BMC Genomics

METHODOLOGY ARTICLE Open Access

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]

Abstract

Background: Count data derived from high-throughput deoxy-ribonucliec acid (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. Commonly used count data normalization approaches like library size scaling/rarefaction/subsampling cannot correct for compositional or any other relevant technical bias that is uncorrelated with library size.

Results: 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.

Conclusions: Compositional bias, induced by the sequencing machine, confounds inferences of absolute abundances. We present a normalization technique for compositional bias correction in sparse sequencing count data, and demonstrate its improved performance in metagenomic 16s survey data. Based on the distribution of technical bias estimates arising from several publicly available large scale 16s count datasets, we argue that detailed experiments specifically addressing the influence of compositional bias in metagenomics are needed.

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 [1–4]. 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].

This remarkable flexibility of sequencing comes with atleast one tradeoff. As noted previously in the literature [11–14] (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

*Correspondence: smuthiah@umiacs.umd.edu
[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
Full list of author information is available at the end of the article

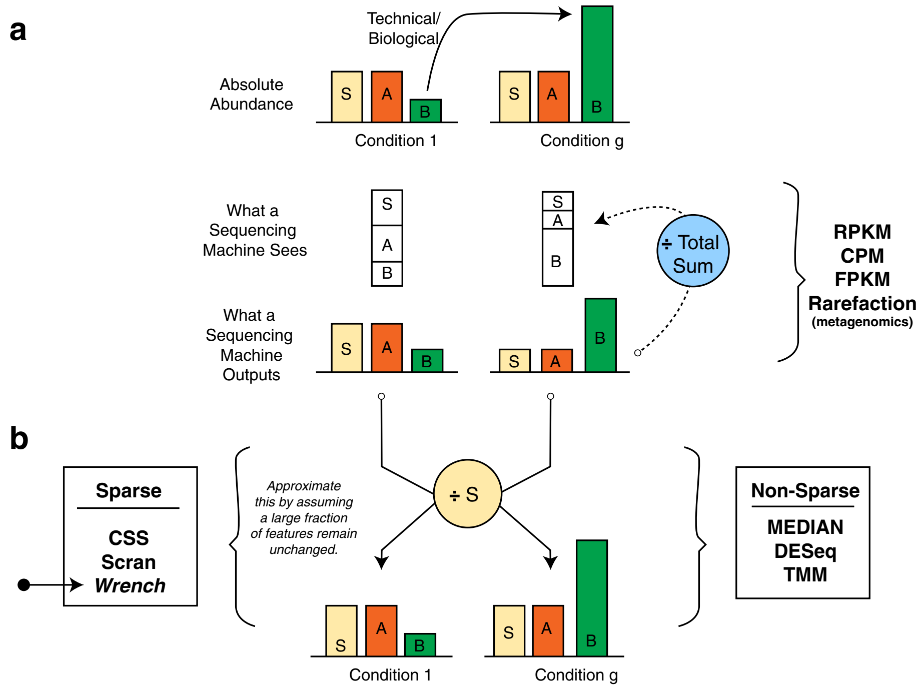


Fig. 1 Scaling normalization approaches from the perspective of compositional correction. **a** Features *S* and *A* have similar absolute abundances in two experimental conditions, while *B* has increased in its absolute abundance in condition *g* due to technical/biological reasons. Because of the 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

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 [13, 15–17], 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 [18–24]. 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 ribonucleic acid sequencing (scRNAseq) 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,

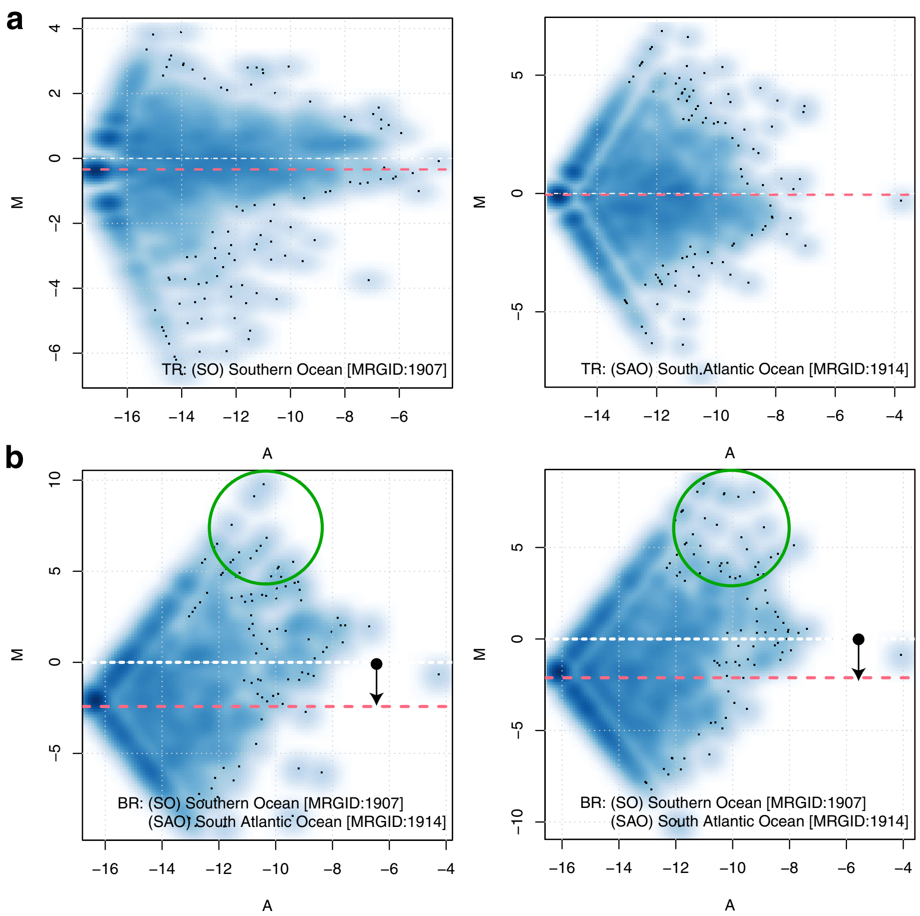


Fig. 2 Importance of compositional bias correction in sparse metagenomic data. **a** M-A pots of 16S reconstructions (from high sequencing depth, whole metagenome shotgun sequencing experiments) from two technical replicates each from the Tara oceans project [10] generated for the Southern and South 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

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 [27–30], compositional bias correction becomes necessary even when analysis is performed on relative abundances. For instance, in metagenomic 16s rRNA surveys, taxonomy specific biases in the count data can arise by variation in rRNA extraction effeciencies [31, 32], PCR primer binding preferences and the target rRNA GC content [33], all which cause differential amplification across the surveyed taxa.

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 [34], rarefication/subsampling in metagenomics [35, 36]) 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 showing that it can lead to reduced false positives and rich annotation discoveries. 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 on average, they would all suffer in analyses of features arising from arbitrary general conditions. In such cases, spike-in based techniques can be effective [37], although methods similar to the External RNA Control Consortium (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.

Figure 3 illustrates a general sequencing experiment and sets up the problem of compositional bias correction. We imagine a set of observations $j = 1 \ldots n_g$ arising from conditions $g = 1 \ldots G$ (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 $T_{gj} = \sum_i X_{gji} = X_{gj+}$, where the $+$ indicates summing over that subscript. 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 $X^0 \rightarrow X$), 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 $X_{gj\cdot}|T_{gj} \sim Multinomial\left(T_{gj}, q_{g\cdot}\right)$, 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} [38–40]). Similarly, we shall assume the observed counts $Y_{gj\cdot}|X_{gj\cdot}, \tau_{gj} \sim Multinomial\left(\tau_{gj}, \frac{X_{gj\cdot}}{T_{gj}}\right)$, τ_{gj} is the corresponding sampling depth. Notice that marginally, $E[Y_{gji}|\tau_{gj}] = q_{gi} \cdot \tau_{gj}$, and hence averaging the observed sample-wise proportions $\hat{q}_{gji} = Y_{gji}/\tau_{gj}$ in group g for feature i yields the marginal expectation $E\left[\overline{\hat{q}_{g+i}}\right] = q_{gi}$. We shall use $E\left[T_{g1}\right]$ to denote the average (across samples) *total* absolute abundance of features in group g at the time of input. Similarly, $E\left[\overline{X_{g+i}}\right]$ 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, $\nu_{gi} = \frac{E[\overline{X_{g+i}}]}{E[\overline{X_{1+i}}]}$, 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:

$$\nu_{gi} = \frac{E\left[\overline{X_{g+i}}\right]}{E\left[\overline{X_{1+i}}\right]} = \frac{E\left[T_{g1}\right] q_{gi}}{E\left[T_{11}\right] q_{1i}} \equiv \Lambda_g \cdot \frac{q_{gi}}{q_{1i}} = \Lambda_g \cdot \frac{E\left[\overline{\hat{q}_{g+i}}\right]}{E\left[\overline{\hat{q}_{1+i}}\right]} \quad (1)$$

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^{-1} can be derived based on an often quoted

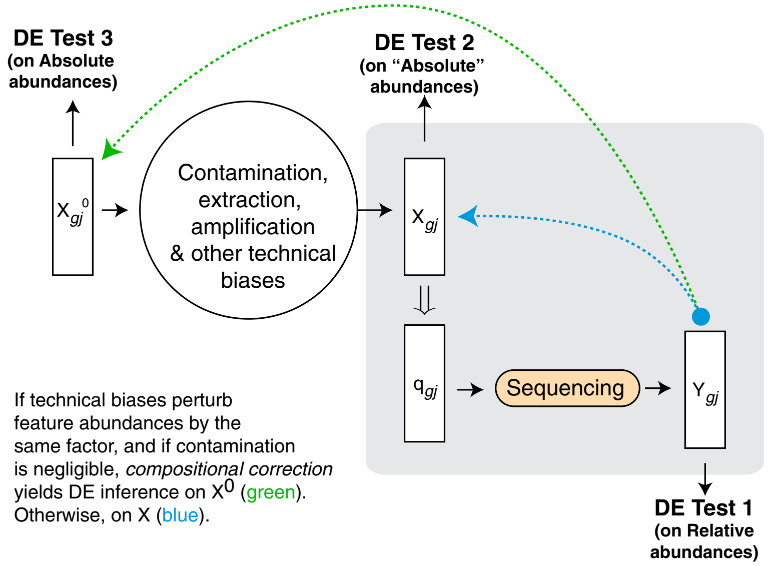

DE Test 3
(on Absolute abundances)
DE Test 2
(on "Absolute" abundances)
X_{gj}^0
Contamination, extraction, amplification & other technical biases
X_{gj}
q_{gj}
Sequencing
Y_{gj}
If technical biases perturb feature abundances by the same factor, and if contamination is negligible, *compositional correction* yields DE inference on X^0 (green). Otherwise, on X (blue).
DE Test 1
(on Relative abundances)

Fig. 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_{gj}^0 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 $q_{gj} = [\ldots x_{gji}/\left(\sum_k x_{gjk}\right)\ldots]$, *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

assumption behind scale normalization techniques [13]: if most features do not change in an experimental condition relative to the control group, Eq. (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}.

So far we have discussed estimating group-specific compositional factors. With this idea in place, a normalization procedure for deriving sample-specific compositional scale factors can be devised. One only needs to carry out the above procedure by pretending that every sample arises from its own experimental group. 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_{gj}^{-1} estimate is derived for every sample *j* based on the ratio of its proportions to that of the reference. 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 (Additional file 1: Section 3, Eq. (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 bad and that other scaling techniques face certain difficulties with sparse data.

Table 1 Scaling normalization approaches derive their technical bias estimates from ratio of proportions

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

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 (Additional file 1: Section 1). Under the most natural adjustments based on the total count (e.g., unaltered reads per kilobase of transcript, per million mapped reads (RPKM)/ fragments Per kilobase of transcript per million mapped reads (FPKM)/ Counts per million (CPM)/subsampling/rarefication based approaches), we found that these conditions can be precisely characterized mathematically and are extremely limited in their applicability in general experimental settings (Additional file 1: Theory Section 1). 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 Additional file 1: 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 [41–43], 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 at high coverage sample depths (similar to bulk RNAseq) 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 (Additional file 1: 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 [44]; the second corresponds to those from a 16S metagenomic dataset [45]. 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, 46]) 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 [47] 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 Bayesshrink 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 [48] and (the very high coverage) Tara oceans [10] 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 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

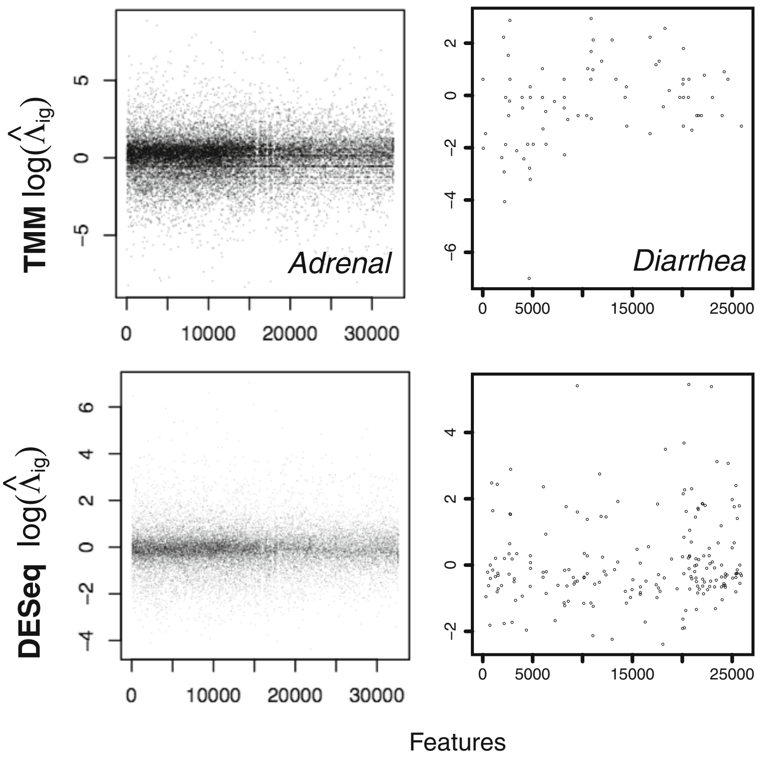


Fig. 4 Estimation of compositional correction scales from sparse count data. On the left column, we plot the feature-wise ratio (Λ_{gji}) estimates adjusted for sample depth from each feature *i* in one of the samples from the Adrenal tissue of the rat body map dataset (bulk RNAseq), and on the right 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

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. For the purpose of metagenomic applications, and analytic convenience, we slighty relax the standard assumption that most features do not change across conditions by assuming that the feature-wise log-fold changes arise from a zero mean Gaussian distribution, a common assumption in differential abundance analysis [26, 49, 50]. 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. To generate Table 2, 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 of scales they can reconstruct. 2) Scran can yield accurate estimators at very large

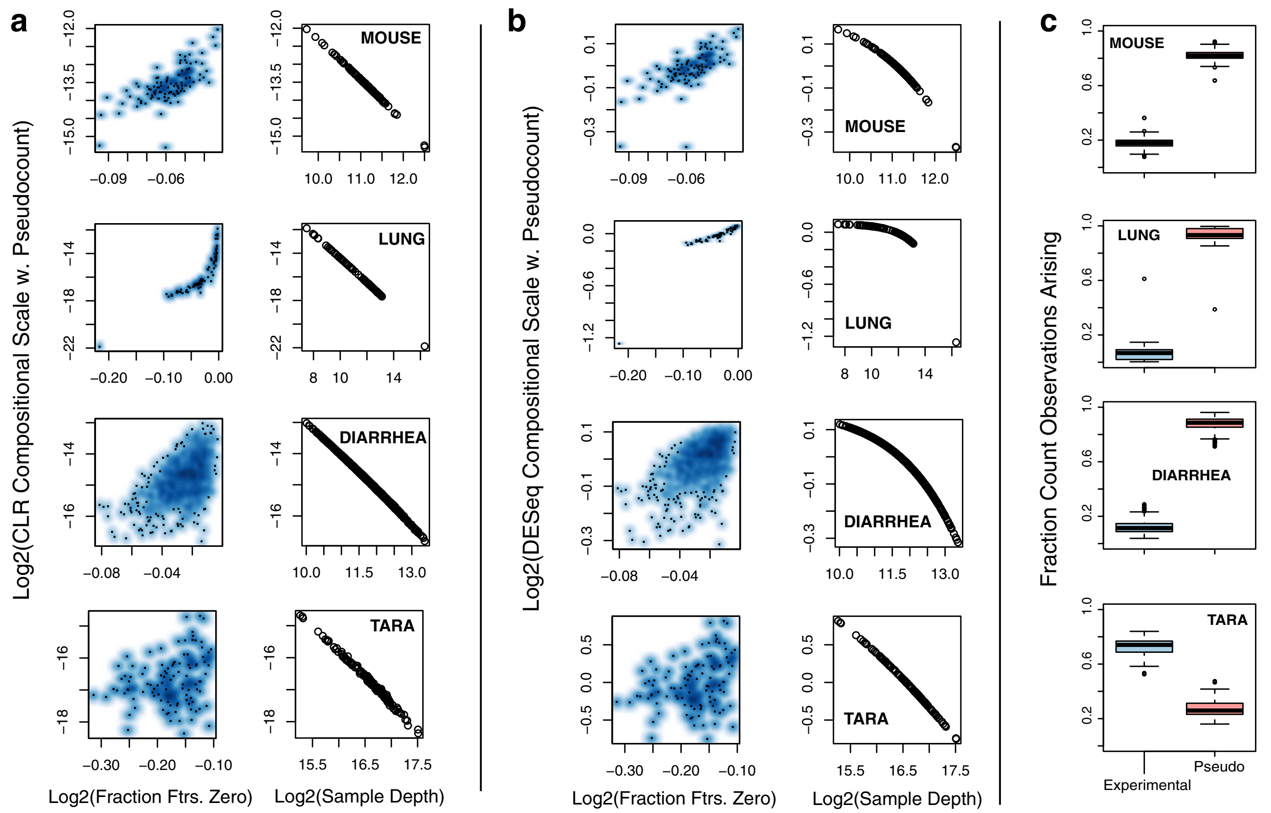


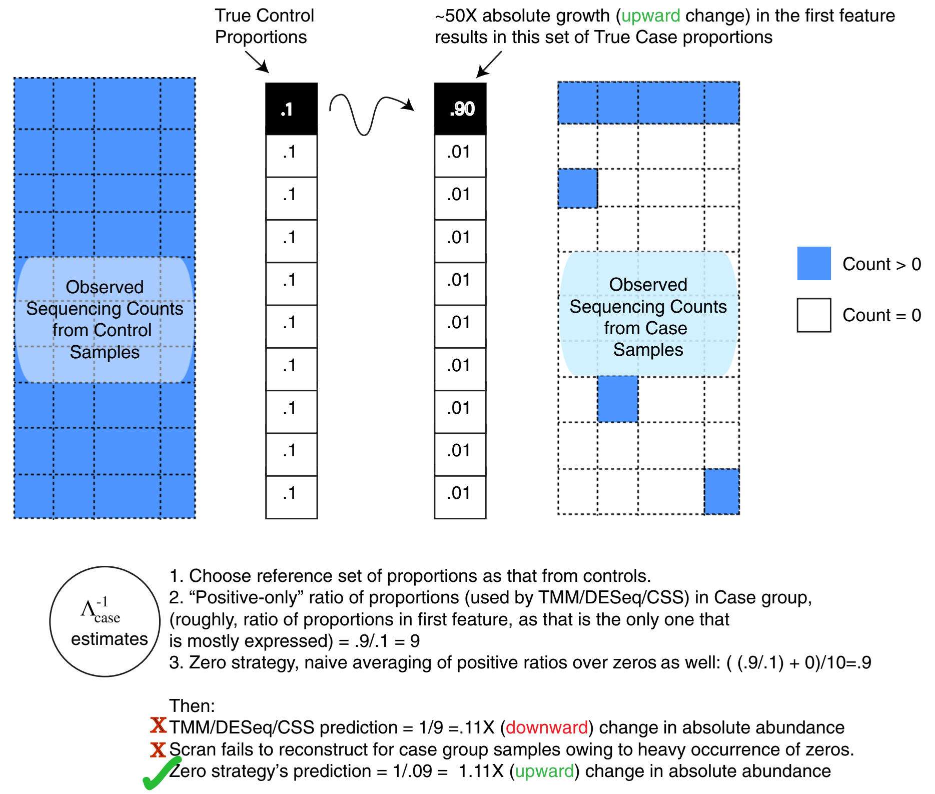
Fig. 5 Adding pseudocounts leads to biased normalization. For each of the four microbiome count datasets (rows: Mouse, Lung, Diarrheal and Tara Oceans), we plot **a** CLR and **b** DESeq compositional scales obtained after adding a pseudo count value of 1, as a function of fraction of features that are 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 Additional file 1: Figure S7 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**

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 against compositional bias at higher coverages. For specific comparisons with pseudocounted CLR, please refer Additional file 1: Figure S9, in which we show the proposed technique (Wrench) performing significantly better. In addition, Additional file 1: Figure S21, and Additional file 1: Figure S22 explore simulation performance as a function of group-wise sample size in balanced and unbalanced designs, where we find the performance to stabilize between roughly 10−20 samples, depending on the fraction of features that change across conditions.

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 has better normalization accuracy in experimental data

Below, we show five different results illustrating the improvements Wrench offers over existing techniques in experimental data. The first two show that Wrench leads


True Control Proportions
~50X absolute growth (upward change) in the first feature results in this set of True Case proportions
.1
.1
.1
.1
.1
.1
.1
.1
.1
.1
.90
.01
.01
.01
.01
.01
.01
.01
.01
.01
Observed Sequencing Counts from Control Samples
Observed Sequencing Counts from Case Samples
🟦 Count > 0
□ Count = 0
$\Lambda^{-1}_{\text{case}}$ estimates

1. Choose reference set of proportions as that from controls.
2. "Positive-only" ratio of proportions (used by TMM/DESeq/CSS) in Case group, (roughly, ratio of proportions in first feature, as that is the only one that is mostly expressed) = .9/.1 = 9
3. Zero strategy, naive averaging of positive ratios over zeros as well: ((.9/.1) + 0)/10=.9

Then:

✗ TMM/DESeq/CSS prediction = 1/9 =.11X (downward) change in absolute abundance
✗ Scran fails to reconstruct for case group samples owing to heavy occurrence of zeros.
✓ Zero strategy's prediction = 1/.09 = 1.11X (upward) change in absolute abundance

Fig. 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

to reduced false positive calls in differential abundance inference, while the other three demonstrate the improved quality of positive associations.

Reduction of false positives We used two approaches to compare the performance of Wrench in reducing false positive calls in differential abundance inference. Each of these analyses was performed across all biological groups with atleast 15 samples in the mouse (2 diet types), Diarrheal (2 groups), Tara (5 oceans), HMP (JCVI, 16 body sites), and HMP (BCM, 16 body sites) and averaged the results across these 41 experimental groups.

We ignored the lung microbiome for these analyses as Scran had particular difficulty making direct comparisons hard. 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. Notice that Wrench not only recovers compositional scales for these samples, but also at magnitudes that were coherent with other samples from similar experimental groups (see next subsection) indicating some validity for the computed normalization factors.

First, a standard resampling analysis was performed. For every given experimental group, two artificial groups are

Table 2 Example simulations illustrate the limitations of current techniques

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

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 54*K* 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

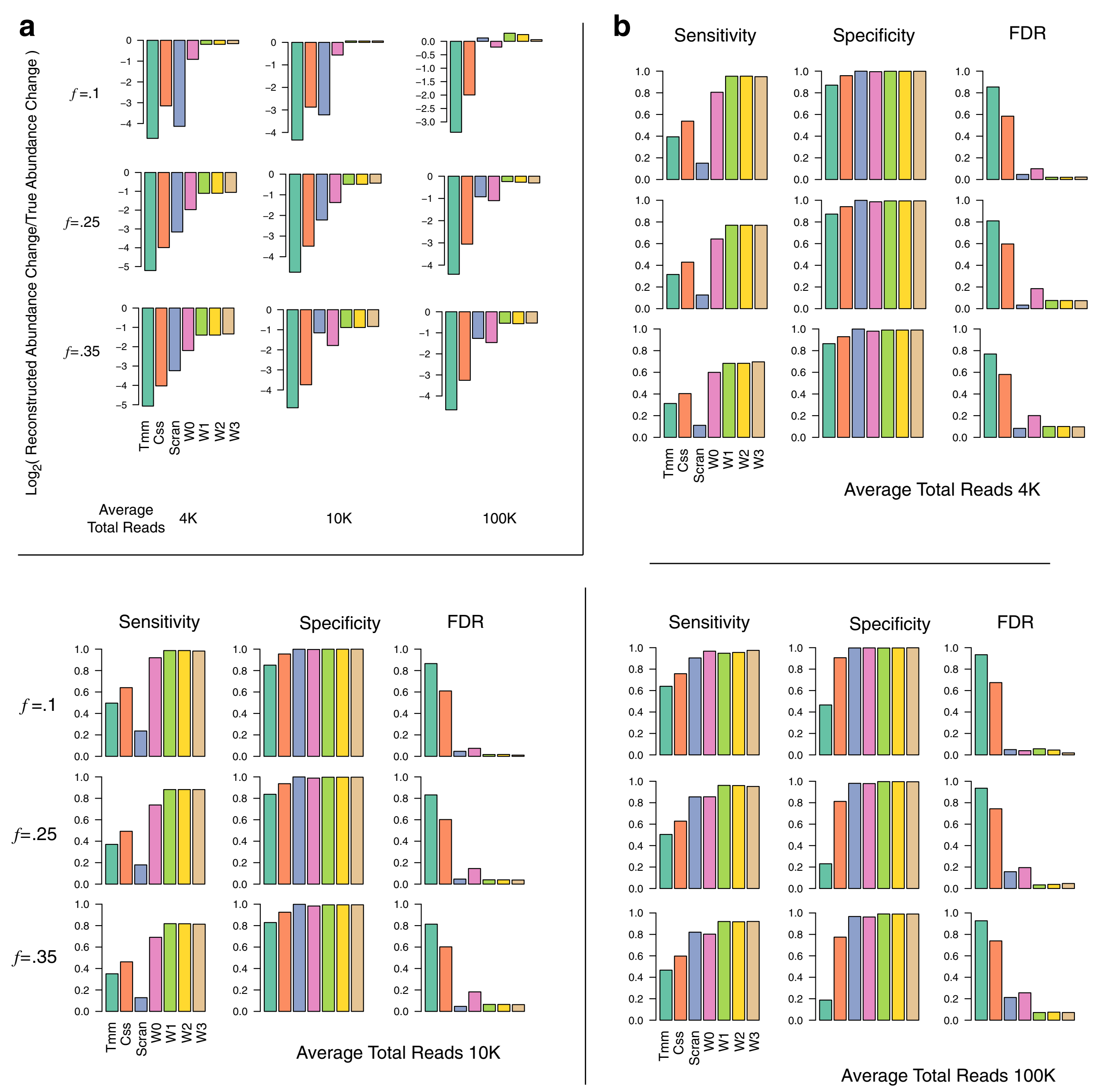


Fig. 7 Wrench scales outperform competing approaches in reconstructing compositional changes and in differential abundance testing. Multiple iterations of two group simulations are simulated with various fractions of features perturbed across conditions (rows, f in figures), total number of reads. Their 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 parametric variations (variances of global and sample-wise fold change distributions, number of samples) of simulations

repeatedly constructed via resampling (without replacement), and the total number of significant calls made during differential abundance analysis is recorded in each repetition. For each iterate, we compute the $\log_2(F_{Other}/F_{Wrench})$ ratio, where F_{Other} is the total number of significant calls made by a competing method (Total Sum / TMM / Scran / CSS) and F_{Wrench} is the total number of significant calls made by Wrench. If Wrench is superior

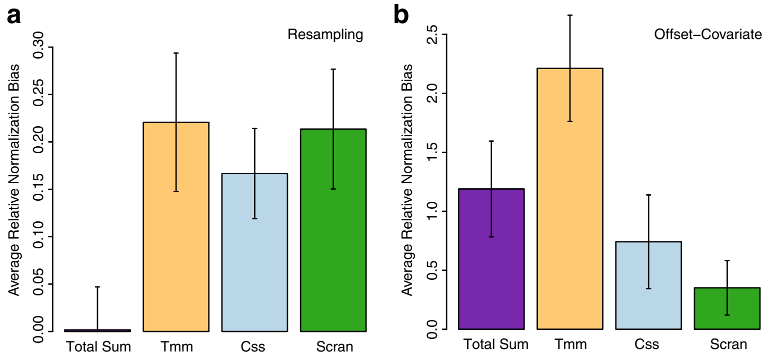


Fig. 8 Wrench scales lead to reduced false positive calls. **a** The average of $\log_2(F_{Other}/F_{Wrench})$ values obtained over artificial two group splits of homogeneous experimental group data is shown and **b** the average of $\log_2(C_{Other}/C_{Wrench})$ values across 41 metagenomic experimental groups are shown. Standard error bars are shown. In both plots, positive values for a method imply reduced accuracy relative to Wrench. F_{Other}: total number of diffferentially abundant features found by a competing method (total sum, TMM, CSS or Scran). F_{Wrench}: total number of differentially abundant features found by Wrench. C_{Other}: total number of features where the covariate term for Wrench normalization factors were found to be significant when competing method is used as offset. C_{Wrench}: total number of features where the covariate term for a competing method's normalization factors were found to be significant, when Wrench is used as covariate

these logged ratios should be $>$ 0. The average of these ratios across all the experimental groups mentioned above is plotted in Fig. 8a, and we find Wrench meeting the goal. Although total sum does not show a significant difference in this analysis, as illustrated next, it is insufficient in capturing the null variation in the data.

We next exploited the offset-covariate approach introduced in [25]. For every feature/OTU within a homogenous experimental group, two generalized linear models are fitted: in model (a) Wrench normalization factors as offset, and those of a competing method as covariate. In model (b), normalization factors from a competing method as offset, and those of Wrench as covariate. The number of features for which the covariate term was called significant is recorded in both (a) and (b). We will denote them respectively as C_{Wrench} and C_{Other}. If Wrench sufficiently captures the variation in data, the number of times the covariate term from a competing method is called significant will be low. That is: the logged ratio $\log_2(C_{Other}/C_{Wrench})$ must be $>$ 0. The average of these values across all the experimental groups mentioned above is plotted in Fig. 8b, and we find Wrench to improve upon other techniques.

Improved association discoveries To compare the quality of associations achieved with the various normalization methods, we re-analyzed the Tara Oceans 16S microbiome dataset.

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 spike-in 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 are advantages to exploiting specialized compositional correction tools even with microbiome datasets teeming with microbes of extraordinary diversity.

We next analyzed the quality of differential abundance inference arising from competing normalization techniques, by performing two sets of enrichment analyses.

In the first procedure, we extracted broad genus-level functional annotations from the Faprotax database [51], and tested for their enrichment in positively associated genera in the deep chlorophyll (DCM) and the mesopelagic layer (MES) samples of the oceans relative to the surface layer. The total number of significantly differentially abundant OTU calls were widely different across techniques: Wrench and Scran made roughly 30% fewer calls compared to total sum, TMM, and CSS. Given the relatively general nature of the annotations, all methods yielded expected annotations in the DCM and MES layers based on previous studies, although there were a few differences (Additional file 2). Nitrite respiration/reduction/anoxygenic phototropy, oil

Table 3 Correlations of compositional scales with orthogonal measurements on absolute abundances/technical biases

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 [44]	Bulk RNAseq	-0.36	0.22	0.16	0.18	.20	.19	.20	.26
Embryonic stem cells [62]	UMI/scRNAseq	-0.70	.70	.67	.67	.71	.70	.70	.68

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. Bland-Altman plots for the data underlying these numbers are presented in Additional file 1: Figure S18–S20, and related discussions in Additional file 1: Section 9

bioremediation were found enriched in mesopelagic layer by all methods, while methanogenesis, a function that is usually associated with mesopelagic and deep sea microbes [10, 51–54] was not found enriched in MES by total sum. Both Wrench and Scran did not find xylanolysis to be enriched in the mesopelagic layer, while other methods did. We were unable to find literature evidence supporting this call, and the result could potentially be due to the higher number of OTUs called differentially abundant by the other methods. Aerobic ammonia/nitrite oxidation and fixation were found to be enriched in DCM by all methods. Total sum and TMM found a methanogenesis related module enriched in DCM, while other methods did not.

To evaluate the methods in a more fine-grained setting, we devised the following validation approach. The design of the Tara oceans experiments - where 16S reconstructions are obtained from whole metagenome shotgun sequencing data - makes the following analysis feasible. Because the Tara project's functional (gene content summarized as Kegg Modules, KMs) and 16S data arise from the same input DNA samples, the same compositional factors should apply for both datatypes. We therefore estimated compositional factors from 16S data using the different normalization methods and applied the resulting estimates to the KM abundance data from the corresponding matched samples. Next, we computed Spearman rank correlation between OTU and KM normalized abundances and annotated OTUs with those KMs which showed correlation of at least 0.75. Finally, we identified OTUs that were positively associated with each layer using differential abundance analysis. With the KM annotations in place, we performed Fisher exact tests to compute the enrichment scores in the identified OTUs. Detailed tables are provided in Additional file 2. In mesopelagic samples, Scran finds enrichment in only 30 KMs, while other methods recovered at least 100 KMs. Specifically, ureolysis, motility, several denitrification/methanogenesis processes and aminoacid biosynthetic/transport mechanisms (functions that have been attributed to microbes in the mesopelagic layer and deep sea) [10, 51, 55, 56], were missed by Scran, while Wrench finds them. On the other hand, Total sum, TMM and CSS found more varied and general processes including various ribosomal, transcription/translation components to be enriched in both MES and DCM layers.

Notice that the first analysis gives a broad sense of the genera identified by the competing methods in light of existing annotations, while the second gives a sense of the quality of annotations one might confer on the OTUs based on the normalized expression levels of OTUs and the measured functional content themselves. In both cases, Wrench is shown to retain relevant information, and the relatively more specific nature of the latter analysis reveals that Wrench demonstrably improves upon other methods.

Inferences following compositional correction show improved coherence with experimental data

We further demonstrate the impact of compositional bias in downstream inference below. The experimental cell density measurements in the Tara Oceans project show a highly significant overall reduction in the mesopelagic samples when compared the surface layer (see Fig. 3 in ref [10]). Thus, we expect an overall negative change in the reconstructed fold changes, when performing a differential abundance analysis of the OTUs across these two ocean layers.

Summing the log-fold changes of significantly associated OTUs (both positive and negative) serves as a measure of a net change experienced by a community. If a given method produces fold change inferences that track the above mentioned empirical cell density measurements, we expect it to yield an overall negative net change value for the significantly differentially abundant OTUs in the mesopelagic community. As illustrated in Fig. 9a, this value for total sum normalized data is +10577.99, while that for Wrench is −8919.65, showing that differential abundances arising from Wrench agrees more appropriately with the underlying community change. Figure 9b and c, show how these values distribute across the major phyla focussed in the Tara oceans article. These plots

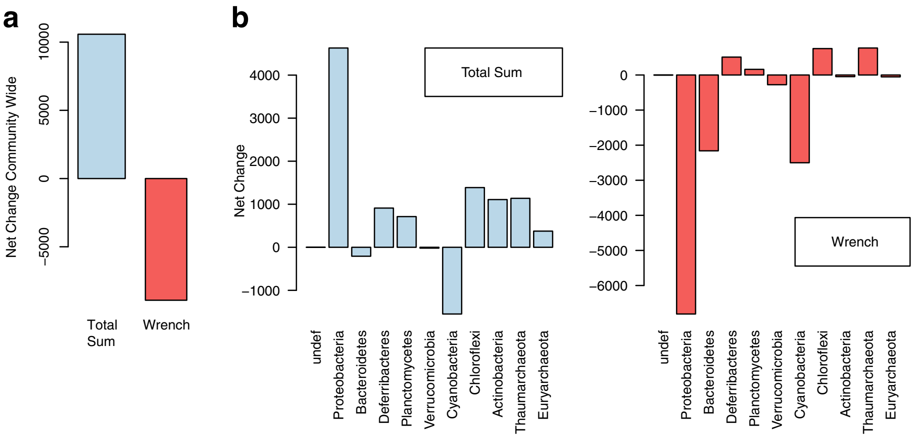


Fig. 9 Wrench normalized data lead to better downstream inferences. **a** The sum of log-fold changes of differentially abundant OTUs is used as a measure of net change experienced by a community. This value is plotted for the differentially abundant OTUs in the mesopelagic ocean layer relative to the surface layer in the Tara oceans 16S data, for Total Sum and Wrench normalization. **b** The same metric plotted for various major phyla of interest in the Tara oceans project

demonstrate that the two approaches lead to markedly different conclusions on the net change experienced by a phylum. In particular, Proteobacteria, Actinobacteria, Euryarchaeota were predicted to have drastically high positive changes by total sum (while Wrench predicts a marked decrease in the negative direction), and sizable differences were apparent in the values obtained with the rest of the phyla.

Compositional scale factor estimates imply substantial technical biases, indicating importance of further experimental studies

We next analyzed the phenotypic integrity of the compositional scales reconstructed by the various methods. In the absence of technical biases, following our discussion in the previous subsection, compositional factors should hover around 1 (upto some arbitrary scaling). This is *not* what we observe in samples from metagenomic datasets. All scale normalization techniques resulted in group-wise integrity in the scales they reconstructed within and across related phenotypic categories, potentially indicating the general importance of correcting for confounding induced by compositional bias in general practice. Total sum normalization is oblivious to these biases, making further experimental studies on compositional bias important. For instance, in the microbiome samples arising from the Human Microbiome Project [48], as shown in Fig. 10a, 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. 10b), behind-ear and stool samples were distinctly located in terms of their compositional scales from the oral and vaginal microbiomes (notice the log scale in these plots). This behavior was also recapitulated in scales reconstructed from other centers. Additional file 1: Figure S10 and S11 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. 10c, and Additional file 1: Figure S12). In the lung microbiome, the lung and oral cavities had roughly similar scales across smokers and non-smokers (Additional file 1: Figure S13), 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. 10d).

For completeness, we also attach similar results from all the 11 organs of the rat body map dataset in the Additional file 1: Figure S15.

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 [14, 57]. 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

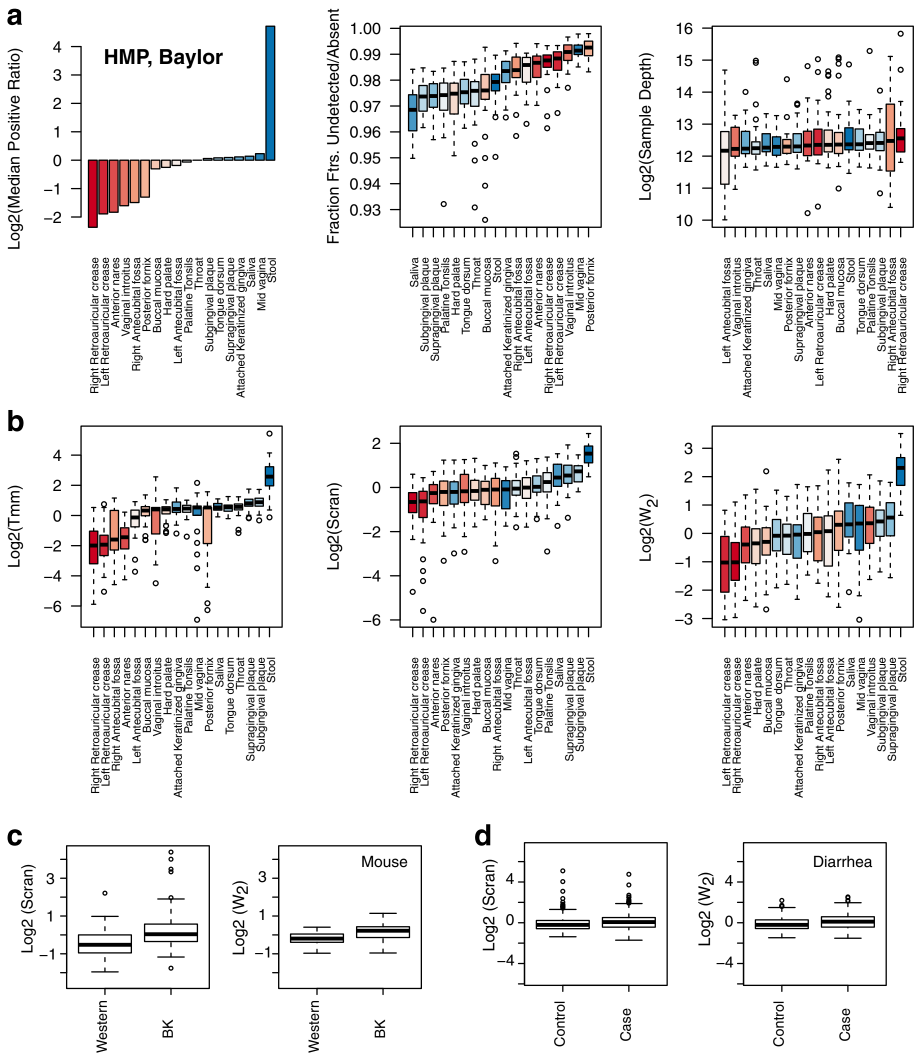


Fig. 10 Wrench retains potential biological information, and indicates importance of compositional correction in general practice. We plot some statistical summaries and the compositional scale factors reconstructed by a few techniques for various Human Microbiome Project samples, sequenced at the Baylor 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 (Additional file 1: Figure S10) and normalization methods. Corresponding CSS scales in Additional file 1: Figure S11. 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

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 normalization 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 arising from metagenomics and single cell RNAseq, which lead us to develop a reference based compositional correction tool (Wrench) to achieve the same. Wrench can be broadly viewed as a generalization of TMM [13] for zero-inflated data. We showed that this procedure, by modeling feature-wise zero generation, reduces the estimation bias associated with other normalization procedures like TMM/CSS/DESeq that ignore zeroes while computing normalization scales. In addition, by recovering appropriate normalization scales for samples even where current state of the art techniques fail, the method avoids data wastage and potential loss of power during differential expression and other downstream analyses (We catalog a few potential ways by which compositional sources of bias can cause sparsity in metagenomic and single cell sequencing count data in Additional file 1: Section 6).

Some practically relevant notes on the application of proposed method to metagenomic datasets follow. First, our choice of methodology and simplifying assumptions were principally determined by the scale and sparsity of the 16s metagenomic datasets and estimation robustness. While fully joint parameter inference algorithms will certainly be more accurate, they are unwieldy and computationally intensive with large scale datasets boasting a large number of features with high sparsity. A case in point is the GAMLSS methodology [58], which improved over our pipeline (Wrench normalization coupled with edgeR differential abundance analysis) in a small scale equimolar miRNA benchmarking dataset (Additional file 1: Figure S23), but could not run to completion even in the simplest of our metagenomic datasets, the mouse gut microbiome. Second, our simulation results indicate that the performance of Wrench stabilizes by $10 - 20$ samples per group depending on sample depth and the fraction of features that change across conditions. While in our experience, this is very well within the limits of practically realized sample sizes in metagenomic experiments, at very low sample sizes and very low sample depths (less than a few thousand reads per sample), some care might be necessary. For instance, coherence of the reconstructed sample-wise compositional scales within groups relative to the experimental design can be checked and deviations from expectations analyzed/corrected. Third, our current implementation exploits categorical group information/factors alone (e.g., cases and controls), and extension to continuous covariates (e.g., age, time) underlying the sampling design are planned for future work. If a continuos covariate is present, converting it to factors by discretizing its range in to non-overlapping windows is an option that the analyst can entertain. Furthermore, because group information is exploited during normalization, our proposed methodology is not immediately applicable for classification purposes. In such applications, immediate extensions of the proposed empirical Bayes formalism by assuming priors on the unknown-sample's group membership (based vaguely, for example, on clustering distances) can be done, and is planned for future work.

A few important insights on compositional bias emerge from our theory, simulation and experimental data analyses. 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 is that most features do not change across conditions (or the closely related assumption that the log-fold change distribution is centered at 0). 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 on the total feature content in the absence of technical biases, the large compositional scale estimates obtained for stool samples (across all normalization techniques) is suspect. Compositional effects can amplify even when a few features experience adverse technical perturbations, 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 (Additional file 1: 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., [59] 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 [60]. 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.

Methods

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

Briefly, our normalization strategy can be described as follows. Based on Eq. (1), for a chosen reference vector $q_{0\cdot}$, accounting for sample depth τ_{gj}, the mean model for the observed positive count of the i^{th} feature can be written as: $E\left[\log Y_{gji} | Y_{gji} > 0\right] = \log\left[q_{gji}\tau_{gj}\right] = \log\left[\frac{q_{gji}}{q_{0i}} q_{0i}\tau_{gj}\right] \equiv \log\left(\theta_{gji} q_{0i}\tau_{gj}\right)$, where $\theta_{gji} = \Lambda_{gj}^{-1}\nu_{gji}$. Thus the true ratio of proportions θ_{gji} encapsulate both the constant Λ_{gj}^{-1} and the absolute fold changes ν_{gji}, and can be viewed as the *net* fold change experienced by feature i in sample j from group g. For the purpose of metagenomic applications, and analytic convenience, we slighty relax the standard assumption that most features do not change across conditions by assuming that the feature-wise log-fold changes $\log \nu_{gji}$ arise from a zero mean Gaussian distribution, a common assumption in differential abundance analysis [26, 49, 50]. It then follows that $\log \theta_{gji}$ follows a Gaussian distribution with a mean parameter $\log \Lambda_{gj}^{-1}$. Thus, a robust location estimate of θ_{gji} for every sample leads us to the desired compositional scale estimate $\hat{\Lambda}_{gji}$. Below, we first illustrate how the θ_{gji} are estimated, and subsequently discuss the robust averaging procedure.

Model We assume the following model for the counts Y_{gji}:

$$\begin{aligned}
Y_{gji} &\sim \begin{cases} 0 & \text{with probability } \pi_{gji} \\ e^{Z_{gji}} & \text{with probability } (1-\pi_{gji}) \end{cases}, \\
Z_{gji} &= \underbrace{\log q_{0i}}_{\text{log-reference}} + \underbrace{\log \tau_{gj}}_{\text{log-sample depth}} \\
&\quad + \underbrace{\log \zeta_{0g} + \mu_{gj} + a_{gji}}_{=\log\theta_{gji},\ \text{log net fold change relative to reference}} + \epsilon_{gji}, \\
a_{gji} &\sim N\left(0, \eta_{0g}^2\right),\ g = 1\ldots G, \\
\epsilon_{gji} &\sim N\left(0, \sigma_{0i}^2\right),\ i = 1\ldots p, \\
\log\left(\frac{\pi_{gji}}{1-\pi_{gji}}\right) &= \beta_{i1} + \beta_{i2}\log \tau_{gj} + \text{possibly other covariates}
\end{aligned} \tag{2}$$

The model assumes the following. For each sample j from group g, the i^{th} feature's count value is sampled from a hurdle log-normal distribution, in which with probability

π_{gji}, a value of 0 is realized; and with probability $1 - \pi_{gji}$ a positive count is observed. The probabilities π_{gji} are determined by sample covariates, including the total sequencing depth. The positive count value is realized as an exponential of a Gaussian random variable Z_{gji} the mean of which is determined (in accordance with the Eq. (1)) by the chosen reference value q_{0i}, sample-depth τ_{gj}, and the *net* fold change $\theta_{gji} = \nu_{gji} * \Lambda_{gj}^{-1}$, the log of which has been modeled in the above equation as a sum of group-wise effect ($\log \zeta_{0g}$), two-way group-sample interaction (μ_{gj}), a three-way group-sample-feature interaction random effect a_{gji} and a noise term.

Estimation of regularized ratios $\hat{\theta}_{gji}$: In the model, the 0 subscripted parameters are considered known, and are determined the following way. $\tau_{gj} = Y_{gj+}$ is the total count of sample *gj*. The reference value for each feature *i*, q_{0i}, is set to the average proportion value $\overline{\hat{q}_{++i}}$, where $\hat{q}_{gji}$ is the observed proportion of feature *i* in sample *gj*, i.e., $\hat{q}_{gji} = Y_{gji}/Y_{gj+} = Y_{gji}/\tau_{gj}$. The mean and variance parameters $\log \zeta_{0g}$ and η_{0g}^2 of the Gaussian prior distribution on the $\log \theta_{gji}$ are determined based on the corresponding moments of the corresponding empirical distribution of the group-wise pooled raw ratios of proportions: $\left\{ r_{gji} = \hat{q}_{gi}/q_{0i} \right\}_{i=1}^{p}$. Here, $\hat{q}_{gi} = Y_{g+i}/Y_{g++}$ i.e., the overall proportion of feature *i* in the samples from the entire group. Specifically, we fix the group-wise compositional scale $\zeta_{0g} = \overline{r_{g+i}}$ i.e., as the average of the raw ratios including the zero values (following discussions in Fig. 6). We set the variance parameter $\eta_{0g}^2 = \frac{1}{\sum_i I_{[Y_{gji}>0]}} \sum_{i:Y_{gji}>0} \left(\log r_{gji} - \overline{\log r_{g+i}} \right)$ i.e., as the empirical variance of the logged-ratios. Finally, the feature-specific expression variances σ_{0i}^2 are fixed with values obtained from Limma/Voom. With the above fixed, the unknown parameters μ_{gj} and a_{gji} are estimated/predicted using standard random effects estimators: $\hat{\mu}_{gj} = \sum_i w_{gji} \left(\log r_{gji} - \log \zeta_{0g} \right)$ with $w_{gji} \propto \frac{1}{\sigma_{0i}^2 + \eta_{0g}^2}$, and $\hat{a}_{gji} = \frac{\sigma_{0i}^2}{\sigma_{0i}^2 + \eta_{0g}^2} \left(\log r_{gji} - \log \zeta_{0g} - \hat{\mu}_{gj} \right)$. The identifiability of these terms is ensured as the other variance components are fixed. The $\hat{\pi}_{gji}$ are estimated with logistic regression. The regularized ratios are then calculated as: $\hat{\theta}_{gji} = exp(\log \zeta_{0g} + \hat{\mu}_{gj} + \hat{a}_{gji})$.

Robust averaging of the $\hat{\theta}_{gji}$: While averaging over the regularized ratios $W_0 =: \frac{1}{p} \sum_i \hat{\theta}_{gji}$ would be one estimation route to Λ_{gj}^{-1}, better control can be achieved by taking the variation in the feature-wise zero generation in to account. We shall notice that $E\left[r_{gji} | r_{gji} > 0 \right] = \theta_{gji} \cdot e^{\sigma_{0i}^2/2}$, and so a robust averaging over $\hat{\theta}_{gji}/e^{\sigma_{0i}^2/2}$, can serve as an estimator of Λ_{gj}^{-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) $\propto 1/(1 - \hat{\pi}_{gji})$ or on the inverse marginal variances ascribed by our model above $\propto \frac{1}{(1-\hat{\pi}_{gji})\left(\hat{\pi}_{gji}+e^{\sigma_{0i}^2+\eta_{0g}^2}-1\right)}$. An estimator that we also found to work well empirically is a weighted average of $\frac{\hat{\theta}_{gji}/e^{\sigma_{0i}^2/2}}{1-\hat{\pi}_{gji}}$with weights proportional to $\frac{1}{\sigma_{0i}^2}$. Additional file 1: 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.

In summary, we explored the performance of the following estimators for sample-wise compositional factors:

$$
\begin{aligned}
W_{0gj} &=: \frac{1}{p} \sum_i \hat{\theta}_{gji} = \overline{\hat{\theta}_{g+j}}, \\
W_{1gj} &=: \frac{1}{p} \sum_i w_{gji} \hat{\theta}_{gji}, \text{ with } w_{gji} \propto 1/(1 - \hat{\pi}_{gji}) \\
W_{2gj} &=: \frac{1}{p} \sum_i w_{gji} \hat{\theta}_{gji}, \text{ with } w_{gji} \propto \frac{1}{\left(1 - \hat{\pi}_{gji}\right) \left(\hat{\pi}_{gji} + e^{\sigma_{0i}^2 + \eta_{0g}^2} - 1\right)} \\
W_{3gj} &=: \frac{1}{p} \sum_i w_{gji} \frac{\hat{\theta}_{gji}}{1 - \hat{\pi}_{gji}}, \text{ with } w_{gji} \propto \frac{1}{\sigma_{0i}^2}
\end{aligned}
\tag{3}
$$

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 [35]. A large scale study of human gut microbiomes ($n = 992$) from diarrhea-afflicted and healthy children from various developing countries [45]. 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 sequence-able 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 [44] was obtained from [61].

The Unique Molecular Identifier (UMI) single cell RNAseq data from Islam et al. [62] was downloaded 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 [46], 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`, DESeq2's `estimateDispersionsGeneEst` and `nbinomLRT` were used to perform differential abundance testing [50]. 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 http://bioconductor.org/packages/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 Additional file 1: Figure S1, 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 $\overline{\nu_{gi}}$ (where the bar indicates this value will now assume the role of an average) 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 $LN\left(\log \overline{\nu_{gi}}, \tilde{\sigma}_\nu^2\right)$ 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. We repeated simulations for 20 iterations.

In both versions of simulations, the total induced abundance change relative to that of the control is $\Lambda_{gj} = \nu_{gj\cdot}^T q_{1\cdot}$, where $\nu_{gj\cdot}$ is the vector of fold changes for sample j in group g, and $q_{1\cdot}$ is the average vector of feature-wise control proportions. As it can be seen from the expression for Λ_{gj}, notice that perturbing features with very low relative abundances do not demonstrably induce compositional bias at low sample depth settings (unless perturbed by very high fold changes). So for every simulation iteration, the fraction f of features that were perturbed in cases were chosen randomly according to their control proportions. We apply the term *compositional correction factor* for Λ_{gj}^{-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 all normalization methods 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.

The offset-covariate analysis followed the procedure in [25]. For resampling analysis, samples from each experimental group (with atleast 15 samples) were split in half randomly to construct two artificial groups. Normalization factors from each method were then used to perform differential abundance analysis, and the total number of differentially abundant calls were recorded. The procedure was repeated for ten iterations for each group, and the results were averaged across 41 experimental groups. Those samples for which Scran fails to reconstruct normalization scales were discarded from differential abundance analyses to avoid any power differences while testing. The normalization scales however, were obtained with all data for each method.

Fisher exact tests were used to perform functional enrichment analyses for positively associated OTUs. A Genus level functional enrichment analysis was first performed by aggregating annotations from the FAPROTAX1.1 database [51] at the Genus level. A more specific OTU level functional enrichment analysis was devised as follows. Because the Tara Oceans Kegg module (KM) abundance data (downloaded from http://ocean-microbiome.embl.de/data/TARA243.KO-module.profile.release.gz) and the 16S reconstructions are obtained from the same input DNA through whole metagenome shotgun, the same compositional factors apply to both datatypes. Each normalization approach's compositional factors for 16S data was used to rescale the KM relative abundance data. This normalized KM data was used to annotate each OTU by (normalized) KMs that Spearman correlate at a value of atleast .75.

Software availability

`Wrench` is available from R/Bioconductor as an R package at the URL: http://bioconductor.org/packages/Wrench.

Endnotes

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

[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 $(1 - \pi)$

Additional files

Additional file 1: Supplementary Note. Presents further discussions on compositional bias, and supplementary results in context. (PDF 17,810 kb)

Additional file 2: Enrichment Analysis Results. The results of enrichment analyses based on faprotax annotations and Kegg modules procedure described in the Methods section is presented. Names in the sheets and their descriptions are as follows: KM.POS.SIG.MES and KM.POS.SIG.DCM show the Kegg module based enrichment analyses for positively associated features in MES and DCM layers respectively. FAPRO.POS.SIG.MES and FAPRO.POS.SIG.DCM show the results of faprotax annotations based enrichment analyses for positively associated features in MES and DCM layers respectively. (XLSX 45 kb)

Abbreviations

CPM: Counts per million; DCM: Deep Chlorophyl maximum; DNA: Deoxy-ribonucleic acid; FPKM: Fragments per kilobase of transcript per million mapped reads; HMP: Human Microbiome project; KM: Kegg module; MES: Mesopelagic layer; OTU: Operational taxonomic units; RNA: Ribonucleic acid; RPKM: Unaltered reads per kilobase of transcript, per million mapped reads; UMI: Unique molecular identifiers

Acknowledgements

MSK thanks Mihai Pop, Tom Goldstein, Joyce Hsiao, Jayaram Kancherla, Charlotte Soneson and Mathieu Almeida for useful discussions, Joseph Paulson for making the mouse and lung 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.

Funding

This work was partially supported by NSF grant 1564785 to SH, by NIH R01 grants GM083084, RR021967/GM103552 and K99HG009007 to SCH, by NIH R01 grants GM114267 and HG005220 to HCB and MSK, and by NIH R1 grants GM107683 to Matthias Chung, and NSF grant IIS 1513615 to Mihai Pop. The funding bodies did not have any role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

`Wrench` is available from GitLab as an R package at the URL: http://bioconductor.org/packages/Wrench.

16S metagenomic count data corresponding to the lung microbiome, mouse diet and the diarrheal studies were obtained from the R/Bioconductor package `metagenomeSeq` [26].

The Tara Oceans project's 16S reconstructions from whole metagenome shotgun sequencing [10] 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 sequence-able DNA material.

The Human Microbiome Project's 16s count data [48] 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 [44] was obtained from [61].

The Unique Molecular Identifier (UMI) single cell RNAseq data from Islam et al., [62] was downloaded from GEO under accession GSE46980.

Authors' contributions

Conceived and designed study: MSK, HCB. Contributed analytical tools/reagents: MSK, EVS, KO, HCB. Data analysis and interpretation: MSK, EVS, KO, HCB. Wrote the paper: MSK, SH, SCH, EVS, HCB. Participated in discussions: All authors. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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.

Received: 3 April 2018 Accepted: 11 October 2018
Published online: 06 November 2018

References

1. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26(10):1135–45. https://doi.org/10.1038/nbt1486. Accessed 09 Mar 2016.
2. Sanger F. Sequences, sequences, and sequences. Ann Rev Biochem. 1988;57:1–28. https://doi.org/10.1146/annurev.bi.57.070188.000245.
3. Hutchison CA. DNA sequencing: bench to bedside and beyond. Nucleic Acids Res. 2007;35(18):6227–37. https://doi.org/10.1093/nar/gkm688.
4. Mardis ER. A decade/'s perspective on DNA sequencing technology. Nature. 2011;470(7333):198–203. https://doi.org/10.1038/nature09796. Accessed 09 Mar 2016.
5. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. R Foundation for Statistical Computing. https://www.R-project.org/.
6. Wooley JC, Godzik A, Friedberg I. A Primer on Metagenomics. PLoS Comput Biol. 2010;6(2):1000667. https://doi.org/10.1371/journal.pcbi.1000667. Accessed 09 Mar 2016.
7. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet. 2009;10(10):669–80. https://doi.org/10.1038/nrg2641. Accessed 09 Mar 2016.
8. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10(1):57–63. https://doi.org/10.1038/nrg2484. Accessed 09 Mar 2016.
9. Tringe SG, Rubin EM. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet. 2005;6(11):805–14. https://doi.org/10.1038/nrg1709. Accessed 09 Mar 2016.
10. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, Djahanschiri B, Zeller G, Mende DR, Alberti A, Cornejo-Castillo FM, Costea PI, Cruaud C, d'Ovidio F, Engelen S, Ferrera I, Gasol JM, Guidi L, Hildebrand F, Kokoszka F, Lepoivre C, Lima-Mendez G, Poulain J, Poulos BT, Royo-Llonch M, Sarmento H, Vieira-Silva S, Dimier C, Picheral M, Searson S, Kandels-Lewis S, Coordinators TO, Bowler C, Vargas CD, Gorsky G, Grimsley N, Hingamp P, Iudicone D, Jaillon O, Not F, Ogata H, Pesant S, Speich S, Stemmann L, Sullivan MB, Weissenbach J, Wincker P, Karsenti E, Raes J, Acinas SG, Bork P. Structure and function of the global ocean microbiome. Science. 2015;348(6237):1261359. https://doi.org/10.1126/science.1261359. Accessed 13 Feb 2017.
11. Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biol Direct. 2009;4(1):14. https://biologydirect.biomedcentral.com/articles/10.1186/1745-6150-4-14. Accessed 12 Apr 2017.
12. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):14. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2010-11-2-r14. Accessed 12 Apr 2017.
13. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11(3):25. https://doi.org/10.1186/gb-2010-11-3-r25. http://genomebiology.com/2010/11/3/R25/abstract. Accessed 09 Jan 2015.
14. Pachter L. Models for transcript quantification from RNA-Seq. arXiv:1104.3889 [q-bio, stat]. 2011. arXiv: 1104.3889. http://arxiv.org/abs/1104.3889. Accessed 21 Jan 2016.
15. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64. http://biostatistics.oxfordjournals.org/content/4/2/249.short. Accessed 02 Jan 2015.
16. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010;11:94. https://doi.org/10.1186/1471-2105-11-94.
17. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):106. https://doi.org/10.1186/gb-2010-11-10-r106.
18. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B (Methodol). 1982139–77. http://www.jstor.org/stable/2345821. Accessed 20 July 2015.
19. Friedman J, Alm EJ. Inferring Correlation Networks from Genomic Survey Data. PLoS Comput Biol. 2012;8(9):1002687. https://doi.org/10.1371/journal.pcbi.1002687. Accessed 30 Oct 2013.
20. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50. https://doi.org/10.1038/nrmicro2832. Accessed 2015-07-28.
21. Fernandes AD, Reid JN, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16s rRNA gene sequencing and selective growth experiments by compositional data analysis. Microbiome. 2014;2:15. https://doi.org/10.1186/2049-2618-2-15. Accessed 21 Jan 2016.
22. Fang H, Huang C, Zhao H, Deng M. CCLasso: correlation inference for compositional data through Lasso. Bioinformatics. 2015;349. https://doi.org/10.1093/bioinformatics/btv349. Accessed 28 July 2015.
23. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: A Valid Alternative to Correlation for Relative Data. PLOS Comput Biol. 2015;11(3):1004075. https://doi.org/10.1371/journal.pcbi.1004075. Accessed 09 Mar 2016.
24. Chen K, Hu Z, Xia Z, Zhao D, Li W, Tyler JK. The overlooked fact: fundamental need of spike-in controls for virtually all genome-wide analyses. Mol Cell Biol. 2015MCB.00970–14. https://doi.org/10.1128/MCB.00970-14. http://mcb.asm.org/content/early/2015/12/15/MCB.00970-14. Accessed 09 Mar 2016.
25. L. Lun AT, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17:75. https://doi.org/10.1186/s13059-016-0947-7. Accessed 13 Aug 2016.
26. Paulson J, Stine OC, Bravo HC, Pop M. Differential abundance analysis for microbial marker-gene surveys. Nat Methods. 2013. http://www.nature.com/nmeth/journal/vaop/ncurrent/full/nmeth.2658.html. Accessed 17 Jan 2015.

27. Huggett JF, Laver T, Tamisak S, Nixon G, O'Sullivan DM, Elaswarapu R, Studholme DJ, Foy CA. Considerations for the development and application of control materials to improve metagenomic microbial community profiling. Accred Qual Assur. 2013;18(2):77–83. https://doi.org/10.1007/s00769-012-0941-z. Accessed 13 Apr 2017.
28. van Dijk EL, Jaszczyszyn Y, Thermes C. Library preparation methods for next-generation sequencing: Tone down the bias. Exp Cell Res. 2014;322(1):12–20. https://doi.org/10.1016/j.yexcr.2014.01.008. http://www.sciencedirect.com/science/article/pii/S0014482714000160. Accessed 13 Apr 2017.
29. O'Sullivan DM, Laver T, Temisak S, Redshaw N, Harris KA, Foy CA, Studholme DJ, Huggett JF. Assessing the Accuracy of Quantitative Molecular Microbial Profiling. Int J Mol Sci. 2014;15(11):21476–91. https://doi.org/10.3390/ijms151121476. Accessed 13 Apr 2017.
30. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, Shakya M, Podar M, Quince C, Hall N. A comprehensive benchmarking study of protocols and sequencing platforms for 16s rRNA community profiling. BMC Genomics. 2016;17:55. https://doi.org/10.1186/s12864-015-2194-9. Accessed 13 Apr 2017.
31. Costea PI, Zeller G, Sunagawa S, Pelletier E, Alberti A, Levenez F, Tramontano M, Driessen M, Hercog R, Jung F-E, Kultima JR, Hayward MR, Coelho LP, Allen-Vercoe E, Bertrand L, Blaut M, Brown JRM, Carton T, Cools-Portier S, Daigneault M, Derrien M, Druesne A, de Vos WM, Finlay BB, Flint HJ, Guarner F, Hattori M, Heilig H, Luna RA, van Hylckama Vlieg J, Junick J, Klymiuk I, Langella P, Le Chatelier E, Mai V, Manichanh C, Martin JC, Mery C, Morita H, O'Toole PW, Orvain C, Patil KR, Penders J, Persson S, Pons N, Popova M, Salonen A, Saulnier D, Scott KP, Singh B, Slezak K, Veiga P, Versalovic J, Zhao L, Zoetendal EG, Ehrlich SD, Dore J, Bork P. Towards standards for human fecal sample processing in metagenomic studies. Nat Biotechnol. 2017;35(11):1069–76. https://doi.org/10.1038/nbt.3960.
32. Olson ND, Morrow JB. DNA extract characterization process for microbial detection methods development and validation. BMC Res Notes. 2012;5:668. https://doi.org/10.1186/1756-0500-5-668.
33. Brooks JP, Edwards DJ, Harwich MD, Rivera MC, Fettweis JM, Serrano MG, Reris RA, Sheth NU, Huang B, Girerd P, Vaginal Microbiome Consortium, Strauss JF, Jefferson KK, Buck GA. The truth about metagenomics: quantifying and counteracting bias in 16s rRNA studies. BMC Microbiology. 2015;15:66. https://doi.org/10.1186/s12866-015-0351-6.
34. Law CW, Chen Y, Shi W, Smyth GK. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15(2):29. https://doi.org/10.1186/gb-2014-15-2-r29. Accessed 17 June 2015.
35. Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI. The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Sci Transl Med. 2009;1(6):6–14. https://doi.org/10.1126/scitranslmed.3000322.
36. McMurdie PJ, Holmes S. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Comput Biol. 2014;10(4):1003531. https://doi.org/10.1371/journal.pcbi.1003531. Accessed 05 Sept 2015.
37. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B. Synthetic spike-in standards for RNA-seq experiments. Genome Res. 2011;21(9):1543–51. http://genome.cshlp.org/content/21/9/1543.short. Accessed 09 Mar 2016.
38. Thattai M. Universal Poisson Statistics of mRNAs with Complex Decay Pathways. Biophys J. 2016;110(2):301–5. https://doi.org/10.1016/j.bpj.2015.12.001. Accessed 25 May 2017.
39. Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: A mathematical analysis. Genomics. 1988;2(3):231–9. https://doi.org/10.1016/0888-7543(88)90007-9. Accessed 29 Aug 2018.
40. Jiang H, Wong WH. Statistical inferences for isoform expression in RNA-Seq. Bioinformatics. 2009;25(8):1026–32. https://doi.org/10.1093/bioinformatics/btp113. Accessed 29 Aug 2018.
41. Schmieder R, Edwards R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. PLoS ONE. 2011;6(3):17288. https://doi.org/10.1371/journal.pone.0017288. Accessed 19 Apr 2017.
42. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, Turner P, Parkhill J, Loman NJ, Walker AW. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87. https://doi.org/10.1186/s12915-014-0087-z. Accessed 19 Apr 2017.
43. Hemme CL, Tu Q, Shi Z, Qin Y, Gao W, Deng Y, Nostrand JDV, Wu L, He Z, Chain PSG, Tringe SG, Fields MW, Rubin EM, Tiedje JM, Hazen TC, Arkin AP, Zhou J. Comparative metagenomics reveals impact of contaminants on groundwater microbiomes. Front Microbiol. 2015;6. https://doi.org/10.3389/fmicb.2015.01205. Accessed 19 Apr 2017.
44. Yu Y, Fuscoe JC, Zhao C, Guo C, Jia M, Qing T, Bannon DI, Lancashire L, Bao W, Du T, Luo H, Su Z, Jones WD, Moland CL, Branham WS, Qian F, Ning B, Li Y, Hong H, Guo L, Mei N, Shi T, Wang KY, Wolfinger RD, Nikolsky Y, Walker SJ, Duerksen-Hughes P, Mason CE, 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. Nat Commun. 2014;5:3230. https://doi.org/10.1038/ncomms4230. Accessed 09 Mar 2016.
45. Pop M, Walker AW, Paulson J, Lindsay B, Antonio M, Hossain MA, Oundo J, Tamboura B, Mai V, Astrovskaya I, Bravo HC, Rance R, Stares M, Levine MM, Panchalingam S, Kotloff K, Ikumapayi UN, Ebruke C, Adeyemi M, Ahmed D, Ahmed F, Alam MT, Amin R, Siddiqui S, Ochieng JB, Ouma E, Juma J, Mailu E, Omore R, Morris JG, Breiman RF, Saha D, Parkhill J, Nataro JP, Stine OC. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. Genome Biol. 2014;15(6):76. https://doi.org/10.1186/gb-2014-15-6-r76. Accessed 09 Mar 2016.
46. Kurtz ZD, Müller C. L., Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Comput Biol. 2015;11(5):1004226. https://doi.org/10.1371/journal.pcbi.1004226. Accessed 13 Aug 2016.
47. Tsilimigras MCB, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol. 2016;26(5):330–5. https://doi.org/10.1016/j.annepidem.2016.03.002.
48. Consortium THMP, Huttenhower C, Gevers D, Knight R, Abubucker S, Badger JH, Chinwalla AT, Creasy HH, Earl AM, FitzGerald MG, Fulton RS, Giglio MG, Hallsworth-Pepin K, Lobos EA, Madupu R, Magrini V, Martin JC, Mitreva M, Muzny DM, Sodergren E, Versalovic J, Wollam AM, Worley KC, Wortman JR, Young SK, Zeng Q, Aagaard KM, Abolude OO, Allen-Vercoe E, Alm EJ, Alvarado L, Andersen GL, Anderson S, Appelbaum E, Arachchi HM, Armitage G, Arze CA, Ayvaz T, Baker CC, Begg L, Belachew T, Bhonagiri V, Bihan M, Blaser MJ, Bloom T, Bonazzi V, Brooks JP, Buck GA, Buhay CJ, Busam DA, Campbell JL, Canon SR, Cantarel BL, Chain PSG, Chen I-MA, Chen L, Chhibba S, Chu K, Ciulla DM, Clemente JC, Clifton SW, Conlan S, Crabtree J, Cutting MA, Davidovics NJ, Davis CC, DeSantis TZ, Deal C, Delehaunty KD, Dewhirst FE, Deych E, Ding Y, Dooling DJ, Dugan SP, Dunne WM, Durkin AS, Edgar RC, Erlich RL, Farmer CN, Farrell RM, Faust K, Feldgarden M, Felix VM, Fisher S, Fodor AA, Forney LJ, Foster L, Francesco VD, Friedman J, Friedrich DC, Fronick CC, Fulton LL, Gao H, Garcia N, Giannoukos G, Giblin C, Giovanni MY, Goldberg JM, Goll J, Gonzalez A, Griggs A, Gujja S, Haake SK, Haas BJ, Hamilton HA, Harris EL, Hepburn TA, Herter B, Hoffmann DE, Holder ME, Howarth C, Huang KH, Huse SM, Izard J, Jansson JK, Jiang H, Jordan C, Joshi V, Katancik JA, Keitel WA, Kelley ST, Kells C, King NB, Knights D, Kong HH, Koren O, Koren S, Kota KC, Kovar CL, Kyrpides NC, Rosa PSL, Lee SL, Lemon KP, Lennon N, Lewis CM, Lewis L, Ley RE, Li K, Liolios K, Liu B, Liu Y, Lo C-C, Lozupone CA, Lunsford RD, Madden T, Mahurkar AA, Mannon PJ, Mardis ER, Markowitz VM, Mavromatis K, McCorrison JM, McDonald D, McEwen J, McGuire AL, McInnes P, Mehta T, Mihindukulasuriya KA, Miller JR, Minx PJ, Newsham I, Nusbaum C, OâLaughlin M, Orvis J, Pagani I, Palaniappan K, Patel SM, Pearson M, Peterson J, Podar M, Pohl C, Pollard KS, Pop M, Priest ME, Proctor LM, Qin X, Raes J, Ravel J, Reid JG, Rho M, Rhodes R, Riehle KP, Rivera MC, Rodriguez-Mueller B, Rogers Y-H, Ross MC, Russ C, Sanka RK, Sankar P, Sathirapongsasuti JF, Schloss JA, Schloss PD, Schmidt TM, Scholz M, Schriml L, Schubert AM, Segata N, Segre JA, Shannon WD, Sharp RR, Sharpton TJ, Shenoy N, Sheth NU, Simone GA, Singh I, Smillie CS, Sobel JD, Sommer DD, Spicer P, Sutton GG, Sykes SM, Tabbaa DG, Thiagarajan M, Tomlinson CM, Torralba M, Treangen TJ, Truty RM, Vishnivetskaya TA, Walker J, Wang L, Wang Z, Ward DV, Warren W, Watson MA, Wellington C, Wetterstrand KA, White JR, Wilczek-Boney K, Wu Y. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486(7402):207–14. https://doi.org/10.1038/nature11234. Accessed 06 Oct 2018.
49. Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 2004;3:3. https://doi.org/10.2202/1544-6115.1027.

50. 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. https://doi.org/10.1186/s13059-014-0550-8.
51. Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. Science. 2016;353(6305):1272–7. https://doi.org/10.1126/science.aaf4507. Accessed 01 Nov 2017.
52. Karl DM, Beversdorf L, Björkman KM, Church MJ, Martinez A, Delong EF. Aerobic production of methane in the sea. Nat Geosci. 2008;1(7):473. https://doi.org/10.1038/ngeo234. Accessed 14 Dec 2017.
53. Borin S, Brusetti L, Mapelli F, D'Auria G, Brusa T, Marzorati M, Rizzi A, Yakimov M, Marty D, Lange GJD, Wielen PVD, Bolhuis H, McGenity TJ, Polymenakou PN, Malinverno E, Giuliano L, Corselli C, Daffonchio D. Sulfur cycling and methanogenesis primarily drive microbial colonization of the highly sulfidic Urania deep hypersaline basin. Proc Natl Acad Sci. 2009;106(23):9151–6. https://doi.org/10.1073/pnas.0811984106. Accessed 14 Dec 2017.
54. Orcutt BN, Sylvan JB, Knab NJ, Edwards KJ. Microbial Ecology of the Dark Ocean above, at, and below the Seafloor. Microbiol Mol Biol Rev. 2011;75(2):361–422. https://doi.org/10.1128/MMBR.00039-10. Accessed 14 Dec 2017.
55. DeLong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N-U, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DM. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. Science. 2006;311(5760):496–503. https://doi.org/10.1126/science.1120250. Accessed 14 Dec 2017.
56. Swan BK, Martinez-Garcia M, Preston CM, Sczyrba A, Woyke T, Lamy D, Reinthaler T, Poulton NJ, Masland EDP, Gomez ML, Sieracki ME, DeLong EF, Herndl GJ, Stepanauskas R. Potential for Chemolithoautotrophy Among Ubiquitous Bacteria Lineages in the Dark Ocean. Science. 2011;333(6047):1296–300. https://doi.org/10.1126/science.1203690. Accessed 14 Dec 2017.
57. Rosa PSL, Brooks JP, Deych E, Boone EL, Edwards DJ, Wang Q, Sodergren E, Weinstock G, Shannon WD. Hypothesis Testing and Power Calculations for Taxonomic-Based Human Microbiome Data. PLoS ONE. 2012;7(12):52078. https://doi.org/10.1371/journal.pone.0052078. Accessed 11 Mar 2016.
58. Argyropoulos C, Etheridge A, Sakhanenko N, Galas D. Modeling bias and variation in the stochastic processes of small RNA sequencing. Nucleic Acids Res. 2017;45(11):104. https://doi.org/10.1093/nar/gkx199.
59. Lovén J., Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. Revisiting global gene expression analysis. Cell. 2012;151(3):476–82. https://doi.org/10.1016/j.cell.2012.10.012.
60. Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16(3):133–45. https://doi.org/10.1038/nrg3833. Accessed 15 Aug 2016.
61. Hicks SC, Okrah K, Paulson J, Quackenbush J, Irizarry RA, Bravo HC. Smooth Quantile Normalization. bioRxiv. 2016085175. https://doi.org/10.1101/085175. Accessed 09 Mar 2017.
62. 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. Nat Methods. 2014;11(2):163–6. https://doi.org/10.1038/nmeth.2772. Accessed 30 Aug 2016.