Nixon *et al. Genome Biology* (2025) 26:139
https://doi.org/10.1186/s13059-025-03609-3

Genome Biology

METHODOLOGY

Open Access

Incorporating scale uncertainty in microbiome and gene expression analysis as an extension of normalization

Michelle Pistner Nixon[1], Gregory B. Gloor[2] and Justin D. Silverman[1,3,4*]

*Correspondence:
jds6696@psu.edu

[1] College of Information Sciences and Technology, Pennsylvania State University, University Park, PA 16802, USA
[2] Department of Biochemistry, University of Western Ontario, London, ON N6A 3K7, Canada
[3] Department of Statistics, Pennsylvania State University, University Park, PA 16802, USA
[4] Department of Medicine, Pennsylvania State University, Hershey, PA 17033, USA

Abstract

Statistical normalizations are used in differential analyses to address sample-to-sample variation in sequencing depth. Yet normalizations make strong, implicit assumptions about the scale of biological systems, such as microbial load, leading to false positives and negatives. We introduce scale models as a generalization of normalizations, which allows researchers to model potential errors in these modeling assumptions, thereby enhancing the transparency and robustness of data analyses. In practice, scale models can drastically reduce false positives and false negatives rates. We introduce updates to the popular ALDEx2 software package, available on Bioconductor, facilitating scale model analysis.

Keywords: Model misspecification, Microbiome, Gene expression, Normalization

Background

Sequence count data (e.g., 16S rRNA-seq or RNA-seq data) are ubiquitous in modern biological research. Statistical methods used to analyze these data often fail to control rates of false positives [1–3]. This phenomenon builds on the broader reproducibility issues in the biomedical sciences [4]. Non-biological differences in sequencing depth between samples can substantially contribute to the occurrence of false positives [2, 3, 5–8]. In brief, sample-to-sample variation in sequencing depth is often driven by the measurement process, rather than by meaningful variation in the scale (size) of the underlying biological system [5, 8]. To address this problem, many tools incorporate some form of statistical normalization. These normalizations are designed to remove technical variation in sequencing depth so analyses can include between-sample comparisons [9]. However, in solving one issue, they can cause another: the choice of normalization can drive inferential results [1, 2, 7, 10]. Unfortunately, researchers can rarely validate the choice of normalization with real data analyses. Recently, we showed that common normalizations imply modeling assumptions about the unmeasured scale

of biological systems [2]. We found false positive rates as high as 80% with only slight errors in these implicit assumptions.

To study this scale issue, we formulated the problem using partially identified statistical models and found simple, intuitive conclusions [2]. When a research question requires knowledge of the system scale but the observed data lacks that information, researchers need to make some modeling assumptions about the system scale. For instance, in differential abundance analysis using 16S rRNA gene sequencing, scale assumptions are often made implicitly through the chosen normalization. However, these assumptions should be an explicit part of the model-building process to enhance the transparency and reproducibility of research. Moreover, statistical methods must incorporate potential errors from scale assumptions to make resulting analyses rigorous. To facilitate such analyses, we introduced a specialized and computationally efficient family of Bayesian partially identified models, called *Scale Simulation Random Variables* (SSRVs). Rather than using a single normalization, SSRVs use a *scale model* to represent uncertainty in the scale of the underlying system. Expert knowledge alone can specify scale models, in which case they generalize standard normalizations, or they can be models of external scale measurements (e.g., qPCR). Moreover, SSRVs are more flexible than prior methods which use sparsity assumptions (e.g., Grantham et al. [11]) because scale models can be designed based on sparsity assumptions, but such assumptions are not required. Through analysis of both real and simulated data, we demonstrated that accounting for scale uncertainty as part of modeling can dramatically reduce both type-I (false positive) and type-II (false negative) error rates [2, 7].

This article presents an intuitive introduction to scale model-based analysis and an update to the popular ALDEx2 library. First, we illustrate the problem with existing normalizations using an explicit example. Next, we review the ALDEx2 library and the origin of its implicit scale assumptions. We describe our updates to the ALDEx2 library which allows users to replace normalizations with scale models, making it the first general-purpose suite of tools for scale model analyses. We highlight a class of scale models that generalize the prior normalizations of ALDEx2 and are guaranteed to reduce false positive rates. Through four case studies, we demonstrate that scale model analysis can drastically decrease false positive rates. In multiple studies, scale model-based analyses control false positive or false discovery rates at nominal levels (e.g., 0.05%) when normalization-based methods such as DESeq2 [12], edgeR [13], baySeq [14], and limma [15] all display rates above 50%. Moreover, by designing scale models based on flow-cytometry data or biological knowledge, we show that scale model analysis can also reduce false negative rates. Due to their remarkable performance improvements and our prior theoretical work, we recommend using scale models in ALDEx2 rather than normalizations in all practical situations.

An illustration of the problem with normalizations

Figure 1 illustrates two microbial communities: one control condition and one treated with a drug of interest. Samples are obtained and sequenced from both communities to estimate the effect of the drug. After sequencing, 2000 reads mapped to a particular taxon in Community A (square taxon), whereas 5000 reads mapped to that same taxon in Community B. From this information, it appears the drug is associated with

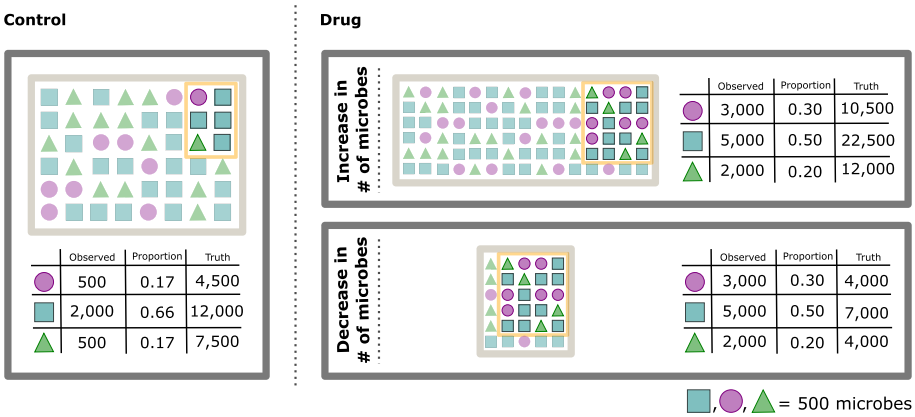

Fig. 1 Scale can confound sequence count data analyses. In this study, we measure counts of three different types of microbes (square, circle, triangle) from samples of a control community and a drug-treated community. The yellow boxes represent our samples ("Observed" columns), while the larger gray boxes represent the entire community ("Truth" columns). In the drug condition, the same observed sample could have come from a large community (top box) or a small community (bottom box). Total Sum Scaling (TSS) normalization estimates the proportional abundance of each type of microbe in the underlying system ("Proportion" columns). However, these proportions are insufficient to determine if a particular taxon increases or decreases in abundance in response to the drug

an *increase* in the taxon's abundance. However, this is not necessarily the case: sequencing depth can alter our conclusion. Here, Community B was sequenced deeper (10,000 total reads) than Community A (3000 total reads). Typically, one uses normalization to remove this confounding technical variation, giving the measurements a common scale.

With Total Sum Scaling (TSS) normalization, we transform the observed data to proportional amounts by dividing the observed counts by the sequencing depth as illustrated in the "Proportion" columns in Fig. 1. The normalized measurements have a common scale: they each sum to one. Based on these proportions, it appears the drug is associated with a *decrease* in the abundance of the square taxon.

At first glance, it seems that normalization has corrected our previous erroneous conclusion, but, unfortunately, TSS normalization is insufficient to determine if the square taxon increases or decreases in response to the drug. In drawing conclusions about the underlying communities from those TSS normalized measurements, we implicitly assumed a constant scale of the underlying communities. That is, we implicitly assumed that the drug and control communities have the exact same microbial load. But, if the scale of the underlying system was larger in Community B than Community A (top right box of Fig. 1), then this taxon would have *increased* in the drug case. Figure 1 shows that the taxon could, in truth, be either increasing or decreasing depending on whether the microbial load in the drug condition is higher or lower than in the control. This disparity arises because TSS normalization does not take into account the scale of the underlying system.

This illustration is an example of differential abundance or expression (DA/DE) analysis, which investigates if any of the D entities are present in different amounts (different abundances or different levels of expression) in two biological conditions (here, control and drug). This goal is often formalized as a problem of estimating log-fold changes

(LFC). The LFC of each entity d is defined as the difference in the average log-transformed amounts of the entity between the two conditions:

$$\theta_d = \underset{n:x_n=1}{\text{mean}} \log W_{dn} - \underset{n:x_n=0}{\text{mean}} \log W_{dn} \tag{1}$$

where x_n denotes the condition for sample n and W_{dn} denotes the amount of entity d in biological system n. For example, in Fig. 1, the LFC of the square taxon is $\log_2 22,500 - \log_2 12,000 = 0.90$ in the increased load case (top box) and $\log_2 7000 - \log_2 12,000 = -0.78$ in the decreased load case (bottom box).

As in our illustration, common tools that use normalization do not estimate LFCs with the true amounts (W_{dn}), but instead use normalized amounts. For example, using TSS normalized amounts, we estimate the LFC of the square taxon as $\log_2 0.5 - \log_2 0.66 = -0.40$ regardless of whether the drug condition has increased or decreased load. In Additional file 1: Section S1, we show that the LFC estimate from TSS normalized data is only correct if the microbial load is exactly equal in the two conditions: $W^{\perp}_{drug} = W^{\perp}_{control}$ where $W^{\perp}_{n}$ denotes the total microbial load in system n. If this condition is unmet, then the estimated LFC will be biased, and resulting hypothesis tests may display elevated rates of false positives, false negatives, or both.

This example illustrates an uncomfortable fact: controlling for technical variation in the scale of data differs from recovering biological variation in the scale of systems. To address this deficiency, some authors supplement sequence count data with external measurements of the system scale, e.g., qPCR, flow cytometry, or DNA spike-ins [6]. While these methods can help, they are not a universal solution for at least three reasons. First, either due to cost or effort, researchers infrequently collect external measurements; hence, public datasets often lack those data. Second, even if those measurements are collected, they can be noisy and often require specialized statistical methods [16]. Finally, these methods may not measure the relevant scientific scale. For example, in studying human microbiota, DNA spike-ins added after DNA extraction may recover variation in DNA concentration within DNA libraries [17]. However, variation within the DNA libraries may still differ from variation in microbial load within the human gut.

Results

An overview of the Scale Reliant Inference approach

Our methods are grounded in a branch of statistics known as Scale Reliant Inference (SRI) [2]. We briefly review SRI from the perspective of microbial differential abundance analysis to illustrate the intuition behind our approach.

Let W_{dn} denote the abundance of taxon d in community n, $W^{\parallel}_{dn}$ the proportional abundance (or *composition*) of taxon d in community n, and $W^{\perp}_{n}$ the total microbial load (or *scale*) in community n. These quantities are linked by the relationships

$$\begin{aligned} W_{dn} &= W^{\parallel}_{dn} W^{\perp}_{n} \\ W^{\perp}_{n} &= \sum_{d=1}^{D} W_{dn} \end{aligned} \tag{2}$$

reflecting the idea that absolute abundance is the product of composition and scale. The goal of differential abundance analysis is to estimate LFCs (Eq. 1). By combining Eqs. (1) and (2), the LFC for taxon d can be written as:

$$\theta_d = \theta_d^{\parallel} + \theta^{\perp}$$

where $\theta_d^{\parallel} = \text{mean}_{n:x_n=1}(\log W_{dn}^{\parallel}) - \text{mean}_{n:x_n=0}(\log W_{dn}^{\parallel})$ and $\theta^{\perp} = \text{mean}_{n:x_n=1}(\log W_n^{\perp}) - \text{mean}_{n:x_n=0}(\log W_n^{\perp})$. In other words, estimating the LFC θ_d requires us to accurately capture how the composition of taxon d changes between conditions ($\theta_d^{\parallel}$) and how the overall scale changes $\theta^{\perp}$. Because estimating θ_d requires knowledge of scale, it is referred to as a *scale reliant estimand* [2].

If we could directly measure absolute abundances (W_{dn}), LFCs would be straightforward to compute. However, sequence count data Y are noisy, imperfect proxies for W. Non-biological variation in sequencing depth cause Y to contain little to no information about $W_n^{\perp}$, making $\theta^{\perp}$ difficult to learn from the data. Moreover, there is also uncertainty about $W_{dn}^{\parallel}$, and therefore $\theta_d^{\parallel}$, due to variability in the sequence counting process. Still, this later source of compositional uncertainty is well-studied and already addressed in common tools like ALDEx2 [18]. As a result, the former, scale uncertainty, remains the most significant challenge.

As mentioned in the prior section and further explored in the next, normalizations impose strong, implicit assumptions on $W^{\perp}$ yielding point estimates $\hat{\theta}^{\perp}$. Even minor violations of those assumptions lead to error (ϵ) in the estimates, such that $\theta^{\perp} = \hat{\theta}^{\perp} + \epsilon$ (see Fig. 2). That error biases estimates of θ_d, resulting in false positives and false negatives in hypothesis testing. Using tools from SRI, we address this issue by moving away from point estimates of $\theta^{\perp}$; instead, we use probability models to capture uncertainty in the system scale. For example, we introduce probability models for $W^{\perp}$ that generalize normalizations and imply uncertainty in the value of $\theta^{\perp}$ (see Fig. 2). By accounting for this uncertainty, we reduce bias and minimize false positives. Beyond generalizing normalizations, our scale models allow researchers to incorporate more biologically plausible assumptions or integrate direct scale measurements (e.g., via flow cytometry), which reduces false negatives.

From normalizations to scale models in ALDEx2

ALDEx2 is a popular tool for DA/DE analysis [18]. While ALDEx2 can perform a wide range of log-linear modeling tasks, we focus on it as a tool for DA/DE analysis and leave a discussion of its more general linear modeling capabilities to Additional file 1. Here, we briefly describe how it works; a more formal description can be found in the "Methods" section.

First, ALDEx2 uses a Bayesian model to simulate proportional amounts ($W^{\parallel}$), taking into account the randomness of the sequencing process reflected in the observed data (Y). Second, ALDEx2 uses a centered log-ratio (CLR) transform to normalize the estimated proportions. Third, ALDEx2 calculates LFCs using the CLR normalized amounts. Finally, for each entity, a summary p-value is computed for a test of the null hypothesis that the LFC of the entity equals zero (no DA/DE). While there were

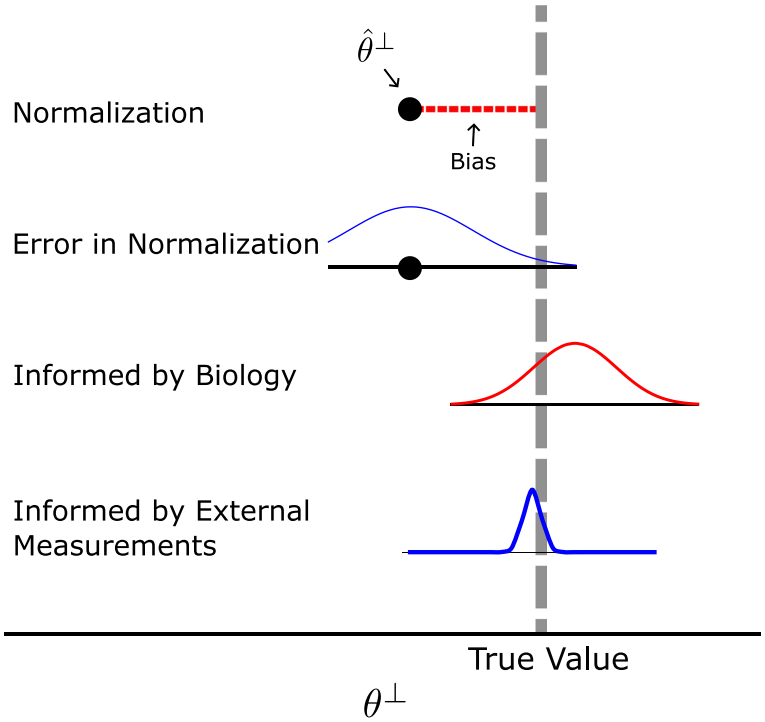

Fig. 2 Modeling scale uncertainty can reduce bias and improve inference. Estimating log-fold-changes (LFCs) requires accounting for the unmeasured scale component ($\theta^{\perp}$). Traditional normalization methods yield point estimates $\hat{\theta}^{\perp}$ with no associated uncertainty, which can introduce bias and lead to false positives or false negatives (top row). In contrast, incorporating uncertainty into normalization assumptions (second row) acknowledges potential errors in these estimates. Furthermore, scale models that leverage prior biological knowledge (third row) or direct external measurements (e.g., via flow cytometry; bottom row) provide a more accurate characterization of scale. These approaches reduce reliance on rigid normalization assumptions, lowering bias and enhancing statistical power. Adapted from McGovern and Silverman [19] with author permission

several technical details we needed to address (see Additional file 1: Sections S2–S4 for details), our principal modification of ALDEx2 is a change to the second step.

Like TSS normalization, the use of the CLR normalization makes an implicit assumption about the system scale. In the "Methods" section, we show the CLR normalization corresponds to an assumption that $W_n^{\perp} = 1/G(W_{1n}^{\parallel}, \ldots, W_{Dn}^{\parallel})$ where G denotes the geometric mean function. That is, ALDEx2 assumes that the system scale can be imputed without error from the proportional amounts of each entity, an assumption contradicted in our illustrative example.

Normalizations like the CLR and TSS have two critical limitations. First, their assumptions about the system scale are implicit and often unrecognized by researchers. Second, these assumptions are strict: resulting LFC estimates and statistical inferences are only valid if the assumptions hold exactly [2]. An intuitive solution to both problems is to make these assumptions an explicit part of the model-building process, then deal with potential errors that arise from those assumptions. Consider a model which generalizes the CLR normalization assumption:

$$p(\log W_n^{\perp}) = -\log G(W_{1n}^{\parallel}, \ldots, W_{Dn}^{\parallel}) + \epsilon_n, \qquad \epsilon \sim N(0, \gamma^2). \tag{3}$$

We express this as a model for $\log W_n^{\perp}$ rather than $W_n^{\perp}$ so that it can be expressed as a normal distribution rather than the lesser known log-normal while still restricting $W_n^{\perp}$ only to take on positive values. When $\gamma^2 \to 0$, then $\epsilon_n \to 0$ and this model is equivalent to the assumption underlying the CLR normalization (on a log-scale $\log W_n^{\perp} = -\log G(W_{1n}^{\parallel}, \ldots, W_{Dn}^{\parallel})$). However, when $\gamma^2 > 0$, the model allows for potential error in that assumption. This is an example of a *scale model*. More generally, a scale model is any probability model for the scale of the system: $p(W_1^{\perp}, \ldots, W_N^{\perp})$. By accounting for potential error in normalization assumptions, scale models can drastically reduce false positives [2]. Moreover, scale models are flexible and allow analysts to specify more biologically realistic assumptions than off-the-shelf normalizations, thereby reducing false negative rates [2]. Finally, scale models can be simple; in the "Methods" section, we discuss various features of DA/DE analysis that reduces the burden in scale model specification. We demonstrate these and other features of scale models in later sections.

Before proceeding, we clarify a frequent source of confusion. When encountering a model parameter (e.g., γ), it may be tempting to estimate it from the observed data. Yet sequence count data lacks information about the system scale and, therefore, the information needed to quantify our uncertainty in scale-related assumptions (e.g., γ). Instead, we discuss the biological interpretation of γ in the "Methods" section and recommend that researchers use this intuition to set the parameter. The key exception is when direct measurements of scale are present (e.g., flow cytometry). We will demonstrate how those data can be used to estimate scale model parameters in later sections.

Scale models can dramatically decrease false discovery rates

To illustrate the advantages of scale models, we reproduced a simulation study developed in Nixon et al. [2] and another based on SparseDOSSA2 [20] developed in McGovern and Silverman [19]. For brevity, we describe the simulation study of [2] here and in the "Methods" section, leaving discussion of the McGovern and Silverman [19] simulation to Additional file 1: Section S5.

Following Nixon et al. [2], we simulated the true abundance of 20 taxa in two conditions (pre- and post-treatment with a narrow-spectrum antibiotic). After treatment, 4 of the 20 taxa decrease in abundance. We simulated the lack of scale information in the observed data by resampling the true abundances to an arbitrary sequencing depth. We benchmarked the resampled data using standard tools for DA/DE analysis including the original ALDEx2 model (with CLR normalization), DESeq2 [12], edgeR [13], baySeq [14], and limma [15] (Fig. 3). All of these methods were unreliable and, at the largest sample sizes, demonstrated over three-times more false positives than true positives. More concerning, the false positive rate for all these methods increased to over 75% with larger sample sizes (Fig. 3). This result contradicts standard statistical wisdom: inferential performance is supposed to improve with more data.

This bizarre phenomenon, where type-I errors increase with more data, is a hallmark of an *unacknowledged bias* [2, 21]. The LFC estimates produced by these methods are biased due to errors in their implicit assumptions about scale. These methods fail to consider such errors; as the sample size increases, these methods become increasingly confident in their incorrect (biased) estimate. Scale models can mitigate this problem by

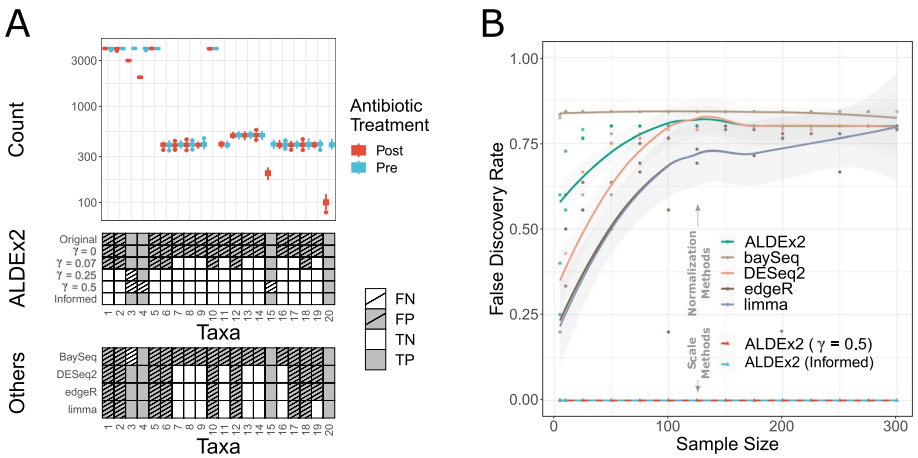

Fig. 3 Scale models can drastically decrease false positive rates. The true abundances of 20 microbial taxa were simulated before and after treatment with a mild, narrow spectrum antibiotic ("Methods" section). **A** The top panel ("Count") shows simulated true counts ($N = 50$ per condition) for each of the 20 taxa: only taxa 3, 4, 15, and 20 change between conditions. The bottom two panels ("ALDEx2" and "Others") shows the true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) for ALDEx2 and many common methods applied to the resampled data. We compared the original normalization-based ALDEx2 model ("Original") to the default scale model for several values of γ and an Informed model mimicking a slight decrease in microbial load after antibiotic administration. **B** The same simulation in Panel A was repeated in triplicate over data sizes ranging from 5 to 300 samples per condition. Only the scale-based ALDEX2 models [ALDEx2 ($\gamma = 0.5$) and ALDEx2 (Informed)] control false discovery rates asymptotically

allowing researchers to consider errors in these assumptions or even make more biologically plausible assumptions.

Our new *default scale model* for ALDEx2 demonstrates how considering errors in modeling assumptions can improve inferences. To stay consistent with prior versions which used the CLR normalization, the default scale model generalizes the CLR transform, as did Eq. (3). The default scale model has better asymptotic performance than Eq. (3) (see the "Methods" section). Like Eq. (3), the default scale model includes one user-defined parameter γ which controls the amount of uncertainty in the CLR assumption. When $\gamma = 0$, we recover the original ALDEx2 model; when $\gamma > 0$, we account for error in the CLR normalization assumption. In fact, for any value of $\gamma > 0$, the new ALDEx2 model will provide better type-I error control than the original ALDEx2 model. As a general guideline, we recommend $\gamma = 0.5$ as a reasonable default value for most cases (see the "Methods" section for explanation based on the interpretation of γ). Figure 3 demonstrates that by incorporating even tiny amounts of scale uncertainty ($\gamma > 0.07$), the false positive rate of ALDEx2 drops precipitously while still revealing true positives.

In Additional file 1: Section S6, sensitivity analyses on this simulation and a real-world data set show how the choice of γ influences false positive and false negative rates. For any biologically reasonable choice of γ (see discussion in Additional file 1: Section S6), the new ALDEx2 model controls the false discovery rate while simultaneously revealing true positives. That section also discusses how sensitivity analyses can facilitate novel and transparent forms of reporting and can sometimes even eliminate the need to choose a single value of γ.

We next designed an *informed* scale model based on our knowledge that the antibiotic is narrow spectrum; we expect only a small decrease in the microbial load (see the "Methods" section). Compared to the default scale model, this informed model reduces the bias of LFC

estimates by reflecting more biologically reasonable beliefs. While the default scale model reduced false positives compared to the original ALDEx2 model, the informed scale model also reduces false negatives (Fig. 3).

Figure 3B shows that of all the methods tested, only those that used scale models mitigated unacknowledged bias and control false discovery rates at a nominal 0.05% as sample sizes increased. All other methods displayed false discovery rates above 75% when given enough data.

Scale uncertainty enhances the reanalysis of a selective growth experiment

We reanalyzed the Selective Growth Experiment (SELEX) study originally published in McMurrough et al. [22] and later highlighted in the original publication of ALDEx2 [18]. Researchers wanted to identify which of 1600 gene variants conferred a growth phenotype upon cell lines exposed to a bacteriostatic toxin. They designed the study so that cell lines with variants capable of removing a toxin increased in abundance when exposed to the toxin while all other cell lines remained unchanged. They took samples from two experimental conditions, one with (selective) and one without the toxin (non-selective). This dataset is useful for two reasons. First, some of the variants have been verified in vitro, giving an objective measure of truth. Second, the directionality of abundance changes between conditions was asymmetric and fixed: any changes in cell line abundance were increases in the selective (rather than non-selective) growth condition. Thus, we can use biological knowledge to design an appropriate scale model.

For this data, the CLR normalization makes the implicit assumption that the absolute scale is approximately 265 times higher in the selective (bacteriostatic toxin) versus non-selective (control) condition (see the "Methods" section). Our biological knowledge supports this direction of change, but the magnitude of the change is uncertain. We use the default scale model to express uncertainty in the CLR assumption.

We used repeated data sub-sampling to investigate how the type-I error (false positive error) rates and sensitivity of different methods varied as a function of sample size. Based on validation experiments ("Methods" section), we knew that only a small fraction (27/1600) of genes confer a growth-promoting phenotype. However, existing methods returned many more genes as significant (e.g., around 1500 genes are returned as significant by DESeq2 at a sample size of 10). The only methods capable of controlling type-I errors are the scale-based ALDEx2 models. At the generally recommended value of $\gamma = 0.5$, ALDEx2 provides loose type-I error control: type-I error only increases above the stated 0.05 level for the largest sample sizes yet still remains near the stated level. We also tested ALDEx2 with an unreasonably large value of $\gamma = 5$ to illustrate performance even with over-estimated uncertainty. At $\gamma = 5$, ALDEx2 displays zero false positives for any sample size and still identifies five true positives. Still, at $\gamma = 5$, sensitivity is lower than that of $\gamma = 0.5$: in the latter, case the sensitivity of ALDEx2 is comparable to other methods with only a fraction of the false positives of other methods (Fig. 4).

Informative scale models can reduce false negatives

The prior two sections showed that false positives can be drastically reduced by integrating potential error in assumptions about scale. Here, we show how false negatives can decrease when scale models better reflect biology.

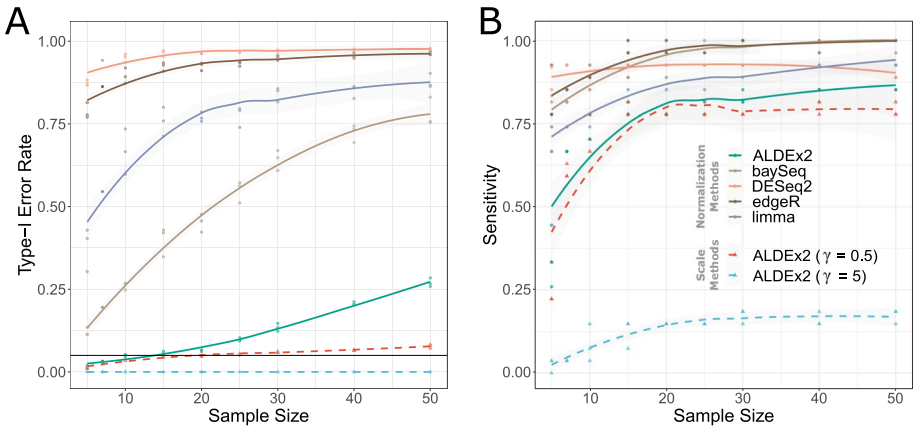

Fig. 4 Incorporating scale uncertainty improves performance over normalizations in a Selective Growth Experiment (SELEX). We reanalyzed the SELEX study at different sample sizes with data resampling (see main text and the "Methods" section). For each resampled dataset, we applied ALDEx2 with the default scale model (with $\gamma = \{0.5, 5\}$). We also applied five other normalization-based methods commonly used for differential expression analysis. **A** Type-I error rates for each tested method. We applied a mean-based smoother for better visualization. For each method, statistical significance was determined at a threshold of $\alpha = 0.05$ based on multiplicity-adjusted *p*-values. Therefore each method should control type-I error at or near a level of 0.05 (black horizontal line). Only methods that account for scale uncertainty achieve this for all sample sizes. **B** The sensitivity for each method with mean-based smoother. While many methods have high sensitivity, they do so with a high rate of false positives. Yet, even at extreme levels of scale uncertainty $\gamma = 5$, ALDEx2 identifies true positives while controlling the number of false positives

We reanalyzed a recent study by Vandeputte et al. [6], who proposed supplementing 16S rRNA microbiome data with flow-cytometry based measurements of fecal microbial concentration. This study compared fecal microbiota between 29 patients with Crohn's disease (CD) and 66 healthy controls. The original authors analyzed these data using a method they called Quantitative Microbiome Profiling (QMP): first, they rarefied the sequence count data to an even sampling depth, and then they multiplied the rarefied counts by the measured flow-cytometry measurements. In the present context, this can be thought of as using Eq. (2) without considering measurement noise in the composition or scale. In contrast, we can account for measurement noise in the sequence count data and in the flow-cytometry by using ALDEx2 with a flow-cytometry-informed scaled model:

$$\log W_n^{\perp} \sim N(\log \mu_n, \gamma^2).$$

In this model, μ_n denotes the flow-cytometry measurement for total cellular concentration in the nth fecal sample, and γ is related to the error in that measurement technique (see the "Methods" section). Treating the results of this model as our gold-standard, we benchmarked a variety of other DA/DE tools, including QMP and ALDEx2 with different scale models.

Echoing the results of Vandeputte et al. [6], normalization-based methods (including the original ALDEx2) demonstrate elevated rates of both false-positives and false negatives (Fig. 5A). Remarkably, QMP missed three bacterial genera that are differentially abundant between groups: *Parabacteroides*, *Sutterella*, and *Anaerobutyricum*. All three of these genera have been previously associated with Crohn's disease [23–25]. Moreover,

Fig. 5B shows that our conclusions about these three taxa are largely insensitive to flow-cytometry measurement error. We suspect all three of these false negatives arise from rarification, which can lead to decreased statistical power by throwing away data [26].

Both QMP and our flow-cytometry informed ALDEx2 model decrease false positive and false negative rates by using supplemental measurements of microbial concentration. As these measurements are often unavailable, we evaluated whether we could design an equally effective *Informed* scale model based only on visual inspection of figures present in an independent study of Crohn's disease [27]. Like Vandeputte et al. [6], Sarrabayrouse et al. [27] estimated total microbial load in patients with CD compared to healthy controls. Unlike Vandeputte et al. [6], which studied a Belgian cohort with flow-cytometry for microbial load measurements, Sarrabayrouse et al. [27] studied a Spanish cohort with quantitative PCR measurements and validated their findings on a Belgian cohort. Biases due to copy-number variation or DNA extraction could make these measurement techniques incomparable. Despite these differences, our *Informed* scale model, built by visual inspection of Fig. 2 of Sarrabayrouse et al. [27] (see the "Methods" section) provided nearly identical results to QMP (Fig. 5A; type-I and type-II error rates of 0% and 9%, respectively). This result highlights that even weak expert knowledge about scale, when expressed as a scale model, can enable dramatic decreases in both false positive and false negative rates compared to normalization-based methods.

We evaluated the CLR normalization and the default scale model in more detail. In short, the CLR normalization does poorly in this case study: the CLR here equates to

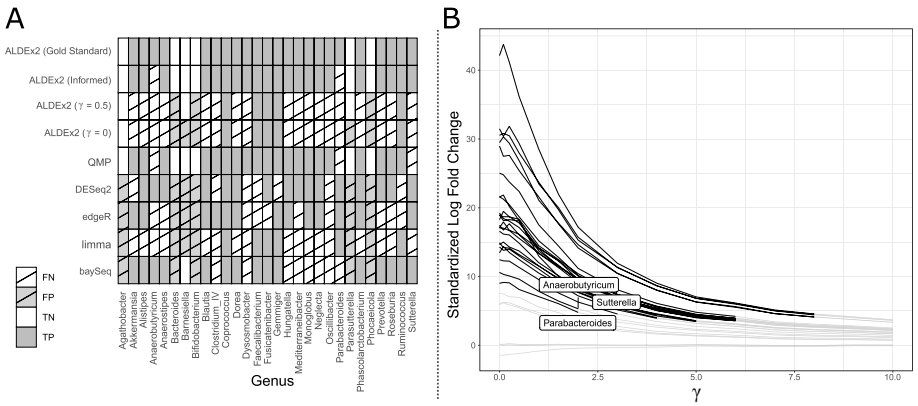

Fig. 5 Scale models built from outside measurements or biological reasoning can reduce both type-I and type-II error rates. **A** The pattern of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) for each differential abundance tool applied to the Vandeputte et al. [6] study data. Only taxa identified as differentially abundant by at least one method are shown. True/False Positives/Negatives were defined based on ALDEx2 with a *Gold Standard* scale model which integrated flow-cytometry measurements of total microbial concentration in fecal samples (see the "Methods" section). The QMP model also had access to these measurements but did not account for compositional uncertainty or measurement error in the flow-cytometry measurements. The i*nformed* scale model was based on visual inspection of an independent study which used real-time PCR to quantify microbial load in healthy versus Crohn's patients (see the "Methods" section). $\gamma = 0$ and $\gamma = 0.5$ refer to the default scale model in ALDEx2. **B** Sensitivity analysis showing how the standardized log fold change (the average LFC across Monte Carlo samples divided by the standard deviation across Monte Carlo samples) varies with different levels of measurement error in the flow-cytometry measurements. Each line corresponds to a single taxon and is gray if $p > 0.05$ and black if $p \leq 0.05$. For reference, the Gold Standard scale model in panel **A** uses $\gamma = 0.7$ based on data available in Vandeputte et al. [6] ("Methods" section). We label the three taxa identified by ALDEx2 with the Gold Standard scale model but not QMP

an assumption that microbial load in CD patients is substantially increased compared to healthy controls when the results of Vandeputte et al. [6] and Sarrabayrouse et al. [27] suggest a slight decrease. These results emphasize the importance of interrogating assumptions about scale as part of sequence count data analyses. If a researcher did want a CLR assumption to analyze these data, they would improve inference by considering scale uncertainty: the default scale model with the default value of $\gamma = 0.5$ achieves the same sensitivity as original ALDEx2 model with the CLR normalization yet has fewer false positives.

In addition to our study of gut microbiota, we observe similar findings when examining oral microbiota. In Additional file 1: Section S7, we reanalyzed a survey of oral microbiota conducted before and after teeth brushing [28] using scale models informed by literature [29, 30]. Consistent with our gut microbiota study, we found that incorporating scale uncertainty reduces false positives compared to normalization-based methods. Furthermore, just like in our gut microbiota analysis, we discovered that scale models derived from existing literature are sufficient to accurately replicate the results obtained from analyses that utilize paired flow-cytometry measurements.

Scale uncertainty can lead to more biologically plausible results in RNA-Seq studies

As a final case study, we reanalyzed a RNA-seq study which has been use to inform sample size selection for gene expression studies [31, 32]. This study contained 48 biological replicates from each of two *Saccharomyces cerevisia* strains: a wild-type (WT) and strain that knocks out the SNF2 gene, a key gene in chromatin remodeling [33]. As reported by Schurch et al. [32], existing tools often report over 70% of genes are differentially expressed in response to this knock-out. We hypothesized that this percentage included a large number of false positives arising from errors in scale assumptions. In Additional file 1: Section S8, we describe this analysis and show that many of these differentially expressed genes are no longer significant when one accounts for even small amounts of scale uncertainty. For example, with ALDEx2's default scale model, in moving from $\gamma = 0$ to $\gamma = 0.25$, the proportion of genes identified as differentially expressed drops from 68% to just 12%. Futhermore, we show how outside results can be used to develop an informed scale model [34]. While we lack a ground truth measure of what genes are differentially abundant, these results suggest that normalization-based methods may have substantial inflation of type-I errors due to a lack of uncertainty in scale assumptions.

Discussion

While implicit modeling assumptions may bias results in the analysis of sequence count data, the choice of normalization can dominate model estimates and obscure biological conclusions [2, 6, 8, 10, 35]. Here, we introduced scale models as a generalization of normalizations so researchers can account for potential errors in their implicit modeling assumptions about scale. We introduced the updated ALDEx2 software package, which provides the first general-purpose suite of tools for scale model analysis. Through case studies, we showed that accounting for potential errors in scale assumptions can drastically reduce false positive rates. Beyond generalizing normalizations, we showed that scale models can be built from prior knowledge or

external scale measurements. By better reflecting biology, such scale models can also reduce false negatives.

Previous research has suggested generalizing beyond a single normalization in analyzing sequence count data. For example, Song et al. [36] introduced a method of combining p-values obtained under different normalizations into an overall p-value. However, such work has fundamental limitations—the assumptions underlying normalizations are often implicit, obscuring which normalizations cover the biologically plausible range of assumptions. There may also be cases where none of the available off-the-shelf normalizations adequately cover actual biology. In contrast, the assumptions underlying scale models are explicit, and standard probability tools can customize scale models to any given study.

Outside of analyzing sequence count data, our work connects to the topic of rigor and reproducibility in statistics and machine learning. To address reproducibility problems in science (e.g., Ioannidis [4]), some authors have suggested increased attention on *stability*: conclusions drawn from data should withstand perturbations of the observed data and the chosen model. For example, in computer vision, researchers often include perturbed data (e.g., rotated images or Gaussian noise) to reduce over-fitting and help models generalize beyond the training set. More recently, some authors have suggested that statistical inference should integrate similar ideas [37]. Our work improves the stability of ALDEx2: the scale model accounts for perturbations to the chosen normalization. Under the new ALDEx2 model, reported p-values and confidence intervals now include a stability guarantee: conclusions drawn based on these quantities are robust to the perturbations encoded in the scale model.

This work demonstrated how to improve the rigor of existing tools by accounting for scale uncertainty. We minimized our changes to ALDEx2 to highlight that scale uncertainty, rather than other changes, drives observed performance improvements. Many avenues exist for future refinement, including developing new scale models. While we included the default scale model, which is reasonable for many cases, it is not a universal solution. Scale models built to generalize normalizations, like the default scale model, can reduce false positive rates, but this may come at the cost of statistical power. Nevertheless, our results suggest that this drop in power may be minimal as false positives tend to be more sensitive to scale uncertainty than true positives. Still, moving beyond normalizations and building scale models that reflect more biologically plausible beliefs will provide the greatest improvements, reducing false positives and improving statistical power.

Beyond ALDEx2, scale models have been integrated into other modeling frameworks. For instance, in Nixon et al. [2], we developed SSRVs by combining regression models, derived from the *fido* software package, with scale models [38]. Additionally, McGovern et al. [7] extended the widely used *Songbird* model [28] to incorporate scale uncertainty. More broadly, Dos Santos et al. [39] applied our scale models to analyze vaginal metatranscriptome data, while Bennett et al. [40] demonstrated how integrating scale uncertainty can help resolve certain analytical challenges in functional glycomics.

Looking ahead, scale models could be incorporated into a wider range of analytical tools. Any method that models system composition could integrate scale uncertainty in a manner similar to ALDEx2: by multiplying the relevant system composition parameter ($W^{\parallel}$) with

samples from a scale model before proceeding with analysis. This approach could enable researchers to extend tools such as *GPMicrobiome* [41], *PhILR* [42], and *propr* [43] to support scale-reliant inference. However, for other tools, the path forward remains unclear. Methods such as *DESeq2* [12], *edgeR* [13], and *limma* [15] lack explicit parameters corresponding to system composition, making it challenging to directly incorporate scale uncertainty.

Overall, our work provides a way to move beyond purely compositional analyses (e.g., estimating $\theta_d^{\parallel}$) to obtain robust estimates of changes in scaled amounts (θ_d). However, there are cases where compositional analyses may be sufficient. For example, if a researcher is only interested in whether a particular taxon d increases in relative abundance in response to treatment, standard tools for estimating $\theta_d^{\parallel}$ may suffice without explicitly considering scale. Nonetheless, we urge caution with this approach. Ignoring scale and focusing solely on $\theta_d^{\parallel}$ does not truly assess whether the treatment has a direct association with taxon d; instead, it may simply detect that taxon d is less associated with the treatment than some other group of taxa in the community. Consequently, researchers should not be surprised if follow-up in vitro studies reveal no actual association between the treatment and taxon d. By incorporating scale models, researchers can avoid such misleading conclusions and ensure their findings better reflect underlying biological processes.

Methods

Problem set-up and notation

We denote a sequence count dataset as a $D \times N$ matrix Y, with elements Y_{dn} denoting the number of sequenced DNA molecules mapping to the dth entity (e.g., taxa or gene) in the nth sample. Following Nixon et al. [2], we think of the observed data as an imperfect measurement of an underlying biological system W called a *scaled system*. We represent the scaled system W as a $D \times N$ matrix whose elements W_{dn} represent the true amount of entity d in the biological system from which the nth sample was taken. The notion of *true amount* depends on both the studied system and the scientific question, e.g., the true amount could represent bacterial cell count, colony-forming units (CFUs), or cellular concentration in a medium in microbiota studies

The term *scaled system* alludes to the fact that W can be uniquely described in terms of its scale (i.e., summed amounts, $W^{\perp}$) and composition (i.e., proportional amounts, $W^{\parallel}$) via:

$$W_{dn} = W_{dn}^{\parallel} W_n^{\perp}$$
$$W_n^{\perp} = \sum_{d=1}^{D} W_{dn}.$$

These relations imply $W^{\parallel}$ is a $D \times N$ matrix with columns summing to one ($\sum_{d=1}^{D} W_{dn} = 1$, e.g., the columns of W are compositional vectors) while $W^{\perp}$ is an N-vector. When we say that sequence count data (Y) lacks information about the system scale, we are referring to the fact that sample-to-sample variation in sequencing depth (i.e., $Y_n^{\perp} = \sum_{d=1}^{D} Y_{dn}$) is driven by the measurement process; such variation is typically unrelated to meaningful biological variation in the scale of the system $W_n^{\perp}$ [6, 8].

We use a $\hat{}$ to distinguish between an estimate of a quantity and its corresponding true value (e.g., $\hat{W}$ vs W). When working with samples of a quantity obtained via computer

simulation, we use super-script $^{(s)}$ to denote the s^{th} sample. When a quantity depends on only composition or scale, we use a superscript $^{\parallel}$ and $^{\perp}$, respectively. Finally, when discussing rows or columns of a matrix we use a subscript "$\cdot$", e.g., $W_{\cdot n}$ refers to the nth column of the matrix W.

The normalization-based ALDEx2 model

The ALDEx2 model consists of four main steps [18]. First, S samples of the system composition are drawn from the posterior of N independent multinomial-Dirichlet models. The posterior of the nth model is given by:

$$\hat{W}_{\cdot n}^{\parallel(s)} \sim \text{Dirichlet}(Y_{\cdot n} + 0.5 \cdot \mathbf{1}_D).$$

where $\mathbf{1}_D$ denotes a D-length vector of 1s. Each posterior sample is then normalized using one of several built-in normalizations (the default is the CLR). Each normalization can be expressed as a sample-wise transformation:

$$\log \hat{W}_{dn}^{(s)} = \log \hat{W}_{dn}^{\parallel(s)} - \phi(\log \hat{W}_{\cdot n}^{\parallel(s)})$$

where ϕ is defined by the chosen normalization. These normalized samples are then used to estimate log-fold-changes (LFCs) for each entity d

$$\hat{\theta}_d^{(s)} = \underset{n:x_n=1}{\text{mean}} \log \hat{W}_{dn}^{(s)} - \underset{n:x_n=0}{\text{mean}} \log \hat{W}_{dn}^{(s)}$$

where $x_n \in \{0, 1\}$ is a binary variable denoting the two conditions (e.g., disease versus health). A parametric or non-parametric test then examines the null hypothesis that $\hat{\theta}_d^{(s)} = 0$. Finally, ALDEx2 summarizes over the S samples, reporting the mean p-value and LFC estimate for each entity. See Additional file 1: Section S3 for a more formal definition of the ALDEx2 model and details of its linear modeling capabilities.

Scale assumption implied by CLR normalization

ALDEx2's normalization step introduces an assumption about the system scale. Note that the decomposition $W_{dn} = W_{dn}^{\parallel} W_n^{\perp}$ presented in the problem set-up can be equivalently stated as $\log W_{dn} = \log W_{dn}^{\parallel} + \log W_n^{\perp}$. Comparing this to the normalization equation:

$$\log W_{dn} = \log W_{dn}^{\parallel} - \phi(\log W_{\cdot n}^{\parallel})$$

reveals the assumption that

$$\log W_n^{\perp} = -\phi(\log W_{\cdot n}^{\parallel}).$$

This assumption says that the system scale can be imputed, without error, as some known function of the system composition. The centered log-ratio (CLR) normalization is defined by $\phi(\log W_{\cdot n}^{\parallel}) = \text{mean}(\log W_{\cdot n}^{\parallel})$, which implicitly assumes that the scale of the system is related the geometric mean of the composition

$$W_n^{\perp} = 1/G(W_{\cdot n}^{\parallel}). \tag{4}$$

A new, scale-based ALDEx2 model

In prior versions of ALDEx2, each estimate of the system's proportions $\hat{W}^{\parallel}$ was normalized by a function ϕ (e.g., the CLR transform) to create an estimate of the absolute amounts: $\hat{W}_{1n}, \ldots, \hat{W}_{Dn} = \phi(\hat{W}^{\parallel}_{1n}, \ldots, \hat{W}^{\parallel}_{Dn})$. In the updated version of ALDEx2, we replace this step with a sample from a scale model which is a probability model for the system scale: $p(W^{\perp}_1, \ldots, W^{\perp}_N)$. Samples from the scale model are now used in lieu of normalization: these scale samples ($W^{\perp(s)}$) are multiplied by the composition ($W^{\parallel(s)}$) to estimate the system:

$$\begin{aligned}\hat{W}^{\perp(s)}_1, \ldots, \hat{W}^{\perp(s)}_N &\sim p \\ \hat{W}^{(s)}_{d1}, \ldots, \hat{W}^{(s)}_{dN} &= \hat{W}^{\parallel(s)}_{d1}\hat{W}^{\perp(s)}_1, \ldots, \hat{W}^{\parallel(s)}_{dN}\hat{W}^{\perp(s)}_N.\end{aligned}$$

That is, absolute amounts ($\hat{W}_{dn}$) are equal to proportional amounts ($\hat{W}^{\parallel}_{dn}$) times scale ($\hat{W}^{\perp}_n$). The resulting estimates $\hat{W}_{dn}$ can be used in the subsequent steps of the ALDEx2 model just as before. This modification turns ALDEx2 into a specialized type of model called a Scale Simulation Random Variable (SSRV) [2] and leads to no perceptible increase in runtime or memory demands compared to the original ALDEx2 software.

Simple scale models for DA/DE analysis

While scale models are flexible and can be arbitrarily complex, they do not need to be. Especially for DA/DE analysis, the structure of the LFC estimand (θ_d) can simplify model specification. Nixon et al. [2] proved that for LFC estimation, scale models only need to be specified up to a global constant c defined by $W^{\perp}_n = c\tilde{W}^{\perp}_n$. This result implies that researchers designing scale models for DA/DE analysis only need to be concerned with how the scale might change between systems. Moreover, for LFC estimation, those authors showed that it often suffices to specify a scale model for a single real-valued quantity $\theta^{\perp}$ called the *Log-Fold-Change in Scales*: defined by

$$\theta^{\perp} = \underset{n:x_n=1}{\text{mean}} \log W^{\perp}_n - \underset{n:x_n=0}{\text{mean}} \log W^{\perp}_n.$$

That is, in many cases, it is sufficient only to model how the average scale might change between conditions. The scale models used in this manuscript use one or both simplifications.

A default scale model for ALDEx2

To ease adoption of the updated ALDEx2 software suite, we developed a scale model that considered potential error in the default CLR normalization. Yet, we avoid using Eq. (3) for this task as, due to the law of large numbers, that model asymptotically assumes that the LFC of scales is equal to the CLR estimate with zero uncertainty (zero variance). Instead, we defined an alternative model with better asymptotic performance that more naturally mimics the linear modeling capabilities of ALDEx2. We present the model in its full form in Additional file 1: Section S3. For DA/DE analyses, the scale model simplifies to

$$\begin{aligned}\log \hat{W}^{\perp(s)}_n &= -\text{mean}\left(\log \hat{W}^{\parallel(s)}_{\cdot n}\right) + \Lambda^{\perp}x_n \\ \Lambda^{\perp} &\sim N(0, \gamma^2).\end{aligned}$$

As in Eq. (3), this scale model reduces to the CLR normalization when $\gamma = 0$ and models error in that assumption for any value of $\gamma > 0$.

The parameter $\Lambda^{\perp}$ represents systematic error in the CLR estimated difference in scales between conditions. More concretely, using Eq. (4), we calculate that the CLR normalization corresponds to an assumption that

$$\hat{\theta}^{\perp} = \underset{n:x_n=1}{\text{mean}}(-\log G(\hat{W}^{\parallel}_{\cdot n})) - \underset{n:x_n=0}{\text{mean}}(-\log G(\hat{W}^{\parallel}_{\cdot n})).$$

The parameter Λ represents potential error in this relationship; the true log-fold-change in scales ($\theta^{\perp}$) is given by $\theta^{\perp} = \hat{\theta}^{\perp} + \Lambda$. Considering the distribution of Λ, this implies a model $\theta^{\perp} \sim N(\hat{\theta}^{\perp}, \gamma^2)$. When choosing γ, one should consider 95% probability intervals of this normal model. According to this model, there is a 95% probability that the true difference in scales between conditions is within a factor of $(2^{-2\gamma}, 2^{2\gamma})$ of the CLR estimate $\hat{\theta}^{\perp}$. Based on this interpretation, we recommend $\gamma = 0.5$ as a reasonable choice: at this value, we consider up to 2-fold errors in the CLR estimate of $\theta^{\perp}$ with 95% probability. Alternatively, rather than interpreting Λ in terms of errors in the CLR estimate, it can be interpreted directly in terms of the log-fold-change of scales. According to the model, there is a 95% probability that the true difference in scales between conditions is within the range $(2^{-2\gamma+\hat{\theta}^{\perp}}, 2^{2\gamma+\hat{\theta}^{\perp}})$. See Additional file 1: Section S3 and S4 for further details on interpretation of the default scale model and advice for choosing γ.

Updates to the summarization of *p*-values in ALDEx2

Upon introducing scale models in ALDEx2, we identified a slight error in how ALDEx2 had previously summarized p-values over the S posterior samples. While unlikely to cause issues in prior versions, this error grew problematic when we introduced scale uncertainty. In Additional file 1: Section S2, we illustrate this problem and describe a solution of summarizing p-values from two one-sided hypothesis tests rather than from a single two-sided test based on the results of [44].

Data analysis details

For all data analyses, p-values reported for ALDEx2 refer to Benjamini-Hochberg corrected p-values of the null hypothesis $\theta_d = 0$ using the Welch's t-test and based on 1000 Monte Carlo replicates. The exception is our analysis of the RNA-seq study of Gierliński et al. [31] presented in Additional file 1: Section S8, where we only used 500 Monte Carlo replicates to accommodate the larger data size. For all analyses, DESeq2, edgeR, limma-voom, and baySeq were fit using recommended defaults. For edgeR, we report the results of the exact test. For clarity, logarithms in the following sections were computed in base 2 to be consistent with ALDEx2.

Mock experiment simulation details

The true abundance of 20 microbial taxa were simulated from $2N$ communities equally split between pre-antibiotic ($x_n = 0$) and post-antibiotic ($x_n = 1$) conditions. Simulations used the following Poisson model:

$$W_{dn} \sim \begin{cases} \text{Poisson}(\lambda_{\text{d,0}}) \text{ if } \text{x}_\text{n} = 0. \\ \text{Poisson}(\lambda_{\text{d,1}}) \text{ if } \text{x}_\text{n} = 1. \end{cases}$$

To simulate a narrow-spectrum antibiotic, 16 of the 20 taxa were specified with $\lambda_{d,0} = \lambda_{d,1}$ (not differentially abundant). Of these 16 non-differentially abundant taxa, 4 had $\lambda_d = 4000$, 3 had $\lambda_d = 500$, and 9 had $\lambda_d = 400$. The four taxa that were differential abundant were $d = \{3, 4, 15, 20\}$. For those taxa, λ was set as: $\lambda_{(3,4,15, \text{ and } 20);0} = \{4000, 4000, 400, 400\}$ and $\lambda_{(3,4,15, \text{ and } 20);1} = \{3000, 2000, 200, 100\}$. Based on these values, the CLR estimate of the LFC of scales was $\hat{\theta}^{\perp} = 0.04$ whereas the true value was $\theta^{\perp} = -0.18$. That is, the CLR assumption implies a slight increase in scales after antibiotic administration; the truth is a moderate decrease in scales after antibiotic administration. Sequencing-based loss of scale information was simulated via multinomial resampling:

$$Y_{\cdot n} \sim \text{Multinomial}\left(\text{M}, \frac{\text{W}_{\cdot\text{n}}}{\sum_{\text{d=1}}^{\text{D}} \text{W}_{\text{dn}}}\right)$$

with a sequencing depth $M = 5000$.

The Informed model was constructed under the assumption that antibiotic administration resulted in a 10% decrease in the total microbial load between conditions:

$$\log W_n^{\perp} \sim N(\mu_{x_n}, 0.25^2)$$

where $\mu_{x_n=0} = \log 1$ (pre-antibiotic) and $\mu_{x_n=1} = \log 0.9$ (post-antibiotic).

SELEX reanalysis

The SELEX experiment is detailed in McMurrough et al. [22]. Preprocessed data from this experiment was obtained from the ALDEx2 Bioconductor package. This data contains 1600 possible sequence variants measured in 14 samples equally split between the selected and non-selected conditions. True positives were identified based on subsequent validation experiments detailed in McMurrough et al. [22].

For this study, the CLR estimate for the log-fold-change of scales $\hat{\theta}^{\perp} = -8.05$ corresponds to an assumption that the average scale in the selected condition is 265 times higher than in the non-selected condition. The default scale model at $\gamma = 0.5$ expresses 95% certainty that the average scale in the selected condition is between 130 and 530 times larger than in the non-selected condition. At $\gamma = 5$, the default scale model expresses 95% certainty that the average scale in the selected condition is between 0.25 and 272,000 times larger than in the non-selected condition. In reality, our true beliefs lie in between these two values of γ. We selected $\gamma = 0.5$ as a reasonable default value and $\gamma = 5$ to highlight performance under unreasonably large amounts of uncertainty.

Data resampling was performed in triplicate for each sample size to address randomness in the resampling process.

Vandeputte reanalysis

Data was obtained from the European Nucleotide Archive with accession code PRJEB21504. A sequence variant table using the DADA2 software processed the raw data, following the software vignette and recommended defaults [45]. Fastq files were filtered with the `filterAndTrim` function setting `maxEE` to 4 as described in online vignettes. We used all default parameters to learn the error rates, run the core DADA2 algorithm, and merge sequence pairs. The consensus method removed chimeras. We used the RDP classifier to assign taxonomy [46], then retained genera present in at least 20% of samples for analysis. QMP was applied by using R code available at https://github.com/raeslab/QMP. The resulting matrix was treated as an estimate of W. A small pseudo-count was added (0.5) prior to log-transformation to mitigate numerical issues associated with taking the logarithm of zero. Two-sided Welch's t-tests were applied to assess significance between CD patients and health controls for each genera. Resulting p-values were adjusted using the Benjamini-Hochberg procedure.

If we assume the flow cytometry measurements are error free, then the microbial load in CD patients decreases by 65% compared to healthy controls ($\theta^{\perp} = -1.49$). However, extended results presented by Vandeputte et al. [6] also suggest that these measurements can have substantial variability leading us to the following (Gold Standard) scale model:

$$\log W_n^{\perp} \sim N(\log \mu_n, \gamma^2)$$

where μ_n denotes the measured flow-cytometry cell count for sample n, and γ^2 is the variance of the measurement noise. We chose γ^2 based on based on Extended Data File 5 of Vandeputte et al. [6] which summaries the mean and standard deviation cell counts from technical replicates of 40 different biological samples. A Taylor expansion was used to estimate standard deviations on log-scale based on the reported means and standard deviations of cell counts. Conservatively, we chose a value of $\gamma^2 = 0.70$ which corresponds to the maximum estimated log-scale variance from the 40 samples studied. This model expresses 95% certainty that the average scale in Crohn's disease patients is between 13 and 94% of the average scale in the healthy controls. Figure 5 depicts the sensitivity of results to this choice of γ^2.

An Informed scale model was designed based on visual inspection of Fig. 2 of Sarrabayrouse et al. [27]. Based on that figure, we estimated an average of 1.5×10^{12} cells per gram of feces for healthy controls compared to 1.0×10^{12} cells per gram of feces for CD patients. Combined with estimates of uncertainty obtained from that figure, we designed a scale model which reflects an assumption of an approximately 30% decrease in microbial load in CD compared to health:

$$\log W_n^{\perp} \sim N(\mu, 0.125^2)$$

where $\mu = \log(1)$ if the sample was from the control condition and $\mu = \log(0.7)$ if the sample was from the CD condition ($\theta^{\perp} = -0.52$). This model expresses 95% certainty that the average scale in the Crohn's disease patients is between 17 and 41% lower than the average scale in the healthy patients.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03609-3.

Additional file 1: Contains derivations for the TSS normalization, information on the updates to the testing procedure, a description of ALDEx2 as a linear model, further details on selecting γ, details on sensitivity analyses, an analysis of an additional simulation, expanded analyses of a RNA-seq data set and an oral microbiome data set, and Supplemental Figures S1-S5.

Acknowledgements
We thank Rachel Silverman and Steve Nixon for their manuscript comments and Yen Duong for her professional editing services. JDS and MPN were supported in part by NIH 1R01GM148972-01. GBG declares no specific funding for this work.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
JDS conceptualized the modification to ALDEx2. MPN, GBG, and JDS contributed to the design and development of the proposed methods. MPN and GBG coded the pertinent updates to the ALDEx2 software, and MPN conducted the data analysis. MPN, GBG, and JDS wrote and edited the manuscript. All authors read and approved the final manuscript.

Funding
JDS and MPN were supported by NIH 1 R01GM 148972-01.

Data availability
All new data and code are available at https://www.github.com/michellepistner/scale-in-aldex2 [47] and https://doi.org/10.5281/zenodo.15321032 [48] under Apache License 2.0. Outside data used in this publication are available at the European Nucleotide Archive under accession codes PRJEB21504 [49], PRJEB29169 [50], and PRJEB5348 [51].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 1 May 2024 Accepted: 7 May 2025
Published online: 22 May 2025

References

1. Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinforma. 2019;20(1):210–21.
2. Nixon MP, McGovern KC, Letourneau J, David L, Lazar NA, Mukherjee S, Silverman JD. Scale reliant inference. 2024;11:2201.03616.
3. Roche KE, Mukherjee S. The accuracy of absolute differential abundance analysis from relative count data. PLoS Comput Biol. 2022;18(7):e1010284.
4. Ioannidis JP. Why most published research findings are false. PLoS Med. 2005;2(8):e124.
5. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224.
6. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative microbiome profiling links gut community variation to microbial load. Nature. 2017;551(7681):507–11.
7. McGovern KC, Nixon MP, Silverman JD. Addressing erroneous scale assumptions in microbe and gene set enrichment analysis. PLoS Comput Biol. 2023;19(11):e1011659.
8. Props R, Kerckhof FM, Rubbens P, De Vrieze J, Hernandez Sanabria E, Waegeman W, et al. Absolute quantification of microbial taxon abundances. ISME J. 2017;11(2):584–7.
9. Srinivasan A, Xue L, Zhan X. Compositional knockoff filter for high-dimensional regression analysis of microbiome data. Biometrics. 2020. https://doi.org/10.1111/biom.13336.
10. Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5:1–18.
11. Grantham NS, Guan Y, Reich BJ, Borer ET, Gross K. MIMIX: A Bayesian mixed-effects model for microbiome data from designed experiments. J Am Stat Assoc. 2020;115(530):599–609.
12. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):1–21.

13. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40.
14. Hardcastle TJ, Kelly KA. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics. 2010;11:1–14.
15. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47.
16. Galazzo G, Van Best N, Benedikter BJ, Janssen K, Bervoets L, Driessen C, et al. How to count our microbes? The effect of different quantitative microbiome profiling approaches. Front Cell Infect Microbiol. 2020;10:403.
17. Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, et al. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome. 2016;4:1–13.
18. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 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(1):1–13.
19. McGovern KC, Silverman JD. Replacing Normalizations with Interval Assumptions Improves the Rigor and Robustness of Differential Expression and Differential Abundance Analyses. bioRxiv. 2024:10 pp. 15.618450
20. Ma S, Ren B, Mallick H, Moon YS, Schwager E, Maharjan S, et al. A statistical model for describing and simulating microbial community profiles. PLoS Comput Biol. 2021;17(9):e1008913.
21. Gustafson P. Bayesian inference for partially identified models: Exploring the limits of limited data. New York: CRC Press. 2015;140.
22. McMurrough TA, Dickson RJ, Thibert SM, Gloor GB, Edgell DR. Control of catalytic efficiency by a coevolving network of catalytic and noncatalytic residues. Proc Natl Acad Sci. 2014;111(23):E2376–83.
23. Wang Y, Gao X, Ghozlane A, Hu H, Li X, Xiao Y, et al. Characteristics of faecal microbiota in paediatric Crohn's disease and their dynamic changes during infliximab therapy. J Crohn's Colitis. 2018;12(3):337–46.
24. Cui Y, Zhang L, Wang X, Yi Y, Shan Y, Liu B, et al. Roles of intestinal Parabacteroides in human health and diseases. FEMS Microbiol Lett. 2022;369(1):fnac072.
25. Suskind DL, Lee D, Kim YM, Wahbeh G, Singh N, Braly K, et al. The specific carbohydrate diet and diet modification as induction therapy for pediatric Crohn's disease: a randomized diet controlled trial. Nutrients. 2020;12(12):3749.
26. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):e1003531.
27. Sarrabayrouse G, Elias A, Yáñez F, Mayorga L, Varela E, Bartoli C, et al. Fungal and bacterial loads: noninvasive inflammatory bowel disease biomarkers for the clinical setting. Msystems. 2021;6(2):10–1128.
28. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Nat Commun. 2019;10(1):2719.
29. Slot D, Wiggelinkhuizen L, Rosema N, Van der Weijden G. The efficacy of manual toothbrushes following a brushing exercise: a systematic review. Int J Dent Hyg. 2012;10(3):187–97. https://doi.org/10.1111/j.1601-5037.2012.00557.x.
30. Funahara M, Yamaguchi R, Honda H, Matsuo M, Fujii W, Nakamichi A. Factors affecting the number of bacteria in saliva and oral care methods for the recovery of bacteria in contaminated saliva after brushing: a randomized controlled trial. BMC Oral Health. 2023;23(1):917.
31. Gierliński M, Cole C, Schofield P, Schurch NJ, Sherstnev A, Singh V, et al. Statistical models for RNA-seq data derived from a two-condition 48-replicate experiment. Bioinformatics. 2015;31(22):3625–30.
32. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016;22(6):839–51.
33. Peterson CL, Tamkun JW. The SWI-SNF complex: a chromatin remodeling machine? Trends Biochem Sci. 1995;20(4):143–6. https://doi.org/10.1016/s0968-0004(00)88990-2.
34. Yoshikawa K, Tanaka T, Ida Y, Furusawa C, Hirasawa T, Shimizu H. Comprehensive phenotypic analysis of single-gene deletion and overexpression strains of Saccharomyces cerevisiae. Yeast. 2011;28(5):349–61.
35. Clausen DS, Willis AD. Evaluating replicability in microbiome data. Biostatistics. 2022;23(4):1099–114. https://doi.org/10.1093/biostatistics/kxab048.
36. Song H, Ling W, Zhao N, Plantinga AM, Broedlow CA, Klatt NR, et al. Accommodating multiple potential normalizations in microbiome associations studies. BMC Bioinformatics. 2023;24(1):1–15.
37. Yu B. Veridical data science. In: Proceedings of the 13th International Conference on Web Search and Data Mining. New York: Association for Computing Machinery New York; 2020. pp. 4-5.
38. Silverman JD, Roche K, Holmes ZC, David LA, Mukherjee S. Bayesian multinomial logistic normal models through marginally latent matrix-T processes. J Mach Learn Res. 2022;23(1):255–96.
39. Dos Santos SJ, Copeland C, Macklaim JM, Reid G, Gloor GB. Vaginal metatranscriptome meta-analysis reveals functional BV subgroups and novel colonisation strategies. Microbiome. 2024;12(1):271.
40. Bennett AR, Lundstrøm J, Chatterjee S, Thaysen-Andersen M, Bojar D. Compositional data analysis enables statistical rigor in comparative glycomics. Nat Commun. 2025;16(1):795.
41. Äijö T, Müller CL, Bonneau R. Temporal probabilistic modeling of bacterial compositions derived from 16S rRNA sequencing. Bioinformatics. 2018;34(3):372–80.
42. Silverman JD, Washburne AD, Mukherjee S, David LA. A phylogenetic transform enhances analysis of compositional microbiota data. Elife. 2017;6:e21887.
43. Quinn TP, Richardson MF, Lovell D, Crowley TM. propr: an R-package for identifying proportionally abundant features using compositional data analysis. Sci Rep. 2017;7(1):1–9.
44. Oosterhoff J. Combination of one-sided statistical tests. Amsterdam: Mathematical Centre; 1969.
45. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.
46. Wang Q, Garrity GM, Tiedje JM, Cole JR. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Appl Environ Microbiol. 2007;73(16):5261–7.
47. Nixon MP. Code for "Incorporating scale uncertainty in microbiome and gene expression analysis as an extension of normalization". Github. https://www.github.com/michellepistner/scale-in-aldex2.

48. Nixon MP. Code for "Incorporating scale uncertainty in microbiome and gene expression analysis as an extension of normalization". Zenodo. 2025. https://doi.org/10.5281/zenodo.15321032.
49. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, et al. Quantitative Microbiome Profiling Study Cohort, Disease cohort and Healthy controls. Eur Nucleotide Arch. 2017. https://www.ebi.ac.uk/ena/browser/view/PRJEB21504.
50. Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Eur Nucleotide Arch. 2020. https://www.ebi.ac.uk/ena/browser/view/PRJEB29169.
51. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, et al. S. cerevisiae WT vs snf2 KO mutant RNA-seq data with 7 technical and 48 biological replicates (336 total) of each condition. Eur Nucleotide Arch. 2016. https://www.ebi.ac.uk/ena/browser/view/PRJEB5348.

Publisher's Note