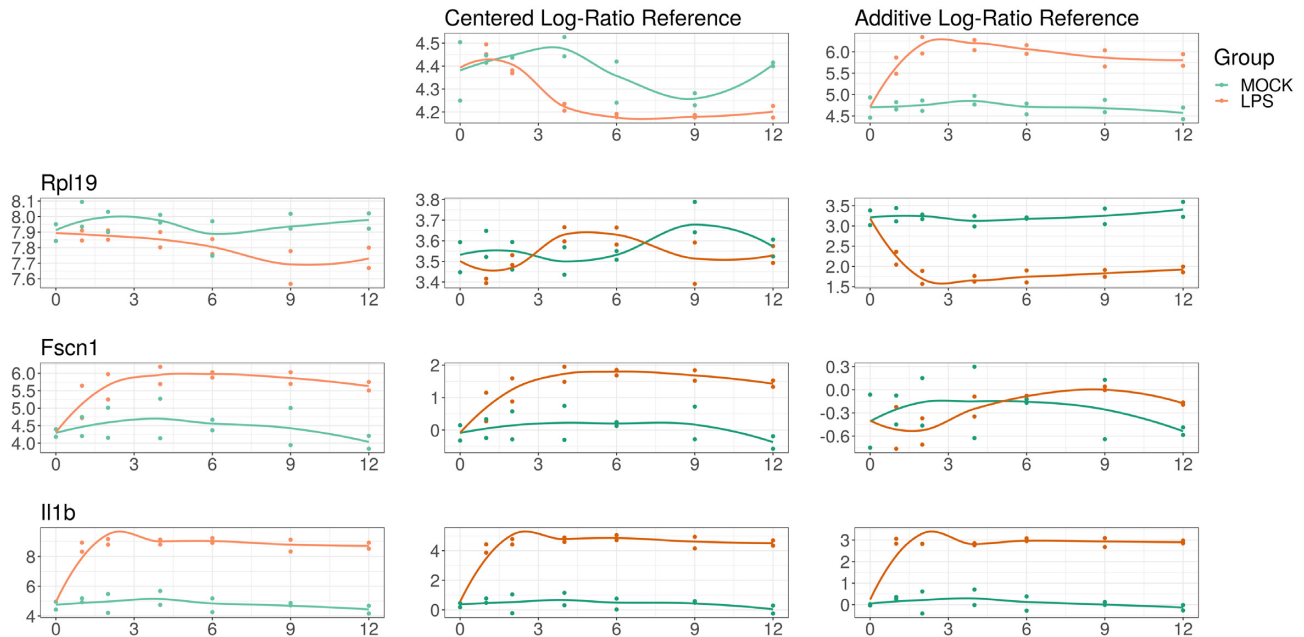
# ALDEx2 outputs a data.frame:
# 'we.eBH': the FDR-adjusted p-value
# 'effect': the effect size
# Below, we get the names of genes
# with relatively more abundance
# in the LPS group
tt.bh05 <- tt[tt$we.eBH < .05,]
up <- rownames(tt.bh05[tt.bh05$effect > 0,])
```

## Proportionality analysis with propr

Proportionality analysis is designed to identify feature coordination in compositional data [44, 45], without assuming sparsity in the association network [46, 47]. The propr package tests for the presence of feature coordination across all samples, irrespective of group label, by calculating 1 of 3 proportionality measures. Two of these have been shown to outperform all 15 competing measures of association in single-cell clustering and network inference tasks across 213 data sets [33]. The default measure,  $\rho_p$ , resembles correlation in that it falls in the range  $[-1, 1]$ . Like DA, proportionality analysis requires a reference.

```
# propr expects:
# 'counts': the data matrix with rows as samples
# 'metric': the proportionality metric to calculate
# 'ivar': the log-ratio transform reference
library(propr)
pr <- propr(counts = rnaseq.no0,
  metric = "rho",
  ivar = "clr")
```

The propr package offers 2 alternatives to zero handling. The `propr::aldex2propr` function will calculate the expected proportionality from the simulated instances generated by ALDEx2, again addressing the uncertainty of low counts [48]. The `alpha` argument will use a zero-handling procedure based on the Box-Cox transform, a pragmatic approach that allows for essential zeros but does not fall under the strict CoDA framework [49]. A Box-Cox transform with  $\alpha = 0.5$  seems to work well in simula-



**Figure 2:** This figure illustrates how the interpretation of differential abundance depends on the reference chosen. On the left margin, we show the log-abundance of 3 genes (RPL19, FSCN1, and IL1B) for the LPS-treated cells (orange) and control (blue). For compositional data, these abundances carry no meaning in isolation because the constrained total imposes a “closure bias.” On the top margin, we show the log-abundance of 2 references: the geometric mean of the samples (a la the clr) and a hypothesis-based reference NF $\kappa$ B (a la the alr). In the middle, we show the abundance of the log-ratio of the left margin feature divided by the top margin reference (equivalent to left margin minus top margin in log space). RPL19 alone appears more abundant in the control but actually has equivalent expression when compared with the geometric mean; however, it has significantly higher expression in the control relative to NF $\kappa$ B. On the other hand, FSCN1 alone appears more highly expressed in the LPS-treated cells, which remains true when compared with the geometric mean; however, it has equivalent expression relative to NF $\kappa$ B (interpreted as NF $\kappa$ B and FSCN1 expression changing similarly in response to LPS stimulation). IL1B alone appears more highly expressed in the LPS-treated cells, which remains true when compared with the geometric mean and with NF $\kappa$ B (interpreted as IL1B expression becomes even higher than NF $\kappa$ B expression in response to LPS stimulation). Choosing a reference makes normalization unnecessary but requires a shift in interpretation.

tions (see Supplementary Analysis S2). For proportionality, we do not calculate parametric  $P$ -values. Instead, we permute the FDR for a given cut-off. From this, we choose the cut-off  $\rho_p > 0.45$  to control FDR below 5%. The package vignette describes several built-in tools for visualizing proportionality. Fig. 3 shows the output of the `getNetwork` function.

```
# We can select a good cutoff for 'rho'
# by permuting the FDR at various cutoffs
# Below, we use [0, .05, ..., .95, 1]
pr <- updateCutoffs(pr, cutoff = seq(0, 1, .05))
pr@fdr

# Let's visualize using a strict cutoff
getNetwork(pr, cutoff = 0.9, col1 = up)
getResults(pr, cutoff = 0.9)
```

Proportionality depends on a log-ratio transformation and must get interpreted with respect to the chosen reference. Although proportionality appears more robust to spurious associations than correlation [30, 44], wrongly assuming that the reference has fixed absolute abundance across all samples could lead to incorrect conclusions [45]. We interpret clr-based proportionality to signify a coordination that follows the general trend of the data. In other words, these proportional genes move together as individuals relative to how most genes move on average.

## Transformation-independent analyses

The methods above depend on a log-ratio transformation to standardize the comparison of 1 gene’s expression (or 1 pair’s coordination) with another. However, by comparing the variance of the log-ratios (VLR) within groups to the total VLR, we do not need a reference to estimate between-group differences in coordination [31, 50]:

$$\text{VLR}^k(\mathbf{x}^g, \mathbf{x}^h) = \text{var} \left( \ln \frac{X_{g,1}}{X_{h,1}}, \dots, \ln \frac{X_{g,N^k}}{X_{h,N^k}} \right), \quad (2)$$

for group  $k$  with  $N^k$  samples, where  $\mathbf{x}^g$  and  $\mathbf{x}^h$  are component vectors. From this equation, we see that any normalization or transformation factor would cancel. The VLR is in the range  $[0, \infty)$ , where zero indicates perfect coordination. Otherwise, VLR lacks a meaningful scale [36]. As such, we cannot compare the VLR of 1 pair to the VLR of another pair (hence why we used proportionality instead) [30, 44]. However, in differential proportionality, we compare the VLR for the same pair across groups [31].

Differential proportionality analysis is designed to identify changes in proportionality between groups [31], interpretable as a change in gene stoichiometry. The `propd` function tests for events where the proportionality factor (i.e., the magnitude of  $x/y$ ) differs between the experimental groups. This is measured



**Figure 3:** A network where edges indicate a high level of coordination between gene expression relative to the per-sample geometric mean. Node color indicates differential expression relative to NF $\kappa$ B. The connections between red nodes indicate genes whose expression increases more than NF $\kappa$ B in a coordinated manner. The connections between white nodes indicate genes whose expression increases the same amount as NF $\kappa$ B in a coordinated manner. The connections between blue nodes indicate genes whose expression either (a) up-regulates less than NF $\kappa$ B, (b) does not change absolutely, or (c) down-regulates, all in a coordinated manner. The high level of connectivity between all nodes suggests a strong coordinated response to LPS. Like correlated pairs, proportional pairs can have any slope in non-log space. Note that this network only shows highly coordinated events (where  $\rho_p > 0.9$ ).

by  $\theta_d$ , which ranges from 0 to 1, where zero indicates a maximal difference between the groups. As above, users can permute the FDR and build a network but can also calculate an exact  $P$ -value from  $\theta_d$  using the `updateF` function [31], with the optional application of `limma::voom` precision weights [51] and  $F$ -statistic moderation [52]. Precision weights eliminate the mean-variance relationship that affects the results for low counts, while the moderated statistic helps avoid false-positive results in the case of few replicates. When testing the significance of multiple log-ratio pairs, it is absolutely necessary to correct the  $P$ -value for multiple testing. In addition, this function implements a zero-handling procedure based on the Box-Cox transform, where  $\alpha = 0.5$  seems to work well in simulations (see Supplementary Analysis S2). Fig. 4 shows significant differentially proportional pairs containing NF $\kappa$ B in the log-ratio. Most of these companion genes were also called (relatively) differentially abundant by ALDEx2.

```
# propd expects:
# 'counts': the data matrix with rows as samples
# 'group': the class labels
library(propd)
pd <- propd(counts = rnaseq.no0,
            group = rnaseq.annot$Treatment)

# Calculate an exact p-value
pd <- updateF(pd)
getResults(pd)
```

## Advanced applications

### Complex study design

Above, we used our pipeline to analyze the data as if samples belonged to 1 of 2 groups. This pipeline can also accommodate complex study designs with multiple covariates. For ALDEx2, we can supply a `model.matrix` R object to find the expectation of a linear model (instead of a  $t$ -test). On the other hand, proportionality is calculated for all samples regardless of class label, and so does not require a new procedure. Differential proportionality measures the difference in the log-ratio abundance between 2 groups. By design, it is an efficient implementation of the 2-group ANOVA expressed by the formula  $[\log(x_g) - \log(x_h)] \sim \text{group}$ , for all combinations of features  $g$  and  $h$ . Thus, we can extend differential proportionality by modeling each pairwise log-ratio outcome as a function of any `model.matrix`. This may become computationally burdensome for high-dimensional data. When testing the significance of multiple log-ratio pairs, it is absolutely necessary to correct the  $P$ -value for multiple testing, e.g., by using the `p.adjust` function in R.

### Vertical data integration

We envision 2 general strategies for the vertical integration of compositional data. First, the “row join” strategy treats other -omics data as additional samples and models the -omics source as a covariate. This requires that all -omics sources map to the same features. For the RNA-Seq and MS data used here, both quantify the relative abundance of gene products. This allows us to use ALDEx2 to find features where mRNA abundance changes more than protein abundance, relative to a common reference (and vice versa). Likewise, we can use proportionality analysis to find feature pairs where genes and proteins both have coordinated expression in response to LPS. Finally, we can use differential proportionality analysis to find feature pairs with stoichiometric differences between a gene pair and its respective protein pair. Fig. 5 shows some examples of differentially proportional pairs.

```
# Get LPS-treated cells only
rna <- rnaseq.no0[rnaseq.annot$Treatment == 'LPS',]
pro <- masshl.no0[masshl.annot$Treatment == 'LPS',]

# Join as single matrix
merge <- rbind(rna, pro)
group <- c(rep('RNA', 14), rep('Protein', 14))

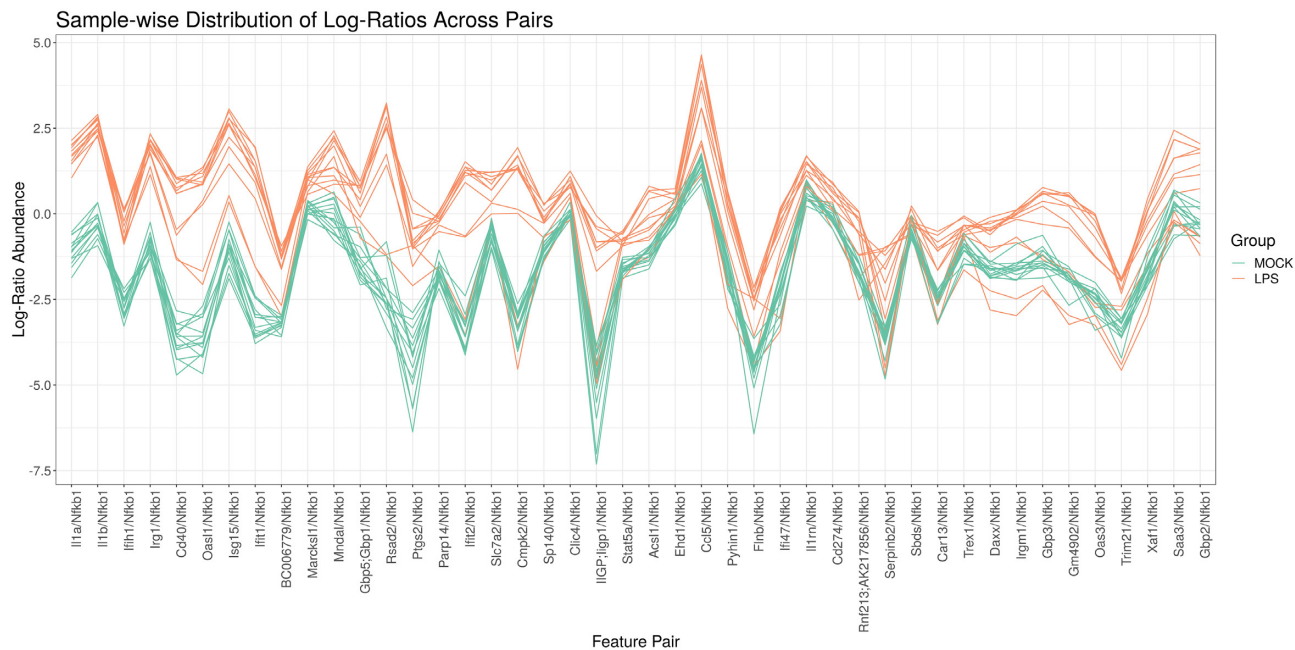
# Run propd analysis
pd.ms <- propd(merge, group)
```

Second, the “column join” strategy treats other -omics data as additional features. This strategy is more complicated because it requires that each -omics source have its own reference. In practice, we should perform differential abundance analysis on each -omics source independently. For proportionality and differential proportionality analysis, we would need to log-ratio transform each -omics source independently, then column join them with `cbind`. Here, any proportionality occurring between features from different sources would be with respect to 2 references and must get interpreted accordingly.

### Horizontal data integration

The term “mega-analysis” describes a single analysis of samples collected across multiple studies [53]. Batch effects pose a major





**Figure 4:** Parallel coordinate plot of the log-ratio abundance (y-axis) of significant differentially proportional pairs that contain NF&B in the log-ratio (x-axis). Each line represents a single sample, colored by group. Gene pairs toward the left of the x-axis have greater differences in the log-ratio means between groups (i.e., smaller  $\theta_d$  values). This plot only shows pairs for which the LPS-stimulated samples have different log-ratio means from the control (with the order of the numerator and denominator chosen such that the LPS average is always greater than the control average). It is not surprising that many of these significant pairs contain the same genes found by differential abundance analysis. Indeed, one can think of differential proportionality analysis as the differential abundance analysis of all pairwise log-ratios. Although pairs toward the right of the x-axis still have large differences in log-ratio abundance on average, some time points deviate from the trend. Indeed, this figure incidentally reveals a time-dependent process that we could test for specifically with models presented in subsection “Complex study design”.

barrier to mega-analyses. Here, we consider 2 types of batch effects. The first affects all genes within a sample proportionally (e.g., due to differences in sequencing depth). A log-ratio transformation will automatically remove this batch effect. The second affects only some genes within a sample (e.g., due to differences in RNA depletion protocols). This requires explicit modification of the corrupted features. If needed, one could apply standard batch correction tools, normally applied to normalized data, to the transformed data instead (cf. the moderated log-link *sva* in [54]).

## Clustering and classification

Most distance measures lack sub-compositional dominance, meaning that it is possible to reduce the distance between samples by adding dimensions [16]. When clustering compositions, methods that rely on distance, like hierarchical clustering, also lack sub-compositional dominance [55]. Instead, one should use the Euclidean distance of clr-transformed compositions (called the Aitchison distance) [55]. Other statistical methods used for clustering, such as principal component analysis and t-distributed stochastic neighbor embedding (t-SNE), also compute distance and should also get clr-transformed prior to analysis. When clustering components, one could use the proportionality metric  $\phi_s$  as a dissimilarity measure [30]. The  $\phi_s$  proportionality metric, like the  $\rho_p$  proportionality metric, is defined for clr-transformed data. If the geometric mean center changes drastically across samples, some proportional pairs may not be proportional in an absolute sense. We refer the reader to the subsection “Proportionality analysis with *propr*” for further explanation.

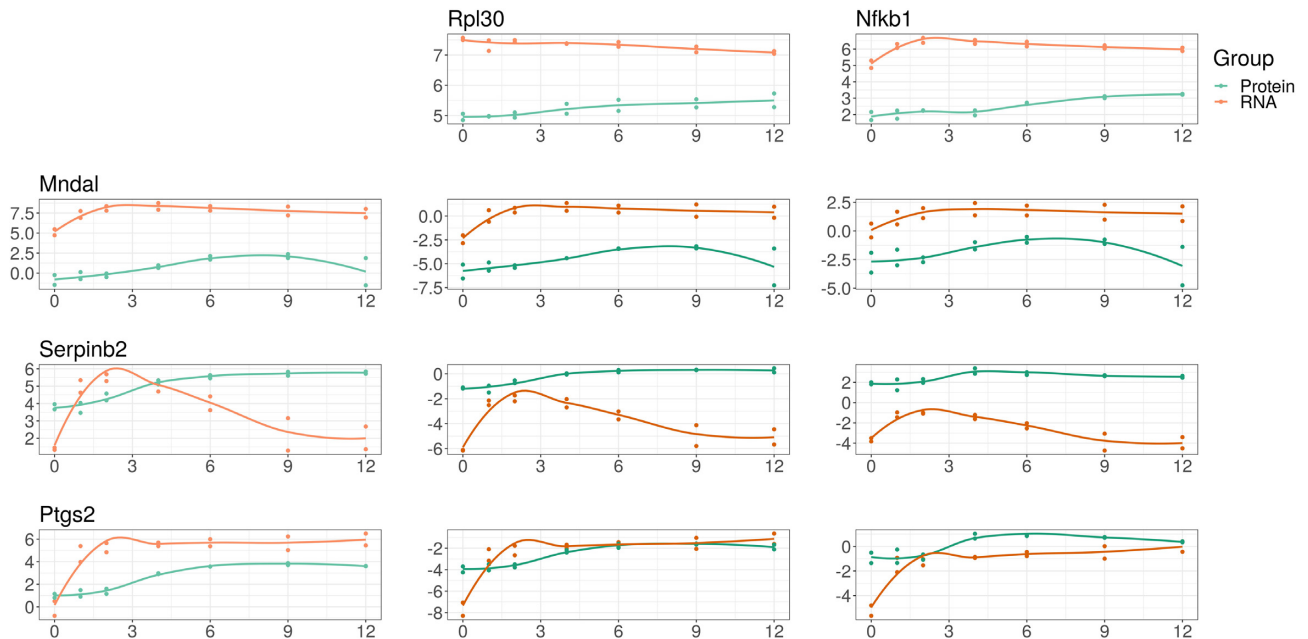
How best to classify compositional data remains an open question, but *ilr*-transforming the data prior to model training would grant the data favorable properties, as done for linear discriminant analysis [56]. Alternatively, one could train models on the log-ratios themselves, although this may not scale to high-dimensional data. Recently, balances have been used for feature selection and classification [57, 58], where they achieve both accuracy and interpretability [59].

## Selected topics

### Closure bias and the implicit reference

NGS count data measure relative abundances because of the arbitrary limit imposed by the cell, the environment, and the sequencer. This is sometimes called the “constant sum constraint” because the sum of the relative abundances must equal a constant. Anything that introduces a constant sum constraint is a kind of “closure”; all closures irreversibly make a data set relative (i.e., “closed”). One could think of a cell (in the case of RNA-Seq) or the environment (in the case of metagenomics) as natural closures, and sequencers as technical closures.

Total library size normalizations, like TPM, are not normalizations at all: they are actually yet another closure, imposing the constant sum constraint of transcripts per million. TPMs do not convert closed sequencing data into an “open” unit such as concentration. Analyzing TPMs as if they were concentrations is theoretically flawed and can substantially affect the modeling of cellular processes. Our own analysis indicates that in Jovanovic et al. [27], mRNA translation rates could have been systematically overestimated due to compositional bias. In Supplementary Analysis S1, we show that at the latest time point, the error compared to normalized data is  $\sim 13\%$  in the control condition,



**Figure 5:** mRNA abundance compared with newly synthesized protein abundance following LPS stimulation, illustrating the vertical integration of multi-omics data under a compositional framework. On the left margin, we show the log-abundance of 3 genes (MNDAL, SERPINB2, and PTGS2) as measured by RNA-Seq (orange) and mass spectrometry (blue). For compositional data, these abundances carry no meaning in isolation because the constrained total imposes a “closure bias.” On the top margin, we show the log-abundance of 2 references: RPL30 (chosen because its abundance is proportional to the geometric mean of the samples) and NFκB (chosen based on the hypothesis). In the middle, we show the abundance of the log-ratio of the left margin feature divided by the top margin reference (equivalent to left margin minus top margin in log space). MNDAL alone appears to exist more as mRNA than protein, which remains true when compared with both references. This suggests that MNDAL is translated with lower efficiency than RPL30 and NKκB. On the other hand, SERPINB2 alone appears to exist as mRNA and protein similarly on average; however, it actually exists more as protein than mRNA when compared with both references. This suggests that MNDAL is translated with greater efficiency than RPL30 and NKκB. PTGS2 alone appears to exist more as mRNA than protein, but this difference is less apparent when compared with both references. This suggests that PTGS2 is translated with a similar efficiency to RPL30 and NKκB. By choosing a reference shared between 2 multi-omics data sets, we can perform an analysis of vertically integrated data without any need for normalization.

reaching 35% in LPS-stimulated samples. This bias is due to the closure operation: if the analyst does not select a reference, the estimates must get interpreted with regard to the unknown and immeasurable “closure bias.” Because the magnitude of this closure bias can be large for samples that range widely in terms of nucleotide synthesis capacity, a reference should always be used when modeling the univariate features of compositional data. If a reference is not chosen, then the closure bias acts as an “implicit reference” that makes interpretation impossible.

### Count compositions and low-count imprecision

Closed count data differ from idealized compositional data because additive variation affects small counts more than large counts [30]. As such, the difference between 1 and 2 counts is not the same as the difference between 1,000 and 2,000 counts. Moreover, NGS experiments often have many more features than samples, leading to severe underestimation of the technical variance; indeed, the technical variance can be much larger than the biological variance at the low-count margin [29]. “Count zero” features are those that are observed as a non-zero value in  $\geq 1$  sample and thus are expected to be observed at or near the margin in other samples. While not intuitive, the distribution of the relative count zero values is quite large and spans many orders of magnitude [60]. In addition, the expected value of a count zero feature must be greater than zero because a value greater than zero was observed in  $\geq 1$  sample.

As mentioned above, the count zero values can be modified to give a point estimate of their expected value, but this leads to

underestimation of their true variance because we are estimating the expected value of the feature. In the approach instantiated in the `aldex.clr` function used by the `ALDEx2::aldex.ttest`, `ALDEx2::aldex.effect`, and `propr::aldex2propr` functions, a distribution of count zero values is determined by sampling from the Dirichlet distribution (i.e., a multivariate generalization of the  $\beta$  distribution). Another way to think about the Dirichlet distribution is a multivariate Poisson sampling with a constant sum constraint. The distribution of relative abundances near the low-count margin can be surprisingly wide, both as estimated by sampling from the Dirichlet distribution and as observed in real data [60]. By sampling from the Dirichlet distribution, we get a set of multivariate probability vectors, each of which is as likely to have been observed from the underlying data as the one actually observed from the sequenced sample. From this, `ALDEx2` and `propr` can account for low-count technical imprecision (which can be much larger than the biological variation) by reporting the expected values of a test statistic instead of the point estimate [29].

### Spike-in “log-ratio normalization”

Transformations are not normalizations because they do not claim to recast the data in absolute terms. However, if one were to choose a set of references with a priori known fixed abundance across all samples, one could use this “ideal reference” to normalize the data (something we call a “log-ratio normalization” [14]). The use of spike-in controls, consisting of multiple synthetic nucleotide sequences with known absolute

abundance, may offer one such option. For RNA-Seq, the External RNA Controls Consortium (ERCC) spike-in set consists of 92 polyadenylated RNA transcripts with varying length (250–2,000 nt) and guanine-cytosine content (5–51%) with a  $10^6$ -fold range in abundance [61]. The spike-in set is added to a standardized amount of purified RNA in equimolar concentrations; then both the spike-in and target transcripts are processed together to create a cDNA library. Because 23 of the ERCC transcripts are designed to have the same absolute abundance, one could use their geometric mean as a reference to recast the data in absolute terms. Similarly, one could spike-in a known quantity of bacteria cells or synthetic plasmids to standardize the abundance of PCR-amplified metagenomics samples [62, 63].

However, 2 important assumptions underlie the use of spike-ins for normalization. First, it is assumed that the spike-in and target sequences have the same capture efficiency of RNA conversion, in that they are both equally affected by the technical biases of cDNA library creation. Second, it is assumed that the spike-ins are calibrated to the number of RNA molecules per cell. In other words, it is assumed that the amount of spike-in is added per molecule of RNA and that each cell yielded the same number of RNA molecules. The latter is a particular issue for bulk RNA-Seq due to the technical difficulty of adding an appropriate amount of spike-in at a cell population level [64]. However, even when technical variation is controlled for, cells may produce less total RNA in 1 of the experimental groups [10] or over time [65]. In this case, standardizing the spike-in to the total amount of input RNA will invalidate this assumption. Without standardizing the spike-in to the total number of cells, it is impossible to reclaim absolute abundances (i.e., in units of transcripts per cell) [66]. Even if it were possible to standardize spike-ins to the total number of cells, the interpretation may be difficult if the cells within a single batch produce varying amounts of total RNA.

Beyond ERCC spike-ins, several other spike-ins have been proposed. For RNA-Seq studies, example spike-ins include sequins [67, 68], control plasmid spiked-in genomes [69], and isoform-specific spike-in RNA variants [70]. For metagenomics studies, example spike-ins include exogenous bacteria [62] and sequins [71]. It is beyond the scope of this field guide to compare and contrast all of the different spike-ins. However, we must emphasize that if the spike-ins are calibrated to the total weight of input RNA, they do not automatically normalize the data to absolute abundances. The reason for this follows logically from how spike-ins work: when spike-ins are added at a fixed proportion to an arbitrary mass of RNA, sequencing will return counts at the same fixed proportion. As such, spike-ins only tell us the amount of RNA sequenced. However, the term “absolute abundances” refers to the amount of RNA present in the biological sample (e.g., in units of transcripts per cell for RNA-Seq or bacteria per liter for metagenomics). Therefore, spike-ins will normalize to absolute abundances if and only if the amount of RNA sequenced is equal to the amount of RNA present in the biological sample. Even if the difference between the absolute RNA and the input RNA—which we call  $\delta$ —is proportional, this  $\delta$  must be the same for all samples. Otherwise, the  $\delta$  becomes yet another a closure bias that could introduce systematic errors. In this case, spike-in “normalization” causes the same problem as TPM “normalization”: the analyst has transformed their old compositions into new compositions under the mistaken belief that the new compositions are absolute concentrations. Before using spike-in normalization, the analyst should critically evaluate their protocol to assess whether they can safely assume that  $\delta$  is fixed for

all samples. On the other hand, a transformation with respect to an internal reference is not affected by global differences in  $\delta$ .

### Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-Seq) resembles bulk RNA-Seq, except that the RNA of individual cells is captured and bar-coded separately prior to building the cDNA library [72]. This RNA capture step involves a non-exhaustive sample of the total RNA, which acts as another closure operation to make the data relative. The sequencer would then reclose the already closed data. Interestingly, if the sequence libraries were then expressed in TPMs, the per-million divisor would act as yet another closure of the data. For these reasons, scRNA-Seq resembles other NGS count data in that each sample is a composition of relative parts. Like other NGS count data, it is impossible to estimate absolute RNA abundance without a per-cell spike-in reference.

scRNA-Seq analysis is described as being more difficult than bulk RNA-Seq analysis for 2 reasons. First, scRNA-Seq library sizes vary more between samples [73]. This is due to differences in the capture efficiency of RNA extraction, sequencing depth, and so-called “doublet” events where 2 cells get captured at once [73]. To address these differences in library size, the data are normalized by effective library size normalization or by reference normalization (via a set of housekeeping or spike-in transcripts). Effective library size normalization assumes that most genes are unchanged; this assumption is especially problematic for scRNA-Seq data because single-cell experiments study heterogeneous cell populations [74]. Reference normalization has limitations too. Housekeeping genes may not have consistent expression at the single-cell level due to transcriptional bursting or tissue heterogeneity [74]. Meanwhile, scRNA-Seq spike-ins imply the same assumptions as bulk RNA-Seq: that the spike-ins and target sequences have the same capture efficiency of RNA conversion and that the spike-ins are calibrated to the number of RNA molecules per cell. The second assumption is problematic for scRNA-Seq because it implies that all cells were similarly affected by the capture efficiency of RNA extraction [74]. Because spike-ins are added to the lysis buffer, spike-in normalization can only reveal how much RNA was captured from the cell, not how much RNA was present in the cell: as such, spike-ins cannot normalize away differences in cell lysis efficiency (which are common, and an important cause of “dropout”) [75]. On the other hand, a transformation with respect to an internal reference is not affected by global differences in cell lysis efficiency. This is analogous to the discussion of  $\delta$  from the preceding subsection.

Second, scRNA-Seq contains many zeros. Although some zeros are described as “biological zeros” (i.e., essential zeros) [76], most are described as “dropout zeros.” For dropout zeros, a zero is a missing value that occurs because the “mRNA molecules are not captured...at the same proportion” for all cells [72]. By this definition, dropout zeros are simply “count zeros” caused by non-exhaustive sampling. Because differences in cell lysis efficiency are an important cause of dropout, spike-ins cannot solve the dropout problem [75]. However, these dropout zeros are really no different than the undersampling zeros found in metagenomics data (which are already handled by our pipeline [29]). However, if an analyst wishes to impute zeros, there exist imputation methods designed specifically for compositional data [77, 78].

## Discussion

CoDA provides a conceptual framework for studying relative data. In this article, we present a collection of software tools designed for NGS count data that together form a pipeline that unifies the analysis of all compositional data, including RNA-Seq, metagenomics, single-cell, and spectrometric peak data. Unlike existing pipelines, ours does not seek to normalize the data to reclaim absolute abundances. Instead, it transforms the data with regard to a reference, allowing the analyst to study any relative data set without invoking the often untestable assumptions underpinning NGS data normalization.

The CoDA framework has evolved independently from much of the alternative techniques currently applied to NGS data. Interestingly, although not explicitly tailored for compositional data, the most rigorous of the NGS methods have converged on similar solutions for handling compositional bias. They rely on effective library size normalizations (and offsets) that make use of the (pseudo-counted) log-transformed data in a manner similar to log-ratio transformations. In CoDA, such transformations are explicitly derived to address the constrained nature of the data. From this perspective, explicit references and pairwise log-ratios apply to a broader range of experiments, including less well-controlled studies where effective library size normalizations may not work. The analysis of count compositions, especially the handling of low-count imprecision, has now reached a state of maturity that allows for NGS analysis without any loss of formal rigor.

An important aspect of CoDA is that it better quantifies the coordination between features than correlation, the latter of which is often spurious when the compositional constraint is ignored. Meanwhile, applying differential abundance analysis with respect to a reference remains valid even across the most widely varying conditions. For clustering and classification, the fully ratio-based Aitchison distance provides a superior inter-sample distance that is still underappreciated in current applications. Last but not least, CoDA opens up new perspectives with respect to the integration of big multi-omics data sets where explicit references may play an important role in the future.

## Availability of source code and requirements

- Project name: CoDa-Protocol
- Project home page: <http://doi.org/10.5281/zenodo.3270954>
- Operating systems: Platform independent
- Programming language: R
- Other requirements: R packages zCompositions, ALDEx2, propr, patchwork, ggplot2, knitr, and plyr
- License: GPLv3

## Availability of supporting data and materials

All data and scripts are publicly available at <http://doi.org/10.5281/zenodo.3270954> [79].

## Additional files

**Supplementary information:** Supplementary Methods and Results are available via the additional file associated with this article.

Supplementary Analysis S1: [supp-1.pdf](#)

Supplementary Analysis S2: [suppZero.pdf](#)

## Abbreviations

alr: additive log-ratio; ANOVA: analysis of variance; cDNA: complementary DNA; clr: centered log-ratio; CoDa: compositional data; CoDA: compositional data analysis; CRAN: Comprehensive R Archive Network; DA: differential abundance; ERCC: External RNA Controls Consortium; FDR: false discovery rate; ilr: isometric log-ratio; iqlr: inter-quartile log-ratio; LPS: lipopolysaccharide; malr: multi-additive log-ratio; mRNA: messenger RNA; MS: mass spectrometry; NF $\kappa$ B: nuclear factor  $\kappa$ B; NGS: next-generation sequencing; OTU: operational taxonomic unit; rclr: robust centered log-ratio; RNA-Seq: RNA sequencing; scRNA-Seq: single-cell RNA sequencing; TPM: transcripts per million; VLR: log-ratio variance.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

T.P.Q. outlined and drafted the field guide. T.P.Q., I.E., G.G., and M.F.R. drafted the Selected Topics section. I.E. prepared Supplementary Analysis S1. T.P.Q. prepared Supplementary Analysis S2. C.N., M.F.R., and T.M.C. supervised the project. All authors revised and approved the final manuscript.

## Acknowledgements

T.P.Q. thanks Larry Croft for helpful discussions.

## REFERENCES

1. Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11(1):31–46.
2. Wooley JC, Godzik A, Friedberg I. A Primer on metagenomics. *PLoS Comput Biol* 2010;6(2):e1000667.
3. Bashiardes S, Zilberman-Schapira G, Elinav E. Use of meta-transcriptomics in microbiome research. *Bioinform Biol Insights* 2016;10:19–25.
4. Park PJ. ChIP-Seq: advantages and challenges of a maturing technology. *Nat Rev Genet* 2009;10(10):669–80.
5. Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. *BioTechniques* 2014;56(2):61–passim.
6. Thorsen J, Brejnrod A, Mortensen M, et al. Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* 2016;4:62.
7. Hawinkel S, Mattiello F, Bijnens L, et al. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief Bioinform* 2019, 20, 1, 210–21.
8. Fernandes AD, Reid JN, Macklaim JM, et al. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* 2014;2:15.
9. van den Boogaart KG, Tolosana-Delgado R. Descriptive analysis of compositional data. In: *Analyzing Compositional Data with R. Use R!* Berlin, Heidelberg: Springer; 2013:73–93.
10. Lovén J, Orlando DA, Sigova AA, et al. Revisiting global gene expression analysis. *Cell* 2012;151(3):476–82.
11. Padovan-Merhar O, Nair GP, Biais AG, et al. Single mammalian cells compensate for differences in cellular volume

- and DNA copy number through independent global transcriptional mechanisms. *Molec Cell* 2015;**58**(2):339–52.
12. Aitchison J. A concise guide to compositional data analysis. In: 2nd Compositional Data Analysis Workshop, Girona, Spain. 2003. <http://www.leg.ufpr.br/lib/exe/fetch.php/pessoais:abtmartins:a.concise.guide.to.compositional.data.analysis.pdf>. Accessed on 16/10/2018.
  13. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, et al. Microbiome datasets are compositional: and this is not optional. *Front Microbiol* 2017;**8**:2224.
  14. Quinn TP, Erb I, Richardson MF, et al. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* 2018;**34**(16):2870–8.
  15. van den Boogaart KG, Tolosana-Delgado R. “compositions”: a unified R package to analyze compositional data. *Comput Geosci* 2008;**34**(4):320–38.
  16. Aitchison J, Barceló-Vidal C, Martín-Fernández JA, et al. Logratio analysis and compositional distance. *Math Geol* 2000;**32**(3):271–5.
  17. Pearson K. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. *Philos Trans R Soc Lond A* 1896;**187**:253–318.
  18. Filzmoser P, Walczak B. What can go wrong at the data normalization step for identification of biomarkers? *J Chromatogr A* 2014;**1362**:194–205.
  19. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 2010;**11**:R25.
  20. Anders S, Huber W. Differential expression analysis for sequence count data. *Genome Biol* 2010;**11**:R106.
  21. Mateu-Figueras G, Pawlowsky-Glahn V, Egozcue JJ. The principle of working on coordinates. In: Pawlowsky-Glahn V, Buccianti A, eds. *Compositional Data Analysis*. Wiley; 2011:29–42.
  22. Greenacre M. Variable selection in compositional data analysis using pairwise logratios. *Math Geosci* 2018;**51**(5):649–82.
  23. Engström PG, Steijger T, Sipos B, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat Methods* 2013;**10**(12):1185–91.
  24. Fonseca NA, Rung J, Brazma A, et al. Tools for mapping high-throughput sequencing data. *Bioinformatics* 2012;**28**(24):3169–77.
  25. Patro R, Duggal G, Love MI, et al. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods* 2017;**14**(4):417.
  26. Trapnell C, Williams BA, Pertea G, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 2010;**28**(5):511–5.
  27. Jovanovic M, Rooney MS, Mertins P, et al. Dynamic profiling of the protein life cycle in response to pathogens. *Science* 2015;**347**(6226):1259038.
  28. Palarea Albaladejo J, Fernández M, Antoni J. zCompositions - R package for multivariate imputation of left-censored data under a compositional approach. *Chemometrics and Intelligent Laboratory Systems* 2015;**143**(1):85–96.
  29. Fernandes AD, Macklaim JM, Linn TG, et al. ANOVA-Like Differential Expression (ALDEx) analysis for mixed population RNA-Seq. *PLoS One* 2013;**8**(7):e67019.
  30. Quinn TP, Richardson MF, Lovell D, et al. propr: an R-package for Identifying proportionally abundant features using compositional data analysis. *Sci Rep* 2017;**7**(1):16252.
  31. Erb I, Quinn T, Lovell D, et al. Differential proportionality - a normalization-free approach to differential gene expression. *Proceedings of CoDaWork 2017, The 7th Compositional Data Analysis Workshop*. bioRxiv 2017, doi:10.1101/134536.
  32. Quinn TP, Crowley TM, Richardson MF. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics* 2018;**19**:274.
  33. Skinnider MA, Squair JW, Foster LJ. Evaluating measures of association for single-cell transcriptomics. *Nat Methods* 2019;**16**(5):381–6.
  34. Martín-Fernández JA, Palarea-Albaladejo J, Olea RA. Dealing with zeros. In: *Compositional Data Analysis*. Wiley-Blackwell; 2011:43–58, doi:10.1002/9781119976462.ch4.
  35. Silverman JD, Roche K, Mukherjee S, et al. Naught all zeros in sequence count data are the same. bioRxiv 2018, doi:10.1101/477794.
  36. Aitchison J. *The Statistical Analysis of Compositional Data*. London, UK: Chapman & Hall; 1986.
  37. Wu JR, Macklaim JM, Genge BL, et al. Finding the centre: corrections for asymmetry in high-throughput sequencing datasets. arXiv 2017: 1704.01841.
  38. Martino C, Morton JT, Marotz CA, et al. A novel sparse compositional technique reveals microbial perturbations. *mSystems* 2019;**4**(1):e00016–19.
  39. Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, et al. Isometric logratio transformations for compositional data analysis. *Math Geol* 2003;**35**(3):279–300.
  40. Silverman JD, Washburne AD, Mukherjee S, et al. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* 2017;**6**, doi:10.7554/eLife.21887.
  41. Washburne AD, Silverman JD, Leff JW, et al. Phylogenetic factorization of compositional data yields lineage-level associations in microbiome datasets. *PeerJ* 2017;**5**:e2969.
  42. Kumar MS, Slud EV, Okrah K, et al. Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* 2018;**19**(1):799.
  43. Pålsson-McDermott EM, O’Neill LAJ. Signal transduction by the lipopolysaccharide receptor, Toll-like receptor-4. *Immunology* 2004;**113**(2):153–62.
  44. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, et al. Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* 2015;**11**(3):e1004075.
  45. Erb I, Notredame C. How should we measure proportionality on relative gene expression data? *Theor Biosci* 2016;**135**:21–36.
  46. Friedman J, Alm EJ. Inferring correlation networks from genomic survey data. *PLoS Comput Biol* 2012;**8**(9):e1002687.
  47. Kurtz ZD, Müller CL, Miraldi ER et al. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol* 2015;**11**(5):e1004226.
  48. Bian G, Gloor GB, Gong A, et al. The Gut microbiota of healthy aged Chinese is similar to that of the healthy young. *mSphere* 2017;**2**(5):e00327–17.
  49. Greenacre M. Measuring subcompositional incoherence. *Math Geosci* 2011;**43**(6):681–93.
  50. Walach J, Filzmoser P, Hron K, et al. Robust biomarker identification in a two-class problem based on pairwise log-ratios. *Chemometr Intell Lab Syst* 2017;**171**:277–85.
  51. Law CW, Chen Y, Shi W, et al. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* 2014;**15**:R29.
  52. Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microar-

- ray experiments. *Stat Appl Genet Mol Biol* 2004;3: Article3.
53. Tseng GC, Ghosh D, Feingold E. Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res* 2012;40(9):3785–99.
  54. Leek JT. svaseq: removing batch effects and other unwanted noise from sequencing data. *Nucleic Acids Res* 2014;42(21):e161.
  55. Martín-Fernández J, Barceló-Vidal C, Pawlowsky-Glahn V, et al. Measures of difference for compositional data and hierarchical clustering methods. In: *Proceedings of IAMG*. vol. 98; 1998:526–531.
  56. Tolosana Delgado R. Uses and misuses of compositional data in sedimentology. *Sediment Geol* 2012;280(S.1):60–79.
  57. Rivera-Pinto J, Egozcue JJ, Pawlowsky-Glahn V, et al. Balances: a new perspective for microbiome analysis. *mSystems* 2018;3(4):e00053–18.
  58. Quinn TP, Erb I. Using balances to engineer features for the classification of health biomarkers: a new approach to balance selection. *bioRxiv* 2019, doi:10.1101/600122.
  59. Calle ML. Statistical analysis of metagenomics data. *Genomics Inform* 2019;17(1):e6.
  60. Gloor GB, Macklaim JM, Vu M, et al. Compositional uncertainty should not be ignored in high-throughput sequencing data analysis. *Austrian J Stat* 2016;45:73–87.
  61. Jiang L, Schlesinger F, Davis CA, et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res* 2011;21(9):1543–51.
  62. Stämmler F, Gläsner J, Hiergeist A, et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* 2016;4(1):28.
  63. Tkacz A, Hortala M, Poole PS. Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* 2018;6:110.
  64. Risso D, Ngai J, Speed TP et al. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotechnol* 2014;32:896.
  65. Marguerat S, Schmidt A, Codlin S, et al. Quantitative analysis of fission yeast transcriptomes and proteomes in proliferating and quiescent cells. *Cell* 2012;151(3):671–83.
  66. Chen K, Hu Z, Xia Z, et al. The overlooked fact: fundamental need for spike-in control for virtually all genome-wide analyses. *Mol Cell Biol* 2016;36(5):662–67.
  67. Hardwick SA, Chen WY, Wong T, et al. Spliced synthetic genes as internal controls in RNA sequencing experiments. *Nat Methods* 2016;13(9):792–8.
  68. Deveson IW, Chen WY, Wong T, et al. Representing genetic variation with synthetic DNA standards. *Nat Methods* 2016;13(9):784–91.
  69. Sims DJ, Harrington RD, Polley EC, et al. Plasmid-based materials as multiplex quality controls and calibrators for clinical next-generation sequencing assays. *J Mol Diagn* 2016;18(3):336–49.
  70. Paul L, Kubala P, Horner G, et al. SIRVs: spike-in RNA variants as external isoform controls in RNA-sequencing. *bioRxiv* 2016, doi:10.1101/080747.
  71. Hardwick SA, Chen WY, Wong T, et al. Synthetic microbe communities provide internal reference standards for metagenome sequencing and analysis. *Nat Commun* 2018;9(1):3096.
  72. AlJanahi AA, Danielsen M, Dunbar CE. An introduction to the analysis of single-cell RNA-sequencing data. *Mol Ther Methods Clin Dev* 2018;10:189–96.
  73. Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. *F1000Res* 2016;5:2122.
  74. Lun ATL, Calero-Nieto FJ, Haim-Vilmovsky L, et al. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. *Genome Res* 2017, doi:10.1101/gr.222877.117.
  75. Kolodziejczyk A, Kim JK, Svensson V, et al. The technology and biology of single-cell RNA sequencing. *Mol Cell* 2015;58(4):610–20.
  76. Van den Berge K, Perraudeau F, Sonesson C, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol* 2018;19(1):24.
  77. Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math Geol* 2003;35(3):253–78.
  78. van den Boogaart KG, Tolosana-Delgado R. Zeroes, missings, and outliers. In: *Analyzing Compositional Data with R*. Use R! Berlin, Heidelberg: Springer; 2013:209–53.
  79. Quinn TP. A field guide for the compositional analysis of any-omics data: Supplemental Scripts. Zenodo 2018. <http://doi.org/10.5281/zenodo.3270954>