Analysis and correction of compositional bias in sparse sequencing count data

M. Senthil Kumar[1,2*], Eric V. Slud[3,4], Kwame Okrah[5], Stephanie C. Hicks[6,7], Sridhar Hannenhalli[2] and Héctor Corrada Bravo[2]

*Correspondence: smuthiah@umiacs.umd.edu
[2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA
Full list of author information is available at the end of the article

Abstract

Count data derived from high-throughput DNA sequencing is frequently used in quantitative molecular assays. Due to properties inherent to the sequencing process, unnormalized count data is compositional, measuring *relative* and not *absolute* abundances of the assayed features. This *compositional bias* confounds inference of absolute abundances. We demonstrate that existing techniques for estimating compositional bias fail with sparse metagenomic 16S count data and propose an empirical Bayes normalization approach to overcome this problem. In addition, we clarify the assumptions underlying frequently used scaling normalization methods in light of compositional bias, including scaling methods that were not designed directly to address it.

Keywords: compositional bias; normalization; empirical Bayes; data integration; count data; metagenomics; absolute abundance; scRNAseq; spike-in

Background

Sequencing technology has played a fundamental role in 21st century biology: the output data, in the form of sequencing reads of molecular features in a sample, are relatively inexpensive to produce [1, 2, 3, 4]. This, along with the immediate availability of effective, open source computational toolkits for downstream analysis [5, 6], has enabled biologists to utilize this technology in ingenious ways to probe various aspects of biological mechanisms and organization ranging from microscopic DNA binding events [7, 8] to large-scale oceanic microbial ecosystems [9, 10].

As noted previously in the literature [11, 12, 13, 14] (illustrated in **Fig. 1**), unnormalized counts obtained from a sequencer only reflect relative abundances of the features in a sample, and not their absolute internal concentrations. When a differential abundance analysis is performed on this data, fold changes of null features, those not differentially abundant in the *absolute* scale, are intimately tied

to those of features that are perturbed in their absolute abundances, making the former appear differentially abundant. We refer to this artifact as *compositional bias*. Such effects are observable in the count data from the large-scale Tara oceans metagenomics project [10], (**Fig. 2**), in which a few dominant taxa are attributable to global differences in the between-oceans fold-change distributions.

Correction for compositional bias can be achieved by re-scaling each sample's count data with its corresponding count of an internal control feature (or "spike-in", **Fig. 1B**). In the absence of such control features, effective correction for compositional bias can still be hoped for, as it can be shown that this correction amounts to resolving a linear technical bias [13]. This fact allows one to exploit several widely used non- spike-in normalization approaches [15, 13, 16, 17], which approximate the aforementioned spike-in strategy by assuming that most features do not change on average across samples/conditions. For the same reason, such an interpretation can also be given to approaches like centered logarithmic transforms (CLR) from the theory of compositional data, which many analysts favor when working with relative abundances [18, 19, 20, 21, 22, 23, 24]. In this paper, we analyze the behavior of these existing scaling normalization techniques in light of compositional bias.

When trying to normalize metagenomic 16S survey data with these methods however, we found that the large fraction of zeroes in the count data, and the relatively low sequencing depths of metagenomic samples posed a severe problem: DESeq failed to provide a solution for all the samples in a dataset of our interest, and TMM based its estimation of scale factors on very few features per sample (as low as 1). The median approach simply returned zero values. CLR transforms behaved similarly. When one proceeds to avoid this problem by adding pseudo-counts, owing to heavy sparsity underlying these datasets, the transformations these techniques imposed mostly reflected the value of pseudocount and the number of features observed in a sample. A recently established scaling normalization technique, Scran [25], tried to overcome this sparsity issue in the context of single cell RNAseq count data – which also entertains a large fraction of zeroes – by decomposing simulated pooled counts from multiple samples. That approach, developed for relatively high coverage single cell RNAseq, also failed to provide solutions for a significant fraction of samples in our datasets (as high as 74%). Furthermore, as we illustrate later, compositional bias affects data sparsity, and normalization techniques that

ignore zeroes when estimating normalization scales (like CSS [26], and TMM) can be severely biased. The relatively low sequencing depth per sample (as low as 2000 reads per sample), large number of features and their diversity across samples thus pose a serious challenge to existing normalization techniques. In this paper, we develop a compositional bias correction technique for sparse count data based on an empirical Bayes approach that borrows information across features and samples

Since we have presented the problem of compositional bias as one affecting inferences on absolute abundances, one might wonder if resolving compositional bias is needed when analyses on *relative* abundances are performed. It is important to realize that compositional bias is infused in the count data, solely due to inherent characteristics of the sequencing process, even before it passes through any specific normalization process like scaling by library size. In practical conditions, because feature-wise abundance perturbations are also driven by technical sources of variation uncorrelated with total library size [27, 28, 29, 30], compositional bias correction becomes necessary even when analysis is performed on relative abundances.

The paper is organized as follows. We first set up the problem of compositional bias correction and with appropriate simulations, evaluate several scaling normalization techniques in solving it. We find that techniques based only on library size (e.g. unaltered RPKM/CPM [31], rarefication/subsampling in metagenomics [32, 33]) are provably bad. Other scaling techniques, while providing robust compositional bias estimates on high coverage data, perform poorly at sparsity levels often observed with metagenomic count data. We then introduce the proposed normalization approach (*Wrench*) and evaluate its performance with simulations and experimental data. We close by discussing the insights obtained by applying Wrench and other scaling normalization techniques to experimental datasets, arguing both for addressing compositional bias in general practice and in benchmarking studies. Because all the aforementioned techniques, including our own proposal, assume that most features do not change across conditions, they would all suffer in analyses of features arising from arbitrary general conditions. In such cases, spike-in based techniques can be effective [34], although methods similar to the ERCC method for bulk RNAseq will not work for the simple reason it starts with an extract, an already compositional data source.

Results

Formalizing compositional bias in differential abundance analysis

Below, we describe the *compositional correction factor*, the quantity we use to evaluate scaling normalization techniques in overcoming compositional bias.

Fig. 3 illustrates a general sequencing experiment and sets up the problem of compositional bias correction. We imagine a set of samples/observations $j = 1 \dots n_g$ arising from conditions $g = 1 \dots G$ (e.g., cases and controls). The true absolute abundances of features in every sample organized as a vector $X^0_{gj\cdot}$, are perturbed by various technical sources of variation as the sample is prepared for sequencing. The end result is a transformed absolute abundance vector $X_{gj\cdot}$, the net total abundance of which is denoted by $T_{gj} = \sum_i X_{gji}$. This is the input to the sequencer, which introduces compositional bias by producing reads *proportional* to the *absolute* feature abundances represented in $X_{gj\cdot}$. The output reads are processed and organized as counts in a vector $Y_{gj\cdot}$, which now retain only *relative* abundance information of features in $X_{gj\cdot}$. The ultimate goal of a normalization strategy is to recover $X^0_{gj\cdot}$ for all g and j.

Our goal is to evaluate existing normalization approaches based on how well they reconstruct X from Y, as it is in this step, that the sequencing process induces the bias we are interested in. We come back to the question of reconstructing X^0 at the end of this subsection. Because we are ignoring all other technical biases inherent to the experiment/technology (i.e., the process from $X^0 \to X$), our discussions apply to RNAseq/scRNAseq/metagenomics and other quantititative sequencing based assays. In this paper, our primary interest will be in the correction of compositional bias for metagenomic marker gene survey data, which are often under-sampled.

Although not strictly necessary, for simplicity, we shall assume that the relative abundances of each feature i is given by q_{gi} for all samples within a group g. It is also reasonable to assume an $X_{gj\cdot}|T_{gj} \sim Multinomial(T_{gj}, q_{g\cdot})$, where $q_{g\cdot}$ is the vector of feature-wise relative abundances. Such an assumption follows for example from a Poisson assumption on the expression of features X_{gji} [35]. Similarly, we shall assume the observed counts $Y_{gj\cdot}|\tau_{gj} \sim Multinomial(\tau_{gj}, q_{g\cdot})$, where τ_{gj} is the corresponding sampling depth. We shall use $E[T_{g1}]$ to denote the average *total* absolute abundance of features for each sample j from group g at the time of input. Similarly, $E[\overline{X_{g+i}}]$ will denote the marginal expectation of absolute abundance of

feature i across samples in group g (number of molecules per unit volume in case of RNAseq / number of distinct 16S fragments per unit volume in an environmental lysate in the case of 16S metagenomics). If we set $g = 1$ as the control group, and define, for every feature i, $\nu_{gi} = \frac{E[\overline{X_{g+i}}]}{E[\overline{X_{g+i}}]}$, then $\log \nu_{gi}$ is the log-fold change of true absolute abundances associated with group g relative to that of the control group. We can write:

$$\nu_{gi} = \frac{E[\overline{X_{g+i}}]}{E[\overline{X_{g+i}}]} = \frac{E[T_{g1}]q_{gi}}{E[T_{11}]q_{1i}} \equiv \Lambda_g \cdot \frac{q_{gi}}{q_{1i}} \tag{1}$$

This indicates that the fold changes based on observed proportions (estimated from Y) from the sequencing machine confounds our inference of the fold changes associated with absolute abundances of features at stage X, through a linear bias term Λ_g. Thus, to reconstruct the average absolute abundances of features in experimental group g, one needs to estimate the *compositional correction factor* Λ_g^{-1}, where for convenience in exposition below, we have chosen to work with the inverse. Note that the compositional correction factor for the control group $\Lambda_1^{-1} = 1$ by definition.

Details on our terminology and how it differs from *normalization factors*, which are compositional factors altered by sample depths, are presented in the Simulations subsection under Methods. Below, we use the terms compositional scale or more simply scale factor interchangeably to refer to compositional correction factors.

The central idea in estimating compositional correction factors For any group g, an effective strategy for estimating Λ_g can be derived based on an often quoted assumption behind scale normalization techniques [13]: if most features do not change in an experimental condition relative to the control group, eqn. 1 should hold true for most features with $\nu_{gi} = 1$. Thus, an appropriate summary statistic of these ratios of proportions could serve as an estimate of Λ_g^{-1}. Indeed, as illustrated in **Table 1**, many scale normalization methods (including the proposal in this work) can be viewed in this light, where some control set of proportions ("reference") is defined, and the Λ_g^{-1} estimate is derived based on the ratio of proportions. This central idea being the same, the robustness of these methods are dependent on how well the assumptions hold with respect to the chosen reference, and the choice of the estimation strategy.

Reconstrucing X^0 from Y It is worth emphasizing that the aforementioned estimation strategy does not restrict compositional factors to only reflect biology-induced global abundance changes; in reality, if feature-wise perturbations (ν_{gi}) are also of technical origin, they can well be correlated with other sources of technical variation, and can be seen to estimate technical variation beyond what is accounted for by sample depth adjustments. Thus, it is interesting to ask under what conditions compositional factors arising from scaling techniques (including our proposed technique in this work) can reconstruct X^0. In the supplementary, we show that in the presence of sequence-able experimentally introduced contaminants, utilizing existing compositional correction tools amounts to applying stricter assumptions than the often-cited assumption of "technical biases affecting all feature the same way". The precise condition is given in the supplement (supplementary section 3, eqn. 6). In the absence of contamination, we find the traditional assumption to be sufficient.

Existing techniques fail to correct for compositional bias in sparse 16S survey data.

In this subsection, we ask how existing techniques fare in estimating compositional correction factors, both in settings at large sample depths and with particular relevance to sparse 16S count data. We will find that library size/subsampling approaches are provably bad and that other scaling techniques face certain difficulties with sparse data. We will also note that the common strategy of deriving normalization factors/data transformations after adding pseudocounts to the original sparse count data transformations also lead to biased estimates of scale factors.

Our analysis below is limited to methods that provide interpretable estimates of fold-changes. We therefore do not consider differential abundance inferences arising from rank-based methods. We also leave the analysis of non-linear normalization techniques for future work.

Library size/Subsampling based approaches To understand the practical importance of resolving confounding caused by compositional bias, we first asked under what conditions, inferences made without compositional correction would continue to reflect changes in absolute abundances in an unbiased manner. We formally analyzed its influence within the framework of generalized linear models, a widely used statistical framework within several count data packages (supplementary sec-

tion 1). Under the most natural adjustments based on the total count (e.g., unaltered R/FPKM/CPM/subsampling/rarefication based approaches), we found that these conditions can be precisely characterized and are extremely limited in their applicability in general experimental settings. It may be tempting to argue that one can resort to total count-based normalization if total feature content is the same across conditions. However, as shown in supplementary section 1, it is easy to see that this assumption is only valid when strict constraints on the levels of technical perturbation of feature abundances and sequence-able contaminants are respected, an assumption that can be very easily violated in metagenomic experiments [36, 37, 38], which usually feature high intra- and inter-group feature diversity.

Reference normalization and robust fold-change estimation techniques We now compare and contrast library size adjustments with a few reference based techniques (reviewed in **Table. 1**) in overcoming compositional bias at *high* sample depths. Furthermore, many widely used genomic differential abundance testing toolkits enforce prior assumptions on reconstructed fold changes, and moderate their estimation. This made us wonder about the robustness of these testing techniques in overcoming the false positives that would otherwise be created without compositional bias correction. With an exhaustive set of simulations generating bulk RNAseq like data with $20M$ reads per sample, by and large, we found that all testing packages behaved the same way, and the key ingredient to overcome compositional bias always was an appropriate normalization technique (supplementary section 2). We also found that reference based normalization procedures outperformed library size based techniques significantly, re-emphasizing the analytic insights we mentioned previously. With sparse 16S data however, such techniques developed for bulk RNAseq faced major difficulties as illustrated next.

In **Fig. 4**, we plot the feature-wise compositional scale estimates (i.e., ratio of sample proportion to that of the reference; third column entries in **Table. 1**), obtained from TMM and DESeq for a sample in two different 16S microbiome datasets. TMM computes a weighted average over these feature-wise estimates, while DESeq proposes the median. The first column corresponds to a bulk RNAseq study of the rat body map [39]; the second corresponds to those from a 16S metagenomic dataset [40]. Strikingly, while a large number of features agree on their scale factors

for a sample arising from bulk RNAseq for both TMM and DESeq strategies, the sparse nature of metagenomic count data makes robust estimation of their scale factors extremely difficult. Furthermore, large variance is also observed across the scale factors suggested by the individual features. Clearly, a moderated estimation procedure is warranted.

One might wonder if adding pseudocounts to the orginal count data (a common procedure in metagenomic data analysis [19, 41]) effectively deals away with the problem. However, as shown in **Fig. 5**, with large number of features absent per sample, these scale factors roughly reflect the value of the pseudocount, and are systematically scaled down in value as sequencing depth, which is strongly correlated with feature presence, increases. This result suggests that addition of pseudocounts to data need not be the right strategy for deriving normalization scales based on CLR [42] or other similar methods, especially when the data is sparse. The alternate idea of only deriving scale factors based on positive values alone, are also associated with problems as we will see later in the text.

Our proposed approach (Wrench) reconstructs precise group-wise estimates, and achieves significantly better simulation performance

To overcome the issues faced by existing techniques, we devised an approach based on the following observations and assumptions. First, group/condition-wise feature count distributions are less noisy than sample-wise feature count distributions, and it may be useful to Bayes-shrink sample-wise estimators towards that of group-wise global estimates. Second, zero abundance values in metagenomic samples are predominantly caused by competition effects induced by sequencing technology (illustrated in **Fig. 1**), and therefore can be indicative of large changes in underlying compositions[1] with respect to a chosen reference. Indeed, ignoring sterile/control samples, the median fraction of features recording a zero count across samples in the mouse, lung, diarrheal, human microbiome project and (the very high coverage) Tara oceans datasets were: .96, .98, .98, .98 and .88. These respectively had median sample depths of roughly $2.2K$, $4.5K$, $3.3K$, $4.4K$ and $100K$ reads. In direct contrast, this value for the high coverage bulkRNAseq rat body map across 11 organs at a median sample depth of $9.7M$ reads, is .33. Large number of features, extreme

[1]the idea being that in the limit $\Lambda_g \to \infty$, feature-wise ratios that reflect Λ_g^{-1}, $\to 0$

diversity, and time-dependent dynamic fluctuations in microbial abundances can result in such high sparsity levels in metagenomic datasets. When working within the fundamental assumption that *most features do not change across conditions*, such extraordinary sparsity levels can then be attributed, by and large, to competition among features for being sequenced. As we illustrate in **Fig. 6**, zero observations in a sample are correlated with compositional changes, and truncated analyses that ignore them (as is done with TMM / DESeq / metagenomic CSS normalization techniques) effectively leads to loss of information and results that are opposite to what is expected.

We now give a brief overview of the technique (Wrench) proposed in this work. More details are presented in the Methods section. With average proportions across a dataset as our reference, we model our feature-wise proportion ratios as a hurdle log-normal model[2], with feature-specific zero-generation probabilities, means and variances. The analytical tractability of the model allows us to standardize the feature-wise values within and across samples, and derive the compositional scale estimates by basing heavy weights on less variable features that are more likely to occur across samples in a dataset. In addition, to make the computed factors robust to low sequencing depths and low abundant features, we employ an empirical Bayes strategy that smooths the feature-wise estimates across samples before deriving the sample-wise factors. Such situations are rather common in metagenomics, and some robustness to overcome heavy sampling variations is desirable.

Table. 2 succinctly illustrates where current state of the art fails, while more comprehensive simulations illustrating the effectiveness of the proposed approach is presented in **Fig. 7**. Roughly, we simulated two experimental groups, with $54K$ features whose proportions were chosen from the lung microbiome data, and let 35% of features change across conditions (see Methods for details on simulations). The net true compositional change resulting from each simulation, and their corresponding reconstructions by the various techniques when the count data are generated at different sequencing depths are shown. The following observations form the theme of these, and the more elaborate simulations summarized in Fig. 7: 1)TMM/CSS, because they focus on positive-valued observations only, are restricted in the range

[2]the random variable assumes a value of zero with probability π and a positive value based on its specific log-normal distribution with probability $(1-\pi)$

of scales they can reconstruct. 2) Scran can yield accurate estimators at very large sequencing depths when high feature-wise coverages are achieved. Unfortunately, this behavior is highly dependent on the underlying feature proportions and their diversity. 3) Wrench estimators offer better alternatives for under-sampled data, and as we shall observe below in their empirical performances, they can still offer robust protection at higher coverages. For specific comparisons with pseudocounted CLR, please refer supplementary Fig. 9, in which we show the proposed technique (Wrench) performing significantly better.

We briefly note a key ingredient about our simulation procedure. Simulating sequencing count data as *independent* Poissons / Negative Binomials – as is commonly done in benchmarking pipelines – does not inject compositional bias into simulated data. From the perspective of performance comparisons for compositional correction, doing so is therefore inappropriate. A renormalization procedure after assigning feature-wise fold-changes is necessary. Alternatively, if absolute abundances are generated, subsampling to a desired sample depth needs to be performed.

Wrench scales correlate with orthogonal measurements of technical biases and maintain group-wise integrity in the reconstructed scales.

Even though the contribution of true compositional changes and other technical biases are not identifiable from the compositional scales without extra information, we asked if the reconstructed scales correlate with orthogonal information on absolute abundances, and other measures of technical biases. The results are summarized in **Table 3**. Interestingly, in the very high coverage Tara Oceans metagenomics project, Wrench and Scran estimators achieve comparable correlations (>50%) with absolute flow cytometry measurements of microbial counts from the Tara Oceans project. Scran failed to reconstruct the scales for 3 samples. TMM and CSS had substantially poor correlations. Similarly, Wrench normalization factors had comparable/slightly better correlations to the total ERCC counts in bulk and single cell RNAseq datasets. In direct contrast, CLR scale factors (the geometric means of proportions) computed with pseudocounts were either uncorrelated or highly anti-correlated with the aforementioned measurements reflecting technical biases. These results reaffirm that there may be advantages to exploiting specialized com-

positional correction tools even with microbiome datasets teeming with microbes of extraordinary diversity.

We applied all the normalization methods discussed so far to a few experimental 16S microbiome datasets, and analyzed the phenotypic integrity of the resulting scales. Owing to the heavy sparsity in these datasets, Scran failed to provide scales for 53 out of 72 samples of the lung microbiome, 10 out of 132 observations of the mouse microbiome, 6 out of 992 samples of the diarrheal dataset. Interestingly, all normalization techniques resulted in group-wise integrity in the scales they reconstructed within and across related phenotypic categories, potentially indicating the general importance of compositional correction in general practice. For instance, in the microbiome samples arising from the Human Microbiome Project, as shown in **Fig. 8A**, we noted systematic body site-specific global deviations in the fold change distributions. This is similar to what was illustrated with the Tara project in Fig. 2. We found the reconstructed compositional scales to largely organize by body sites, across normalization techniques (**Fig. 8B**), behind-ear and stool samples were distinctly located in terms of their compositional scales from the oral and vaginal microbiomes. This behavior was also recapitulated in scales reconstructed from other centers. Supplementary Figs. 10 and 11 present similar results on samples arising from the J. Craig Venter Institute. In the case of the mouse microbiome samples, most normalization techniques predicted a mild change in differential feature content across the two diet groups (**Fig. 8C**, and supplementary Fig. 12). In the lung microbiome, the lung and oral cavities had roughly similar scales across smokers and non-smokers (supplementary Fig. 13), while scales from the probing instruments had relatively higher variability, which we found to directly correlate with the high variability of feature presence in the count data arising from these samples. In the diarrheal datasets of children, however, no significant compositional differences were found across the various country/health-status populations (**Fig. 8D**).

From the perspective of differential abundance results in these metagenomic datasets, although there were differences in which taxonomic units were called as being differentially abundant (mostly in the low and moderate abundance ranges), we did not find any specific taxonomic categories ("species") uniquely identified as being differentially abundant with the normalization methods compared above. This is likely due to a large number of features being assigned similar species definitions.

Finally, we noted that the rat body map samples also showed systematic tissue-specific global deviations in the expressed features' fold change distribution. We include this analysis of low sparsity samples for completeness. **Fig. 9** shows this result and the general behavior of compositional scales across various methods compared and a few related statistics of the dataset. Given that these samples arise from a well designed series of experiments, and that the similarity in the scales within and across related tissues, and across normalization methods, is striking, the observed trend in the reconstructed scales could indeed reflect underlying true compositional differences for the most part. TMM and CSS ascribe substantially deviated scales to muscle, heart and liver tissues, when compared to Scran and Wrench estimators. This effect may be due to the truncated estimation strategy which biases the scales for a relatively fewer but highly expressed genes in these tissues. Nevertheless, these results indicate potentially heavy compositional bias injected into downstream differential abundance analysis that compare tissues of different types. These results suggest that compositional bias can be costly not only in metagenomics, but even in common bulk-RNAseq studies.

Discussions

For some researchers, statistical inference of differential abundance is a question of differences in relative abundances; for others, it is a matter of characterizing differences in absolute abundances of features expressed in samples across conditions [43, 14]. In this work, we took the latter view and aimed to characterize the compositional bias injected by sequencing technology on downstream statistical inference of absolute abundances of genomic features.

It is clear that the probability of sequencing a particular feature (ex: mRNA from a given gene or 16S RNA of an unknown microbe) in a sample of interest is not just a function of its own fold change relative to another sample, but inextricably linked to the fold changes of the other features present in the sample in a systematic, statistically non-identifiable manner. Irrevocably, this translates to severely confounding the fold change estimate and the inference thereof resulting from generalized linear models. Because the onus for correcting for compositional bias is transferred to the normalization and testing procedures, we reviewed existing spike-in protocols from the perspective of compositional correction, and analyzed several widely used nor-

malization approaches and differential abundance analysis tools in the context of reasonable simulation settings. In doing so, we also identified problems associated with existing techniques in their applicability to sparse genomic count data like that of metagenomics and single cell RNAseq, which lead us to develop a reference based compositional correction tool (Wrench) to achieve the same. (We catalog a few potential ways by which compositional sources of bias can cause sparsity in metagenomic and single cell sequencing count data in Supplementary section 6). A few important insights emerge.

In our simulations, we found reference based normalization approaches to be far superior in correcting for sequencing technology-induced compositional bias than library size based approaches. From a more practically relevant perspective, we found that in all the tissues from the rat body map bulk RNAseq dataset, the scale factors can be robustly identified. We expect that in other bulk RNAseq datasets, the assumptions underlying compositional correction techniques to hold well. These results reinforce trust in exploiting such scaling practices for other downstream analyses of sequencing count data apart from differential abundance analysis; for example, in estimating pairwise feature correlations. In the regimes where assumptions underlying these techniques are met, an analyst need not be restricted to scientific questions pertaining to relative abundances alone. The fundamental assumption behind all the aforementioned techniques (including our own) is that most features do not change across conditions. As we illustrated, these assumptions appear to hold rather well in bulk RNAseq. Do we expect these to hold in arbitrary microbiome datasets as well? This question is not easy to address without more experiments, but the relatively high correlations obtained with orthogonal measurements of technical biases, the similarity in the compositional scales obtained within samples arising from biological groups, and their sometimes highly significant shifts preserved across normalization techniques and across sequencing centers in large scale studies certainly reinforce the critical importance of characterizing compositional biases, if any, in metagenomic analyses by establishing carefully designed spike-in protocols. In particular, given the inverse dependence of compositional correction factors and the total feature content, the large compositional scale estimates obtained for stool samples (across all normalization techniques) is highly suspect. Compositional effects can amplify even when a few features experience adverse technical pertur-

bations, and only carefully designed experiments can isolate these effects to inform further normalization approaches. Finally, our results also emphasize the tremendous care one needs to exercise before applying the most natural normalizations based on total sequencing depth or by applying pseudocounts when the data is excessively sparse (CLR, RPKM, CPM, rarefication are a few examples).

This brings us to the question of how effective spike-in strategies are in enabling us to overcome compositional bias. It is immediately clear that the widely used ERCC recommended spike-in procedure for RNAseq cannot help us in overcoming confounded inference due to compositional bias for the simple reason that it already starts with an extract, a compositional data source (supplementary section 2). If one is able to add the spike-in quantities at a prior stage during feature extraction, we would have some hope. Lovén et al., [44] demonstrate a procedure for RNAseq that precisely does this, in which the spike-ins are added at the time when the cells are lysed and suspended in solution [45]. One can perhaps extend these solutions to metagenomics, where we may expect confounding due to compositionality to be heavy by adding barcoded 16S RNAs during feature extraction. We expect similar problems to arise in other genomic and epigenetic measurement techniques that exploit sequencing technology, and the need for the development of appropriate spike-in procedures should be addressed.

Finally, it is imperative that we enforce new tools and techniques for normalization and differential abundance analysis of sequencing count data be benchmarked for compositional bias at least in the simulation pipelines. Data analyses based on large-scale integrations of different data types for predicting clinical phenotypes is increasingly common, and care should be taken to include effective normalization techniques to overcome compositional bias. We hope the results and ideas presented and summarized in our paper enables a researcher to do just that.

Conclusions

Compositional bias, a linear technical bias, underlying sequencing count data is induced by the sequencing machine. It makes the observed counts reflect relative and not absolute abundances. Normalization based on library size/subsampling techniques cannot resolve this or any other practically relevant technical biases that are uncorrelated with total library size. Reference based techniques developed for

normalizing genomic count data thus far, can be viewed to overcome such linear technical biases under reasonable assumptions. However, high resolution surveys like 16S metagenomics are largely undersampled and lead to count data that are filled with zeroes, making existing reference based techniques, with or without pseudocounts, result in biased normalization. This warrants the development of normalization techniques that are robust to heavy sparsity. We have proposed a reference based normalization technique (Wrench) that estimates the overall influence of linear technical biases with significantly improved accuracies by sharing information across samples arising from the same experimental group, and by exploiting statistics based on occurrence and variability of features. Such ideas can also be exploited in projects that integrate data from diverse sources. Results obtained with our and other techniques, suggest that substantial compositional differences can arise in (meta)genomic experiments. Detailed experimental studies that specifically address the influence of compositional bias and other technical sources of variation in metagenomics are needed, and must be encouraged.

Materials and Methods

An approach (Wrench) for compositional correction of sparse, genomic count data

We now present the theory behind our technique. Our approach is based on the following observations: 1) that group/condition-wise feature count distributions are less noisy than sample-wise feature count distributions, and it may be useful to Bayes-shrink sample-wise ratios of proportions towards that of group-wise proportions. 2) A relatively large fraction of zero expression values in a sample are caused due to competition effects induced by the sequencer (under-sampling), and therefore can be indicative of large changes in the underlying compositions. Some discussions on the potential sources of sparsity are presented in supplementary section 6.

We start off with the following assumptions, commonly used in quantitative genomic analyses i) that feature-wise fold changes within a group/condition are log-normally distributed, and ii) that feature-wise expression values follow a log-normal model. We further note that feature abundances, due to low/moderate sequencing depths or complete absence of features, can entertain zero values. With g, i and j indexing groups/conditions, features and samples respectively, let Y_{gji} indicate the count of feature i in sample j and $\tau_{gj} = \sum_i Y_{gji}$ indicate its total feature

count/sample depth. Based on eqn. 1, and accounting for sample depth, we can write the mean model when a feature is expressed with a positive count value with respect to a chosen reference q_i^* as: $E[\log Y_{gji}|Y_{gji} > 0] = \log\left(\Lambda_j^{-1}\nu_{gji}\tau_{gj}q_i^*\right) \equiv \log\left(\theta_{gji}\tau_{gj}q_i^*\right)$. Our goal is to estimate Λ_j^{-1}, the compositional correction factor for sample j. We assume the following hurdle lognormal (LN) hierarchy:

$$\begin{aligned}
&Y_{gji} \sim \pi_{gji}\delta_0 + (1-\pi_{gji})\cdot\theta_{gji}\cdot\tau_{gj}q_i^*\cdot LN(0,\sigma_i^2), \quad i = 1\ldots p,\ j = 1\ldots n \\
&\theta_{gji} \sim \zeta_g \cdot LN(0,\sigma_{\nu g}^2),\ g = 1\ldots G \\
&\log\left(\frac{\pi_{gji}}{1-\pi_{gji}}\right) = \beta_1 \log\tau_{gj} + \text{possibly other covariates}
\end{aligned} \tag{2}$$

We will assume sample depth τ_{gj} as being fixed constants and shall estimate $\{\pi_{gji}, \theta_{gji}, \zeta_g, \sigma_i^2, \sigma_{\nu g}^2\}$. Logistic regression is used to fit the hurdle parameters π. Our estimation strategy for the variance components is as follows: notice that $\log\zeta_g$ and $\sigma_{\nu g}^2$ have the interpretations of the mean and variance parameters of the group-wise log-fold change distribution of features expressed in group g. Thus, we first generate the within group pooled ratios q_i^g/q_i^* with $q_i^g = \sum_{j\in g} w_j q_{gji}$, where $w_j \propto \tau_j$, and q_i^* set to the average proportion vector across all the samples in the dataset. Group specific fold-change variances $\sigma_{\nu g}^2$ is then estimated as the variance of the logged-positive part of the same ratio distribution for each g. The scale parameter ζ_g is obtained as the average of these ratio values (including those that are zeroes).

Feature specific σ_i^2 is estimated using group information with linear model fits (lmFit, Limma [46]) on the logged-positive count values. For those features expressed in only one sample, these estimates are obtained using a non-parameteric variance function fitted using a procedure similar to that of voom [31].

Each positive-valued feature specific ratio of proportions from a sample $r_{gji;*} = q_{gji}/q_i^*$ is then rescaled by $\hat{\zeta}_g$ to yield $h_{gji;*} = r_{gji;*}/\hat{\zeta}_g$, which assumes a $LN(0, \sigma_i^2 + \sigma_{\nu g}^2)$ distribution according to the model above. For stability, we re-write its conditional mean as $E[\log h_{gji;*}|h_{gji;*} > 0, a_{gi}] = \log\mu_j + \log a_{gi}$. The intercept term $\log\mu_j$ is estimated by the generalized least squares estimator given by the weighted mean of the positive-valued rescaled ratios, with weights proportional to $\frac{1}{\hat{\sigma}_i^2+\hat{\sigma}_{\nu g}^2}$, and predict $\log a_{gji}$ in a sample specific manner with $\frac{\sigma_{\nu g}^2}{\sigma_i^2+\sigma_{\nu g}^2}(\log h_{gji;*} - \log\hat{\mu}_j)$, the best linear unbiased predictor. The sample-wise ratio of proportions are then regularized as: $r_{gji;*} = \frac{q_{gji}}{q_i^*} \longrightarrow \tilde{r}_{gji;*} = \hat{\mu}_j\hat{a}_{gji}\hat{\zeta}_g$.

At this point, we are ready to derive the summary statistics on the regularized ratios: while averaging over the regularized ratios $W_0 =: \frac{1}{p}\sum_i \tilde{r}_{gji;*}$ would be one estimation route to Λ_j^{-1}, better control can be achieved by taking the variation in the feature-wise zero generation in to account. We shall notice that $E[r_{gji;*}|r_{gji;*} > 0] = \tilde{r}_{gji;*} \cdot e^{\sigma_i^2/2}$, and so a robust averaging over $\tilde{r}_{gji}/e^{\sigma_i^2/2}$, can serve as an estimator of Λ_j^{-1}. One might choose the weights for averaging to be proportional to that of the inverse hurdle/inclusion probabilities (as is done in survey analysis) $\propto 1/(1-\pi_{ij})$ or on the inverse marginal variances ascribed by our model above $\propto \frac{1}{(1-\pi_{ij})(\pi_{ij}+e^{\sigma_i^2+\sigma_{\nu_g}^2}-1)}$. Supplementary section 7 sketches the derivations.

An advantage of these weights (and hence the model) is that the weighting strategies proceed smoothly for features with zero expression values as well, unlike the binomial weights employed in the TMM procedure. Furthermore, when constructing averages, the weights have a favorable property of downweighting zeroes at higher sample depths relative to those in samples at lower sample depths.

An estimator that we also found to work well empirically is a weighted average of $\frac{\tilde{r}_{gji}/e^{\sigma_i^2/2}}{1-\pi_{ij}}$ with weights proportional to $\frac{1}{\sigma_i^2}$.

In summary, we explored the performance of the following estimators:

$$
\begin{aligned}
W_0 &=: \frac{1}{p}\sum_i \tilde{r}_{gji;*} = \overline{\tilde{r}_{g+j;*}}, \\
W_1 &=: \frac{1}{p}\sum_i w_{ij}\tilde{r}_{gji;*}, \text{ with } w_{ij} \propto 1/(1-\pi_{ij}) \\
W_2 &=: \frac{1}{p}\sum_i w_{ij}\tilde{r}_{gji;*}, \text{ with } w_{ij} \propto \frac{1}{(1-\pi_{ij})\left(\pi_{ij}+e^{\sigma_i^2+\sigma_{\nu_g}^2}-1\right)} \\
W_3 &=: \frac{1}{p}\sum_i w_{ij}\frac{\tilde{r}_{gji;*}}{1-\pi_{ij}}, \text{ with } w_{ij} \propto \frac{1}{\sigma_i^2}
\end{aligned}
\tag{3}
$$

We have found W_1, W_2 and W_3 to work comparably well in simulations and empirical comparisons, and W_0 slightly less so at high sparsity levels at low sample depths. We prefer W_2 as it systematically integrates both the hurdle and positive component variations. In our software implementation, users have the option for other weighted variants, and whether weighted averaging over zeroes is necessary as they see fit. Software documentation and supplementary material embark on further discussions on these ideas.

Finally, with this framework setup, extensions for *batch correction* can be immediately made; this work is being planned for a forthcoming submission.

Data We principally demonstrate our results with five datasets from metagenomic surveys. A smoking study ($n = 72$) where the lung microbiome of smokers and non-smokers were surveyed (along with the instruments that were used to sample the individual). A diet study in which the gut microbiomes ($n = 139$) of carefully controlled laboratory mice fed plant-based or western diets were sequenced [32]. A large scale study of human gut microbiomes ($n = 992$) from diarrhea-afflicted and healthy children from various developing countries [40]. 16S metagenomic count data corresponding to all these studies were obtained from the R/Bioconductor package `metagenomeSeq` [26]. The Tara Oceans project's 16S reconstructions from whole metagenome shotgun sequencing ($n = 139$) was downloaded from The Tara Oceans project website under http://ocean-microbiome.embl.de/data/miTAG.taxonomic.profiles.release.tsv.gz. The flow cytometry counts for autotrophs, bacteria, heterotrophs, picoeukaryotes were obtained from TaraSampleInfo_OM.CompanionTables.txt from the same website and summed to serve as a rough measure of total cell count that correlates with sequenceable DNA material. The Human Microbiome Project count data were downloaded from http://downloads.hmpdacc.org/data/HMQCP/otu_table_psn_v35.txt.gz, and the associated metadata are from v35_map_uniquebyPSN.txt.bz2 under the same website.

The processed bulk-RNAseq data corresponding to the rat body map from [39] was obtained from [47].

The UMI single cell RNAseq data from Islam et al., [48] was downladed from GEO under accession GSE46980.

Implementation of normalization and differential abundance techniques All analysis and computations were implemented with the R 3.3.0 statistical platform. EdgeR's `compNormFactors` for TMM, DESeq's `estimateSizeFactors`, Scran's `computeSumFactors` (with `positive=TRUE` in sparse datasets) and `metagenomeSeq`'s `calcNormFactors` for CSS were used to compute the respective scales. Implementation of CLR factors used a pseudo-count of 1 following [41], and were computed as the denominator of column 3 in table 1. Limma's eBayes in

combination with lmFit, edgeR's `estimateDisp`, `glmFit` and `glmLRT`, DESeq's `estimateDispersionsGeneEst` and `nbinomLRT` were used to perform differential abundance testing [17]. Welch's t-test results were obtained with `t.test`.

Implementation of Wrench Wrench is implemented in R, and is available through the `Wrench` package at https://gitlab.umiacs.umd.edu/smuthiah/Wrench.

Simulations Given a set of control proportions q_{1i} for features $i = 1 \ldots p$, and the fraction of features that are perturbed across the two conditions f, we sample the set of true log fold changes ($\log \nu_{gi}$) from a fold change distribution (fold change distribution) for those randomly chosen features that do change. The fold change distribution is a two-parameter distribution chosen either as a two-parameter Uniform or a Gaussian. Based on the expressions from the first subsection of the results section, the target proportions were then obtained as $q_{gi} = \frac{\nu_{gi} q_{1i}}{\sum_k \nu_{gk} q_{1k}}$. Conditioned on the total number of sequencing reads τ, the sequencing output $Y_{gi\cdot}$ for all i were obtained as a multinomial with proportions vector $q_{g\cdot} = [|q_{gi}|]_{i=1}^{p}$. We set the control proportions from various experimental datasets (specifically, mouse, lung and the diarrheal microbiomes). With this setup, we can vary f, and the two parameters of the fold change distribution, and ask, how various normalization and testing procedures compare in terms of their performance. For bulk RNAseq data, as illustrated in supplementary figure 1, we simulated $20M$ reads per sample.

For comparison of Wrench scales with other normalization approaches, we altered the above procedure slightly to allow for variations in internal abundances of features in observations arising from a group g. We used ν_{gi} generated above as a prior fold change for observation-wise fold change generation. That is, for all samples $j \in 1 \ldots n_g$ for all g, where n_g represents the number of samples in group g, for all i (including the truly null features), sample ν_{gji} from $LN(\log \nu_{gi}, \tilde{\sigma}_\nu^2)$ for a small value of $\tilde{\sigma}_\nu^2 = .01$. This induces sample specific variations in the proportions within groups. Notice that this makes the problem harder and more realistic, as feature marginal count distributions now arise from a mixture of distributions. Based on empirically observed MA plots for our metagenomic datasets, we set the mean and standard deviation of prior log-fold change distribution to 0 and 3 respectively. For generating 16S metagenomic-like datasets, logged sample depths were sampled from a log-normal distribution with logged-standard deviation of .25 and logged-means

corresponding to $\log(4K)$, $\log(10K)$ and $\log(100K)$ reads. These parameters were chosen based on comparisons with MA plots, the sparsity levels and total sample depths observed in current experimental datasets.

In both versions of simulations, the total induced abundance change relative to that of the control is $\Lambda_j \propto \nu_j^T q_1$, where ν_j is the vector of fold changes for sample j, and q_1 is the vector of control proportions. We apply the term *compositional correction factor* for Λ_j^{-1} and the term *normalization factor* for a sample as the product of its compositional correction factor with something that is proportional to that of its sample depth. Thus, all technical artifacts like total abundance changes, but sample depth, are incorporated into the definition of compositional factors.

Performance comparisons: For simulations, we used edgeR as the workhorse fitting toolkit. The compositional scale factors provided by CSS, TMM, Scran and Wrench were provided to edgeR as offset factors. We define detectable differential abundance in our simulated count data as follows. For each simulation, as we know the true compositional factors, we input them as normalization factors in edgeR, and the detectable differences in abundances are recorded. All the performance metrics are then defined based on this ground truth. Because we are interested in fold changes and their directions, the performance metrics we report are redefined as follows: Sensitivity as the ratio of the number of detectable true-positives with true sign over the total number of positives, False discovery as the ratio of the number of detectable true positives with false sign and false positives, over the total number of significant calls made.

Software availability.

`Wrench` is available from GitLab at the URL: https://gitlab.umiacs.umd.edu/smuthiah/Wrench.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

Conceived and designed study: MSK, HCB.
Contributed analytical tools/reagents: MSK, EVS, KO, HCB.
Wrote the paper: MSK, SH, HCB.
Participated in discussions: All authors.

Acknowledgements

MSK thanks Mihai Pop, Tom Goldstein, Joyce Hsiao and Mathieu Almeida for useful discussions, Joseph Paulson for making the metagenomic count data used in this paper available through R/Bioconductor, the Tara

Oceans and the HMP teams for making their processed count data easily available. This work was partially supported by NSF grant 1564785 to SH, by NIH R01 grants GM083084, RR021967/GM103552 and K99HG009007 to SCH, and by NIH R01 grants GM114267 and HG005220 to HCB and MSK.

Author details

[1]Graduate Program in Bioinformatics, University of Maryland, College Park, MD, USA. [2]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD, USA. [3]Department of Mathematics, University of Maryland, College Park, MD, USA. [4]Center for Statistical Research and Methodology, U.S Census Bureau, Suitland, MD, USA. [5]GRED Oncology Biostatistics, Genentech, San Francisco, CA, USA. [6]Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Harvard University, Boston, MA, USA. [7]Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, USA.

References

1. Shendure, J., Ji, H.: Next-generation DNA sequencing. Nature Biotechnology **26**(10), 1135–1145 (2008). doi:10.1038/nbt1486. Accessed 2016-03-09
2. Sanger, F.: Sequences, sequences, and sequences. Annual Review of Biochemistry **57**, 1–28 (1988). doi:10.1146/annurev.bi.57.070188.000245
3. Hutchison, C.A.: DNA sequencing: bench to bedside and beyond. Nucleic Acids Research **35**(18), 6227–6237 (2007). doi:10.1093/nar/gkm688
4. Mardis, E.R.: A decade/'s perspective on DNA sequencing technology. Nature **470**(7333), 198–203 (2011). doi:10.1038/nature09796. Accessed 2016-03-09
5. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2016). R Foundation for Statistical Computing. https://www.R-project.org/
6. Wooley, J.C., Godzik, A., Friedberg, I.: A Primer on Metagenomics. PLOS Comput Biol **6**(2), 1000667 (2010). doi:10.1371/journal.pcbi.1000667. Accessed 2016-03-09
7. Park, P.J.: ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics **10**(10), 669–680 (2009). doi:10.1038/nrg2641. Accessed 2016-03-09
8. Wang, Z., Gerstein, M., Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics **10**(1), 57–63 (2009). doi:10.1038/nrg2484. Accessed 2016-03-09
9. Tringe, S.G., Rubin, E.M.: Metagenomics: DNA sequencing of environmental samples. Nature Reviews Genetics **6**(11), 805–814 (2005). doi:10.1038/nrg1709. Accessed 2016-03-09
10. Sunagawa, S., Coelho, L.P., Chaffron, S., Kultima, J.R., Labadie, K., Salazar, G., Djahanschiri, B., Zeller, G., Mende, D.R., Alberti, A., Cornejo-Castillo, F.M., Costea, P.I., Cruaud, C., d'Ovidio, F., Engelen, S., Ferrera, I., Gasol, J.M., Guidi, L., Hildebrand, F., Kokoszka, F., Lepoivre, C., Lima-Mendez, G., Poulain, J., Poulos, B.T., Royo-Llonch, M., Sarmento, H., Vieira-Silva, S., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Coordinators, T.O., Bowler, C., Vargas, C.d., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Jaillon, O., Not, F., Ogata, H., Pesant, S., Speich, S., Stemmann, L., Sullivan, M.B., Weissenbach, J., Wincker, P., Karsenti, E., Raes, J., Acinas, S.G., Bork, P.: Structure and function of the global ocean microbiome. Science **348**(6237), 1261359 (2015). doi:10.1126/science.1261359. Accessed 2016-03-09
11. Oshlack, A., Wakefield, M.J.: Transcript length bias in RNA-seq data confounds systems biology. Biology direct **4**(1), 14 (2009). Accessed 2017-04-12
12. Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A.: Gene ontology analysis for RNA-seq: accounting for selection bias. Genome biology **11**(2), 14 (2010). Accessed 2017-04-12
13. Robinson, M.D., Oshlack, A.: A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology **11**(3), 25 (2010). doi:10.1186/gb-2010-11-3-r25. Accessed 2015-01-09
14. Pachter, L.: Models for transcript quantification from RNA-Seq. arXiv:1104.3889 [q-bio, stat] (2011). arXiv: 1104.3889. Accessed 2016-01-21
15. Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U., Speed, T.P.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics **4**(2), 249–264 (2003). Accessed 2015-01-02

16. Bullard, J.H., Purdom, E., Hansen, K.D., Dudoit, S.: Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC bioinformatics **11**, 94 (2010). doi:10.1186/1471-2105-11-94
17. Anders, S., Huber, W.: Differential expression of RNA-Seq data at the gene level–the DESeq package. Heidelberg, Germany: European Molecular Biology Laboratory (EMBL) (2012). Accessed 2015-07-28
18. Aitchison, J.: The statistical analysis of compositional data. Journal of the Royal Statistical Society. Series B (Methodological), 139–177 (1982). Accessed 2015-07-20
19. Friedman, J., Alm, E.J.: Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol **8**(9), 1002687 (2012). doi:10.1371/journal.pcbi.1002687. Accessed 2013-10-30
20. Faust, K., Raes, J.: Microbial interactions: from networks to models. Nature Reviews Microbiology **10**(8), 538–550 (2012). doi:10.1038/nrmicro2832. Accessed 2015-07-28
21. Fernandes, A.D., Reid, J.N., Macklaim, J.M., McMurrough, T.A., Edgell, D.R., Gloor, G.B.: Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome **2**, 15 (2014). doi:10.1186/2049-2618-2-15. Accessed 2016-01-21
22. Fang, H., Huang, C., Zhao, H., Deng, M.: CCLasso: correlation inference for compositional data through Lasso. Bioinformatics, 349 (2015). doi:10.1093/bioinformatics/btv349. Accessed 2015-07-28
23. Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S., Bähler, J.: Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Comput Biol **11**(3), 1004075 (2015). doi:10.1371/journal.pcbi.1004075. Accessed 2016-03-09
24. Chen, K., Hu, Z., Xia, Z., Zhao, D., Li, W., Tyler, J.K.: The overlooked fact: fundamental need of spike-in controls for virtually all genome-wide analyses. Molecular and Cellular Biology, 00970–14 (2015). doi:10.1128/MCB.00970-14. Accessed 2016-03-09
25. L. Lun, A.T., Bach, K., Marioni, J.C.: Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biology **17**, 75 (2016). doi:10.1186/s13059-016-0947-7. Accessed 2016-08-13
26. Paulson, J.N., Stine, O.C., Bravo, H.C., Pop, M.: Differential abundance analysis for microbial marker-gene surveys. Nature methods (2013). Accessed 2015-01-17
27. Huggett, J.F., Laver, T., Tamisak, S., Nixon, G., O'Sullivan, D.M., Elaswarapu, R., Studholme, D.J., Foy, C.A.: Considerations for the development and application of control materials to improve metagenomic microbial community profiling. Accreditation and Quality Assurance **18**(2), 77–83 (2013). doi:10.1007/s00769-012-0941-z. Accessed 2017-04-13
28. van Dijk, E.L., Jaszczyszyn, Y., Thermes, C.: Library preparation methods for next-generation sequencing: Tone down the bias. Experimental Cell Research **322**(1), 12–20 (2014). doi:10.1016/j.yexcr.2014.01.008. Accessed 2017-04-13
29. O'Sullivan, D.M., Laver, T., Temisak, S., Redshaw, N., Harris, K.A., Foy, C.A., Studholme, D.J., Huggett, J.F.: Assessing the Accuracy of Quantitative Molecular Microbial Profiling. International Journal of Molecular Sciences **15**(11), 21476–21491 (2014). doi:10.3390/ijms151121476. Accessed 2017-04-13
30. D'Amore, R., Ijaz, U.Z., Schirmer, M., Kenny, J.G., Gregory, R., Darby, A.C., Shakya, M., Podar, M., Quince, C., Hall, N.: A comprehensive benchmarking study of protocols and sequencing platforms for 16s rRNA community profiling. BMC Genomics **17**, 55 (2016). doi:10.1186/s12864-015-2194-9. Accessed 2017-04-13
31. Law, C.W., Chen, Y., Shi, W., Smyth, G.K.: voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology **15**(2), 29 (2014). doi:10.1186/gb-2014-15-2-r29. Accessed 2015-06-17
32. Turnbaugh, P.J., Ridaura, V.K., Faith, J.J., Rey, F.E., Knight, R., Gordon, J.I.: The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Science Translational Medicine **1**(6), 6–14 (2009). doi:10.1126/scitranslmed.3000322
33. McMurdie, P.J., Holmes, S.: Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol **10**(4), 1003531 (2014). doi:10.1371/journal.pcbi.1003531. Accessed 2015-09-05
34. Jiang, L., Schlesinger, F., Davis, C.A., Zhang, Y., Li, R., Salit, M., Gingeras, T.R., Oliver, B.: Synthetic

spike-in standards for RNA-seq experiments. Genome research **21**(9), 1543–1551 (2011). Accessed 2016-03-09

35. Thattai, M.: Universal Poisson Statistics of mRNAs with Complex Decay Pathways. Biophysical Journal **110**(2), 301–305 (2016). doi:10.1016/j.bpj.2015.12.001. Accessed 2017-05-25
36. Schmieder, R., Edwards, R.: Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLOS ONE **6**(3), 17288 (2011). doi:10.1371/journal.pone.0017288. Accessed 2017-04-19
37. Salter, S.J., Cox, M.J., Turek, E.M., Calus, S.T., Cookson, W.O., Moffatt, M.F., Turner, P., Parkhill, J., Loman, N.J., Walker, A.W.: Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biology **12**, 87 (2014). doi:10.1186/s12915-014-0087-z. Accessed 2017-04-19
38. Hemme, C.L., Tu, Q., Shi, Z., Qin, Y., Gao, W., Deng, Y., Nostrand, J.D.V., Wu, L., He, Z., Chain, P.S.G., Tringe, S.G., Fields, M.W., Rubin, E.M., Tiedje, J.M., Hazen, T.C., Arkin, A.P., Zhou, J.: Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. Frontiers in Microbiology **6** (2015). doi:10.3389/fmicb.2015.01205. Accessed 2017-04-19
39. Yu, Y., Fuscoe, J.C., Zhao, C., Guo, C., Jia, M., Qing, T., Bannon, D.I., Lancashire, L., Bao, W., Du, T., Luo, H., Su, Z., Jones, W.D., Moland, C.L., Branham, W.S., Qian, F., Ning, B., Li, Y., Hong, H., Guo, L., Mei, N., Shi, T., Wang, K.Y., Wolfinger, R.D., Nikolsky, Y., Walker, S.J., Duerksen-Hughes, P., Mason, C.E., Tong, W., Thierry-Mieg, J., Thierry-Mieg, D., Shi, L., Wang, C.: A rat RNA-Seq transcriptomic BodyMap across 11 organs and 4 developmental stages. Nature Communications **5**, 3230 (2014). doi:10.1038/ncomms4230. Accessed 2016-03-09
40. Pop, M., Walker, A.W., Paulson, J., Lindsay, B., Antonio, M., Hossain, M.A., Oundo, J., Tamboura, B., Mai, V., Astrovskaya, I., Bravo, H.C., Rance, R., Stares, M., Levine, M.M., Panchalingam, S., Kotloff, K., Ikumapayi, U.N., Ebruke, C., Adeyemi, M., Ahmed, D., Ahmed, F., Alam, M.T., Amin, R., Siddiqui, S., Ochieng, J.B., Ouma, E., Juma, J., Mailu, E., Omore, R., Morris, J.G., Breiman, R.F., Saha, D., Parkhill, J., Nataro, J.P., Stine, O.C.: Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome Biology **15**(6), 76 (2014). doi:10.1186/gb-2014-15-6-r76. Accessed 2016-03-09
41. Kurtz, Z.D., Müller, C.L., Miraldi, E.R., Littman, D.R., Blaser, M.J., Bonneau, R.A.: Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Comput Biol **11**(5), 1004226 (2015). doi:10.1371/journal.pcbi.1004226. Accessed 2016-08-13
42. Tsilimigras, M.C.B., Fodor, A.A.: Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Annals of Epidemiology **26**(5), 330–335 (2016). doi:10.1016/j.annepidem.2016.03.002
43. Rosa, P.S.L., Brooks, J.P., Deych, E., Boone, E.L., Edwards, D.J., Wang, Q., Sodergren, E., Weinstock, G., Shannon, W.D.: Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLOS ONE **7**(12), 52078 (2012). doi:10.1371/journal.pone.0052078. Accessed 2016-03-11
44. Lovén, J., Orlando, D.A., Sigova, A.A., Lin, C.Y., Rahl, P.B., Burge, C.B., Levens, D.L., Lee, T.I., Young, R.A.: Revisiting global gene expression analysis. Cell **151**(3), 476–482 (2012). doi:10.1016/j.cell.2012.10.012
45. Stegle, O., Teichmann, S.A., Marioni, J.C.: Computational and analytical challenges in single-cell transcriptomics. Nature Reviews Genetics **16**(3), 133–145 (2015). doi:10.1038/nrg3833. Accessed 2016-08-15
46. Smyth, G.K.: Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology **3**, 3 (2004). doi:10.2202/1544-6115.1027
47. Hicks, S.C., Okrah, K., Paulson, J.N., Quackenbush, J., Irizarry, R.A., Bravo, H.C.: Smooth Quantile Normalization. bioRxiv, 085175 (2016). doi:10.1101/085175. Accessed 2017-03-09
48. Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., Lönnerberg, P., Linnarsson, S.: Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods **11**(2), 163–166 (2014). doi:10.1038/nmeth.2772. Accessed 2016-08-30

Figures

Technique	Proposed Abundance Measure, Scale factor	Signal for Compositional Scale in
Total Sum	$\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = 1$	
TMM	$\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = e^{\left[\sum_{i: y_{ij} > 0 \ \cap \ i \in \textbf{trimmed set for j}} w_{ij} \log\left(\frac{q_{gji}}{q_{1ji}}\right)\right]}$	$\frac{q_{gji}}{q_{1ji}}$, ratio of proportions
DESeq	$\frac{y_{gji}}{C \cdot \tau_{gj} \cdot s_{gj}} \propto \frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = \text{median}_i \ \frac{q_{gji}}{[\prod_k q_{ik}]^{\frac{1}{n}}}$	$\frac{q_{gji}}{[\prod_k q_{ik}]^{\frac{1}{n}}}$, ratio of proportions
Median	$\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = \text{median}_i \ q_{gji} \propto \text{median}_i \ \frac{q_{gji}}{1/p}$	$\frac{q_{gji}}{1/p}$, ratio of proportions
Upper quartile	$\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = \text{upper quartile}_i \ q_{gji} \propto \text{upper quartile}_i \ \frac{q_{gji}}{1/p}$	$\frac{q_{gji}}{1/p}$, ratio of proportions
CLR Transformation	$\log\left(\frac{y_{gji}}{[\prod_i y_{gji}]^{\frac{1}{p}}}\right) \equiv \log\left(\frac{q_{gji}}{[\prod_i q_{gji}]^{\frac{1}{p}}}\right) \equiv \log\left(\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}\right)$, with $s_{gj} = [\prod_i q_{gji}]^{\frac{1}{p}} \propto \left[\prod_i \frac{q_{gji}}{1/p}\right]^{\frac{1}{p}}$	$\frac{q_{gji}}{1/p}$, closely tracks Median factors above; ratio of proportions
Scran	$\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = \text{fit linear models to } \left\{\frac{q_{1ji}}{\overline{q_{++i}}}, \ldots, \frac{q_{nji}}{\overline{q_{++i}}}\right\}_{i=1}^{p}$	$\frac{q_{gji}}{\overline{q_{++i}}}$, ratio of proportions
Wrench	$\frac{y_{gji}}{\tau_{gj} \cdot s_{gj}}$, $s_{gj} = \frac{1}{p} \sum_i w_{ij} \frac{q_{gji}}{\overline{q_{++i}}}$	$\frac{q_{gji}}{\overline{q_{++i}}}$, ratio of proportions

Table 1 Scaling normalization approaches derive their technical bias estimates from ratio of proportions. For each scaling normalization technique (rows of the table, named in the first column), we present the transformation they apply to the raw count data (second column) to produce normalize counts. The third column shows how all techniques use statistics based on ratio of proportions (third column) to derive their scale factors. In the table, $i = 1 \ldots p$ indexes features (genes/taxonomic units), and each sample is considered to arise from its own singleton group: $g = 1 \ldots n$ and $j = 1$, τ_{gj} the sample depth of sample j, q_{gji} the proportion of feature i in sample j, w_{ij} represents a weight specific to each technique, and $\overline{q_{++i}}$ is the average proportion of feature i across the dataset. In the second column, the first row in each cell represents the transformation applied on the raw count data by the respective normalization approach. They all adjust a sample's counts based on sample depth (τ_{gj}) and a compositional scale factor s_{gj}, **denoted as Λ_{gj}^{-1} in text**. As noted in the third column, the estimation of s_{gj} is based on the ratio of sample-wise relative abundances/proportions (q_{gji}) to a reference that are all some robust measures of central tendency in the count data. The logarithmic transform accompanying CLR should not worry the reader about its relevance here, in the following sense: the log-transformation often makes it possible to apply statistical tests based on normal distributions for the rescaled data; this is in-line with applying log-normal assumptions on the rescaled data obtained with the rest of the techniques. $C = \left[\prod_j \tau_{gj}\right]^{-1/n}$ is a constant factor independent of sample, and its presence does not matter. For the same reason, Median and Upper Quartile scalings and CLR transforms, can be thought to base their estimates on a reference that assigns equal mass to all the features or if the reader wishes, a more complicated reference that behaves proportionally. When most features are zero, values arising from classical scale factors can be severely biased or undefined as we shall illustrate in the rest of the paper.

Net Compositional Change (Λ_g)	Average Sample Depth	CLR	TMM	CSS	Scran	W_0	W_1	W_2	W_3
36.86X	1M	1.36	1.45	5.41	22.57	19.32	31.44	30.65	32.01
7.75X	10K	.95	3.05	1.47	12.08 (14/40 samples failed)	5.30	6.32	6.31	6.70

Table 2 Example simulations illustrate the limitations of current techniques. Shown are the group-wise true and reconstructed compositional scales from the methods compared on two simulated examples, each at different sequencing depths and at different total true absolute abundance changes for a roughly $54K$ features with control group proportions derived from the Lung microbiome. Low-coverage and/or high compositional changes are problematic for current techniques due to the sparsity they cause in the count data. $W_1, \ldots W_3$ are Wrench estimators proposed in the Methods section that adjust the base estimator W_0 for feature-wise zero-generation properties. All are presented here for comparison purposes. Our default estimator is W_2.

Dataset	Type	CLR	TMM	CSS	Scran	W_0	W_1	W_2	W_3
Tara Oceans [10]	16s (from Whole Metagenome)	0 (-2.65×10^{-6})	0.26	0.15	0.52	.58	.54	.53	.53
Rat BodyMap [39]	Bulk RNAseq	-0.36	0.22	0.16	0.18	.20	.19	.20	.26
Embryonic Stem Cells [48]	UMI/scRNAseq	-0.70	.70	.67	.67	.71	.70	.70	.68

Table 3 Correlations of compositional scales with orthogonal measurements on absolute abundances/technical biases. Correlations of logged reconstructed abundance factors (1/compositional correction factor) with logged total flow cytometry cell counts is shown for the Tara project. Correlations of logged normalization factors with logged total ERCC counts are shown in the case of the rat body map and embryonic stem cells datasets. Given the high sparsity in these datsets, CLR factors computed by adding pseudocounts, essentially had no information on technical biases. $W_1, \ldots W_3$ are estimators proposed in the Methods section that adjust the base estimator W_0 for feature-wise zero-generation properties. All are presented here for comparison purposes. The default Wrench estimator (W_2) compares well at low and high coverage settings. For more details on these and the distinction in terminology between compositional correction factors and normalization factors, refer Materials and Methods.

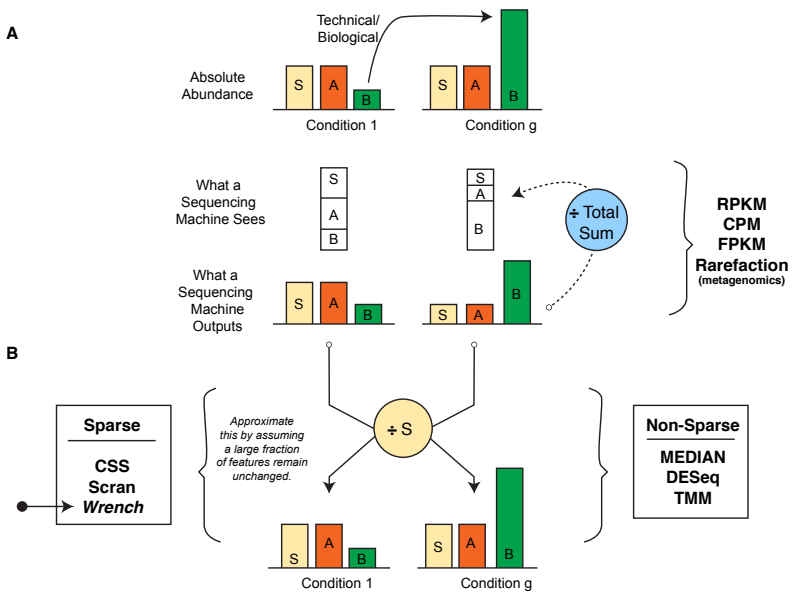

Figure 1 Scaling normalization approaches from the perspective of compositional correction. (**A**) Features S and A have similar absolute abundances in two experimental conditions, while B has increased in its absolute abundance in condition g due to technical/biological reasons. Because of the proportional nature of sequencing, increase in B leads to reduced read generation from others (compositional bias). An analyst would reason A and S to be significantly reduced in abundance, while, in reality they did not. (**B**) Knowing S is expressed at the same concentration in both conditions allows us to scale by its abundance, resolving the problem. DESeq and TMM, by exploiting rerefence strategies across feature count data (described below), approximate such a procedure, while techniques that are based only on library size alone like RPKM and rarefication/subsampling can lead to unbiased inference only under very restrictive conditions. Approaches available for sparse settings are indicated. *Wrench* is the proposed technique in this paper.

Additional Files

Additional file 1 — Supplementary Note

Presents further discussions on compositional bias, and supplementary results in context.

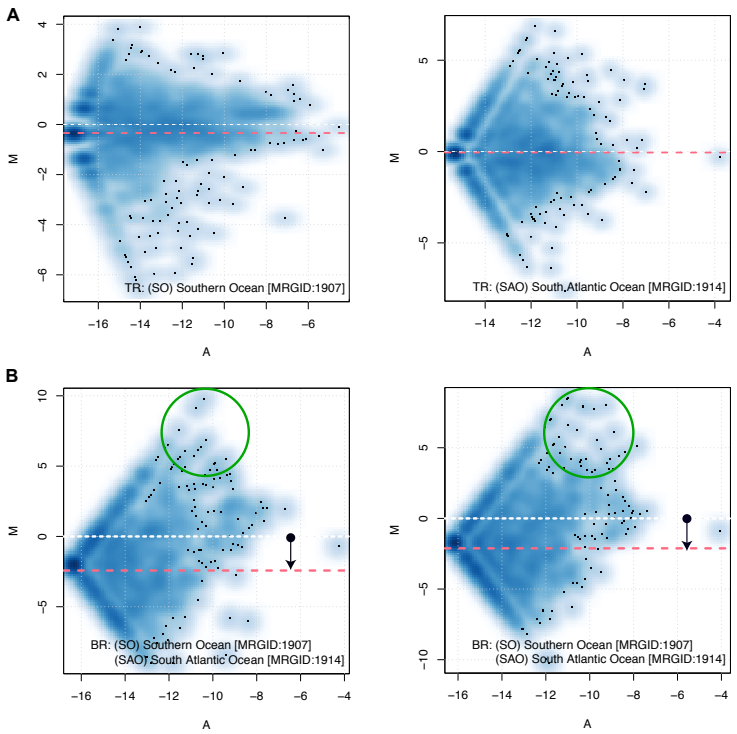

Figure 2 Importance of compositional bias correction in sparse metagenomic data. (A) M-A pots of 16S reconstructions (from high sequencing depth, whole metagenome shotgun sequencing experiments) from two technical replicates each from the Tara oceans project [10] generated for the Southern and South Atlantic Oceans. In all subplots, x-axis plots for each feature, its average of the logged proportions in the two compared samples; y-axis plots the corresponding differences. The red dashed line indicates the median log fold change, which is 0 across the technical replicates. (B) M-A plots of the same replicates but plotted across the two oceans. The median of the log-fold change distribution is clearly shifted. A few dominant taxa in the South Atlantic Ocean (circled) are attributable for driving this overall apparent differences in the observed fold changes. The Tara 16s dataset, reconstructed from very deep whole metagenome shotgun experiments of oceanic samples, albeit boasting of an average 100,000 16S contributing reads per sample, still encourages a median 88% feature absence per sample.

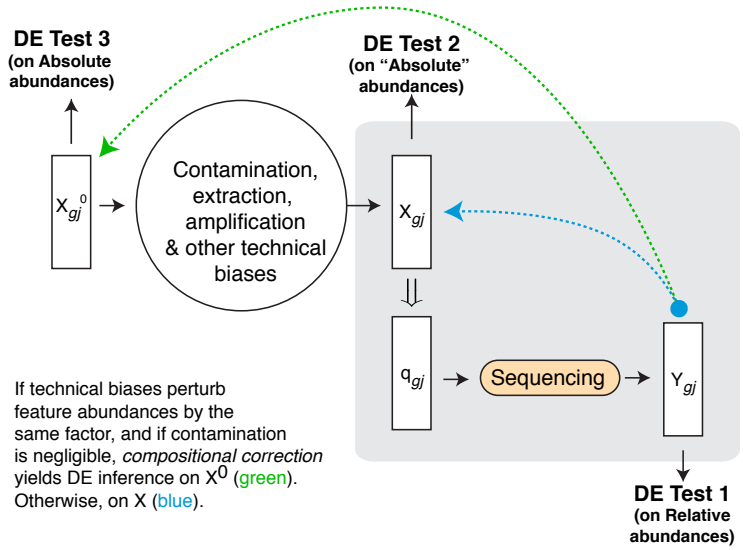

Figure 3 Compositional bias introduced by sequencing technology. As a sample j from group g of interest is prepared for sequencing, its true internal feature concentrations (organized as a vector) X^0_{gj} is transformed by various technical biases to X_{gj}. A sequencing machine introduces compositional bias by generating counts Y_{gj} proportional to the input absolute abundances in X_{gj} according to proportions $q_{gj} = [\ldots x_{gji}/(\sum_k x_{gjk})\ldots]$, i and k indexing features. Directly performing a differential abundance test on Y (DE Test 1), by using normalization factors proportional to that of total sequencing output (ex: R/FPKM/subsampling in metagenomics) amounts to testing for changes in *relative* abundances of features in X, in general (*not* X^0). For inferring differences in absolute abundance, we need to reconstruct X^0 from Y to perform our inference (DE Test 3). For compositional bias correction in particular, we care about reconstructing X_j from Y (DE Test 2). We show more formally later that compositional correction can reconstruct X^0 if technical biases perturb all feature abundances by the same factor, and that the presence of sequence-able contaminants induces more stricter assumptions behind their application.

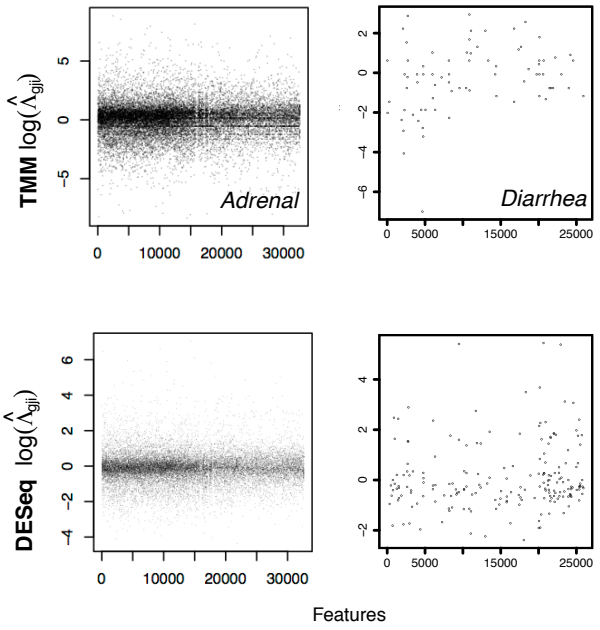

Figure 4 Estimation of compositional correction scales from sparse count data. On the left column, we plot the feature-wise ratio (Λ_{gji}) estimates adjusted for sample depth from each feature i in one of the samples from the Adrenal tissue of the rat body map dataset (bulk RNAseq), and on the right column, we plot the same values arising from a sample in the Diarrheal dataset (16S metagenomics). The top and bottom rows correspond to the scales estimated using TMM and DESeq respectively. In the case of bulk RNAseq data, large numbers of individual feature estimates agree on a compositional scale factor. Simple averaging, or some robust averaging would help us obtain the scale factor exactly. A similar robust behavior is observed with all the tissues available in the bodymapRat dataset (considered later in text). On the second column, we plot the feature-wise ratio values from a metagenomic 16S marker gene survey of infant gut microbiota. There is no general agreement among the features on the scale factors, and simple averaging will not work. We note that what we have shown are fairly good cases. Several samples entertain only a few tens of shared species with an arbitrary reference sample within the dataset. In this work, we aimed to model this variability and estimate the scale factors robustly by borrowing information across features and samples.

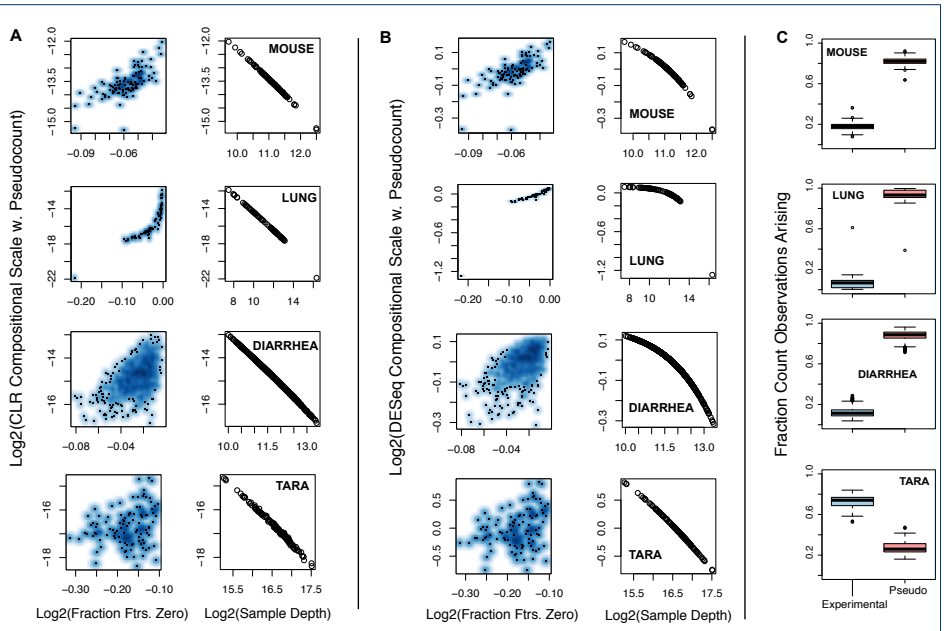

Figure 5 Adding pseudocounts leads to biased normalization. For each of the four microbiome count datasets (rows: Mouse, Lung, Diarrheal and Tara Oceans), we plot (**A**) CLR and (**B**) DESeq compositional scales obtained after adding a pseudo count value of 1, as a function of fraction of features that are zero in the samples (first column) and the sample depth (second column). The observed behavior was not sensitive to the value of pseudocount used. Refer supplementary Fig. 7 for the same plot for a pseudocount value of 10^{-7}. (**C**) shows the total number of pseudocounts added, which is essentially the number of features observed in a dataset, and the total actual counts observed in the dataset divided by their sum i.e., the total implied sequencing depth after pseudocounts addition. A large fraction of sequencing depth in the new pseudocounted dataset is now arising from pseudocounts than the true experimental counts, when the data is excessively sparse. Indeed, if the pseudocount value is altered to a very low positive fraction value, the boxplots will reflect reversed locations, but this plot is only used to stress the level of alteration made to a dataset. Only in the Tara Oceans project, where the sample depth is 100K reads, do the boxplots shift. However, at a roughly median 90% features absent, that data when altered by pseudocounts, also leads to biased scaling factors as seen in (A) and (B).

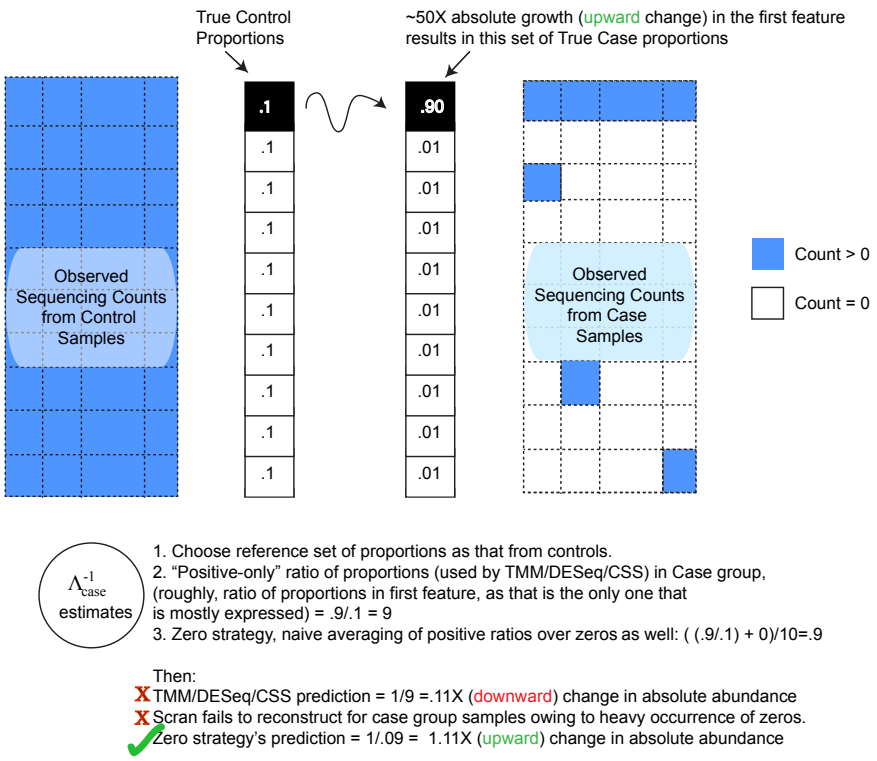

Figure 6 Ignoring zeroes can introduce bias in normalization, when zeroes predominantly arise from under-sampling. An artificial example with 10 features and two groups ("controls" and "cases"), when one of the features undergoes a roughly 50X expansion (a $\log_2$ fold change of 5.64) in cases compared to controls. This drives the relative abundances of the rest of the 9 features relatively low in the case group. As a result features that are largely present in the controls are not observed in the case group at moderate sequencing depths. Scaling normalization strategies that derive scales based only on the positive count values, can underestimate compositional changes as shown.

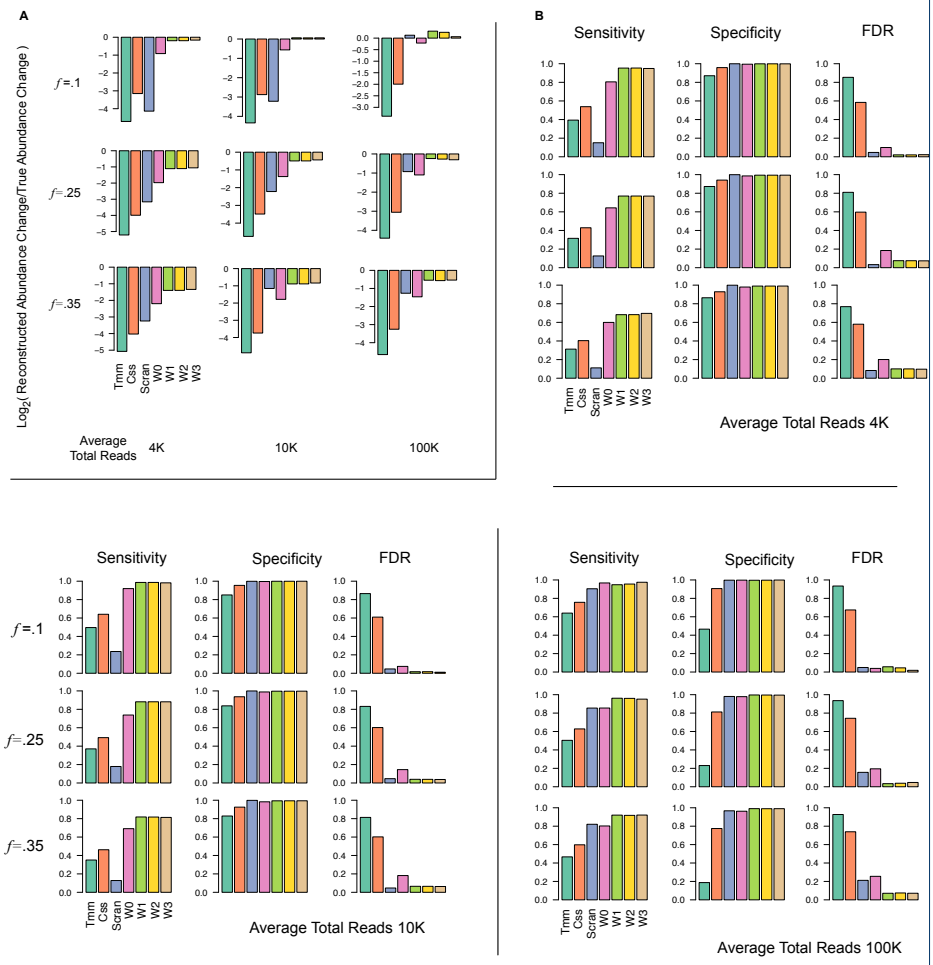

Figure 7 Wrench scales outperform competing approaches in reconstructing compositional changes and in differential abundance testing. Multiple iterations of two group simulations are simulated with various fractions of features perturbed across conditions (rows, f in figures), total number of reads. Their average accuracy metrics in reconstruction and differential abundance testing are plotted. The control proportions were set to those obtained from the mouse microbiome dataset. (A) Average log ratios of reconstructed to true absolute abundance changes. Each row corresponds to a particular setting of f, and each column a particular setting of average sequencing depth. Scran also suffered from being unable to provide scales for samples in each simulation set (sometimes as high as 60% of the samples at 4K and 10K average reads). (B) Average sensitivity, specificity and false discoveries at FDR .1 of detecting true differential absolute abundances. W_0 is the regularized Wrench estimator without sparsity adjustments and $W_1, ..W_3$ are various adjusted estimators compared here. For details on this and simulations, see Methods. Behavior was similar for other parameteric variations (variances of global and sample-wise fold change distributions, number of samples) of simulations.

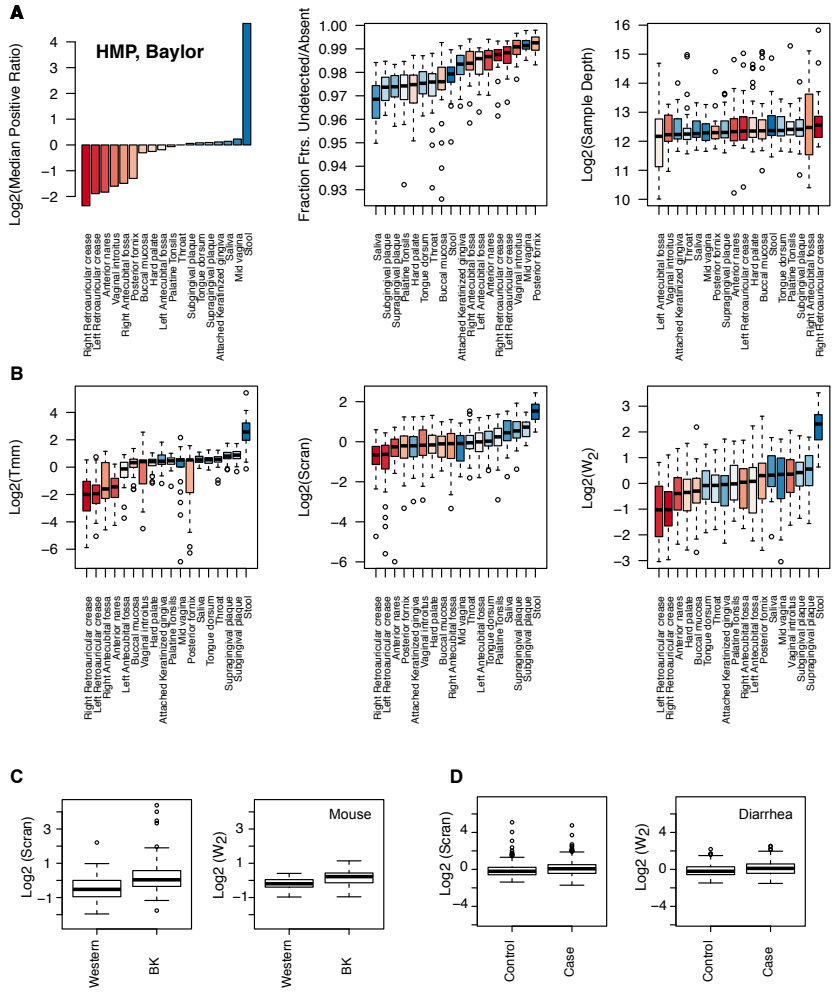

Figure 8 Wrench retains potential biological information, and indicates importance of compositional correction in general practice. We plot some statistical summaries and the compositional scale factors reconstructed by a few techniques for various Human Microbiome Project samples, sequenced at the Baylor College of Medicine. (**A**) On the top-left, we plot the logged median of the positive ratios of group-averaged proportions to that of *Throat* chosen as the reference group. Stool samples show considerable deviation from the rest of the samples despite having comparable fraction of features detected and sample depths to other body sites. Notice the log scale. (**B**) The similarity in the reconstructed scales across techniques (second row) for closely related body sites are striking; although minor variations in the relative placements were observed across centers potentially due to technical sources of variation, the overall behavior of highly significant differences in the scales of behind-ear and stool samples were similar across sequencing centers (supplementary Fig. 10) and normalization methods. Corresponding CSS scales in supplementary Fig. 11. These techniques predict a roughly 4X-8X (ratio of medians)inflation in the Log2-fold changes when comparing abundances across these two body sites. (**C**) Wrench and scran compositional scale factors across the plant-based diet (BK) and Western diet (Western) mice gut microbiome samples. (**D**) Compositional scale factors for healthy (Control) and diarrhea afflicted (Case) children. Slight differences in the compositional scales are predicted in the diet comparisons with t-test p-values $<$ 1e-3 for all methods except TMM, but not as much in the diarrheal samples.

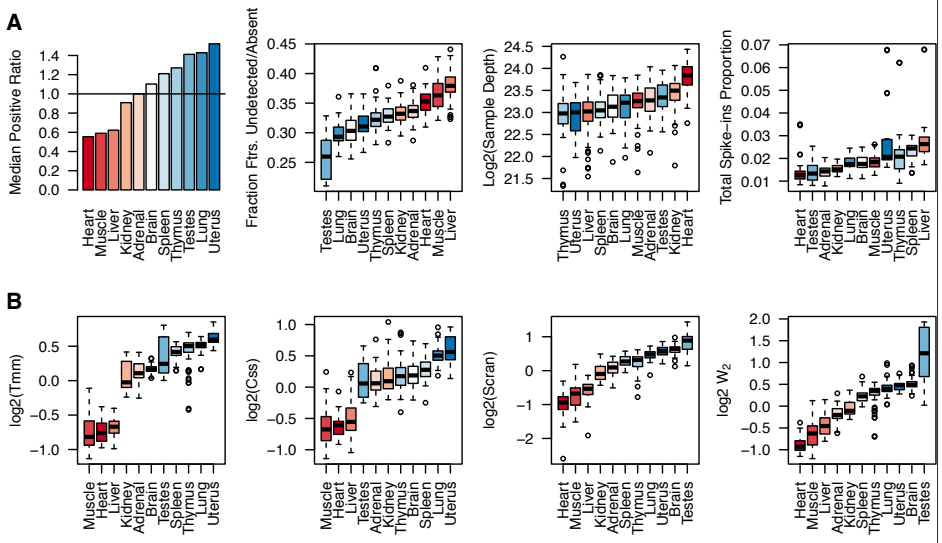

Figure 9 Importance of compositional correction in common bulk RNAseq studies. (**A**) Application of scaling techniques to the rat body map data across tissues. Median positive ratio: median of the positive ratios of group-averaged proportions to that of Adrenal chosen as the reference. Subsequent figures in the top row indicate higher sparsity levels in the heart, muscle and liver samples, although at sequencing depths that are comparable/slightly higher to those from other tissue groups. (**B**) Reconstructed scales from several normalization techniques. If one were to perform a differential expression analysis between Testes and Heart, the fold changes are roughly 4X (ratio of medians) inflated as predicted by Scran/Wrench, which can lead to high false positive rates especially if most features are not changed across the two tissues.Notice the similarity in scales for closely related tissues, across techniques; for these tissues, the influence of compositional bias in the related differential abundance tests will be low.