Supplementary Materials to "Incorporating scale uncertainty in microbiome and gene expression analysis as an extension of normalization"

Michelle Pistner Nixon, Gregory B. Gloor, Justin D. Silverman

S1 Implied Scale Assumption of the TSS Normalization

In this section, we show that Total Sum Scaling normalization (TSS) leads to the implicit assumption that scale remains constant between conditions. As in *Methods*, let W_{dn} denote the amount of entity d in sample n. Recall that this amount can be decomposed as the product of the proportional abundance of entity d, $W^{\parallel}_{dn}$, and the total scale of the system n, $W^{\perp}_{n}$:

$$W_{dn} = W^{\parallel}_{dn} W^{\perp}_{n}.$$

The log-fold-change (LFC) of entity d can be decomposed into two additive components, one depending only on $W^{\parallel}$ and one on $W^{\perp}$:

$$\theta_d = \underset{n:x_n=1}{\text{mean}} \log W^{\parallel}_{dn} W^{\perp}_{n} - \underset{n:x_n=0}{\text{mean}} \log W^{\parallel}_{dn} W^{\perp}_{n}$$

$$= \underbrace{\underset{n:x_n=1}{\text{mean}} \log W^{\parallel}_{dn} - \underset{n:x_n=0}{\text{mean}} \log W^{\parallel}_{dn}}_{\theta^{\parallel}_d} + \underbrace{\underset{n:x_n=1}{\text{mean}} \log W^{\perp}_{n} - \underset{n:x_n=0}{\text{mean}} \log W^{\perp}_{n}}_{\theta^{\perp}}$$

In TSS, the observed data Y is normalized and used as estimates of $W^{\parallel}$, yet LFCs are calculated directly from these normalized estimates:

$$\hat{W}^{\parallel}_{dn} = \frac{Y_{dn}}{\sum_{d=1}^{D} Y_{dn}}$$
$$\hat{\theta}_d = \underset{n:x_n=1}{\text{mean}} \hat{W}^{\parallel}_{dn} - \underset{n:x_n=0}{\text{mean}} \hat{W}^{\parallel}_{dn}.$$

From $\hat{\theta}_d$ above, it is clear that this is simply an estimate of the compositional term $\theta^{\parallel}$. In other words, using TSS normalized data for LFC estimation is equivalent to an implicit assumption that $\theta_d = \theta^{\parallel}_d$ or equivalently that $\theta^{\perp} = 0$. The assumption that $\theta^{\perp} = 0$ occurs when the average scale in one condition is exactly equal to the average scale in the other, i.e.:

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

S2 An updated hypothesis test for ALDEx2

While adding scale to ALDEx2, we discovered an error in how ALDEx2 had been summarizing p-values over the S simulation samples. While this error could lead to counter-intuitive results, we expect that it would rarely, if ever, cause issues in a normalization-based ALDEx2, yet the error may become problematic when scale models are added.

For each entity d, ALDEx2 computes an average p-value across the S Monte Carlo replicates. Originally, a p-value, p_{ds}, was computed for each entity $d \in 1, ..., D$ and Monte Carlo sample $s \in \{1, ..., S\}$ using a two-sided alternative, i.e. in the case of a

two sample t-test:

$$p_{ds} = \Pr(|T| > |t_{ds}| \ \mid H_0) \tag{1}$$

$$p_d = \frac{1}{S} \sum_{s=1}^{S} p_{ds} \tag{2}$$

where t_{ds} denotes the observed value of the test statistic and T denotes the the distribution of the test statistic. Since ALDEx2 is averaging over multiple p-values, this specification can be problematic if the signs of the test statistic vary over the S replicates for a given entity d. To see this concretely, consider a simple example where there are two conditions with 10 samples per condition ($N = 20$). Suppose that, for a given entity, two Monte Carlo samples are drawn, resulting in the test statistics $t_{d\cdot} = \{-3, 3\}$. The two-sided p-values are $p_{d\cdot} = \{0.008, 0.008\}$ (18 degrees of freedom) which corresponds to $p_d = 0.008$. At face value, the aggregated p-value would suggest that there is a significant effect, yet the two Monte Carlo samples disagree on the sign of the effect.

This problem is well-known in the statistical literature of combining p-values from multiple hypothesis tests [1]. It can largely be avoided by careful selection of the tested hypotheses. In line with recommendations in the literature [1], we propose a slight modification of the original approach in ALDEx2 and average over p-values obtained from one-sided tests (p_{ds}^{mod}) instead of two-sided tests:

$$p_{ds}^{\text{mod}} = Pr(T > t_{ds} \ |H_0) \tag{3}$$

$$p_d^{\text{mod}} = \frac{1}{S} \sum_{s=1}^{S} p_{ds}^* \tag{4}$$

$$p_d^{\text{mod}} = 2 \times \min(p_d^*, 1 - p_d^*). \tag{5}$$

where T is defined as before and p_d^{mod} denotes the p-value obtained from this modified procedure for entity d. In words, we compute the one-sided p-value for a right-tailed alternative for all Monte Carlo samples (p_{ds}^{mod}), average over all Monte Carlo samples (p_d^{mod}), and select the direction that leads to the lowest overall p-value (p_d^{mod}) after multiplying by a factor of two to convert two a two-sided p-value. Equation (5) holds due to the fact that the left-tailed and right-tailed p-value must sum to one due to complimentary events.

Practically, the new procedure requires that each Monte Carlo replicate offers evidence towards both the magnitude and direction for an entity to be returned as significant. For example, $p_d^{\text{mod}} = 1$ in our toy example. More generally, the modified p-value (p_d^{mod}) will be larger or equal to the original p-value (p^d) produced by ALDEx2. If all Monte Carlo replicates agree on sign, the two will be equal ($p_d^{\text{mod}} = p^d$), otherwise $p_d^{\text{mod}} > p^d$.

S3 ALDEx2 as a Linear Model

While the results in the main text were described for differential abundance or expression, they apply more generally to linear models. Note that differential abundance and expression is equivalent to a linear model with an intercept and a binary covariate denoting the condition.

To view the results of the main text as a linear model, let $\hat{\theta}_{dr}$ denote the estimated linear relationship between covariate r and entity (taxa or gene) d: i.e., an estimate of the parameter in a linear model of the form $\log W_{dn} = \alpha_d + \sum_r \theta_{dr} X_{rn} + \epsilon_{dn}$. In words, θ_{dr} captures the association between changes in the r-th covariate and changes in the amount of entity d. When viewed as a linear model, estimates produced by ALDEx2 suffer from the same phenomenon described in the main text: $\hat{\theta}_{dr}$ captures the association between changes in the covariate r and changes in the *normalized* amount

of entity d. This leads to the same phenomenon of unacknowledged bias described in the main text.

Procedurally, ALDEx2 remains virtually identical to the method for differential abundance and expression with one exception: for each taxa d and sample s, the t-test is replaced by a linear model with mean (i.e., expectation, $\mathbb{E}$):

$$\mathbb{E}[\log \hat{W}_{dn}^{(s)}] = \theta_{d0} + \theta_{d1} X_{1n} + \cdots + \theta_{dR} X_{Rn},$$

resulting in estimates $\hat{\theta}_{dr}^{(s)}$ and p-values $p_{dr}^{(s)}$ corresponding to the null hypothesis $\theta_{dr} = 0$ (using a t-test or Wald test). Corrected p-values using either the Holm or Benjamini-Hochberg procedure are also returned. These estimates and p-values are summarized over the S samples in an analogous manner to the procedure described for differential abundance or expression analysis.

S3.1 Default Scale Model in Linear Models

In the case of linear models, the new default scale model is:

$$\log \hat{W}_{n}^{\perp(s)} = -\log \phi\left(\hat{W}_{\cdot n}^{\parallel(s)}\right) + \sum_{r=1}^{R} \Lambda_r^{\perp} x_{rn}$$
$$\Lambda_r^{\perp} \sim \; N(0, \gamma^2).$$

where x_{rn} denotes the user specified covariates X_{rn} re-scaled so that continuous covariates have mean zero and standard deviation 1, and binary covariates are coded as 0 or 1. This scaling ensures the resultant model is invariant to arbitrary recodings of the covariates, e.g., it is invariant to whichever binary encoding is used for X.

Since the default scale model is an extension of the one presented in the main text for differential abundance and expression, it retains the design features described previously: namely, it requires only a single input and will reduce type-I error rates

compared to the CLR normalization. In addition, this scale model has a third property for linear models that we find appealing: the term $\sum_{r=1}^{R} \Lambda_r^{\perp} x_{rn}$ scales with the number of covariates R. The rational is as follows: if each covariate could change the scale of the system under study (e.g., a drug could kill microbes), then our uncertainty in the system scale should also scale with the number of covariates.

S4 Advice on Choosing γ

For simplicity, we state this advice in terms of the default scale model for differential abundance. Extending this intuition to the default scale model for linear models is straightforward. As discussed in *Methods*, the parameter γ controls the amount of uncertainty in the scale assumption implied by the CLR normalization. Similarly, $\Lambda^{\perp}$ is a factor representing how the scale changes between biological conditions with larger values of γ corresponding to more scale uncertainty. Thus, the default scale model assumes that, with 95% certainty, the log-fold-change of scales between conditions is between $(2^{-2\gamma+\hat{\theta^{\perp}}}, 2^{2\gamma+\hat{\theta^{\perp}}})$.

This interpretation can be used to find a reasonable value of γ in a given context. For example, consider the simulation presented in the main text. Suppose we know the antibiotic is narrow spectrum and expect that scale changes minimally between conditions (most taxa are not affected): we expect that the antibiotic is unlikely to kill more than 20% of microbial load. We can use the Dirichlet samples to get an approximation of the CLR estimate ($\hat{\theta}^{\perp} = 0.04$), which means the CLR assumption is estimating a slight increase in microbial load after antibiotic treatment. As this seems counter-intuitive, if we want to still use the default scale model we should choose a suitably large value of γ to cover our beliefs and account for this bias in $\hat{\theta}^{\perp}$. Equating our assumption that the antibiotic is unlikely to kill more than 20% of the microbial load with the lower bound of our 95% probability region of the default scale model, we can calculate the value of γ by solving the equation:

$$2^{-2\gamma+0.04} = 0.8, \tag{6}$$

yielding a value of $\gamma = 0.18$:

$$\gamma = \frac{\log_2 0.8 - 0.04}{-2} \tag{7}$$

$$\gamma = 0.18. \tag{8}$$

S5 Analysis of a SparseDOSSA Simulation

To show the applicability of our methods in other simulation frameworks, we re-created the simulation study of McGovern and Silverman [2], which imitates the effects of a potent pre-biotic that increases the abundance of over half of the taxa in the human gut. Those authors created this simulation using SparseDOSSA2, which is trained on real data to produce realistic simulations of microbiome data [3]. McGovern and Silverman [2] created this simulation to show how the CLR normalization can fail and how scale reliant inference (SRI) methods can address the issue. Specifically, they designed the study such that $\theta^{\perp} = 1.46$ yet the CLR-based point estimate was approximately $\hat{\theta}^{\perp} = -0.01$. Those authors then showed that the limitations of the CLR normalization can be addressed using a right-skewed scale model that reflected weak prior knowledge that total microbial load increased after treatment. Here, we recreate those results and illustrate how, with large enough values of γ, the default scale model can also address the limitations of the CLR normalization, albeit at the cost of statistical power. Full details of this simulation can be found in McGovern and Silverman [2].

We applied three scale-based methods and four normalization-based methods to each simulated dataset. For scale-based methods, we applied the default scale model with $\gamma = 0.5$, the default scale model with $\gamma = 1$, and a right-skewed scale model

developed by McGovern and Silverman [2]. We compared these scale-based methods to the original ALDEx2 model [4], DESeq2 [5], edgeR [6], and limma [7]. Default values were used for these methods. False positive rates and power versus sample size for each method are plotted in Figure S1.

Only the default scale model (with $\gamma = 1$) and the right-skewed scale model controlled false positive rates at nominal levels. Remarkably, even at the smallest sample sizes, all normalization-based methods display false positive rates above 5%. As expected, the default scale model does decrease false positive rates compared to the original ALDEx2, but false positive control (at $\leq$5%) only occurs with larger values of γ (e.g., $\gamma = 1$). This behavior is expected given the simulation was designed to violate the assumptions behind the CLR normalization, which the default scale model is based on. As a result, large amounts of uncertainty are required to overcome the bias of the default scale model. However, so much uncertainty does come at the price of statistical power. In contrast, the more plausible right-skewed scale model controls FDR at all sample sizes while simultaneously having high statistical power.

S6 Sensitivity Analyses

To facilitate the choice of γ, we developed and implemented a special sensitivity analysis for the default scale model. Simply, this sensitivity analysis repeats our method over a user-specified range of γ, allowing for an easy comparison of results over different scale assumptions. This algorithm was made efficient using the specialized memorization algorithm developed in Nixon et al. [8], and the sensitivity analysis is implemented in the `aldex.senAnalysis` function in the `ALDEx2` package. A special plotting function is available to visualize these results (`plotGamma`); an example of the visualized output is in Figure S2.

To show how these visualizations can be helpful, we conducted a sensitivity analysis of the simulation (Figure S2) and the selected growth experiment (Figure S3) from the

main text. In both instances, true negatives are typically more sensitive to errors in scale assumptions than true positives. For example, consider the simulation example in Figure S2: there is a large cluster of taxa that are only significant (black) for tiny values of γ. This cluster contains all of the false positives. Including even small amounts of scale uncertainty (choosing small but non-zero γ) is enough to remove these false positives. Even if we did not know which were true versus false positives, seeing a cluster of taxa so sensitive to even slight amounts of error in scale assumptions should be sufficient to allow researchers to avoid drawing conclusions about these taxa and would suggest that γ should be set to be greater than 0.1. Similarly, we conducted a sensitivity analysis in the selected growth experiment example (Figure S3). Our results show that the true positives are much less sensitive to errors in scale assumptions than true negatives (e.g., all significant genes are true positives for $\gamma > 1$). Although higher levels of scale uncertainty does cause some reduction in power, these effects are minimal at realistic levels of uncertainty (e.g., $\gamma < 3$).

In addition, sensitivity analyses can eliminate the need to choose γ and facilitate a novel, reproducible, and transparent reporting form. Rather than specifying a value of γ a priori, a researcher would report the maximum value of γ under which a result holds, e.g., based on Figure S2, taxa 20 is differentially decreased after antibiotic treatment ($p < 0.05$) so long as $\gamma < 0.70$. Overall, we would expect that such a reporting would enhance the transparency and reproducibility of sequence count data analyses, allowing readers to interpret the robustness of reported conclusions themselves.

S7 An Analysis of an Oral Microbiome Dataset

To further highlight how literature can be used to create biologically plausible scale models, we reanalyzed an oral microbiome data set originally published by Morton et al. [9] and recently reanalyzed by McGovern and Silverman [2]. The latter authors found that existing normalization-based models had high false positives and negative

rates, yet a novel form of *interval assumption* addressed those problems. As discussed in that article, those interval assumptions complement – and are inspired by – the scale model methods we introduce. Here, we show that the interval assumptions built in that article can also be implemented as scale models. Both lead to the same conclusions and decreases in false positive and negative rates when compared to normalization-based methods.

This dataset consists of $N = 32$ pairs of salivary samples collected before and after teeth brushing. Each sample was profiled using 16S rRNA-seq and flow cytometry. Here, we take the results of McGovern and Silverman [2], obtained using interval assumptions built from the paired flow-cytometry models, as our gold standard. In contrast, we built a scale model that does not use those flow cytometry measurements but only uses results previously reported in the literature. Slot et al. [10] reported up to 99% reductions in plaque index scores in *in vitro* studies whereas Funahara et al. [11] found increases on the order of 25% in the bacterial load after brushing. To mimic the wide range of results, we used a uniform scale model that ranges between a 99% reduction to a 25% increase in bacterial loads after brushing.

Results for each method are depicted in Figure S4. Without exception, each of the normalization-based methods displays false positives. DESeq2 and the original ALDEx2 even display false negatives as well. In contrast, our biologically informed scale model perfectly recreates the gold-standard results. This is remarkable given that our biologically informed scale model used only prior literature, whereas the gold standard had access to the paired flow cytometry results.

While our default scale model does not perform as well as the biologically informed model, it does improve upon the results of the normalization-based methods. Adding scale noise removes two of the false positives seen in the normalization-based models. Yet it leaves the two false negatives observed in the original ALDEx2 model.

S8 Analysis of an RNA-Seq Dataset

We reanalyzed an RNA-seq study originally published in Gierliński et al. [12] and reanalyzed in Schurch et al. [13] which contained 48 biological replicates from either wild-type (WT) or a SNF2-knockout (SNF2) strain of *Saccharomyces cerevisiae*. Raw reads for this data set was obtained from the European Nucleotide Archive repository (ENA, accession code , http://www.ebi.ac.uk/ena/data/view/ERX425102) and processed as in Schurch et al. [13]. Aggregated data is also available at https://github.com/ggloor/datasets. To mimic Schurch et al. [13], the data set was restricted to only those replicates that passed cleaning checks described in Gierliński et al. [12], resulting in 44 Δsnf2 biological replicates and 42 wild-type replicates. Entities were filtered such that all entities with zero counts across all of the restricted replicates were removed, resulting in 6,327 entities. Full experimental details can be found in Gierliński et al. [12] and Schurch et al. [13].

Prior analyses of these data have shown that standard differential expression tools typically report about 70% of genes as being differentially expressed at the full data size [13]. These results seem implausible. While SNF2 is a key gene involved in chromatin remodeling [14], we think it unlikely that this single gene knockout leads to change in the expression of over 70% of expressed genes. Instead, we suspect that this might be a problem of unacknowledged bias, especially as prior authors have noted that the positive identification rate drastically rises as a function of sample size [13]. While we do not know which genes are truly differentially expressed within this study, we can investigate our hypothesis by quantifying the sensitivity of differential expression to modeling assumptions about scale. Central to our argument is the idea that a scientific conclusion is suspect if it only holds under a very restrictive set of unrealistic modeling assumptions.

We compared the results from ALDEx2 using two different scale models: the default scale model and an *informed* scale model built based on consideration of the experimental design. This model was specified using the results of Yoshikawa et al. [15] which suggested a 10% decrease in the SNF2 deficient strains compared to wild-type. Echoing this belief, we used this as a basis for our informed scale model:

$$\log(W_{\mathrm{n}}^{\perp(s)}) \sim N(\mu, \gamma^2) \tag{9}$$

where $\log(W_{\mathrm{n}}^{\perp(s)})$ indicates the s^{th} draw from the scale model for sample n, $\mu = \log(1)$ for the wild-type strain, and $\mu = \log(0.9)$ for the SNF2 strain. Within the context of this study, the CLR assumption corresponds to an implicit assumption that total RNA expression in the WT strain is about 20% higher than in the SNF2 strain. Note that the assumed direction is the same as the custom scale model. For both the default scale model and the user-specified model, we set $\gamma = \{0, .25, .5, .75, 1, 1.25, 1.5, 1.75, 2\}$, sampled 500 Monte Carlo replicates, applied Welch's t-test, and selected significant entities according to a corrected p-value threshold of 0.05. The results are summarized in Figure S5.

Echoing the results of Schurch et al. [13], at the full data size, ALDEx2 with the CLR assumption (default scale model with $\gamma = 0$) finds that approximately 68% of genes are differentially expressed between conditions. As hypothesized, this result drops dramatically with even low levels of scale uncertainty. For example, at $\gamma = 0.25$, this figure drops from 68% to just 12%. Similar results are seen for the ALDEx2 with the informed scale model. For the smallest value of γ, the two scale models generally agree, but, for larger values of γ, the informed scale model returns more genes as significant than the default scale model, representing potential loss of power if the default scale model is biased. Additionally, we note that the knockout gene (YOR290C)

is significant for either scale model for all values tested and has the largest effect size for both models when $\gamma = 2$.

Overall, these sensitivity analyses suggest that less than 70% of genes are differentially expressed when even small amounts of scale are included in the model, and that prior reports may have been biased due to error caused by inappropriate assumptions about scale.

References

[1] Oosterhoff J. Combination of one-sided statistical tests. Amsterdam: Mathematical Centre; 1969.

[2] McGovern KC, Silverman JD. Replacing Normalizations with Interval Assumptions Improves the Rigor and Robustness of Differential Expression and Differential Abundance Analyses. bioRxiv. 2024;p. 2024–10.

[3] 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 computational biology. 2021;17(9):e1008913.

[4] 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.

[5] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology. 2014;15(12):1–21.

[6] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–140.

[7] 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 Research. 2015;43(7):e47–e47.

[8] Nixon MP, McGovern KC, Letourneau J, David L, Lazar NA, Mukherjee S, et al. Scale reliant inference. arXiv pre-print arXiv:220103616. 2024;p. 2024–11.

[9] Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, et al. Establishing microbial composition measurement standards with reference frames. Nature Communications. 2019;10(1):2719.

[10] Slot D, Wiggelinkhuizen L, Rosema N, Van der Weijden G. The efficacy of manual toothbrushes following a brushing exercise: a systematic review. International Journal of Dental Hygiene. 2012 Jun;10(3):187–197. https://doi.org/10.1111/j.1601-5037.2012.00557.x.

[11] 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.

[12] 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):3625–3630.

[13] 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):839–851.

[14] Peterson CL, Tamkun JW. The SWI-SNF complex: a chromatin remodeling machine? Trends in Biochemical Sciences. 1995 Apr;20(4):143–6. https://doi.

org/10.1016/s0968-0004(00)88990-2.

[15] 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):349–361.

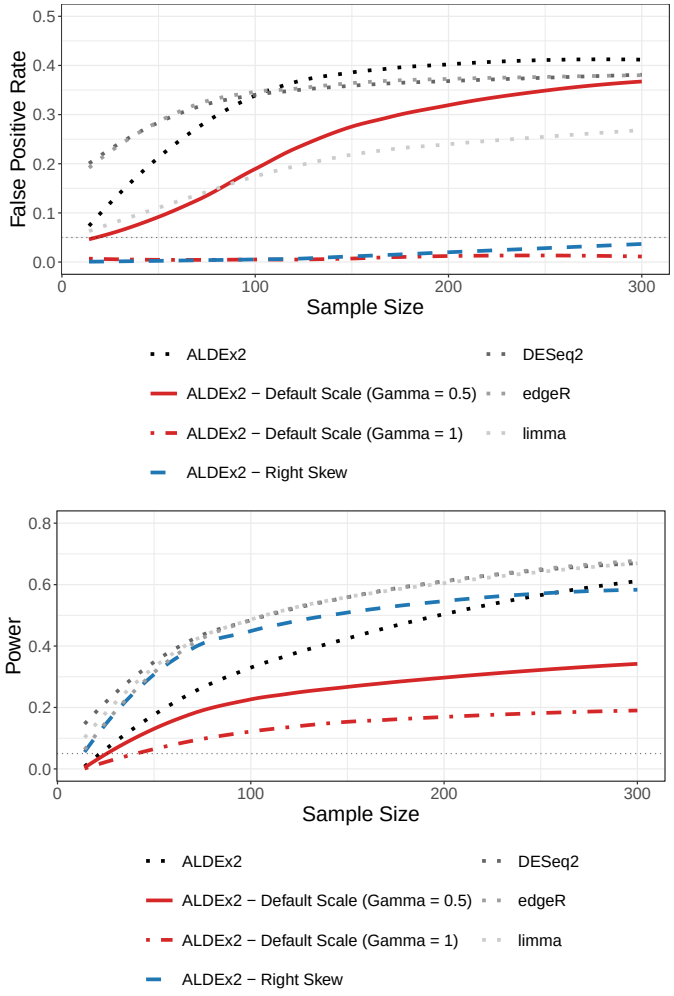

Fig. S1 False positive rate and power for the SparseDOSSA simulation developed by McGovern and Silverman [2].

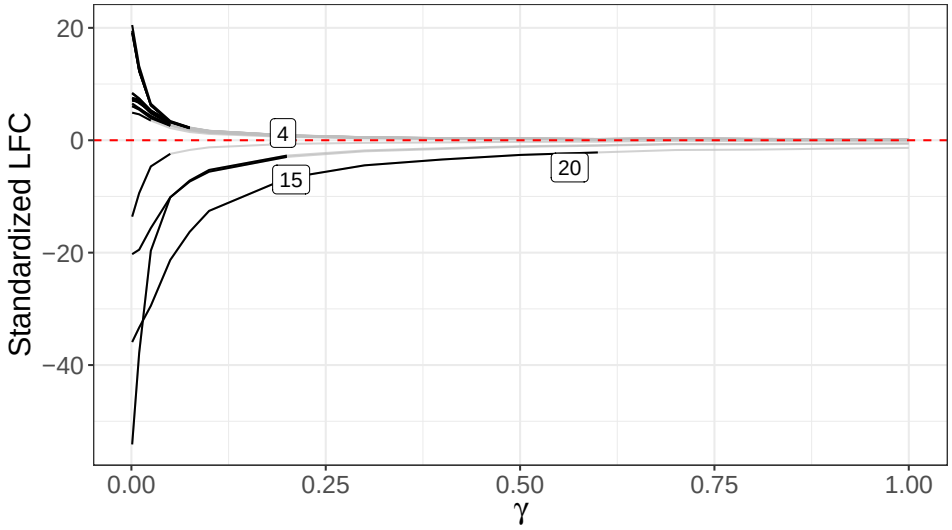

Fig. S2 Sensitivity analysis for ALDEx2 with varying levels of γ. We conducted a sensitivity analysis of the simulation data using the default scale model and fitting procedures as described in *Methods*, varying the amount of uncertainty added to the default scale model ($\gamma \in [0, 1]$). At each value of γ and for each entity, we computed the log fold change (LFC) for each Monte Carlo sample and calculated the mean and standard deviation across these samples. Then, we estimated a standardized value of the LFC by dividing the mean by its corresponding standard deviation. Standardized LFCs as a function of γ were plotted, and lines were marked in black if the p-value for that entity was less that 0.05. Even as γ increases, several of the true positive results remain, e.g. taxa 4, 15, 20.

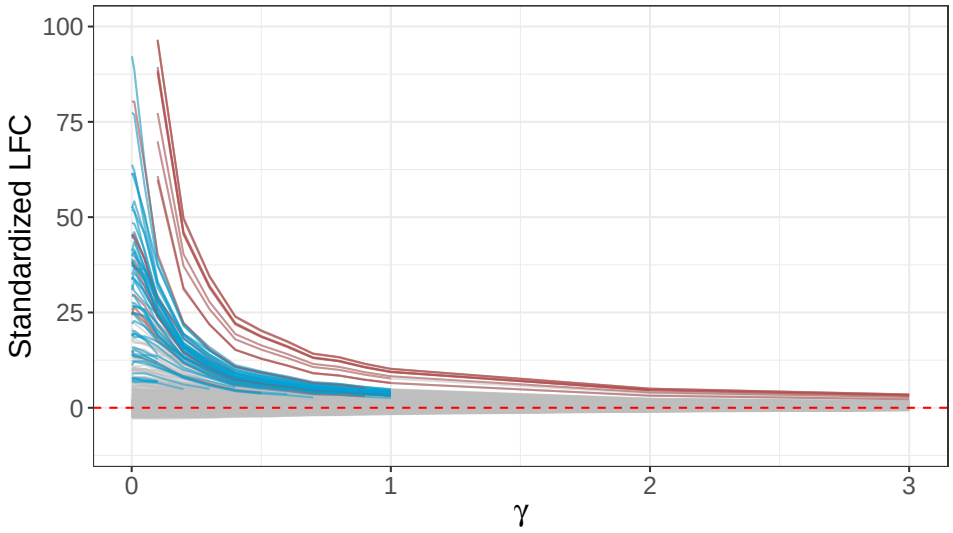

Fig. S3 Sensitivity analysis for the selected growth experiment (SELEX) data with varying levels of γ. We conducted a sensitivity analysis of the SELEX data using the default scale model and fitting procedures as described in *Methods*, varying the amount of uncertainty added to the default scale model ($\gamma \in [0, 3]$). At each value of γ and for each entity, we computed the log fold change (LFC) for each Monte Carlo sample and calculated the mean and standard deviation across these samples. Then, we estimated a standardized value of the LFC by dividing the mean by its corresponding standard deviation. Standardized LFCs as a function of γ were plotted. Lines were marked in red if the p-value for that entity was less than 0.05 and the entity was truly changing (true positives), in blue if the p-value for that entity was less than 0.05 and the entity was not truly changing (false positives), and in gray if the p-value for that entity was greater than 0.05 (non-significant entities). In aggregate, the true positives are much less sensitive to the added noise than the true negatives.

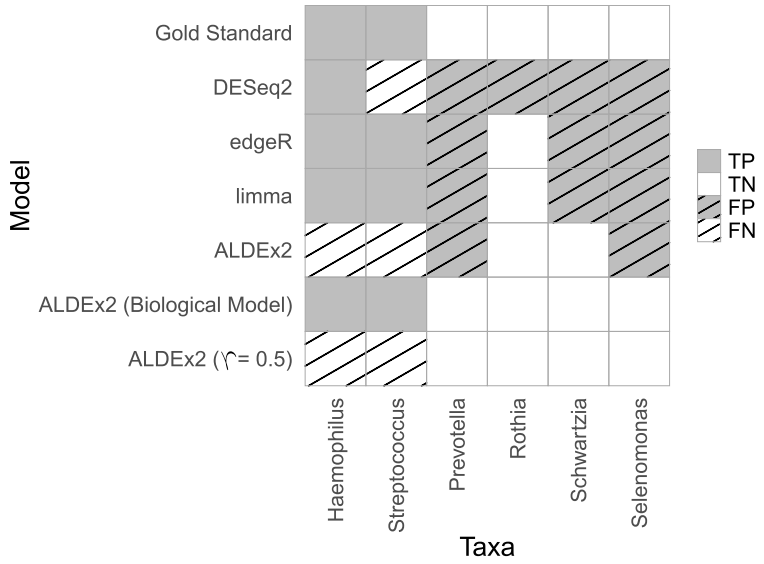

Fig. S4 Scale models enhance the analysis of oral microbiome data.

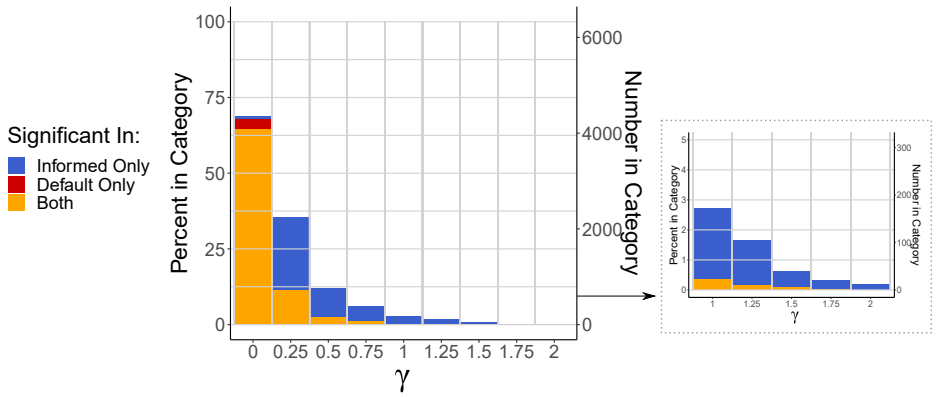

Fig. S5 A sensitivity analysis for the RNA-seq data. We analyzed the RNA-seq data set using the updated ALDEx2 model as outlined in Section S8. We tested two scale models, the default scale model and the Informed model, varying the amount of uncertainty added (γ). For each model and value of γ tested, we recorded which entities were significant and compared these results to the opposing scale model. As scale uncertainty increases, the number of positives drops regardless of scale model (e.g., approximately 65% of entities are positive at $\gamma = 0$ compared to 1.5% of entities at $\gamma = 0.5$ for the default scale model).