The accuracy of absolute differential abundance analysis from relative count data

Kimberly E. Roche[1*], Sayan Mukherjee[123]

1. Program in Computational Biology and Bioinformatics, Duke University, Durham, NC 27708, United States

2. Departments of Statistical Science, Mathematics, Computer Science, Biostatistics & Bioinformatics, Duke University, Durham, NC 27708, United States

3. Center for Genomic and Computational Biology, Duke University, Durham, NC 27708, United States

* kimberly.roche@duke.edu

Abstract

Concerns have been raised about the use of relative abundance data derived from next generation sequencing as a proxy for absolute abundances. In the differential abundance setting compositional effects are hypothesized to contribute to increased rates of spurious differences (false positives). However in practice, partial reconstruction of total abundance can be imputed through renormalization of observed per-sample abundance. Given the renormalized data differential abundance need not be called on relative counts themselves but on estimates of absolute counts. We use simulated data to explore the consistency of differential abundance calls made on these adjusted relative abundances and find that while overall rates of false positive calls are low substantial error is possible. Conditions consistent with microbial community profiling are the most at risk of error induced by compositional effects. Increasing complexity of composition (i.e. increasing feature number) is generally protective against this effect. In real data sets drawn from 16S metabarcoding, expression array, bulk RNA-seq, and single-cell RNA-seq experiments, results are similar: though median accuracy is high, microbial community profiling and single-cell transcriptomic data sets can have poor outcomes. However, we show that problematic data sets can often be identified by summary characteristics of their relative abundances alone, giving researchers a means of anticipating problems and adjusting analysis strategies where appropriate.

Introduction

Warnings about the consequences of compositional effects in sequence count data have been published repeatedly in the decades since the technology's advent and its application to a host of biological problems. The issue relates to a loss of scale information during sample processing, which renders counts of genes, transcripts, or bacterial species as relative abundances. No consensus solution for this problem exists. In this work, we use simulated and real data on differential abundance calling to quantify the discrepancy between differential abundance estimates made on relative versus "absolute" abundances. Our simulations show that methods which heuristically rescale sample abundances are often highly consistent across relative and absolute count data and we confirm that the low complexity case, roughly corresponding to bacterial community profiling, is the most problematic. Further, we show that data sets which are especially susceptible to distortion by compositional effects can often be predicted on the basis of "signatures" of this distortion.

Compositionality in sequence count data

Compositionality refers to the nature of sequence count data as containing relative abundance information only. In the differential abundance setting, several authors [1, 2, 3] have described the problem this poses: whereas researchers would like to interpret change in absolute abundances, compositional effects mean using change in relative abundances as a proxy can lead to false discoveries. A few authors have cited instances of these false discoveries in real data. Coate and Doyle [4, 5] discussed the issue of transcriptome size variation in plants and other systems and the impact of this on accurate transcriptome profiling. Nie *et al* and Lin *et al* [6, 7] documented the phenomenon of widespread "transcription amplification" by the transcription factor *c-Myc* and Lovén *et al* [8] used *c-Myc* data and parallel RNA quantification assays to show that substantial differences in total abundance between control and elevated *c-Myc* conditions resulted in very different interpretations of apparent differential expression.

Common to these studies of transcriptomes is a recommendation that, where feasible, researchers leverage RNA spike-ins as controls against which changes in observed abundance can be scaled [9, 10]. But this practice has fallen short of widespread adoption. While several papers have expressed confidence in the utility of spike-ins [11, 12, 13, 14, 15], the doubt cast by reports of widespread batch effects [16] and technical noise [17] have had the effect of reducing researcher confidence in their use. Further, the introduction of spike-ins is not practical on all platforms.

Box 1: Measuring *relative* abundances

Sequence counting has become widespread as a means of census-taking in microscopic biological systems. Genomic material, typically RNA, is captured and quantified at the component level. Sampled cells are lysed, messenger RNA is captured and fragmented, transcribed into cDNA, sequenced, classified, and quantified. The results are relative abundances of gene products in the cell (in the case of single-cell RNA-seq) or tissue (in bulk RNA-seq). In another instance, whole bacterial communities are profiled by barcoding of the 16S subunit of the ribosome. Ribosomal RNA associated with this piece of translation machinery is ubiquitously present across the bacterial kingdom but variations in the genetic sequence of this component can uniquely identify bacteria to the species or strain level in well-characterized systems, allowing a researcher to profile bacterial community composition. Absent measurements of microbial load or transcriptome size, however, the observed sequence counts in all these cases represent relative abundances.

Sequence count data is compositional due to steps in sample processing. Across domains, samples are typically normalized to some optimal total amount of genetic material prior to sequencing in accordance with manufacturer recommendations for best performance. This step removes variation in total abundance across samples. Saturation of sequencing has been cited [18] as another mechanism by which abundances are rendered relative: a finite amount of reagent means there is an upper limit on biological material which can be captured; rare components can be forced out by a ”competition” to be sampled. These factors withstanding, observed total abundances would likely still be noisy. Repeated subsampling of small amounts of material and variation in the efficiency of library preparation steps can distort observed totals.

In transcriptomics and in microbial community profiling, residual variation in observed total abundances across samples is generally taken to be technical noise and most analytics pipelines involve steps to renormalize observed abundances. The simplest of these is the counts per million (CPM) transformation which converts observed counts to relative abundances, then scales by 1 million.

Where approaches that rely on spike-ins are undesirable or infeasible, sample renormalization procedures have proliferated. These methods typically assume the existence of a stable set of features and attempt to normalize compositions in such a way as to recover this stable set across samples. In fact, in transcriptomics, these methods predominate.

In the setting of microbial community profiling, the prevailing assumption is that typical compositions are too simple for renormalization methods to work well (although results in benchmarking studies have been mixed [19, 20]). Competing approaches have been developed for dealing with compositionality in microbial sequence count data. Quantitative microbiome profiling [21] and similar approaches combine relative abundances with complementary measurements of microbial load to reconstruct absolute abundances. In contrast, so-called compositional methods are also utilized. These involve log relative representations which can give approximate log-normality, such that workhorse statistical methods for continuous data may be applied. However, interpretation of these quantities can be challenging (e.g. as with the isometric logratio [22]).

Though there is evidence from simulated and real data that *scale* - i.e. increasing complexity of composition in terms of numbers of genes, transcripts, or bacterial sequence variants - mitigates the problem of compositionality [2, 20], it remains unclear when it is practical to substitute relative abundances for absolute abundances. Several fields could benefit from clarity on the nature of the boundary between ”safe” and problematic scales.

In this work, we quantify the discrepancy in differential abundance calling on simulated and real data sets representative of 16S metabarcoding, bulk RNA-seq, and single-cell RNA-seq experiments. We show that discrepancy in differential abundance calls is low across a very wide range of settings. The lowest complexity cases are associated with the greatest error but under these conditions a tradeoff becomes apparent between specificity and sensitivity: the rate of false discoveries can be kept low at the cost of lesser sensitivity. Using real data sets with substantial absolute and compositional change, we show that false positive rates in differential abundance calling in real data tend to be low and that these outcomes can be predicted using signatures of compositional distortion derived from summaries of sparsity and feature-level change.

Results

We evaluated the consistency of differential abundance calls between observed (relative) abundances and absolute abundances. Three methods were used to make differential abundance calls: ALDEx2 [23], DESeq2 [24], and scran [25]. Using thousands of simulated data sets, we found that consistency of calls was high overall. A microbial metabarcoding-like setting was associated with the lowest consistency but high rates of false positive calls were possible for all settings evaluated. The ”transcriptomic” or high complexity setting was not inherently protective against false positives induced by compositional effects. In this setting discrepancy was a function of the number of differentially abundant features; the scale of the overall change in abundance between conditions was of less importance. In real data sets representative of more problematic possible cases, we saw moderate-to-high sensitivity - generally highest for scran and lowest for ALDEx2. Specificity was also high overall. We developed a method for predicting these outcomes from observed data alone and demonstrate its utility in identifying especially problematic combinations of data set and differential abundance calling method. We discuss these findings in detail in the sections below where we first address results in simulated, then in real data.

Results in simulation

We simulated differentially abundant count data in paired sets of absolute and relative abundances, where the relative data were derived from a resampling procedure. We explored ranges of complexity in composition and amount of differential abundance, grouping simulations into three partially overlapping settings: a Microbial setting, characterized by low complexity and high differential abundance; a Bulk transcriptomic setting with high complexity and low differential abundance; and an intermediate Cell transcriptomic setting. The full results from almost 6000 simulations are shown in Figure 1. We present the same results in terms of increasing complexity of composition (expressed as increasing feature number) in Figure 2. We evaluated the consistency of a small set of popular differential abundance testing methods, each of which attempts to renormalize per-sample total abundances, generally by rescaling these relative to a reference quantity. Details on the simulation procedure and analysis methods evaluated are given in Methods.

We report outcomes in terms of sensitivity (true positive rate) and specificity (100% - false positive rate). Perfect concordance of differential abundance calls made on observed versus absolute counts would yield 100% sensitivity and 100% specificity. Sensitivity drops as more differentially abundant features are ”missed” and specificity drops as more erroneous differential calls are made. We highlight key observations made on simulated data below.

High false positive rates are possible for all methods

We see false positive rates well in excess of single digits for all methods. Low complexity simulations are the most problematic in this respect. In our lowest complexity setting (100 features), almost a quarter of simulated data sets exceed a 5% false positive rate for scran and DESeq2. For ALDEx2 this proportion was over half, at 57%.

Larger feature number yields more predictable outcomes for all methods and reduces false positives for ALDEx2

Per-method performance does not inevitably improve with higher complexity of composition. However, the ultimate drivers of performance become more predictable at scale and we discuss these in the next section. While large feature number simulations tend toward predictable outcomes, simulations of low-complexity compositions are volatile: small changes in feature abundance have larger relative effects. Specificity for ALDEx2 improves markedly as feature number increases. This is likely the result of mean feature abundance - the quantity ALDEx2 uses to rescale observed sample abundances - stabilizing as compositions grow larger.

An increasing proportion of differential features drives increased false positive rates in renormalization-based methods

In Supplemental Figure S4 we summarize results on data sets with a large fold change across simulated conditions driven by a strict minority of differentially abundant features. Specificity is substantially higher for these simulated data sets than overall (and often very high when using scran). In other words, renormalization-based methods perform well where assumptions about large, relatively stable sets of features are valid, irrespective of the scale of change in total abundance across conditions.

Notably this is not true of methods which do not attempt renormalization. The poor performance of metagenomeSeq relative to other methods in Hawinkel *et al.* [19] and Calgaro *et al.* [20] is presumably for this reason.

Renormalization-based methods are robust to missing information about changes in scale

The methods we evaluated often performed well where fold change between simulated conditions exceeded 5-fold, so long as this change was driven by a minority of features in the composition. In particular, scran had a median specificity of 91% under these conditions, suggesting that distortion by compositional effects can be powerfully mitigated by available renormalization approaches.

Data consistent with microbial community profiling is the most sensitive to compositional effects and bulk transcriptomic data, the least

We present results for ”settings” corresponding broadly to several experimental modalities in Figure 1. See Methods for details. These are the Microbial, Bulk transcriptomic, and Cell transcriptomic settings.

Though the median specificity across all methods and settings was very high (98%), the proportion of data sets which gave false positive rates in excess of 10% by one or more methods was high as well. In the Microbial setting, 19% of data sets exceeded this ”high” false positive rate threshold and in the Bulk and Cell transcriptomic settings, the figure was 11% and 14%. Method-specific performance was as follows. Differential abundance calling via ALDEx2 in the Microbial setting had the worst performance, with fully one third of simulated data sets yielding ”high” (¿10%) false positive rates when differential abundance calls in observed counts were compared to calls made on absolute counts. In the same setting, the top performer was scran, with 9% of data sets yielding false positive rates in excess of this ”high” threshold. In the stabler Bulk transcriptomic setting, specificity was improved for ALDEx2 and scran: only 6% of all data sets exceeded a 10% false positive rate for either method versus almost 20% of data sets when DESeq2 was applied. The percent of ”high” false positive rate experiments with respect to scran never exceed single-digits, ALDEx2 performed poorly but improved with scale, and DESeq2’s performance remained largely consistent across settings.

Sparsity and feature-level change predict outcomes in random forest models

We next predicted sensitivity and specificity from observed data. It is possible to imagine characteristics of relative abundances which might indicate the presence of distortion by compositional effects: for example, an increase in the percentage of rare features from one condition to the next - in effect, dropouts. While we might not expect any single characteristic capable of predicting compositional distortion, composites of such characteristics might be. We generated ”signatures” of change in the form of combinations of summary features for each of our thousands of simulated data sets. These summary features included estimates of the relative sparsity in each of the simulated conditions, the maximum change in relative abundance between conditions, measures of uniformity of change in feature abundance between conditions, and many others outlined in Supplemental Table S4. Models were trained on 80% of our simulated data and performance was evaluated on the held-out 20% of simulations.

We used a measure of feature importance to identify the most informative features for each model. In almost all cases, the prediction of sensitivity was most improved by a set of features which captured information about the percent of simulated sequence variants with very low abundance. A large number of low abundance features was associated with a lower sensitivity. For the prediction of outcomes from ALDEx2 and DESeq2, features which estimated the apparent correlation of simulated sequence variants also had strong predictive value.

Specificity prediction was most improved by features encoding information about the percent of sequence variants with large apparent fold change between conditions. In general, data sets having a large number of features with apparent increases in abundance between conditions were associated with higher false positive rates (and thus, lower specificity). A summary of the top features for each model is given in Supplemental Tables S1 and S2.

False positive rates are ”ballpark” predictable for ALDEx2 and scran

Accuracy varied by method but was occasionally striking. Outcomes for DESeq2 were the easiest to predict, with R^2 values for observed versus predicted sensitivities and specificities of 86% and 74% respectively. For ALDEx2 and scran, sensitivity prediction was generally successful, with R^2 values of 80% and 92% respectively. Specificities were difficult to predict for these methods, however, and observed versus predicted outcomes had R^2 values of around 50% in both cases, indicating characteristics of the observed data can only give ”ballpark” predictions of false positive rates for these methods. See Supplemental Table S3 for full results on simulated data.

Results in real data

Next, we examined a variety of real data sets across several experimental settings in order to sketch a picture of outcomes in real data. We collected publicly available data from eight studies [26, 27, 28, 29, 30, 31, 32, 33] and attempted to reconstruct absolute abundances by normalizing observed total abundances against reference quantities provided in the same published materials. In most cases, these reference quantities were external RNA spike-in sequences. In others, reconstructed absolute abundances had already been estimated, as in [26] through quantitative microbiome profiling or QMP [21]. In one case [32], we normalized against *Gapdh*, a stable, highly expressed housekeeping gene [34]. We acknowledge the difficulty in reconstructing absolute abundances and caution that these estimates are partial approximations. These data sets are summarized in Table 1 and Figure 3 and were selected because we consider compositional effects possible for all: each case exhibits substantial change between conditions in terms of overall abundance and composition.

Low sensitivity is possible for all data modalities

Sensitivity was high overall, at 86%, but cases of low sensitivity were observed in every data type: the 16S metabarcoding data of Vieira-Silva *et al.* [26], the bulk RNA-seq of Hagai *et al.* [30], and the single-cell data of Klein *et al.* [32].

Specificity on real data is high

Median specificity was 95% indicating that, inasmuch as these data sets are representative samples of potential problematic cases, the concordance of calls made on observed versus absolute counts is high. That said, one of the lowest observed specificities occurred with ALDEx2 on the single-cell data set of Yu *et al.* [33], at 69%. In absolute terms, that amounts to over 1800 false positives - genes which were not confidently differentially expressed between conditions according to the absolute abundance data but which were differential from the perspective of the observed counts. Such cases of low specificity appeared to be method-specific as with DESeq2's high false positive rate on the data of Vieira-Silva *et al.* [26], a microbial experiment where relatively few bacterial genera made up the bulk of a highly dynamic composition.

The observed values for sensitivity and specificity for all methods are given in Figures 4 and 5 along with predicted ranges for the same quantities.

The noisy Microbial setting challenges all methods

All methods performed poorly on the data of Vieira-Silva et al. [26] in terms of either sensitivity, specificity, or both. This experiment featured low compositional complexity (70 unique bacterial genera) and substantial within-condition variation in composition across subjects. Results were notably improved on the microbial data of Barlow *et al.* [27], where composition was relatively stable within conditions.

Sensitivity is underpredicted where sequencing depth is low

Predicted intervals were accurate in most cases. Underprediction of sensitivity was an issue with the data sets of Barlow *et al.* [27] and Owens *et al.* [31], where observed total abundances were less than half their ”true” values. In other words, the resolution of observed counts was low relative to absolute counts and sensitivity was underpredicted accordingly.

Problems with specificity prediction are method-specific

As noted above, specificity was high overall and was generally well-predicted for all methods and data sets. Exceptions were the combination of DESeq2 on the 16S data set of Vieira-Silva *et al.* [26], which featured a small number of highly variable bacterial sequence variants, and ALDEx2 as applied to the data of Yu *et al.* [33], where two genes made outsized contributions to overall differences in abundance between tissue types.

Discussion

We have evaluated the accuracy of estimates of differential abundance in absolute count data from relative abundances. While the potential of compositional effects to drive differential abundance has repeatedly been described in the literature, uncertainty remains about the scope of this problem. Previous authors [19, 20] have shown that rates of false positives calls can be high. However, while this problem is typically discussed in terms of estimates made from relative abundances, sophisticated renormalization techniques are frequently applied to the data, yielding partial approximations of absolute abundances. We were interested in the performance of these renormalization-based methods applied over a range of data settings. Our results, in line with those of others [2, 10], indicate that high rates of false discoveries driven by compositional effects are certainly possible for these methods but that accuracy, especially with respect to false positive rates, can nonetheless be high overall.

We find that scale (i.e. complexity of composition) does matter. The Microbial setting as we have characterized it - having simple, volatile compositions - is the most challenging for all methods. But we observed both low and high accuracy on representative real data sets, suggesting there are tolerable regimes within this most-challenging setting.

At the higher level of compositional complexity of 1000 features, most methods are fairly consistent in their calls on relative versus absolute count data. Of almost 2000 simulations in this higher-complexity setting, only 6.4% of data sets evaluated with scran exhibited false positive rates exceeding 10%. In fact, DESeq2 and scran were remarkably consistent across all settings. In the differential abundance calling methods we have considered, the strongest single predictor of consistency of results was the number of differentially abundant features not the scale of the change between conditions. For renormalization-based methods, a large fold change in overall abundance driven by a minority of features can often be rescued, in contrast to the case of relative abundance data.

Our simulated results mirror observations from real data. While the worst false positive rates in real data were over 25%, these were exceptions to a theme of generally high specificity. Sensitivity, on the other hand, varied as a function of sequencing depth.

Problematic data sets can often be identified from observed (relative) count data alone: some intuitive characteristics of the observed count data can forecast problems in differential abundance calling, in particular sparsity and the prevalence of variable features. The most common defect in prediction was an underestimate of sensitivity.

From these results, we can establish expectations about whole experimental domains. In metagenomics experiments utilizing 16S barcoding, a small number of microbial strains often dominates a composition and feature number can be quite low. In this setting, accuracy can be poor but workarounds exist. Firstly, a tradeoff seems to exist between sensitivity and specificity [19, 20]. Low rates of false positives on differential abundance calls made on relative abundances can be achieved at the cost of sensitivity. Also, viable methods for estimating changes in overall abundance have been developed in the field of microbial ecology [21, 9], meaning it may not be necessary to work from relative abundances at all.

Expression profiling is likely less burdened by the effects of compositionality. The relatively deep sequencing and complex, stable compositions of bulk RNA-seq data mean accuracy is overall high for renormalization-based methods. Where large changes in total mRNA or widespread differential expression is possible, control quantities like RNA spike-ins, though controversial, probably have real utility as renormalization references.

Differential expression from single-cell expression profiling may be reasonably accurate as well, given its transcriptomic scale, although results from real data indicate that choice of method is key here. Further, we note that some single-cell platforms generate library sizes (i.e. total per-sample observed abundances) which are already roughly proportional to absolute abundances. Brief examples of this are explored in the Supplement. This is in line a view that abundances in deeply sequenced UMI-barcoded single cells are likely to be a good proxy for absolute abundances [12, 14]. The effect of compositionality may be negligible on these platforms.

Several parts in this work invite further exploration. In attempting to reconstruct absolute abundances in real data sets, we generally utilize spike-ins as control quantities against which to scale total sample abundances. The resulting reconstruction is undoubtedly noisy and only an approximation of real change in the system. Further, we note that in our real data sets, choices about the inclusion or omission of very low abundance features were observed to affect the outcomes of differential abundance testing. The absence of an optimal strategy for filtering out ”uninteresting,” near-zero-abundance features is a deficiency.

While our simulated data sets averaged a zero composition of about 20%, in 16S metabarcoding and single-cell data sets especially, this proportion can be much higher. The filtering procedure referenced above rendered real data comparable to our simulations in this respect but a more thorough treatment of the subject would explore a wider range of sparsity in simulation. Also, for simplicity, we refrained from exploring the effect of varying numbers of samples per conditions. This no doubt affects results through the certainty of estimates. Lastly, a further important investigation in the spirit of recent work by Lloréns-Rico *et al.* [35] would assess the usefulness of partial reconstructions of abundance (e.g. by quantitative microbiome profiling or spike-in renormalization) by quantifying the direct effect information restored on accuracy of downstream analyses.

Methods

We simulated general purpose molecular count data. These abundances are interpretable as a variety of biological quantities, for example, transcript abundance in a cell or bacterial species abundance in a microbial community. These counts undergo a sampling step intended to loosely replicate the process of measurement itself and, crucially, the approximate normalization of total abundance across samples during that step, giving a second set of count data. We refer to the first set of count data as ”absolute” counts and the second, resampled set as ”observed” counts. We quantitatively explored the degree to which this loss of information about changes in total abundance alters the results of a mock differential abundance analysis by simulating a huge range of settings in our data, where key characteristics like complexity of composition (e.g. gene number) and fold change across simulated conditions are allowed to vary widely. Though we report results related to differential abundance testing, we expect our findings to generalize to other types of analyses.

Simulation model

We designed a simulation framework to generate count data corresponding to two arbitrarily different conditions, denoted by superscripts in the equations below. First, for $p = 1, \ldots, P$ features in the first condition, a set of log mean abundances was drawn as

$$\theta_p^{(1)} \sim \mathrm{N}(m, S^2)$$

where hyperparameters m and S tune the mean and standard deviation of baseline log abundances. A correlation matrix was drawn as

$$\Omega \sim \text{Inverse-Wishart}(n, Q)$$

where scale matrix Q was supplied as either the identity matrix (for a minority of simulations) or a dense correlation matrix with net positive elements. The matrix Ω is subsequently re-scaled to a correlation matrix and used to draw correlated feature perturbations in a second condition as

$$\theta_p^{(2)} \sim \text{MVN}(\theta_p^{(1)}, \Omega \cdot a)$$

where the hyperparameter a exists in order to tune the overall scale of the correlated log perturbations. Mean abundances on the scale of sequence counts for each condition are calculated as

$$\gamma_p^{(1)} = \exp(\theta_p^{(1)}), \quad \gamma_p^{(2)} = \exp(\theta_p^{(2)})$$

A desired proportion of differentially abundant features c is obtained as follows: features are selected as differentially abundant with probability c. For those selected features only, the perturbed $\gamma_p^{(2)}$ serves as the mean abundance in the second condition; for all other features, the mean abundance in both the first and second conditions is given by $\gamma_p^{(1)}$. Let these new vectors be $\mu_p^{(1)}, \mu_p^{(2)}$. These represent mean the abundances of P features in two conditions, some of which differ across conditions, others of which are identical. Replicates $i = 1, \ldots, 10$ are then generated for each condition as follows. A fixed dispersion parameter for absolute counts is defined as $d_{\text{abs}} = 1000$ and those counts are drawn as

$$y_{i,p}^{(1)} \sim \text{NegBinom}(\mu_p^{(1)} \cdot \delta, 1000)$$

where

$$\delta \sim \max(0.1, \text{N}(1, g))$$

(Note that the dispersion parameter has been chosen such that the resulting counts are only barely overdispersed with respect to a Poisson.) The purpose of the truncated, per-sample multiplier δ is to re-scale all abundances within a given sample by some factor of approximately 1 but by increasing the scale of hyperparameter g, increasing replicate noise can be added within a condition. This process is repeated for the second condition to give a set of absolute counts $y_p^{(2)}$.

A new average observed total abundance (or library size) is randomly drawn as

$$u \sim \text{Unif}(5000, 2 \times 10^6)$$

and realized library sizes for each of the samples are then obtained by sampling

$$w_i^{(1)} \sim \text{NegBinom}(u, 100), \quad w_i^{(2)} \sim \text{NegBinom}(u, 100)$$

Finally, observed abundances z are generated through a multinomial resampling procedure similar to that of [10, 20, 23], using these new library size analogs. Where i and k index different samples prior to resampling, i' indexes the sample i after resampling, and total counts for sample i prior to resampling are given by $n_i = \sum y_i$, we have

$$z_{i'} \sim \mathrm{Mult}(\pi_{i'} = y_i/n_i, n_{i'} = n_k)$$

where superscripts have been suppressed as this procedure is identical across simulated "conditions." The resulting P-length vector of counts for a given sample contains relative but not absolute abundance information. These vectors are collapsed into a $P \times 20$ count matrix containing 10 replicate samples for each of two simulated conditions. In order to evaluate the discrepancy of differential abundance calling on observed versus absolute counts, we apply differential abundance methods to count matrices Z and Y respectively and score the differences.

Breadth of simulations

In order to generate simulations with a wide variety of characteristics, we swept in a grid over all our hyperparameters. Feature number P was stepped through values 100, 1000, and 5000. A maximum feature number of 5000 was chosen as txhese simulations were computationally intensive and major trends had become apparent at that scale. The degree of feature correlation was encoded in five realizations of scale matrices Q, encoding fully independent features at one extreme and 50% strongly positively correlated features at the other extreme. Log mean abundance (m) and the log variance (S) were independently incremented through low to high values. Likewise, the average log perturbation size (a) was swept from low to high in five steps, as a proportion of log mean abundance.

Replicate noise g varied from low to high in three steps. And finally, the proportion of differentially abundant features ranged across 20%, 30%, 50%, 70%, and 90%. Note that because many "perturbations" were very small, detectable differential abundance was generally only a fraction of the parameterized amount and most data sets contain a minority of differentially abundant features. Overall this 5625 simulated data sets were generated with almost continuous variation characteristics of interest.

We suggest that ranges of these parameter settings approximately represent different data modalities. We term the Microbial setting that with low to moderate feature number ($P \leq 1000$) and largest average perturbation, in accordance with a belief that bacterial communities are often simple (in terms of sequence variants with more than negligible abundance) and that they are highly variable even at short time scales

[36].

We designate the Bulk transcriptomic setting as that with the largest feature number ($P = 5000$) and having a lower average perturbation, the rationale being that transcriptomes sampled in aggregate over many cells are complex but largely stable compositions. Similarly, we define the intermediate Cell transcriptomic setting, approximately representative of single-cell RNA-seq data, to comprise simulations with moderate to large feature numbers ($P \geq 1000$) and moderate perturbation sizes. These categories are intended as rough outlines and we note that within these settings the realized data varies in terms of 1) degree of feature correlation, 2) overall abundance, 3) (un)evenness of composition, and 4) within-condition variation in totals.

Calling differential abundance

Three differential abundance calling methods were used in this study; each relies upon the use of a reference quantity to renomalize sample total abundances. For ALDEx2 this reference quantity is a trimmed version of per-sample mean log abundance. DESeq2 rescales samples using a sophisticated Bayesian model. Implicit in DESeq2's procedure is an assumption that a large set of stable features exists against which the observed changes in other features can be adjusted. scran's procedure follows a similar rationale, but employs a local rescaling, within clusters of like samples.

For simplicity, we omit from consideration models which lack a rescaling. We also omit zero-inflation models. Although these are popular in single-cell mRNA sequencing data, the debate continues about whether these models are appropriate for these data [37, 38].

Finally, differential abundance calls made on observed counts must be evaluated against a reference. At least two such references are obvious: calls made by the same method on absolute abundances or calls made by an independent method (a pseudo-gold standard or ”oracle”) on absolute abundances. While the second scenario (method vs. oracle), allows us to evaluate the performance of all methods relative to common standard, discrepancy in calls made under these conditions will be at least partially driven by differences in noise modeling and sensitivity across methods, complicating their interpretation of results. In the alternative, ”model vs. self” scenario, discrepancy in calls should be largely driven by differences in the count data itself. In general, we choose to report performance outcomes in this way (model vs. self) but we include a summary of performance in the model vs. oracle scenario in the Supplement and show that results are similar across these settings. Details of the simulation and testing procedure and an overview of simulated data sets are also provided there.

Accuracy (in terms of sensitivity and specificity) was calculated between differential abundance calls made on absolute abundances and observed abundances. Three methods were evaluated: ALDEx2, DESeq2, and scran. ALDEx2 was called using the aldex() from the associated R package [23] using the interquartile logratio reference. DESeq2 was called using the Seurat wrapper (FindMarkers(test.use = ...)) [24, 39]. scran was called using the method described in the vignette associated with its Bioconductor package [25]. Unadjusted p-values were collected from all methods and multiple test correction applied via p.adjust in R using Benjamini-Hochberg method.

Predictive modeling

In total, we trained 6 random forest models over 57 summary features, for each combination differential abundance calling method (ALDEx2, DESeq2, or scran) and accuracy measure (sensitivity or specificity). All such predictive models were fit with the randomForest package in R [40]. A random forest is an ensemble of decision trees and this tree-based approach was chosen because, while extensive, our simulations were not exhaustive. We anticipated that learning sets of decision rules might generalize well to unseen conditions, in particular feature numbers larger than those we explored in simulation. Feature importance was measured as "gain," or the relative increase in predictive accuracy achieved by the inclusion in the model of a given feature, as computed by the caret package in R [41].

We show in Supplemental Figure S5 that both simulated and real data sets exhibit comparable variation in terms of these features. Predictive models built from these summary features attempted to estimate sensitivity and specificity values explicitly. Details on these models are given in the Methods. All models were trained on 80% of the simulated data and their predictive accuracy was assessed on the reserved 20%.

Acknowledgements

The authors would like to acknowledge conversations with Justin Silverman, Lawrence David, and Tim Reddy for useful comments. The authors would like to acknowledge partial funding from HFSP RGP005, NSF DMS 17-13012, NSF BCS 1552848, NSF DBI 1661386, NSF IIS 15-46331, NSF DMS 16-13261, as well as high-performance computing partially supported by grant 2016-IDG-1013 from the North Carolina Biotechnology Center and a Forge Fellowship through Duke University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of any of the funders.

Code availability

All R code related to this study is available on Github at https://github.com/kimberlyroche/codaDE.

References

[1] Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017;8:57.

[2] Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. Brief Bioinform. 2018;19(5):776–792.

[3] Vieth B, Parekh S, Ziegenhain C, Enard W, Hellmann I. A systematic evaluation of single cell RNA-seq analysis pipelines. Nat Commun. 2019;10(1):4667.

[4] Coate JE, Doyle JJ. Quantifying whole transcriptome size, a prerequisite for understanding transcriptome evolution across species: an example from a plant allopolyploid. Genome Biol Evol. 2010;2:534–546.

[5] Coate JE, Doyle JJ. Variation in transcriptome size: are we getting the message? Chromosoma. 2015;124(1):27–43.

[6] Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell. 2012;151(1):68–79.

[7] Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012;151(1):56–67.

[8] Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting global gene expression analysis. Cell. 2012;151(3):476–482.

[9] Athanasiadou R, Neymotin B, Brandt N, Wang W, Christiaen L, Gresham D, et al. A complete statistical model for calibration of RNA-seq counts using external spike-ins and maximum likelihood theory. PLoS Comput Biol. 2019;15(3):e1006794.

[10] McGee WA, Pimentel H, Pachter L, Wu JY. Compositional data analysis is necessary for simulating and analyzing RNA-Seq data; 2019.

[11] Wu AR, Neff NF, Kalisky T, Dalerba P, Treutlein B, Rothenberg ME, et al. Quantitative assessment of single-cell RNA-sequencing methods. Nat Methods. 2014;11(1):41–46.

[12] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133–145.

[13] Lun ATL, Calero-Nieto FJ, Haim-Vilmovsky L, Göttgens B, Marioni JC. Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data. Genome Res. 2017;27(11):1795–1806.

[14] Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14(4):381–387.

[15] Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. Mol Cell. 2017;65(4):631–643.e4.

[16] Tung PY, Blischak JD, Hsiao CJ, Knowles DA, Burnett JE, Pritchard JK, et al. Batch effects and the effective design of single-cell gene expression studies. Sci Rep. 2017;7(1):1–15.

[17] Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):896–902.

[18] Quinn TP, Erb I, Richardson MF, Crowley TM. Understanding sequencing data as compositions: an outlook and review. Bioinformatics. 2018;34(16):2870–2878.

[19] Hawinkel S, Mattiello F, Bijnens L, Thas O. A broken promise: microbiome differential abundance methods do not control the false discovery rate. Brief Bioinform. 2019;20(1):210–221.

[20] Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. Genome Biol. 2020;21(1):191.

[21] 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):507–511.

[22] Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. Isometric Logratio Transformations for Compositional Data Analysis. Math Geol. 2003;35(3):279–300.

[23] Fernandes AD, Macklaim JM, Linn TG, Reid G, Gloor GB. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. PLoS One. 2013;8(7):e67019.

[24] Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

[25] Lun ATL, McCarthy DJ, Marioni JC. A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor. F1000Res. 2016;5:2122.

[26] Vieira-Silva S, Sabino J, Valles-Colomer M, Falony G, Kathagen G, Caenepeel C, et al. Quantitative microbiome profiling disentangles inflammation- and bile duct obstruction-associated microbiota alterations across PSC/IBD diagnoses. Nat Microbiol. 2019;4(11):1826–1831.

[27] Barlow JT, Bogatyrev SR, Ismagilov RF. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. Nat Commun. 2020;11(1):2590.

[28] Song SG, Kim S, Koh J, Yim J, Han B, Kim YA, et al. Comparative analysis of the tumor immune-microenvironment of primary and brain metastases of non-small-cell lung cancer reveals organ-specific and EGFR mutation-dependent unique immune landscape. Cancer Immunol Immunother. 2021;70(7):2035–2048.

[29] Monaco G, Lee B, Xu W, Mustafah S, Hwang YY, Carré C, et al. RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. Cell Rep. 2019;26(6):1627–1640.e7.

[30] Hagai T, Chen X, Miragaia RJ, Rostom R, Gomes T, Kunowska N, et al. Gene expression variability across cells and species shapes innate immunity. Nature. 2018;563(7730):197–202.

[31] Owens NDL, Blitz IL, Lane MA, Patrushev I, Overton JD, Gilchrist MJ, et al. Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development. Cell Rep. 2016;14(3):632–647.

[32] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. Cell. 2015;161(5):1187–1201.

[33] Yu Y, Fuscoe JC, Zhao C, Guo C, Jia M, Qing T, et al. A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. Nat Commun. 2014;5(1):1–11.

[34] Padovan-Merhar O, Nair GP, Biaesch AG, Mayer A, Scarfone S, Foley SW, et al. Single mammalian cells compensate for differences in cellular volume and DNA copy number through independent global transcriptional mechanisms. Mol Cell. 2015;58(2):339–352.

[35] Lloréns-Rico V, Vieira-Silva S, Gonçalves PJ, Falony G, Raes J. Benchmarking microbiome transformations favors experimental quantitative approaches to address compositionality and sampling depth biases. Nat Commun. 2021;12(1):3562.

[36] Johnson AJ, Vangay P, Al-Ghalith GA, Hillmann BM, Ward TL, Shields-Cutler RR, et al. Daily Sampling Reveals Personalized Diet-Microbiome Associations in Humans. Cell Host Microbe. 2019;25(6):789–802.e5.

[37] Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. Comput Struct Biotechnol J. 2020;18:2789–2798.

[38] Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. Nat Genet. 2021;53(6):770–777.

[39] Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):296.

[40] Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.

[41] Kuhn M. Building predictive models in R using the caret package. J Stat Softw. 2008;.

[42] Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17:77.

Figures and Tables

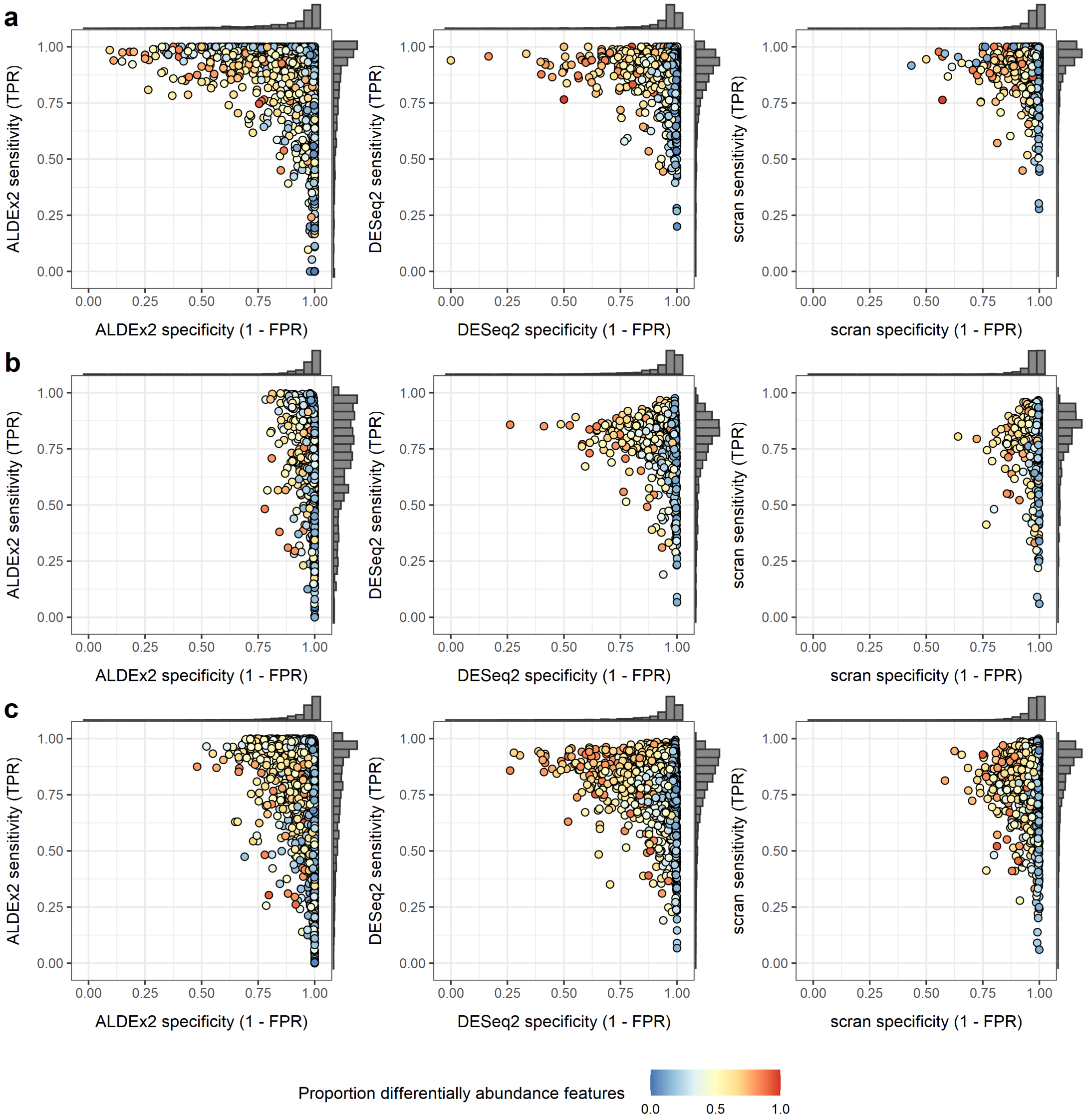

Figure 1: Sensitivity and specificity for three differential abundance calling methods in three experimental settings. Data sets are labeled by proportion of features with at least a 50% increase or decrease in abundance between conditions. Results shown are for all methods in the **a)** Microbial, **b)** Bulk transcriptomic, and **c)** Cell transcriptomic settings.

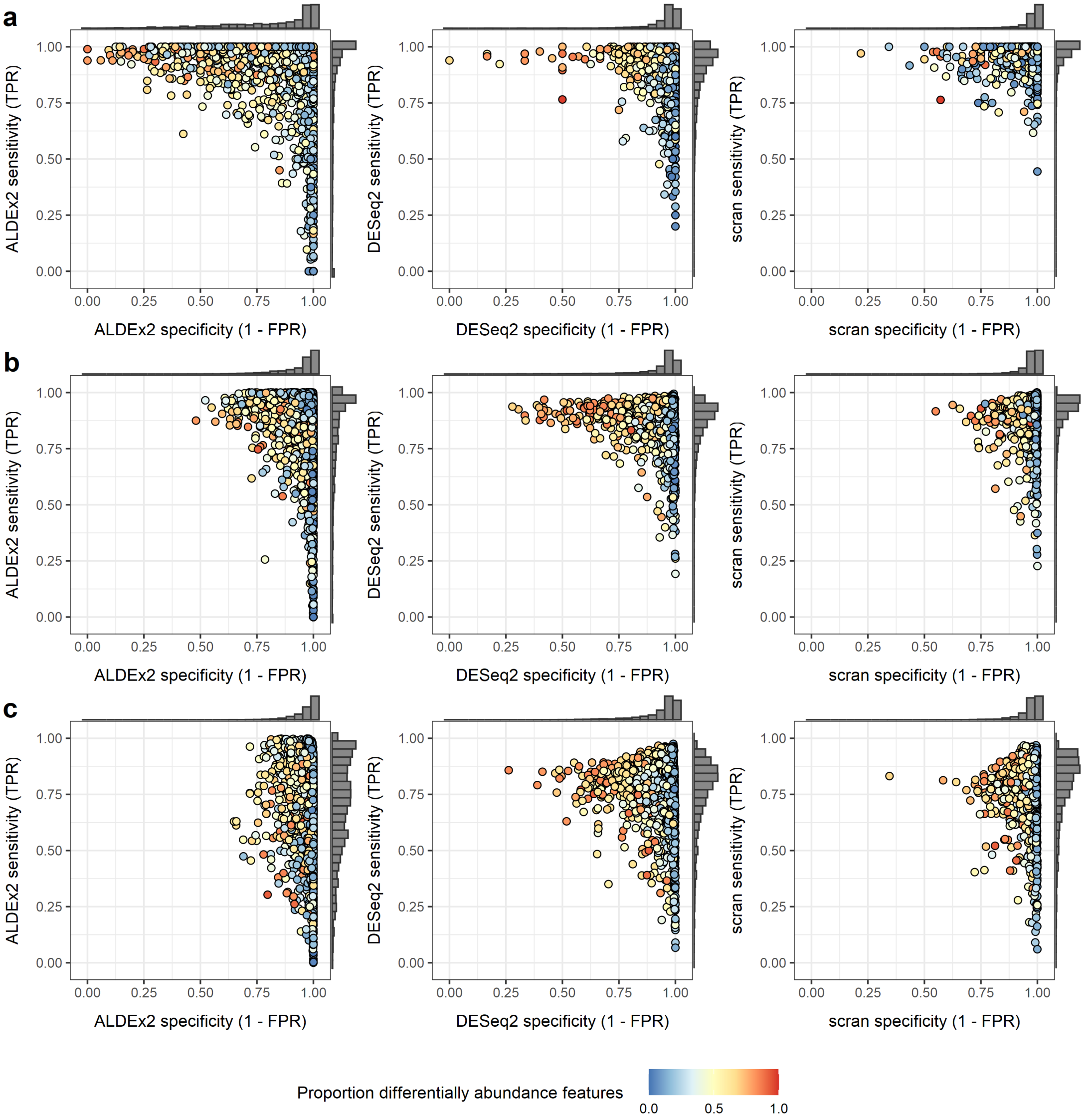

Figure 2: Sensitivity and specificity for three differential abundance calling methods in three compositional complexity settings. Data sets are labeled by proportion of features with at least a 50% increase or decrease in abundance between conditions. Results shown are for all methods over **a)** 100-feature simulations, **b)** 1000-feature simulations, and **c)** 5000-feature simulations.

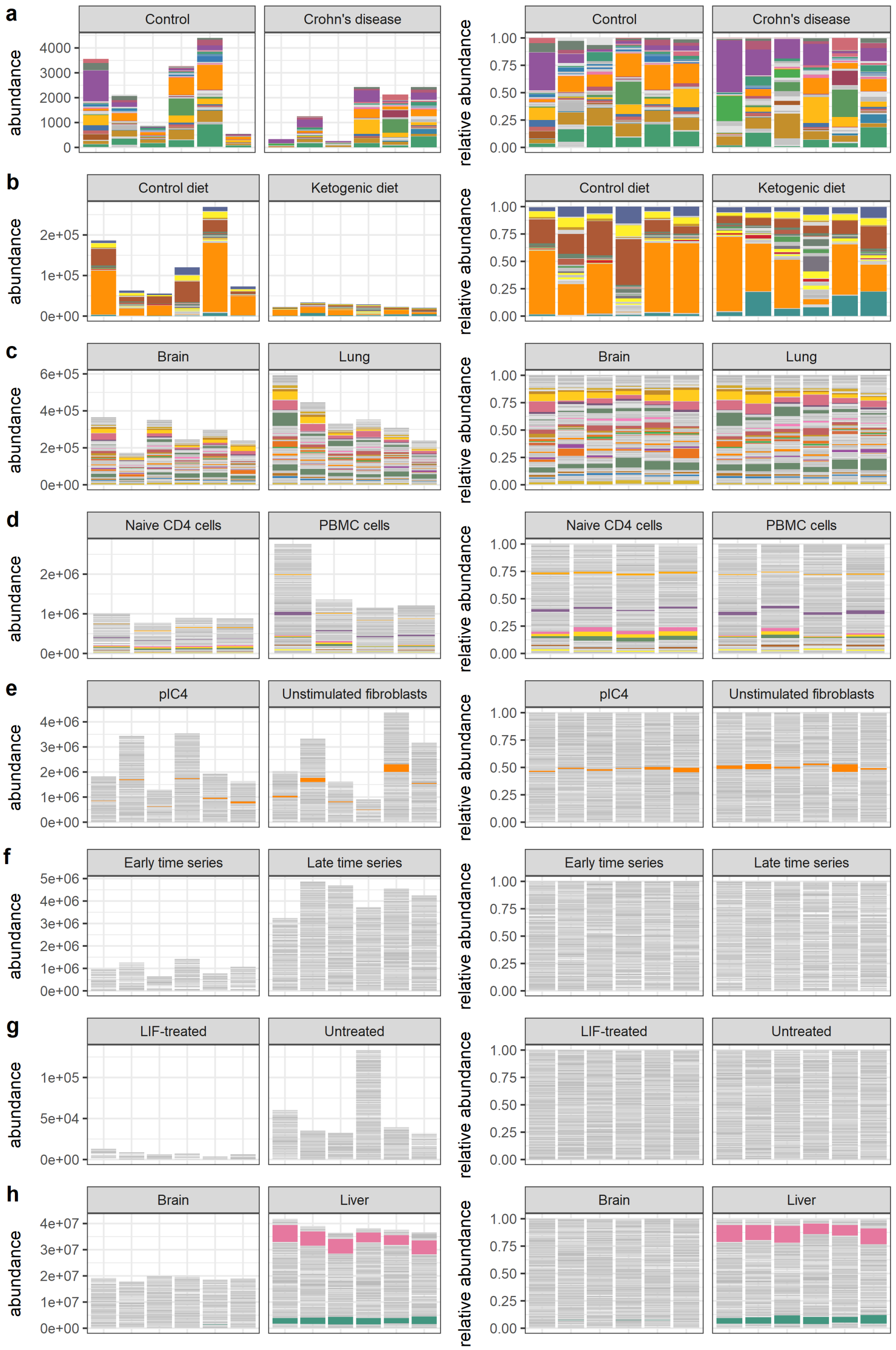

Figure 3: Visual summaries of eight real sequence count data sets: metagenomics data from **a)** Vieira-Silva et al. and **b)** Barlow et al.; nCounter array data from **c)** Song et al.; bulk RNA-seq data from **d)** Monaco et al. and **e)** Hagai et al.; single-cell RNA-seq data from **f)** Owens et al., **g)** Klein et al., and **h)** Yu et al. Left panels show absolute abundances for subsets of samples across two experimental conditions. Right panels show relative abundances for the same samples and conditions. Features (genes or bacterial sequence variants) with at least 1% relative abundance across all samples are colored; all other features are gray.

Source	Description	No. sequence variants	Approx. fold change	Approx. percent differential features
Vieira-Silva et al. (2019)	16S metagenomics from human gut samples of control and Crohn's disease patients	70	2.5	24
Barlow et al. (2020)	16S metagenomics from ketogenic diet and control mice	78	3.4	19
Song et al. (2021)	nCounter array of human primary lung cancer vs. brain metastases	773	1.4	45
Monaco et al. (2019)	immune cell profiling in human humans via bulk RNA-seq	17,261	3.4	28
Hagai et al. (2018)	bulk RNA sequencing of both unstimulated and mock-viral infected mouse fibroblasts	13,937	1.5	39
Owens et al. (2016)	single cell sequencing of zebrafish embryos; early vs. late time course samples drawn	40,476	3.7	23
Klein et al. (2015)	single cell RNA-sequencing of normally developing and leukemia inhibitory factor-treated mouse ESCs	2,928	2.9	12
Yu et al. (2014)	single cell expression profiling of rat brain and liver tissue	26,871	2.0	36

Table 1: Real 16S metabarcoding, bulk RNA-seq, and single cell RNA-seq data sets corresponding to the abundances shown in Figure 1. The approximate proportion of differentially abundant features is estimated by simple thresholding: reconstructed "absolute" features with an average increase or decrease between conditions of 50% abundance are considered differential here.

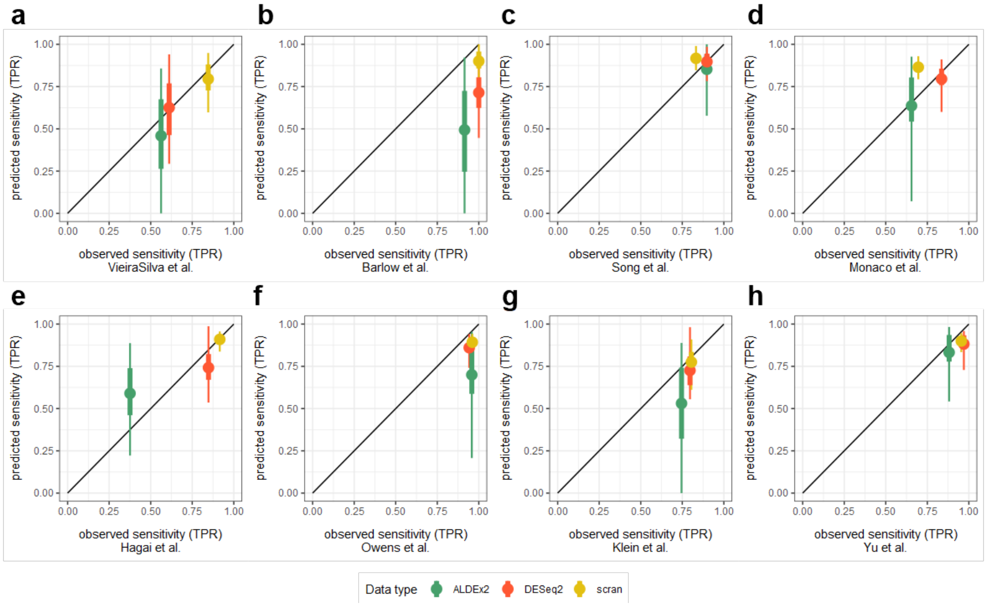

Figure 4: Predicted and observed sensitivity on eight real data sets for three differential abundance testing methods. **a)** Vieira-Silva et al., **b)** Barlow et al. **c)** Song et al., **d)** Monaco et al., **e)** Hagai et al., **f)** Owens et al., **g)** Klein et al., and **h)** Yu et al. Prediction intervals are enabled by the ensemble of decision trees in the "forest" and correspond to 50% (thicker line) and 90% (thinner line) consensus intervals over the predictions of individual trees in the forest.

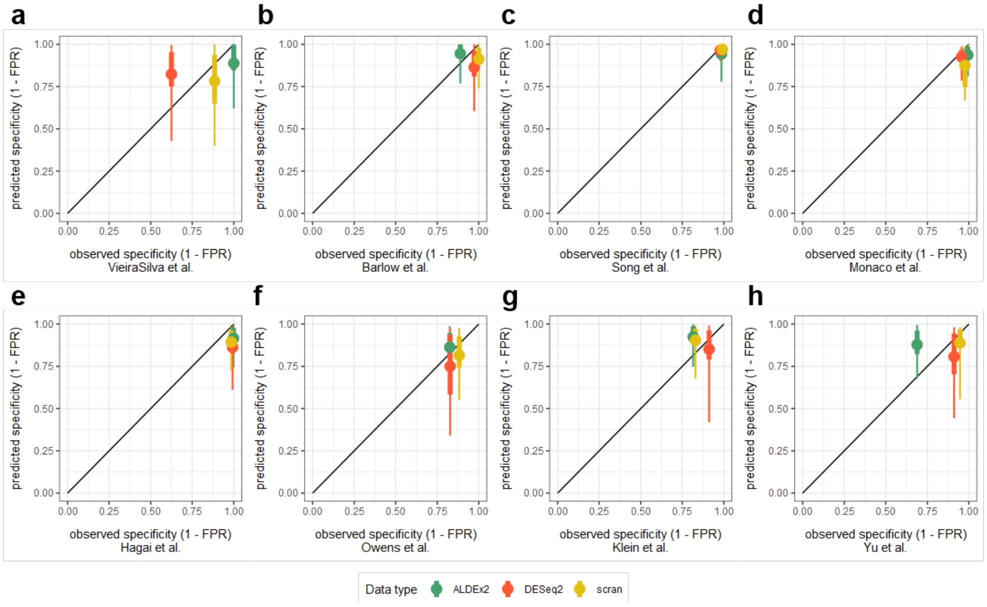

Figure 5: Predicted and observed specificity on eight real data sets for three differential abundance testing methods. **a)** Vieira-Silva et al., **b)** Barlow et al. **c)** Song et al., **d)** Monaco et al., **e)** Hagai et al., **f)** Owens et al., **g)** Klein et al., and **h)** Yu et al. Intervals are as described in Figure 4.

Supplement

Characteristics of the simulated data

The properties of simulated data sets in terms of percent differentially abundant features, change in abundance across conditions, and percent zero counts are shown in Supplemental Figure S1. Most simulations featured a minority of differentially abundant features and the distribution of the proportion of differentially abundant features was very similar across feature number settings (100, 1000, and 5000 features). Fold change in total abundance between conditions was similar across settings as well. The percent zero counts in our simulations increased as the number of features increased. This is because, on average, simulations with a larger number of features had larger overall absolute total abundances and were more likely to be downsampled when ”observed” relative abundances were drawn, yielding dropouts.

Renormalization-based methods mostly concur with a simple negative binomial model of differential abundance

Differential abundance was simulated in a continuous way: for a given feature, count data were drawn from a negative binomial model in accordance with a per-condition mean abundance for that feature. Per-feature differences in mean abundance across conditions were often quite small.

We point out in the main text that other alternative references for “true” differential abundance in absolute count data could exist, including a negative binomial generalized linear model (NBGLM). We show in Supplemental Figure S2 that the number of differentially abundant features in absolute abundance data detected by a NBGLM generally accords with the number of differentially abundant features identified by ALDEx2, DESeq2, and scran but that renormalization-based methods called a larger number of features differential overall.

Using NBGLM calls as a pseudo-ground truth increases the rate of false positives

We also evaluated the sensitivity and specificity of differential abundance calls made by ALDEx2, DESeq2, and scran on observed abundances against differential abundance calls made by a negative binomial model on absolute abundances. We did this in order to compare the relative performance of these methods against a common reference. In all cases, false positive rates were higher when the NBGLM calls were used as a baseline. (Compare results in Supplemental Figure S3 to Figure 1 in the main text.) This derives from the lesser sensitivity of the NBGLM model. Differentially abundant features which are too "noisy" to be significantly different in the NB model are often significantly different after adjustment by ALDEx2, DESeq2, and scran.

Results on simulations with a minority of differentially abundant features

Results on a subset of simulations having a minority of differentially abundant features (as evaluated by the NBGLM) and an average log fold change between conditions of at least 2.5 are shown in Supplemental Figure S4. We highlight these as a set of extreme simulations - with respect to the scale of absolute change between conditions - which should best adhere to the assumptions of the methods evaluated. In each of these simulated data set a (proportionally) large, reasonably stable subset of features exists. Median specificity is high under for all methods, across all feature number settings, at 2.5%, though especially for ALDEx2 in the low-feature number setting, a large number of data sets exhibit high false positive rates.

Estimating absolute and relative abundances in real data sets

All real data sets were downloaded from the public repositories indicated in the published article, except where noted. For the Vieira-Silva *et al.* [26] data set, absolute abundances estimated by quantitative microbiome profiling were available from the authors' website. These data were rescaled such that the lowest non-zero count was one. Relative abundances for the same data were simulated by shuffling the observed library sizes across all samples. We performed this shuffling of total abundances in order to guarantee any correlation between observed and reconstructed absolute abundances would be eliminated, giving a "worst case" scenario: relative abundances with no information about changes in scale either within or between conditions.

Absolute abundances in the Barlow *et al.* [27] study were estimated by those authors via digital droplet PCR and relative abundances were provided in the form of proportions. These relative abundances were scaled up such that the minimum non-zero abundance was one.

nCounter array expression profiles from the Song et al. [28] study were obtained and we noted that the abundances of positive controls (in the form of ERCC spike-ins) correlated well with observed total abundances. For that reason, we treated the observed data as absolute counts and derived relative abundances from these by shuffling the observed total abundances across all samples and resampling, as in our simulation method.

The Monaco et al. [29] data were published in transcripts per million (TPM) format. Absolute abundances were derived by estimating a scaling factor from ERCC spike-ins present in this data. Estimation of the scaling factor from a set of reference quantities (e.g. spike-in abundances) simply involved computing the mean of all references in each sample to give a per-sample multiplier, then scaling this multiplier to have a mean of one. The original, TPM-format data were then rescaled on a sample-to-sample basis using these centered scaling factors. The TPM-format data were used as the relative (observed) counts. The Hagai et al. [30] bulk RNA-seq data were treated using methods already described above. A per-sample scaling factor was computed from spike-in sequences and used to rescale samples to give approximate absolute abundances. Relative abundances were derived using the technique outlined for the Song et al. data - a shuffling of observed total abundances and subsequent resampling. The Owens et al. [31] data were treated in exactly the same fashion, with absolute counts given by an ERCC spike-in rescaling and relative counts derived from a resampling procedure. Distinct "conditions" were manufactured from the data by selecting early and late samples from this time course data as differential conditions A and B, respectively.

The expression of the gene *Gapdh* was used as a rescaling factor for the Klein et al. [32] single-cell data. This per-sample factor was estimated by the procedure described above and used to rescale samples, giving absolute abundances. Relative counts were resampled by the previously outlined procedure.

Finally, brain and liver tissue samples were obtained from the Yu et al. [33] expression atlas. The existing log total abundances correlated well with log ERCC spike-in abundances and those unaltered data were used as absolute abundances. Relative abundances were derived by resampling.

In general, we removed features with a mean abundance of less than a single count from each of the data sets before calling differential abundance.

Simulated data "resembles" real data with respect to features of interest

We visualize simulated and real data together in Supplemental Figure S5 in the space of the summary features used for the prediction of sensitivity and specificity outcomes. Simulated and real data largely inhabit the same space of variation with respect to these features.

Variable importance in predictive models

In Supplemental Tables S1 & S2, we show the top several most important features (as scored by gain in accuracy upon feature inclusion) for each of the six predictive models over sensitivity and specificity for each of ALDEx2, DESeq2, and scran.

In all models over sensitivity, features summarizing the prevalence of low-count elements in the composition were most informative. For ALDEx2 and DESeq2, an estimate of the correlation of centered logratio features - summarizing shared change relative to the mean - was the single most informative feature. Specificity was more difficult to predict and models leveraged a variety of features for this. The most informative features were those indicating the proportion of relative abundance, log abundances, or centered logratio abundances undergoing apparent change across conditions.

Performance of regression models for prediction of sensitivities and specificities on simulated data

Results from the prediction of sensitivity and specificity are shown in Supplemental Table S3 for simulated data sets held out from model training. Prediction of sensitivity was broadly accurate for all methods. Specificity proved more challenging: DESeq2's specificities are reasonably predictable from the characteristics of observed data alone but specificity prediction for ALDEx2 and scran only achieved an R^2 of around 0.5.

Library sizes in single-cell data can be informative

We note in the main text that in single cell RNA-seq data observed total abundance (unnormalized library size) often correlates well with proxies of transcriptome size or total mRNA. For example, using *Gapdh* expression as a proxy for total mRNA, we see good correspondence between the abundance of this putatively stable gene product and observed library size (labeled "total abundance" in Supplemental Figure S6) in the Klein et al. [32] data. The R^2 for *Gapdh* on total abundance in both the stimulated and unstimulated cells is greater than 0.5. This positive association is also found between mean spike-in abundance and library size in the CEL-seq2 data generated by Hashimshony et al. [42] ($R^2 > 0.5$ across phases of the cell cycle in this experiment).

Supplemental Figures and Tables

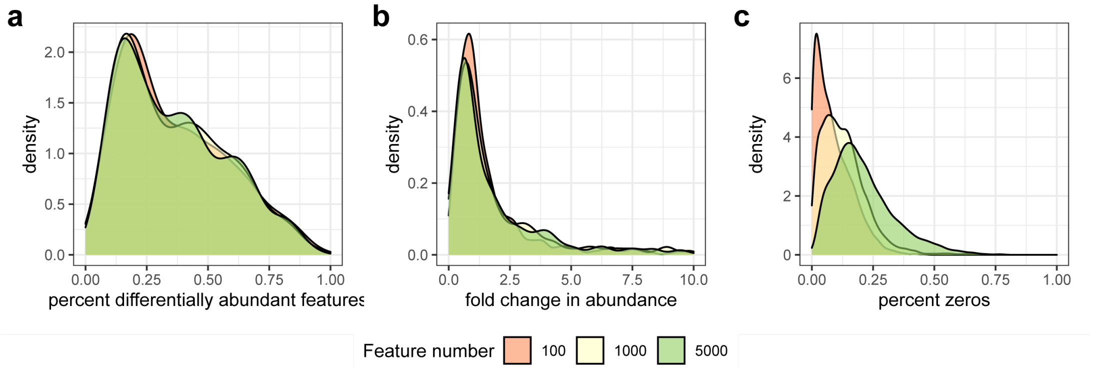

Figure S1: Distributions associated with three characteristics of the 5625 simulated data sets: **a)** percent differentially abundant features, **b)** fold change in total abundance across conditions, and **c)** percent zeros.

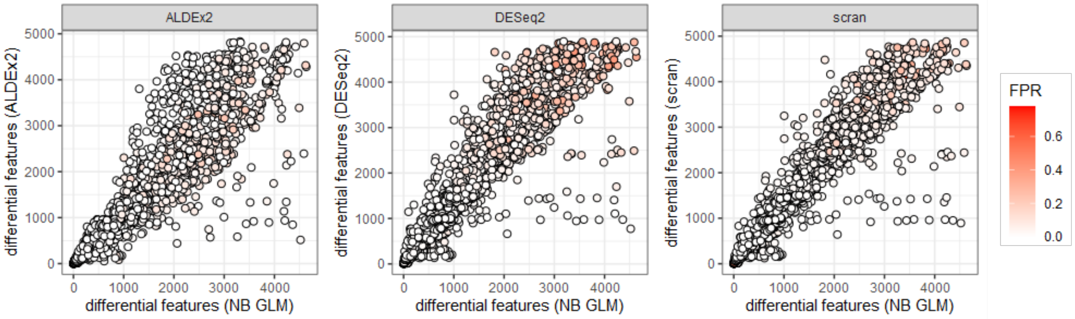

Figure S2: Differential abundance calls on absolute abundances made by ALDEx2, DESeq2, and scran are similar to those made by a negative binomial generalized linear model, although these methods are less conservative than the NBGLM. R^2 estimates for the number of differential features called by ALDEx2, DESeq2, and scran versus the NBGLM are 0.88, 0.91, and 0.92 respectively.

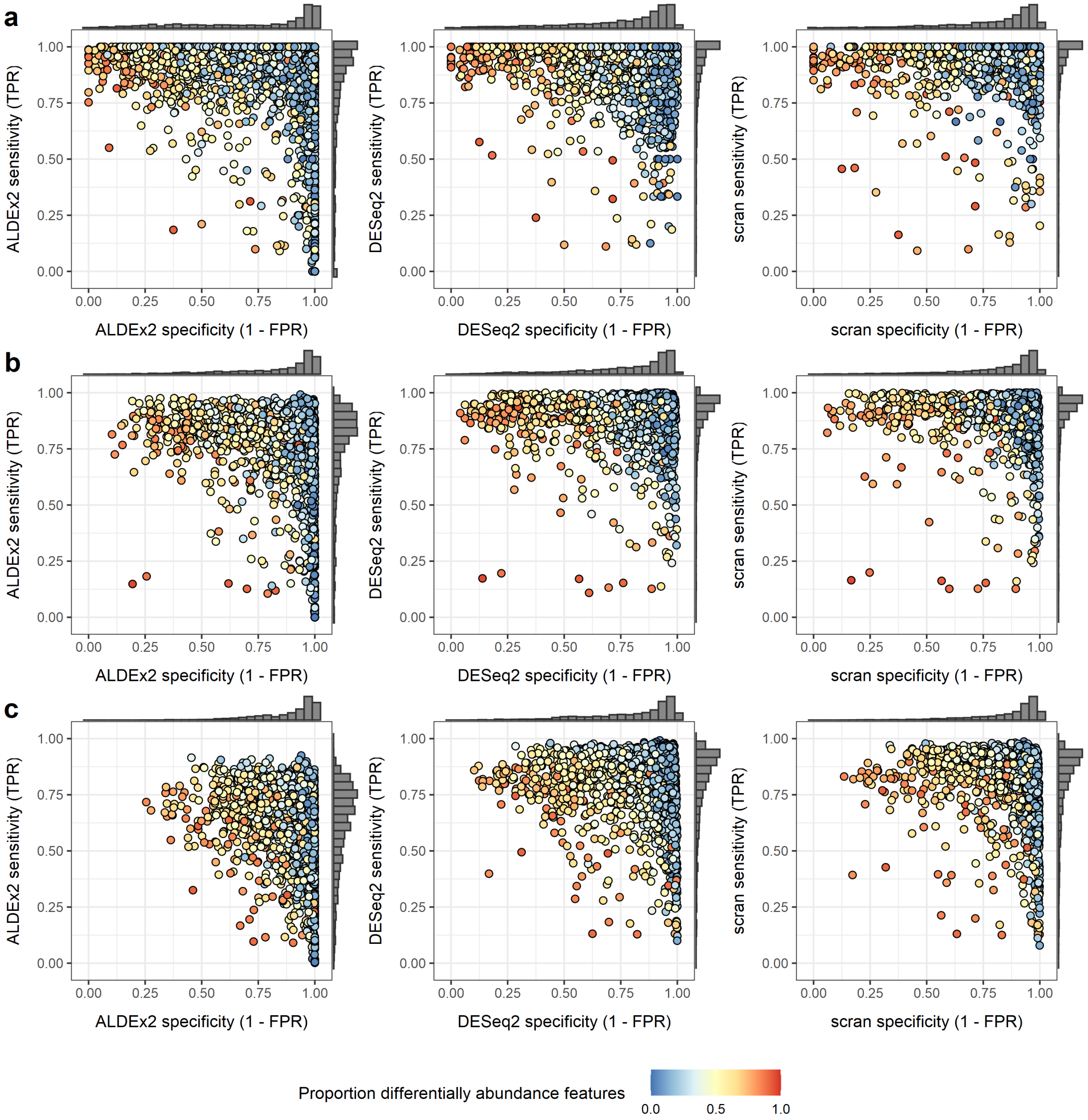

Figure S3: Sensitivity and specificity results for differential abundance calls made on observed abundances versus calls made on absolute abundances. Here, calls made on observed abundances using ALDEx2, DESeq2, or scran are compared to calls made on absolute abundances using a negative binomial GLM. Results shown are for **a)** 100-feature simulations, **b)** 1000-feature simulations, and **c)** 5000-feature simulations. Data sets having the lowest specificies are generally those with the largest proportion of differentially abundant features.

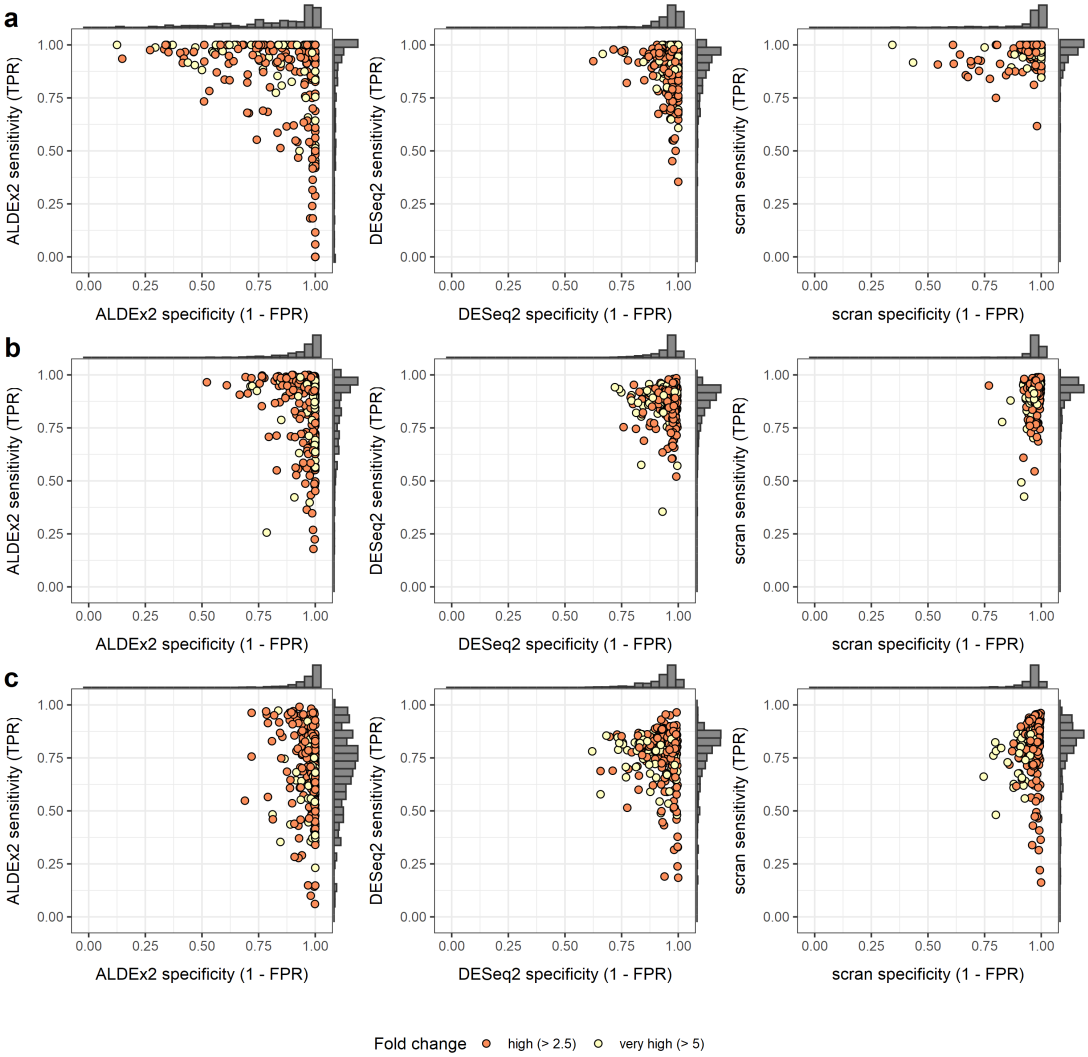

Figure S4: Sensitivity and specificity results for differential abundance calls made on observed abundances versus calls made on absolute abundances. Only simulations with several fold change in abundance between conditions and having less than 50% of features differentially abundant are shown. Data sets are labeled for the scale of fold change realized between simulated conditions. Results shown are for **a)** 100-feature simulations, **b)** 1000-feature simulations, and **c)** 5000-feature simulations.

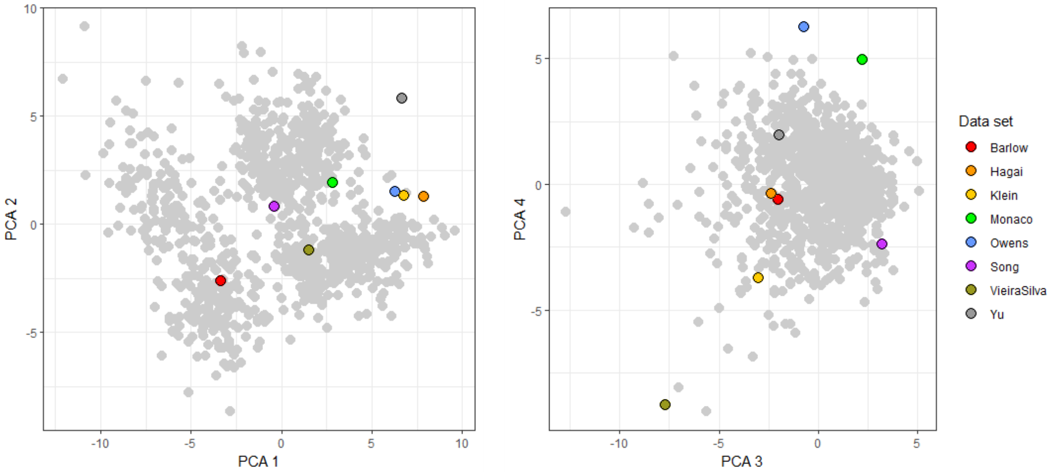

Figure S5: A subset of simulated data sets (gray) and real data sets (colored) are plotted in the top 4 principle components associated with their predictive features.

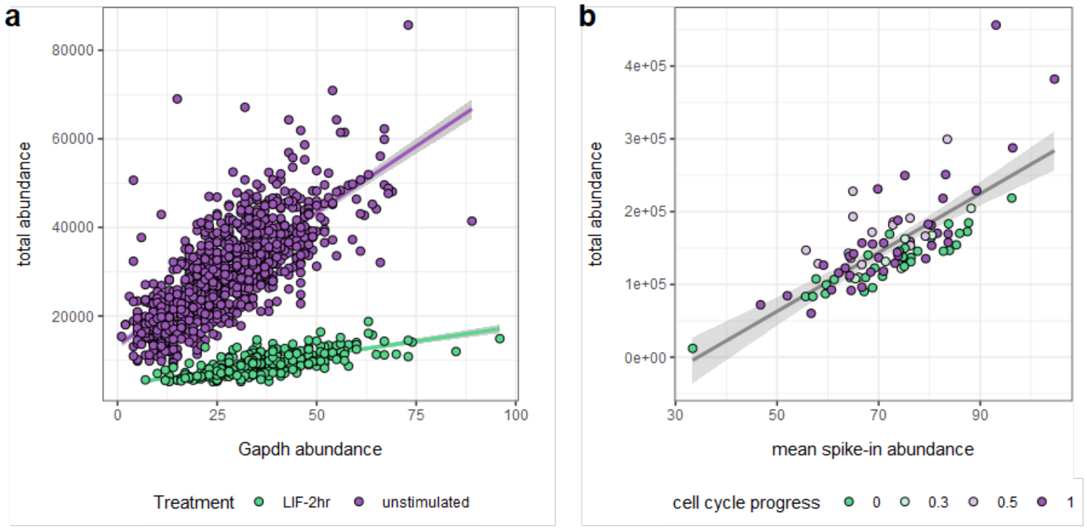

Figure S6: Correspondence between observed total abundances and proxies of absolute abundance: **a)** Gapdh abundance positively correlates with observed total abundance in each treatment group in the Klein et al. data; **b)** mean ERCC spike-in abundance positively correlates with observed total abundance in the Hashimshony et al. data set.

Method	Predictive feature	Feature rank
ALDEx2	median correlation of CLR features	1
ALDEx2	skew correlation of CLR features	2
ALDEx2	percent features $= 1$ in B	3
ALDEx2	percent features $= 0$ in B	4
ALDEx2	percent features ≤ 5 in B	5
DESeq2	median correlation of CLR features	1
DESeq2	percent features $= 0$ in B	2
DESeq2	percent features ≤ 5 in B	3
DESeq2	percent features $= 1$ in A	4
DESeq2	percent features $= 1$ in B	5
scran	percent features $= 1$ in B	1
scran	percent features $= 1$ in A	2
scran	percent features ≤ 5 in B	3
scran	percent features ≤ 5 in A	4
scran	percent features $= 0$ in B	5

Table S1: Predictive features and their importance rank (1 = most important) in the prediction of sensitivity.

Method	Predictive feature	Feature rank
ALDEx2	percent features ≤ 5 in B	1
ALDEx2	percent features with ≤ 0.5 FC in CLR	2
ALDEx2	SD change in log + PC counts	3
ALDEx2	SD change in relative abundance	4
ALDEx2	median correlation of CLR features	5
DESeq2	percent features with ≤ 0.5 FC in log + PC counts	1
DESeq2	percent features with ≤ 0.5 FC in CLR	2
DESeq2	SD change in log + PC counts	3
DESeq2	percent features with ≤ 2 FC in CLR	4
DESeq2	median change in log + PC counts	5
scran	percent features with ≤ 0.5 FC in log + PC counts	1
scran	percent features with ≤ 0.5 FC in CLR	2
scran	median change in log + PC counts	3
scran	SD change in log + PC counts	4
scran	median relative abundance in B	5

Table S2: Predictive features and their importance rank (1 = most important) in the prediction of specificity.

Method	Sensitivity R^2	Specificity R^2
ALDEx2	80%	51%
DESeq2	86%	74%
scran	92%	50%

Table S3: R^2 values for observed versus predicted values of sensitivity and specificity for six random forest models over three differential abundance calling methods.

Feature symbol	Description	Category
P	number of features	general
TOTALS_C_FC	absolute fold change in mean totals (A vs. B)	totals
TOTALS_C_D	absolute change in mean totals	totals
TOTALS_C_MAX_D	max delta in totals	totals
TOTALS_C_MED_D	median delta in totals	totals
TOTALS_C_SD_D	SD in totals	totals
CORR_RA_MED	median correlation of relative abundances	feature correlation
CORR_RA_SD	SD correlation of relative abundances	feature correlation
CORR_RA_SKEW	skew correlation of relative abundances	feature correlation
CORR_LOG_MED	median correlation of log + PC counts	feature correlation
CORR_LOG_SD	SD correlation of log + PC counts	feature correlation
CORR_LOG_SKEW	skew correlation of log + PC counts	feature correlation
CORR_CLR_MED	median correlation of CLR features	feature correlation
CORR_CLR_SD	SD correlation of CLR features	feature correlation
CORR_CLR_SKEW	skew correlation of CLR features	feature correlation
COMP_C_P0_A	percent features == 0 in A	composition
COMP_C_P0_B	percent features == 0 in B	composition
COMP_C_P1_A	percent features == 1 in A	composition
COMP_C_P1_B	percent features == 1 in B	composition
COMP_C_P5_A	percent features <= 5 in A	composition
COMP_C_P5_B	percent features <= 5 in B	composition
COMP_RA_P01_A	percent features < 0.1% relative abundance in A	composition
COMP_RA_P01_B	percent features < 0.1% relative abundance in B	composition
COMP_RA_P1_A	percent features < 1% relative abundance in A	composition
COMP_RA_P1_B	percent features < 1% relative abundance in B	composition
COMP_RA_P5_A	percent features < 5% relative abundance in A	composition
COMP_RA_P5_B	percent features < 5% relative abundance in B	composition
COMP_RA_MAX_A	max relative abundance in A	composition
COMP_RA_MED_A	median relative abundance in A	composition
COMP_RA_SD_A	SD relative abundance in A	composition
COMP_RA_SKEW_A	skew relative abundance in A	composition
COMP_RA_MAX_B	max relative abundance in B	composition
COMP_RA_MED_B	median relative abundance in B	composition

COMP_RA_SD_B	SD relative abundance in B	composition
COMP_RA_SKEW_B	skew relative abundance in B	composition
COMP_C_ENT_A	entropy in A	composition
COMP_C_ENT_B	entropy in B	composition
FW_RA_MAX_D	max change in relative abundance	feature-wise change
FW_RA_MED_D	median change in relative abundance	feature-wise change
FW_RA_SD_D	SD change in relative abundance	feature-wise change
FW_RA_PPOS_D	percent features with + change in relative abundances	feature-wise change
FW_RA_PNEG_D	percent features with - change in relative abundances	feature-wise change
FW_RA_PFC05_D	percent features with < 0.5 FC in relative abundance	feature-wise change
FW_RA_PFC1_D	percent features with < 1 FC in relative abundance	feature-wise change
FW_RA_PFC2_D	percent features with < 2 FC in relative abundance	feature-wise change
FW_LOG_MAX_D	max change in log + PC counts	feature-wise change
FW_LOG_MED_D	median change in log + PC counts	feature-wise change
FW_LOG_SD_D	SD change in log + PC counts	feature-wise change
FW_LOG_PPOS_D	percent features with + change in log + PC counts	feature-wise change
FW_LOG_PNEG_D	percent features with - change in log + PC counts	feature-wise change
FW_LOG_PFC05_D	percent features with < 0.5 FC in log + PC counts	feature-wise change
FW_LOG_PFC1_D	percent features with < 1 FC in log + PC counts	feature-wise change
FW_LOG_PFC2_D	percent features with < 2 FC in log + PC counts	feature-wise change
FW_CLR_MAX_D	max change in CLR	feature-wise change
FW_CLR_MED_D	median change in CLR	feature-wise change
FW_CLR_SD_D	SD change in CLR	feature-wise change
FW_CLR_PPOS_D	percent features with + change in CLR	feature-wise change
FW_CLR_PNEG_D	percent features with - change in CLR	feature-wise change
FW_CLR_PFC05_D	percent features with < 0.5 FC in CLR	feature-wise change
FW_CLR_PFC1_D	percent features with < 1 FC in CLR	feature-wise change

Table S4: R^2 values for observed versus predicted values of sensitivity and specificity for six random forest models over three differential abundance calling methods.