Assessment of single cell RNA-seq statistical methods on microbiome data

Matteo Calgaro[1], Chiara Romualdi[2], Levi Waldron[3], Davide Risso[4,5,*], Nicola Vitulo[1,*]

1. Department of Biotechnology, University of Verona, Italy
2. Department of Biology, University of Padova, Padova, Italy.
3. Graduate School of Public Health and Health Policy, City University of New York, New York, NY, USA.
4. Department of Statistical Sciences, University of Padova, Padova, Italy.
5. Division of Biostatistics and Epidemiology, Department of Healthcare Policy and Research, Weill Cornell Medicine, New York, NY, USA.

* co-last authors. Correspondence should be addressed to davide.risso@unipd.it and nicola.vitulo@univr.it.

Abstract

The correct identification of differentially abundant microbial taxa between experimental conditions is a methodological and computational challenge. Recent work has shown that commonly used methods do not control the false discovery rate due to the peculiarity of these data (e.g. high sparsity), leading to an abundance of false positive results.

Since single-cell RNA-seq shares some of these peculiarities, we apply methods developed for single cell differential expression to microbiome data. We compare these approaches to methods developed for bulk RNA-seq and microbiome data, in terms of suitability of distributional assumptions, ability to control false discoveries, consistency, replicability, and power. We benchmark these methods using 100 manually curated datasets from 16S and whole metagenome shotgun sequencing. A simulation framework is developed to assess the impact of experimental design in power analysis.

Our analyses suggest that DESeq2 and limma-voom show the best performance. We recommend a careful exploratory data analysis prior to application of any inferential model and we present a framework to help scientists make an informed choice of analysis methods in a dataset-specific manner.

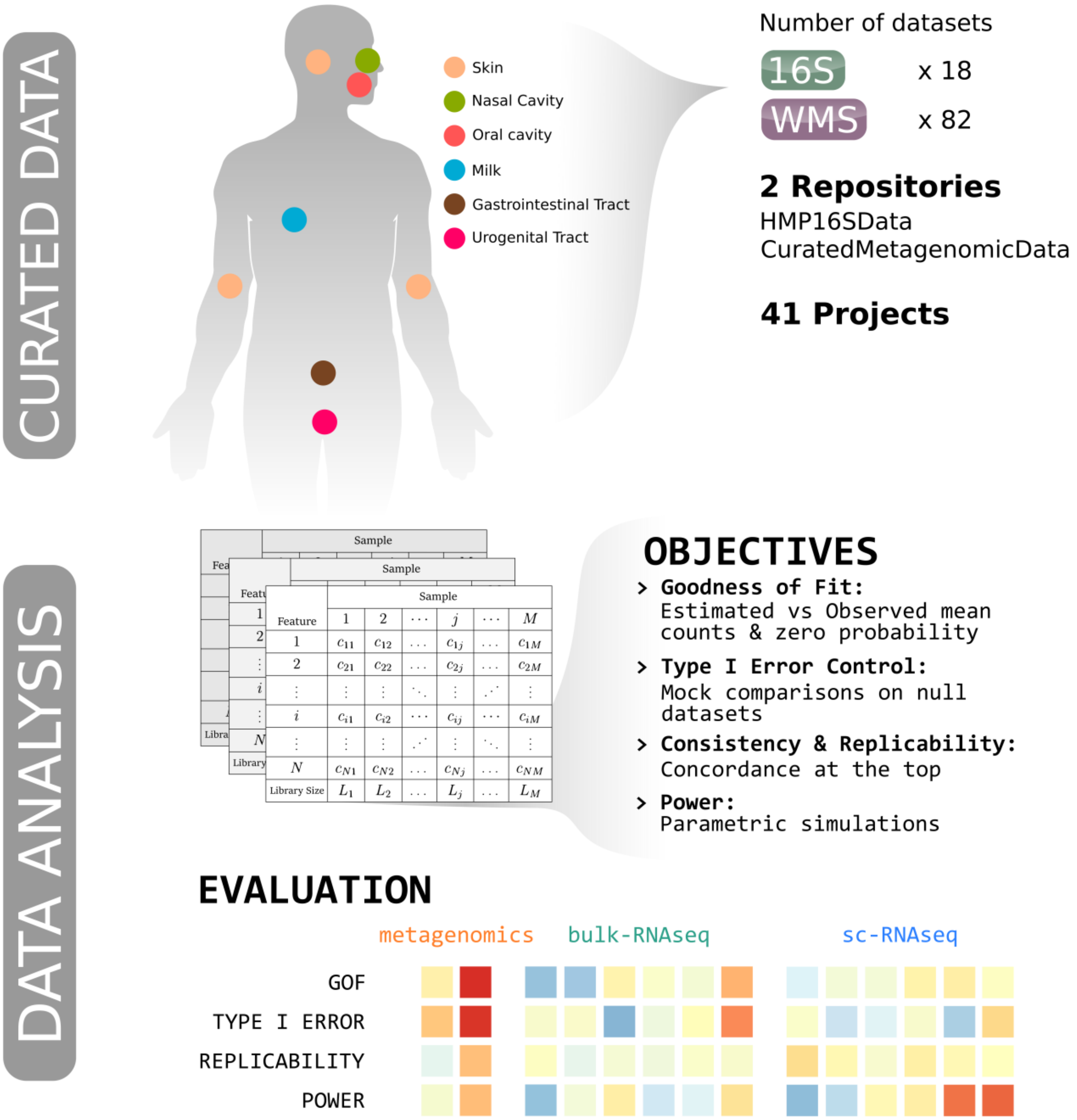
Introduction

Study of the microbiome, the uncultured collection of microbes present in most environments, is a novel application of high-throughput sequencing that shares certain similarities but important differences from other applications of DNA and RNA sequencing. Common approaches for the microbiome studies are based on the deep sequencing of amplicons of universal marker-genes, such as the 16S rRNA genes, or on whole metagenome shotgun sequencing (WMS). Community taxonomic composition can be estimated from microbiome data by assigning each read to the most plausible microbial lineage using a reference annotated database, with a higher taxonomic resolution in WMS than in 16S[1,2]. The final output of such analyses usually consists of a large, highly sparse taxa per samples count table.

Differential abundance (DA) analysis is one of the primary approaches to identify differences in the microbial community composition between samples and to understand the structures of the microbial communities and the associations between the microbial compositions and the environment. DA analysis has commonly been performed using methods adapted from RNA-seq analysis; however, the peculiar characteristics of microbiome data make differential abundance analysis challenging. Compared to other high-throughput sequencing techniques such as RNA sequencing (RNA-seq), metagenomic data are sparse, i.e., the taxa count matrix contains many zeros. This sparsity can be explained by both biological and technical reasons: some taxa are very rare and present only in a few samples, while others are very lowly represented and cannot be detected because of an insufficient sequencing depth or other technical reasons.

In recent years, single-cell RNA-seq (scRNA-seq) has revolutionized the field of transcriptomics, providing new insight on the transcriptional program of individual cells, shading light on complex, heterogeneous tissues, and revealing rare cell populations with distinct gene expression profiles[3–6]. However, due to the relatively inefficient mRNA capture rate, scRNA-seq data are characterized by dropout events, which leads to an excess of zero read counts compared to bulk RNA-seq data[7,8]. Thus, with the advent of this technology new statistical models accounting for dropout events have been proposed. The similarities with respect to sparsity observed in both scRNA-seq and metagenomics data led us to pose the question of whether statistical methods developed for the differential expression of scRNA-seq data perform well on metagenomic DA analysis.

Some benchmarking efforts have compared the performance of methods[9–12] both adapted from bulk RNA-seq and developed for microbiome DA[13,14]. While some tools exist to guide researchers[15], a general consensus on the best approach is still missing, especially regarding the methods' capability of controlling false discoveries. In this study, we benchmark several statistical models and methods developed for metagenomics[13,14], bulk RNA-seq[16–18] and, for the first time, single-cell RNA-seq[7,8,19–21] on a collection of manually curated 16S[22] and WMS[23] real data as well as on a comprehensive set of simulations. We also consider use of the Dirichlet Multinomial Distribution for explicit compositional analysis (i.e., reflecting the fact that counts represent a proportion of the total rather than absolute counts), and the use of geometric mean normalization for reducing the impact of compositionality[24]. The novelty of our benchmarking efforts are two-fold. First, we include in the comparison novel methods recently developed in the scRNA-seq literature; second, unlike previous efforts, our conclusions are based on several performance metrics that range from type I error control and goodness of fit to replicability across datasets and internal consistency among methods.



	Sample					
Feature	1	2	$\cdots$	j	$\cdots$	M
1	c_{11}	c_{12}	$\ldots$	c_{1j}	$\ldots$	c_{1M}
2	c_{21}	c_{22}	$\ldots$	c_{2j}	$\ldots$	c_{2M}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\iddots$	$\vdots$
i	c_{i1}	c_{i2}	$\cdots$	c_{ij}	$\cdots$	c_{iM}
$\vdots$	$\vdots$	$\vdots$	$\iddots$	$\vdots$	$\ddots$	$\vdots$
N	c_{N1}	c_{N2}	$\ldots$	c_{Nj}	$\ldots$	c_{NM}
Library Size	L_1	L_2	$\ldots$	L_j	$\ldots$	L_M

OBJECTIVES

> **Goodness of Fit:**
Estimated vs Observed mean counts & zero probability

> **Type I Error Control:**
Mock comparisons on null datasets

> **Consistency & Replicability:**
Concordance at the top

> **Power:**
Parametric simulations

Figure 1: Starting from 41 Projects collected in 2 manually curated data repositories (*HMP16SData* and *curatedMetagenomicData* Bioconductor packages), 18 16S and 82 WMS datasets were downloaded. Biological samples belonged to several body sites (e.g. oral cavity), body subsites (e.g. tongue dorsum) and conditions (e.g. healthy vs disease).
Feature per sample count tables were used in order to evaluate several objectives: goodness of fit (GOF) for 5 parametric distributions, type I error control, consistency, replicability, and power for 14 differential abundance detection method. Methods, developed in metagenomics, bulk-RNAseq or sc-RNAseq, were ranked using empirical evaluations of the above cited objectives.

Results

We benchmarked a total of 14 approaches (Supplementary Table 2) on 100 real (Supplementary Table 1) and 28,800 simulated datasets (Supplementary Table 3), evaluating goodness of fit, type I error control, consistency, replicability, and power (Figure 1). The benchmarked methods include both DA methods specifically proposed in the metagenomic literature and methods proposed in the single-cell and bulk RNA-seq fields. The manually curated real datasets span a variety of body sites and characteristics (e.g., sequencing depth, alpha and beta diversity). The diversity of the data allowed us to test each method on a variety of circumstances, ranging from very sparse, very diverse datasets, to less sparse, less diverse ones.

We first analyzed 18 16S, 82 WMS and 28 scRNA-seq public datasets in order to assess whether scRNA-seq and metagenomic data are comparable in terms of sparsity. We observed overlap in the fractions of zero counts between the scRNA-seq, WMS, and 16S, but with scRNA-seq datasets having a lower distribution of sparsities (ranging from 12% to 75%) as compared to 16S (ranging from 55% to 83%) and WMS datasets (ranging from 35% to 89%) whose distributions of zero frequencies were not significantly different from each other (Wilcoxon test, W = 734, p = 0.377, Fig. S1a-b). To establish whether this difference was due to a different number of features and samples, which are intrinsically related to sparsity, we explored the role of library size and experimental protocol (Fig. S1c). scRNA-seq datasets showed a marked difference in terms of number of features and sparsity degree, as they are derived from different experimental protocols. Full-length data (e.g., Smart-seq) are on average sparser than droplet-based data (e.g., Drop-seq) but both are less sparse than 16S and WMS.

These results indicate that metagenomic data is even more sparse than scRNA-seq, and thus that DA models designed for scRNA-seq could at least in principle have good performance in a metagenomic context.

Goodness of fit

As different methods rely on different statistical distributions to perform DA analysis, we started our benchmark by assessing the goodness of fit (GOF) of the statistical models underlying each method on the full set of 16S and WMS data. For each model, we evaluated its ability to correctly estimate the mean counts and the probability of observing a zero (Fig. 2). We evaluated five distributions: (1) the negative binomial (NB) used in edgeR[16] and DeSeq2[17], (2) the zero-inflated negative binomial (ZINB) used in ZINB-WaVE[20], (3) the truncated Gaussian Hurdle model of

MAST[7], (4) the zero-inflated Gaussian (ZIG) mixture model of metagenomeSeq[13], and (5) the Dirichlet-Multinomial (DM) distribution underlying compositional approaches. The truncated Gaussian Hurdle model was evaluated following two data transformations, the default logarithm of the counts per million (logCPM) and the logarithm of the counts rescaled by the median library size (see Methods). Similarly, the ZIG distribution was evaluated considering the scaling factors rescaled by either one thousand (as implemented in the metagenomeSeq Bioconductor package) and by the median scaling factor (as suggested in the original paper). We assessed the goodness of fit for each of these models using the stool samples from the Human Microbiome Project (HMP) as a representative dataset (Fig. 2a - d); all other datasets gave similar results (Supplementary Fig. S2). A useful feature of this dataset is that a subset of samples was processed both with 16S and WMS and hence can be used to compare the distributional differences of the two data types. Furthermore, this dataset includes only healthy subjects in a narrow age range, providing a good testing ground for covariate-free models.

The NB distribution showed the lowest root mean square error (RMSE, see Methods) for the mean count estimation (MD), followed by the ZINB distribution (Fig. 2a-b). This was true both for 16S and for WMS data, in most of the considered datasets (Supplementary Fig. S1). Moreover, for both distributions, the difference between the estimated and observed means were symmetrically distributed around zero, indicating that the models did not systematically under- or over-estimate the mean abundances (Fig. 2a-b; Supplementary Fig. S2). Conversely, the ZIG distribution consistently underestimated the observed means, both for 16S and WMS and independently on the scaling factors (Fig. 2a-b). The Hurdle model was sensitive to the choice of the transformation: rescaling by the median library size rather than by one million reduced the RMSE in both 16S and WMS data (Fig. 2a-b). This was particularly evident in 16S data (Fig. 2a), in which the default logCPM values resulted in a substantial overestimation of the mean count, while the median library size scaling lead to under-estimation. Given clear problems with logCPM, we only used the median library size for MAST and the median scaling factor for metagenomeSeq in all subsequent analyses. The DM distribution overestimated observed means for low-mean count features and underestimated observed values for high-mean count features. This overestimation effect was more evident in WMS than in 16S.

Concerning the ability of models to estimate the probability of observing a zero (referred to as zero probability difference, ZPD), we found that Hurdle models provided good estimates of the observed zero proportion for 16S (Fig. 2c) and WMS datasets (Fig. 2d). The NB and ZINB distributions, on the other hand, tended to overestimate the zero probability for features with a low observed proportion of zero counts in 16S (Fig. 2c). In WMS data, the ZINB distribution perfectly fitted the observed proportion of zeros, while the NB and DM models tended to underestimate it (Fig. 2d). Finally, the ZIG distribution always underestimated the observed proportion of zeros, especially for highly sparse features (Fig. 2c-d).

In summary across all datasets, the best fitting distributions were the NB and ZINB: the NB distribution seemed to be particularly well-suited for 16S datasets, while the ZINB distribution seemed to better fit WMS data (Fig. 2e). We hypothesize that this is due to the different sequencing depths of the two platforms. In fact, while our 16S datasets have an average of 4891 reads per sample, in WMS the mean depth is $3.6\text{x}10^8$ ($3\text{x}10^8$ for HMP). To confirm this observation, we carried out a simulation experiment by down-sampling reads from deep-sequenced WMS samples (rarefaction): while the need for zero inflation seemed to diminish as we got closer to the number of reads typical of the corresponding 16S experiments, the profile did not completely match between approaches (Supplementary Fig. S4b). This suggests that, while sequencing depth is an important contributing factor, it is not enough to completely explain the distributional differences between the two platforms.

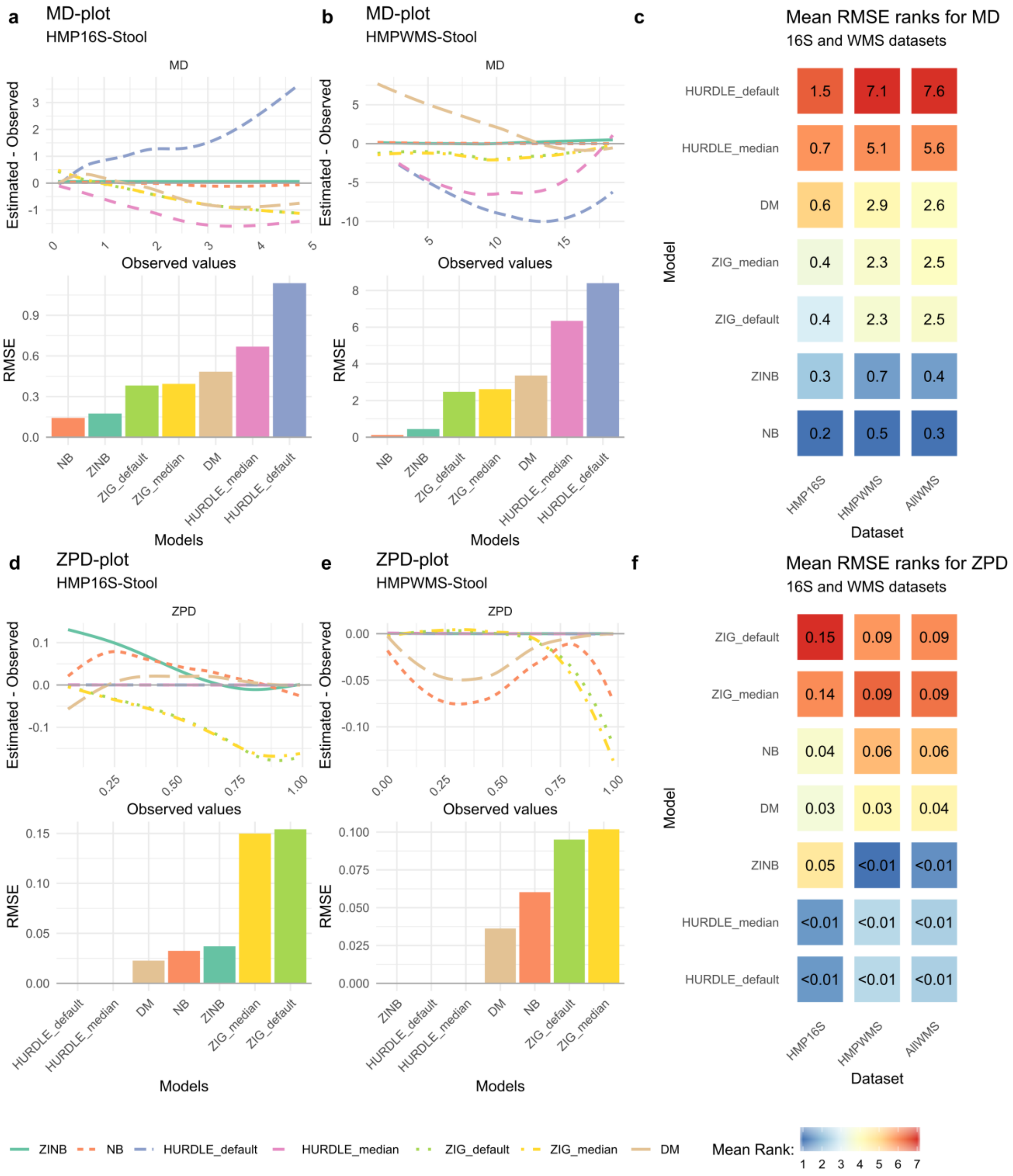


Figure 2: a. Mean-Differences (MD) plot and Root Mean Squared Errors (RMSE) for HMP 16S Stool samples. **b.** MD plot and RMSE for HMP WMS Stool samples. **c**. Average rank heatmap for MD performances in HMP 16S datasets, HMP WMS datasets and all other WMS project datasets. The value inside each tile refers to the average RMSE value on which ranks are computed. **d.** Zero Probability-Differences (ZPD; see Methods) plot and RMSE for HMP 16S Stool samples. **e.** ZPD plot and RMSE for HMP WMS Stool samples. **f**. Average rank heatmap for ZPD performances in HMP 16S datasets, HMP WMS datasets and all other WMS project datasets. The value inside each tile refers to the average RMSE value on which ranks are computed.

Type I error control

We next sought to evaluate type I error rate control of each method, i.e., the probability of the statistical test to call a feature DA when it is not. To do so, we considered mock comparisons between the same biological Stool HMP samples (using the same Random Sample Identifier in both 16S and WMS), in which no true DA is present. Briefly, we randomly assigned each sample to two experimental groups and compared them, repeating the process ten times (see Methods for additional details). In this setting, the p-values of a perfect test should be uniformly distributed between 0 and 1 (ref.[25]) and the false positive rate (FPR or observed α), which is the observed proportion of significant tests, should match the nominal value (e.g., $\alpha = 0.05$).

To evaluate the impact of both the normalization step and the estimation and testing step in bulk RNA-seq inspired methods, we included in the comparison both edgeR with its default normalization (TMM), as well as with DESeq2 recommended normalization ("poscounts", i.e., the geometric mean of the positive counts) and vice versa (Table S2). Similarly, because the zinbwave observational weights can be used to apply several bulk RNA-seq methods to single-cell data[21], we have included in the comparison edgeR, DESeq2, and limma-voom with zinbwave weights.

The qq-plots and Kolmogorov-Smirnov (KS) statistics in Figure 3 show that most methods achieved a p-value distribution reasonably close to the expected uniform. The notable exceptions in the 16S experiment were edgeR with TMM normalization and robust dispersion estimation (edgeR_TMM_robustDisp), metagenomeSeq, and ALDEx2 (Fig. 3a-b). While the former two appeared to employ liberal tests, the latter was conservative in the range of p-values that are typically of interest (0 - 0.1). In the WMS data, departure from uniformity was observed for metagenomeSeq and edgeR_TMM_robustDisp, which employed liberal tests, and Wilcoxon and ALDEx2, which appeared to be conservative in the range of interest (Fig. 3b-c). We note that in the context of DA, liberal tests will lead to many false discoveries, while conservative tests will control the type I error at a cost of reduced power, potentially hindering true discoveries.

We next recorded the FPR by each method (by definition all discoveries are false positives in this experiment) and compared it to its expected nominal value. This analysis confirmed the tendencies observed in Figures 3a-b and 3c-d. In particular, edgeR_TMM_robustDisp and metagenomeSeq were very liberal in both 16S (Fig. 3e) and WMS data (Fig. 3f); in the case of metagenomeSeq, as much as 30% of the features were deemed DA in the 16S datasets when claiming a nominal FPR of 5% (Fig. 3e). ALDEx2, scde and MAST, albeit conservative, were able to control type I error. In between these two extremes, edgeR, DESeq2 and limma showed an observed FPR slightly higher than its nominal value. In particular, DESeq2-based methods were very close to the nominal FPR for 16S (Fig. 3e), while limma-voom was the closest to the nominal value in the WMS data (Fig. 3f). The zinbwave weights showed mixed results: edgeR with zinbwave weights was always better than the unweighted versions, while the weights did not help DESeq2 and limma in controlling the type I error rate. Taken together, these results suggest that the majority of the methods does not control the type I error rate, both in 16S and WMS data, confirming previous findings[10,12]. However, for most approaches, the observed FPR is only slightly higher than its nominal value, making the practical impact of this result unclear.

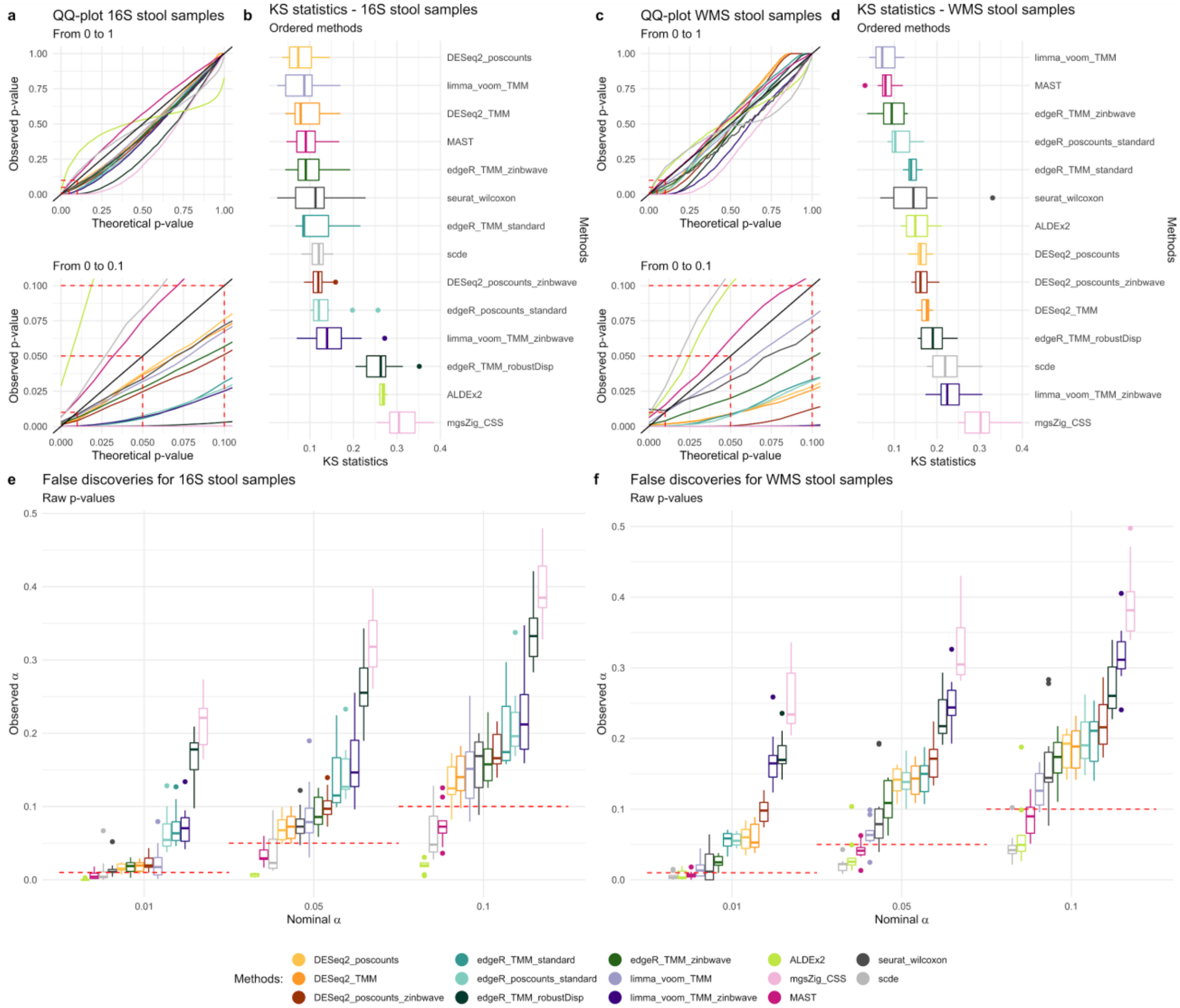


Figure 3: a. Quantile-quantile plot from 0 to 1 and 0 to 0.1 zoom for DA methods in 41 16S HMP stool samples. Average curves for mock comparisons reported. **b.** Kolmogorov-Smirnov statistic boxplots for DA methods in 41 16S HMP stool samples. **c.** Quantile-quantile plot from 0 to 1 and 0 to 0.1 zoom for DA methods in 41 WMS HMP stool samples. Average curves for mock comparisons reported. **d.** Kolmogorov-Smirnov statistic boxplots for DA methods in 41 WMS HMP stool samples. **e.** Boxplots for the proportion of raw p-values lower than 0.01, 0.05, 0.1 values of the commonly used thresholds of nominal α for 41 16S stool samples. **f.** Boxplots for the proportion of raw p-values lower than 0.01, 0.05, 0.1 values of the commonly used thresholds of nominal α for 41 WMS stool samples.

Consistency

To measure the ability of each method to produce consistent or replicable results in independent data, we looked at six datasets[22,23,26–28] (Supplementary Table S3), with different alpha and beta diversity, as well as different amounts of DA between two experimental conditions (Supplementary Figure S5). Each dataset was randomly split in two equally sized subsets and each method was separately applied to each subset. The process was repeated ten times (see Methods for details). To assess the ability of methods to return consistent results from independent samples, we employed the Concordance At the Top[29] (CAT) measure to assess between-method concordance (BMC) by comparing the list of DA features across methods in the subset and reporting the average value. We used BMC to (i) group methods based on their degree of agreement, and (ii) identify those methods sharing the largest amount of discoveries with the majority of the other methods. Although consistency is not a guarantee of validity, it is a requirement of validity, so methods sharing the largest amount of discoveries with the majority of other methods may be more likely to also be producing valid results.

Concordance analysis performed on 16S Tongue Dorsum vs Stool dataset (Fig. 4a) showed that the methods clustered within two distinct groups: the first comprising all methods that include a TMM normalization step, and the second containing all the other approaches (Fig. 4a). Even within the second group, methods segregated by normalization, as can be seen by the tight clustering of all the methods that include a poscount normalization step (Fig. 4a). This indicates that, in 16S data, the choice of the normalization has a pronounced effect on inferential results, even more so than the choice of the statistical test. A similar result was previously observed in bulk RNA-seq data[30]. The use of observational weights to account for zero inflation did not seem to matter in these data, and in general, scRNA-seq methods did not agree with each other (Fig. 4a).

A different picture emerged from the analysis of the WMS data (Fig. 4b. Here, methods clustered by the testing approach. The top cluster comprised the bulk RNA-seq methods with the inclusion of the Wilcoxon nonparametric approach, SCDE, and metagenomeSeq. The bottom cluster consisted of the scRNA-seq methods, ALDEx2, and edgeR robust. Overall, the methods based on NB generalized linear models showed the highest BMC values. When observational weights were added to those models, the BMC decreased, but still a good level of concordance was observed with their respective unweighted version.

We noted that the BMC is highly dataset-specific and depends on the amount of DA between the compared groups. Indeed, BMC decreased with the beta diversity of the dataset, and the role of normalization became less clear (Supplementary Fig. S6).

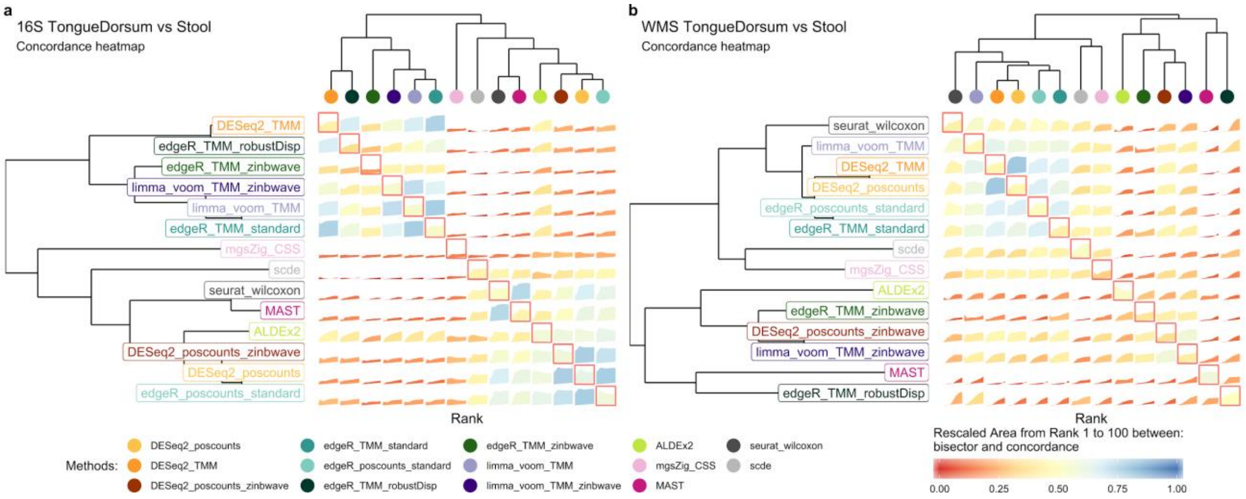


Figure 4: a. BMC and WMC (main diagonal) averaged values from rank 1 to 100 for DA methods evaluated in replicated 16S Tongue Dorsum vs Stool comparisons. **b.** BMC and WMC (main diagonal) averaged values from rank 1 to 100 for DA methods evaluated in replicated WMS Tongue Dorsum vs Stool comparisons.

Replicability

The replicability of methods between subsets was evaluated using again the CAT metric but assessing the within-method concordance (WMC).

WMC was clearly dataset-dependent, showing high levels of concordance in datasets with high differential signal (e.g., tongue vs. stool, Fig. 5a) and low concordance in datasets with low differential signal (e.g., supragingival vs. subgingival, Fig. 5e). Overall, the reproducibility of the results in WMS studies was slightly higher than that of 16S datasets. In terms of method comparison, MAST showed relatively high reproducibility in all WMS datasets and lower reproducibility in all 16S datasets (Fig. 5). Similarly, the addition of zinbwave weights to edgeR and DESeq2 did not always help: it was sometimes detrimental, e.g., in the schizophrenia dataset (Fig. 5d), and sometimes led to an improvement in reproducibility, e.g., in the CRC dataset (Fig. 5f). The schizophrenia dataset had the lowest numerosity among all the datasets evaluated, suggesting that sample size may play an important role in estimating zinbwave weights. While this analysis confirmed the unsatisfactory performance of metagenomeSeq (Fig 5a,f), ALDEx2, which was very conservative in terms of type I error control (Fig. 3), showed overall good performance (Fig. 5a,b).

Summarizing, both BMC and WMC are highly dependent on the amount of DA observed in the dataset: higher DA leads to a higher concordance. Moreover, WMC was similar among the compared methods, indicating that the reproducibility of the DA results depends more on the strength of DA than on the choice of the method (Figure 5).

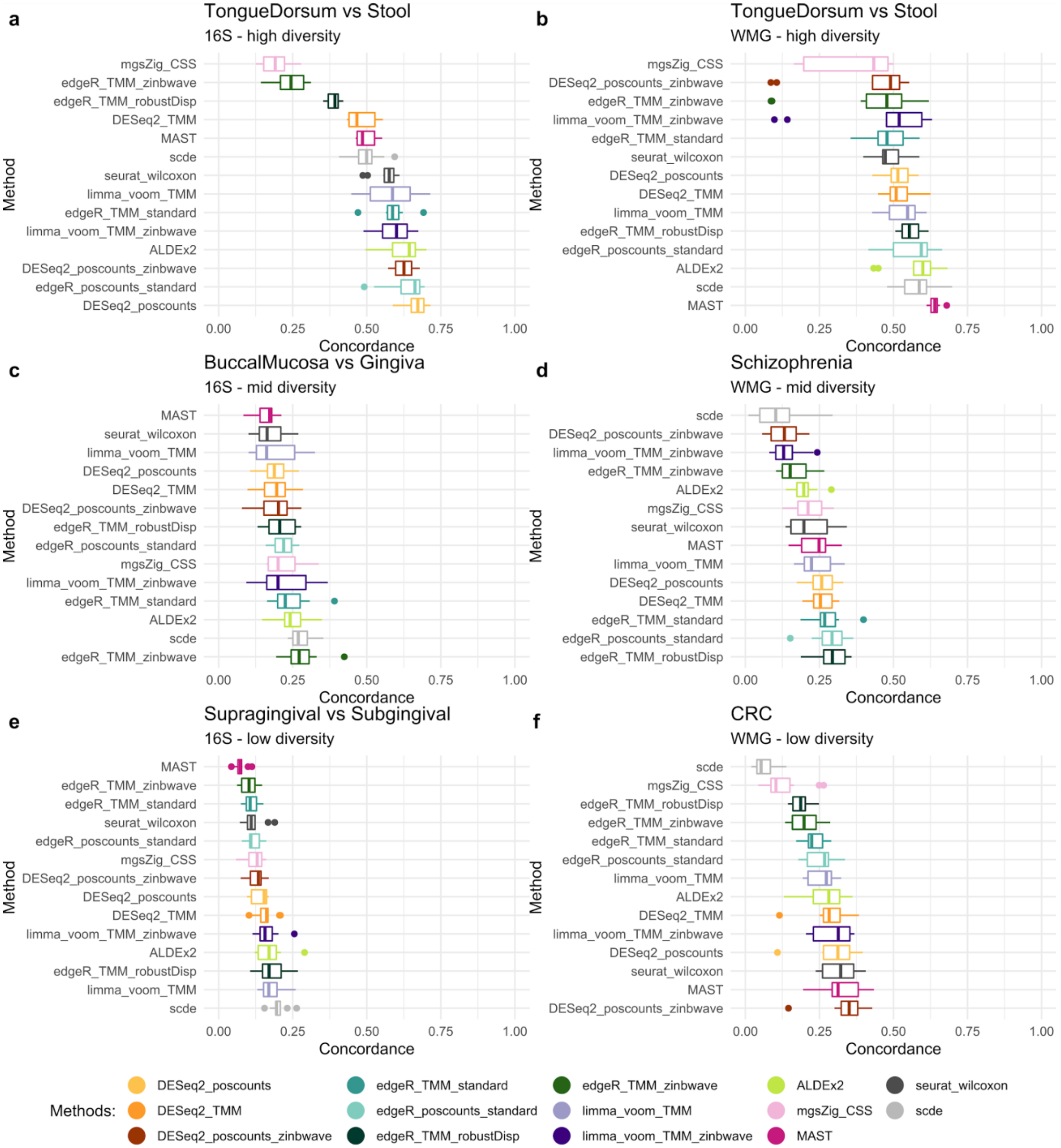


Figure 5: a. Boxplot of WMC on high level diversity, Tongue Dorsum vs Stool, 16S datasets. **b.** Boxplot of WMC on high diversity, Tongue Dorsum vs Stool, WMS datasets. **c.** Boxplot of WMC on mid level diversity, Buccal Mucosa vs Attached Keratinized Gingiva, 16S datasets. **d.** Boxplot of WMC on mid level diversity, Schizophrenic vs Healthy Control saliva samples, WMS datasets. **e.** Boxplot of WMC on low level diversity, Supragingival vs Subgingival plaque, 16S datasets. **f.** Boxplot of WMC on low level diversity, Colon Rectal Cancer patient vs Healthy Control stool samples, WMS datasets.

Parametric simulations

While the results of the experimental datasets are best for assessing model fit and consistency of discoveries, the lack of ground truth makes it impossible to assess the validity of discoveries. For this reason, we turned to simulated data to explore the properties of the methods in more detail. Here, we specifically asked whether it was important to model zero inflation and, given the results of our GOF analysis (Fig. 2), we only used the NB and ZINB distributions to simulate the data. Briefly, for each distribution we simulated 7200 and 19200 scenarios respectively, mimicking both 16S and WMS data, and varying the sample size, the proportion of DA features, and the amount of the effect. We also varied the proportion of zeros and whether there was an interaction between the amount of zeros and DA (sparsity effect, see Methods for details).

Figure 6a summarizes the performance of all methods according to all the different variables involved in the simulation procedure. To condense all the results into a single figure, we ranked the methods summarizing the pAUROC performance independently for each simulation parameter. Importantly, this summary ignored many interaction effects (e.g., whether the fold effect influences the results differently depending on the technology). Albeit simplified, this summary is nonetheless useful to get an overview of each method’s performance. Supplementary File 1 shows the joint effects of selected variables. We used the partial Area Under the Receiver Operating Characteristic Curve (pAUROC) between 0 and 0.1 of False Positive Rate (FPR) values as an indicator of the method performances, since it only considers the range of FPR values that are important in practice and measures the ability of methods to correctly detect true differential abundant features. A method-specific pattern is clearly visible, indicating the robustness and coherence of each method across different simulation scenarios (Fig. 6a).

Briefly, edgeR with TMM normalization (with and without zinbwave weights) and DESeq2 with poscount normalization and zinbwave weights were the overall best methods (Fig. 6a). The other DESeq2-based methods were close second. Unsurprisingly, the parametric distribution that generated the data had great influence on the method performances. Indeed, ZINB generated datasets showed lower mean values for all methods because of the increase in sparsity. All methods’ performances increased as the sample size and/or the fold effect increased (Fig. 6a). Focusing on the amount of zero counts, we observed that the mean performance increased when the sparsity effect increased from 0.05 to 0.15, not only for edgeR and DESeq2 based method, but for limma-voom, ALDEx2 and Wilcoxon (Fig. 6a).

Confirming our real data results, metagenomeSeq, scde, and edgeR robust performed poorly. On the other hand, MAST, which showed mixed results in real data, did not behave in simulations, partly because of the misspecified model with respect to the data generating distribution.

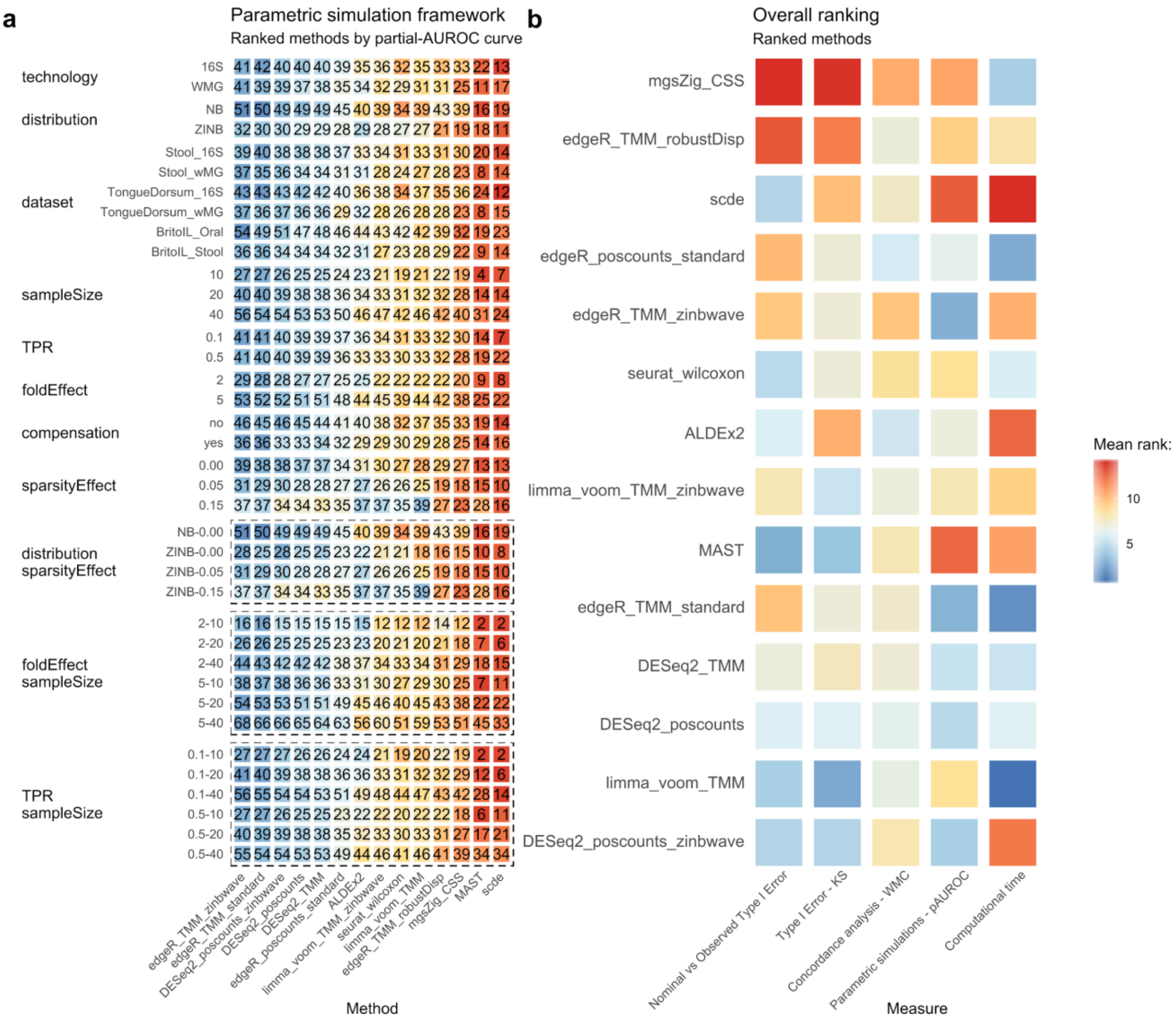


Figure 6: a. Comprehensive univariate and bivariate (dotted line boxes) evaluation of the pAUROC curves from FPR 0 to 0.1 (For a full multivariate method comparison see Supplementary File 1). Average ranks range from 1 to 14, lower values correspond to better performances. The value inside each tile refers to the average rescaled partial AUROC value multiplied per 100, on which ranks are computed. Higher pAUROC values correspond to better performance. **b.** Overall method ranking regarding to 5 evaluation criteria. Average ranks range from 1 to 14, lower values correspond to better performances.

Discussion

We have investigated different theoretical and practical issues related to the analysis of metagenomic data. The main objective of the study was to compare several DA detection methods adapted from bulk RNA-seq, single-cell RNA-seq, or specifically designed for metagenomics. Unsurprisingly, there is no single method that outperforms all others in all the tested scenarios. As often is the case in high-throughput biology, the results are data-dependent and careful data exploration is needed to make an informed decision on which workflow to apply to a specific dataset. We recommend applying our explorative analysis framework to gain useful insights about the assumptions of each method and their suitability given the data at hand. To this end, we provide all the R scripts to easily reproduce the analyses of this paper on any given dataset (see code availability).

Our GOF analysis highlighted the advantages of using count models for the analysis of metagenomics data. The need for modeling zero inflation seemed dependent on whether the data come from 16S or WMS experiments. The difference between these two approaches translates to different count data structures: while for WMS many features are characterized by a clearly visible bimodal distribution (with a point mass at zero and another mass, quite far from zero, at the second positive mode), 16S data are as sparse as or even more sparse than WMS data, presenting for many features a less clearly bimodal distribution (Supplementary Fig. S4a). This difference is probably due to a mix of factors: primarily sequencing depth, but also different taxonomic classification between technologies (entire metagenomic sequences versus clusters of similar amplicon sequences), bioinformatics methods for data preprocessing, etc. Further analyses are needed to inspect this unsolved issue and related efforts are ongoing in the single-cell RNA-seq literature, where similar differences are observed between protocols with and without unique molecular identifiers[31,32].

After recognizing the better fit of NB and ZINB over other distributions, we expected a better performance for those methods based on the count distributions. Indeed, DESeq2 (with and without zinbwave weights) and limma-voom were the methods with the most consistent performance (Fig. 6b). It is curious that the methods based on edgeR, in which the distribution of the counts is still hypothesized NB, gave more variable results.

The parametric simulation framework is useful to inspect how individual characteristics of the data-generating distribution impact the sensitivity and specificity of the methods. As the entire analysis was supported by real data, we decided to focus only on a very simple but easily reproducible implementation of the NB and ZINB distributions for the simulations. The choice was justified by our GOF analysis on real datasets. Unsurprisingly, the sample size and the effect size were the characteristics that had the most impact on method performances. This translates into an evident suggestion for experimental design: large sample sizes are needed when dealing with low effect sizes. Our simulation framework can in principle be used for power calculations in the context of DA analysis.

As shown in Figure 6a, all methods displayed drastically lowered performance for datasets with increased sparsity. A way to decrease sparsity is to filter rare features or to impute the zero counts: in our simulations, for the sake of simplicity, we decided to keep the features with more than ten reads in at least two samples. Other works[13] or analysis pipelines[33], suggest several different filters that may have an impact on the results. Beyond the filtering choice, for a method that does not treat zero counts in any special way, it is easier to detect differentially abundant features when zero rates are clearly different between experimental groups (Fig. 6a). Indeed, the change in the proportion of zeros will affect the mean and hence the DA statistic. The same situation can be tricky for methods that downweight the contribution of zero counts (e.g., zinbwave; Fig. 6a). On the other hand, when zero counts are equally present in both groups, downweighting them is favorable (Fig. 6a). These two opposite situations are similar to the difference between treating the zero counts as “biological” or “technical” and more research is needed to understand whether zero-inflated model can help classify the two.

Metagenomic data are inherently compositional, but whether compositionality should be taken into account in the statistical model is a debated topic in the literature[9,13,24,34,35]. Here, we found that compositional methods (using the Dirichlet Multinomial distribution) did not outperform non-compositional methods designed for count data. This can be explained by two considerations. First, compositional methods assume that the data arise from a multinomial distribution, with *n* trials (reads) and a vector *p* indicating the probability of the reads to be mapped to each OTU. Note that in metagenomic studies, we have a large *n* (number of sequenced reads) and small *p* (since there are many OTUs, the probability of each read to map to any given OTU is small). In this setting, the Poisson distribution is a good approximation of the multinomial. Similarly, the negative binomial is a good approximation of the Dirichlet-Multinomial[31]. Secondly, some normalizations, such as the geometric mean method implemented in DESeq2 or the trimmed mean of M-values of edgeR, have size factors mathematically equivalent or very similar to the compositional log-ratio proposed by Aitchison[24,36]. This has been shown to reduce the impact of compositionality on DA results[37]. We did not test the ANCOM package[38] because it was too slow for assessment in the simulation studies. Nonetheless we believe that the above comparison is an adequate assessment of compositional vs non-compositional approaches.

The most time-consuming methods were scde, ALDEx2, MAST and zinbwave. In simulated datasets with 40 samples per condition and less than 1000 features we observed an average elapsed time of around 5 minutes for scde; ALDEx2 took an average of 30 seconds, MAST took an average of 14 seconds, DESeq2 less than 8 seconds, and edgeR and limma-voom took less than 1 second, although observational weights estimation took an average of 18 seconds (Supplementary Table S5 and Supplementary Fig. S7).

Concluding, as already noted in recent publications[10–12], the perfect method does not exist. However, taken together, our analyses suggest that DESeq2 and limma-voom show the most consistent performance across all datasets (Fig. 6b). In general, we recommend a careful exploratory data analysis and we present a framework that can help scientists make an informed choice in a dataset-specific manner. In this study, we did not find evidence that bespoke differential abundance methods outperform methods developed for the differential expression analysis of RNA-seq data. However, new directions in DA method development, e.g., leveraging the phylogenetic tree, are promising[39,40].

Methods

Datasets

The *HMP16SData*[22] (v1.2.0) and *CuratedMetagenomicData*[23] (v1.12.3) Bioconductor packages are used to download high-quality, uniformly processed, and manually annotated human microbiome profiles for thousands of people, using 16S and Whole Metagenome shotgun sequencing technologies respectively. *HMP16SData* comprises the collection of 16S data from the Human Microbiome Project (HMP), while *CuratedMetagenomicData* contains data from several projects. Gene-level counts for a collection of public scRNA-seq datasets are downloaded from *scRNAseq* (v 1.99.8) Bioconductor package.

While the latter datasets are used only for a comparison between technologies, the formers are widely used for all the analyses. A complete index with dataset usage is reported in Supplementary Table S1.

HMP16SData is split by body subsite in order to obtain 18 separated datasets. Stool and Tongue Dorsum datasets are selected for example purposes thanks to their high sample size. The same is done on CuratedMetagenomicData HMP dataset, obtaining 9 datasets. Moreover, for the evaluation of type I error control, 41 stool samples with equal RSID, in both 16S and WMS, are used to compare DA methods. For each research project, CuratedMetagenomicData is split by body site and treatment or disease condition, in order to create homogeneous sample datasets. A total of 82 WMS datasets are created.

A total of 100 datasets are evaluated, however for the CAT analysis, non-split by condition or body subsite datasets are evaluated (e.g. Tongue Dorsum vs Stool in HMP, 2012 for both 16S and WMS).

To consider the complexity and the variety of several experimental scenarios, an attempt to select a wide variety of datasets for the analysis is done. The datasets are chosen based on several criteria: the sample size, the homogeneity of the samples or the availability of the same RSID for both technologies.

Statistical Models

The following distributions were fitted to each dataset, either by directly modeling the read counts, or by first applying a logarithmic transformation:

- Negative Binomial (NB) model, as implemented in the *edgeR* (v3.24.3) Bioconductor package (on read counts);

- Zero Inflated Negative Binomial (ZINB), as implemented in the *zinbwave* (v1.4.2) Bioconductor package (on read counts);
- Truncated Gaussian hurdle model, as implemented in the *MAST* (v1.8.2) Bioconductor package (on log count);
- Zero Inflated Gaussian (ZIG), as implemented in the *metagenomeSeq* (v1.24.1) Bioconductor package (on log count).
- Dirichlet-Multinomial (DM), as implemented in the *MGLM* (v0.2.0) CRAN R package.

Negative Binomial (NB). We used the implementation of the NB model of the *edgeR* Bioconductor package. In particular, normalization factors were calculated with the Trimmed Mean of M-values (TMM) normalization[41] using the *calcNormFactors* function; common, trended and tagwise dispersions were estimated by *estimateDisp*, and a negative binomial generalized log-linear model was fit to the read counts of each feature, using the *glmFit* function.

Zero-Inflated Negative Binomial (ZINB). We used the implementation of the ZINB model of the *zinbwave* Bioconductor package. We fitted a ZINB distribution using the *zinbFit* function. As explained in the original paper, the method can account for various known and unknown, technical and biological effects[20]. However, to avoid giving unfair advantages to this method, we did not include any latent factor in the model (K = 0). We estimated a common dispersion for all features (common_dispersion = TRUE) and we set the likelihood penalization parameter epsilon to 1e10 (within the recommended set of values[21]).

Truncated Gaussian Hurdle Model. We used the implementation of the *MAST* Bioconductor package. After a log2 transformation of the reascaled counts with a pseudocount of 1, a zero-truncated Gaussian distribution was modeled through generalized regression on positive counts, while a logistic regression modeled feature expression/abundance rate. As suggested in the MAST paper[7], cell detection rate (CDR) which is computed as the proportion of positive count features for each sample, was added as a covariate in the discrete and continuous model matrices as normalization factor.

Zero-Inflated Gaussian. *metagenomeSeq* Bioconductor package was used to implement a ZIG model for log2 transformed counts with a pseudocount of 1, rescaled by the median of all normalization factors or by 1e03 which gives the interpretation of "count per thousand" to the offsets. The *CumNormStat* and *CumNorm* functions were used to perform Cumulative Sum

Scaling (CSS) normalization, which accounts for specific data characteristics. Normalization factors were included in the regression through the *fitZig* function.

Note that both *MAST* and *metagenomeSeq* are applied to the normalized, log-transformed data. We evaluated both models, using their default scale factor $\log_2\left(\frac{counts \cdot 10^6}{libSize}+1\right)$ for *MAST* and $log_2\left(\frac{normFacts}{1000}+1\right)$ for *metagenomeSeq,* as well as by rescaling the data to the median library size[13], $log_2\left(\frac{counts{\cdot}median(libSize)}{libSize}+1\right)$ and $log_2\left(\frac{normFacts}{median(normFacts)}\right)$, respectively.

Dirichlet-Multinomial. The *MGLM* package was used to fit a Dirichlet-Multinomial regression model for counts. The *MGLMreg* function with *dist = "DM"*, allowed the implementation of the above model and the estimation of the parameter values.

Goodness of Fit (GOF)

To evaluate the goodness of fit of the models, we computed the mean differences between the estimated and observed values for several datasets.

For each model, we evaluated two distinct aspects: its ability to correctly estimate the mean counts (plotted in logarithmic scale with a pseudo-count of 1) and its ability to correctly estimate the probability of observing a zero, computed as the difference between the probability of observing a zero count according to the model and the observed zero frequencies (Zero Probability Difference, ZPD). We summarized the results by computing the Root Mean Squared Error (RMSE) of the two estimators. The lower the RMSE, the better the fit of the model.

This analysis was repeated for 100 datasets available in *HMP16SData* and *CuratedMetagenomicData* (Table S1 and Supplementary Figure S2).

Assuming homogeneity between samples inside the same body subsite or study condition, we specified a model consisting of only an intercept, or including a normalization covariate.

Differential abundance detection methods

DESeq2. The *DESeq2* (v1.22.2) Bioconductor package fits a negative binomial model for count data. DESeq2 default data normalization is the so-called Relative Log Expression (RLE) based on scaling each sample by the median ratio of the sample counts over the geometric mean counts across samples. As 16S and WMS data sparsity may lead to a geometric mean of zero, it is replaced by n-th root of the product of the non-zero counts (which is the geometric mean of the positive count values) as proposed in *phyloseq* package[33] and implemented in the DESeq2 *estimateSizeFactors* function with option *type=”poscounts”*. We also tested DESeq2 with TMM normalization (see below). Moreover, as proposed in Van den Berge et al.[21], observational weights are supplied in the *weights* slot of the *DESeqDataSet* class object to account for zero inflation. Observational weights were computed by the *ComputeObservationalWeights* function of the *zinbwave* package. To test for DA, we used a Likelihood Ratio Test (LRT) to compare the reduced model (intercept only) to the full model with intercept and group variable. The p-values were adjusted for multiple testing via the Benjamini-Hochberg (BH) procedure. Some p-values were set to NA via the *cooksCutoff* argument that prevents rare or outlier features from being tested.

edgeR. The *edgeR* Bioconductor package fits a negative binomial distribution, similarly to DESeq2. The two approaches differ mainly in the normalization, dispersion parameter estimation, and default statistical test. We examined different procedures by varying the normalization and the dispersion parameter estimation: *edgeR_TMM_standard* involves TMM normalization and tagwise dispersion estimation through the *calcNormFactors* and *estimateDisp* functions respectively (with default values). Analogously to DESeq2, “*poscounts”* normalization was used in addition to TMM in *edgeR_poscounts_standard* to investigate normalization impact. We also evaluated the impact of employing a robust dispersion estimation, accompanied with a quasi-likelihood F test through the *estimateGLMRobustDisp* and *glmQLFit* functions respectively (*edgeR_TMM_robustDisp)*. As with DESeq2, *zinbwave* observational weights were included in the *weights* slot of the *DGEList* object in *edgeR_TMM_zinbwave* to account for zero inflation, through a weighted F test. Benjamini-Hochberg correction was used to adjust p-values for multiple testing.

Limma-voom. The *limma* Bioconductor package (v3.38.3) includes a *voom* function that (i) transforms previously normalized counts to logCPM, (ii) estimates a mean-variance relationship and (iii) uses this to compute appropriate observational-level weights[18]. To adapt *limma-voom* framework to zero-inflations, *zinbwave* weights have been multiplied by *voom* weights as done previously[21]. The residual degrees of freedom of the linear model were adjusted before the empirical Bayes variance shrinkage and were propagated to the moderated statistical tests. Benjamini-Hochberg correction method was used to correct p-values.

ALDEx2. *ALDEx2* is a Bioconductor package (v1.14.1) that uses a Dirichlet-multinomial model to infer abundance from counts[14]. The *aldex* method infers biological and sampling variation to calculate the expected False Discovery Rate, given the variation, based on several tests. Technical variation within each sample is estimated using Monte-Carlo draws from the Dirichlet distribution. This distribution maintains the proportional nature of the data while scale-invariance and sub-compositionally coherence of data, is ensured by centered log-ratio (CLR). This removes the need for a between sample normalization step. In order to obtain symmetric CLRs, the *iqlr* argument is applied, which takes, as the denominator of the log-ratio, the geometric mean of those features with variance calculated from the CLR between the first and the third quantile. Statistical testing is done through Wilcoxon Rank Sum test, even if Welch's t, Kruskal-Wallis, Generalized Linear Models and correlation tests were available. Benjamini-Hochberg correction method was used to correct the p-values for multiple testing.

metagenomeSeq. *metagenomeSeq* is a Bioconductor package designed to address the effects of both normalization and under-sampling of microbial communities on disease association detection and testing feature correlations. The underlying statistical distribution for $log_2(count + 1)$ is assumed to be a zero-inflated Gaussian mixture model. The mixture parameter is modeled through a logistic regression depending on library sizes, while the Gaussian part of the model is a generalized linear model with a sample specific intercept which represent the sample baseline, a sample specific offset computed by Cumulative Sum Scaling (CSS) normalization and another parameter which represents the experimental group of the sample. We opted for the implementation suggested in the original publication[13], where CSS scaling factors are divided by the median of all the scaling factors instead of dividing them by 1000 (as done in the Bioconductor package). An EM algorithm is performed by *fitZig* function to estimate all parameters. An empirical Bayes approach is used for variance estimation and a moderated t-test is performed to identify differentially abundant features between conditions. Benjamini-Hochberg correction method was used to account for multiple testing.

MAST. *MAST* is a Bioconductor package for managing and analyzing qPCR and sequencing-based single-cell gene expression data, as well as data from other types of single-cell assays. The package also provides functionality for significance testing of differential expression using a Hurdle model. Zero rate represents the discrete part, modelled as a binomial distribution while $log_2\left(\frac{counts_{i,j}\cdot median(libSize)}{libSize_j} + 1\right)$ where i and j represents the i-th feature and the j-th sample respectively, is used for the continuous part, modelled as a Gaussian distribution. The kind of data considered, different from scRNA-seq, doesn't allow the usage of the adaptive thresholding procedure suggested in the original publication[7]. Indeed, because of the amount of feature loss if adaptive thresholding is applied, the comparison of MAST with other methods would be unfair. However, a normalization variable is included in the model. This variable captures information about each feature sparsity related to all the others; hence, it helps to yield more interpretable results and decreases background correlation between features. The function *zlm* fits the Hurdle model for each feature: the regression coefficients of the discrete component are regularized using a Bayesian approach as implemented in the *bayesglm* function; regularization of the continuous model variance parameter helps to increase the robustness of feature-level differential expression analysis when a feature is only present in a few samples. Because the discrete and continuous parts are defined conditionally independent for each feature, tests with asymptotic χ^2 null distributions, such as the Likelihood Ratio or Wald tests, can be summed and remain asymptotically χ^2, with the degrees of freedom of the component tests added. Benjamini-Hochberg correction method was used to correct p-values.

Wilcoxon Rank Sum Test. *Seurat* (v2.3.4) R package is a scRNA-Seq data analysis toolkit for the analysis of single-cell RNA-seq[19]. Briefly, counts were scaled, centered and LogNormalized. Wilcoxon Rank-Sum test for detecting differentially abundant features was performed via the *FindMarkers* function. Rare features, which are present in a fraction lower than 0.1 of all samples, and weak signal features, which have a log fold change between condition lower than 0.25, are not tested. Benjamini-Hochberg correction method was used to correct p-values.

SCDE - Single Cell Differential Expression. The *scde* Bioconductor package (v1.99.1) with *flexmix* package (v2.3-13) implements a Bayesian model for scRNA-seq data[8]. Read counts observed for each gene are modeled using a mixture of a negative binomial (NB) distribution (for the amplified/detected transcripts) and low-level Poisson distribution (for the unobserved or background-level signal of genes that failed to amplify or were not detected for other reasons). The *scde.error.models* function was used to fit the error models on which all subsequent calculations rely. The fitting process is based on a subset of robust genes detected in multiple cross-cell comparisons. Error model for each group of cells were fitted independently (using two different sets of “robust” genes). Translating in a metagenomic context, cells corresponds to samples and genes to OTU or amplicon sequence variants. Some adjustments were needed to calibrate some function default values such as the minimum number of features to use when determining the expected abundance magnitude during model fitting. This option, defined by the *min.size.entries* argument, set by default at 2000, was too big for many 16S or WMS experiment scenarios: as we usually observe around 1000 total features per dataset (after filtering out rare ones), we decided to replace 2000 with the 20% of the total number of features, obtaining a dataset-specific value. Particularly poor samples may result in abnormal fits and were removed as suggested in the *scde* manual. To test for differential expression between the two groups of samples a Bayesian approach was used: incorporating evidence provided by the measurements of individual samples, the posterior probability of a feature being present at any given average level in each subpopulation was estimated. To moderate the impact of high-magnitude outlier events, bootstrap resampling was used and posterior probability of abundance fold-change between groups was computed.

Type I error control

For this analysis, we used the collection of HMP Stool samples in *HMP16SData* and *CuratedMetagenomicData*. The multidimensional scaling (MDS) plot of the beta diversity did not show patterns associated with known variables (Supplementary Fig. S3), hence we assumed no differential abundance. All samples with the same Random Subject Identifier (RSID) in 16S and WMS were selected in order to easily compare the two technologies. 41 biological samples were included.

Starting from the 41 samples, we randomly split the samples in two groups: 21 assigned to Group 1 and 20 to Group 2. We repeated the procedure 10 times. We applied the DA methods to each randomly split dataset. Every method returned a p-value for each feature. DESeq2 and Seurat_Wilcoxon returned some NA p-values. This is due to feature exclusion criteria, based on distributional assumptions, performed by these methods (see above).

We compared the distribution of the observed p-values to the theoretical uniform distribution, as no differential abundant features should be present. This was summarized in the qq-plot where the bisector represents a perfect correspondence between observed and theoretical quantiles of p-values. For each theoretical quantile, the corresponding observed quantile was obtained averaging the observed p-values' quantiles from all 10 datasets. Departure from uniformity was evaluated with a Kolmogorov-Smirnov statistic. P-values were also used to compare the number of false discoveries with 3 common thresholds: 0.01, 0.05 and 0.1.

Consistency and replicability

We used the Concordance At the Top (CAT) to evaluate the consistency and replicability of each differential abundance methods. Starting from two lists of features ranked by p-values, the CAT statistic was computed in the following way. For a given integer *i*, concordance is defined as the cardinality of the intersection of the top *i* elements of each list, divided by *i*, i.e. $\frac{\#\{L_{1:i} \cap M_{1:i}\}}{i}$, where *L* and *M* represent the two lists. This concordance was computed for values of *i* from 1 to *R*.

Depending on the study, only a minority of features may be expected to be differentially abundant between two experimental conditions. Hence, the expected number of differentially abundant features is a good choice as the maximum rank *R*. In fact, CAT displays high variability for low ranks as few features are involved, while concordance tends to 1 as *approaches* the total number of features, becoming uninformative. We set *R* = 100, considering this number biologically relevant and high enough to permit an accurate concordance evaluation. In our filtered data, the total number of features was close to 1000, and 100 corresponds to 10% of total taxa.

We used CAT for two different analyses:

- Between Method Concordance (BMC), in which a method was compared to other methods in the same dataset to evaluate consistency;
- Within Method Concordance (WMC), in which a method is compared to itself in random splits of the datasets to evaluate replicability.

To summarize this information for all pairwise method comparisons, we computed the Area Under the Curve, hence giving a better score to two methods that are consistently concordant for all values of *i* from 1 to 100.

We selected several datasets, with different alpha and beta diversity, for our concordance analysis. Table S3 describes the six datasets used. For each dataset, the same sample selection step, described next, was used.

The concordance evaluation algorithm can be easily summarized by the following steps:

1. Each dataset was randomly divided in half to obtain two subsets (Subset1 and Subset2) with two balanced groups;
2. DA analysis between the groups was performed with all evaluated methods independently on each subset;

3. For each method, the list of features ordered by p-values obtained from Subset1 was compared to the analogous list obtained from Subset2 and used to evaluate WMC;
4. For each method, the list of features ordered by p-values obtained from Subset1 was compared to the analogous list obtained from Subset1 by all the other methods and used to evaluate BMC for Subset1. The same was done in Subset2.
5. Steps 1-4 were repeated 10 times;
6. WMC and BMC were averaged across the 10 values (and between Subset1 and Subset2 for BMC) to obtain the final values.

Sample selection step. For each dataset, a subset was chosen in order to have a balanced number of samples for each condition. In lower diversity studies (e.g. Subgingival vs Supragingival Plaque) different biological samples from the same subject may be strongly correlated. Hence, we selected only one sample per individual, no matter the condition. To further increase the homogeneity of the datasets, we selected only samples from the same sequencing center.

Parametric simulations

Several real datasets were used as templates for the simulations:

- 41 Stool samples available for both 16S and WMS from HMP;
- 208 16S samples and 90 WMS samples of Tongue Dorsum body subsite from HMP.
- 67 Stool and 56 Oral cavity WMS data of Fijian adult women from BritoIl_2016.

Each dataset was filtered to obtain only a sample per individual. 16S and WMS samples were pruned to keep sequencing runs with library sizes of more than 10^3 and 10^6, respectively. Moreover, only features present in more than 1 sample with more than 10 reads were kept. After the data filtering step, the simulation framework was established, by specifying the parametric distribution and other data characteristics, described in Supplementary Table S4.

For each combination of parameters, we simulated 50 datasets, yielding a total of 28,800 simulations. Variables to be included in the simulation framework were chosen based on the role they may play in the analysis of a real experiment.

NB and ZINB are simple parametric distributions, easy to fit on real data through a reliable Bioconductor package and above all, seemed to fit 16S and WMS data better than other statistical models (see Figure 2). The *zinbSim* function from *zinbwave* Bioconductor package easily allows the user to generate both NB and ZINB counts after the *zinbFit* function estimates model parameters from real data. The user can set several options in *zinbFit*, we used *epsilon=1e10*, *common_dispersion=TRUE*, and *K=0*.

Generating two experimental groups requires the specification of enough samples for each condition and a more or less substantial biological difference between them.

Sample size is a crucial parameter: many pilot studies start with 10 or even fewer samples per condition, while clinical trials and case-control studies may need more samples in order to achieve the needed power. We included 10, 20 and 40 samples per condition in our simulation framework.

We considered two different scenarios for the number of features simulated as DA: 10%, representing a case where the majority of the features are not DA, a common assumption made by analysis methods; and 50%, a more extreme comparison. Similarly, we simulated a fold change difference for the DA features of 2 or 5. This is obviously a simplification, since in reality a continuum gradient of fold effects is present. Nevertheless, it allowed us to characterize the role of the effect size in the performance of the methods. For the DA features, the fold change between conditions was applied to the mean parameter of the ZINB or NB distributions, with or without “compensation” as introduced by Hawinkel et al[10]. Without compensation, the absolute abundance of a small group of features responds to a physiological change. This simple procedure modifies the mean relative abundances of all features, a microbiologist would only want to detect the small group that initially reacted to the physiological change. For this reason, significant results for other features will be considered as false discoveries. Compensation prevents the changes in DA features to influence the other, non-DA, features. The procedure comprises the following steps:

1. The relative mean for each feature is computed using estimated mean parameter of NB;
2. 10% or 50% of features are randomly sampled;
3. If there is no compensation, half of their relative means are multiplied by *foldEffect* while the remainings are divided by *foldEffect* generating up and down regulated features respectively.

   If there is compensation, 1/(1+*foldEffect*) of the selected feature relative means are multiplied by *foldEffect* while the remaining ones are multiplied by *(a/b)*(1-foldEffect)+1*, where *a* is the sum of the relative means of the features that will be up-regulated while *b* is the sum of the features that will be down-regulated.

4. The resulting relative means are normalized to sum to 1.

Sparsity is a key characteristic of metagenomic data. The case in which a bacterial species presence rate varies between conditions was emulated in the simulation framework via the so called *sparsityEffect* variable. Acting on the mixture parameter of the ZINB model it is possible to exacerbate down-regulation and up-regulation of a feature, adding zeros for the former and reducing zeroes for the latter. This scenario provided by 0 (no sparsity change at all), 0.05 and 0.15 of sparsity change should help methods to identify more differentially abundant features. As the mixing parameter can only take values between 0 and 1, when the additive sparsity effect yielded a value outside this range, it was forced to the closer limit.

The previously described DA methods were tested in each of the simulated datasets (50 for each set of simulation framework parameters) and the adjusted p-values were used to compute the False Positive Rate (FPR = 1 - Specificity) and the True Positive Rate (TPR = Sensitivity). Partial area under the Receiver Operating Characteristic (pAUROC) curve with an FPR from 0 to 0.1 values were computed and then averaged in order to obtain a single value for each set of variables.

Computational complexity

On the Stool 16S and WMS dataset, one simulated dataset for each set of the simulation framework variables (192 simulated datasets out of 9200) is used in order to measure each method’s computational complexity. Time evaluation is performed on a single core for each dataset where all methods are tested sequentially and then properly averaged. The methods’ performance evaluations on the 28800 total parametric simulations are performed in the same way, equally dividing the simulated datasets across 30 cores. The working machine is a Linux x86_64 architecture server with: 2 Intel® Xeon® Gold 6140 CPU with 2.30 GHz for a total of 72 CPUs and 128 GB of RAM.

Data availability

The real datasets used in this article are available in the *HMP16SData* Bioconductor package, available at http://bioconductor.org/packages/HMP16SData, and in the *CuratedMetagenomicData* Bioconductor packages, available at http://bioconductor.org/packages/CuratedMetagenomicData.

Code availability

The scripts to reproduce all analyses and figures of this article are available at https://github.com/mcalgaro93/sc2meta

Acknowledgements

MC was supported by “Fondo Unico della Ricerca” (FUR) Biotechnology Department, University of Verona, 2018-2019. CR was supported by the Italian Association of Cancer (AIRC n. 21837). DR was supported by “Programma per Giovani Ricercatori Rita Levi Montalcini” granted by the Italian Ministry of Education, University, and Research. We thank the "Centro Piattaforme Tecnologiche" of the University of Verona for the computing resources.

Author contributions

DR, CR, NV conceived the project, LW co-developed the evaluation strategies, MC and DR drafted the manuscript, LW CR NV reviewed and edited the manuscript, MC performed the data analyses and curated the code repository. All Authors read and approved the final manuscript.

References

1. Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Env. Microbiol* **73**, 5261–5267 (2007).
2. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat Methods* **12**, 902–903 (2015).
3. Zhu, S., Qing, T., Zheng, Y., Jin, L. & Shi, L. Advances in single-cell RNA sequencing and its applications in cancer research. *Oncotarget* **8**, 53763–53779 (2017).
4. Wagner, A., Regev, A. & Yosef, N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol* **34**, 1145–1160 (2016).
5. Papalexi, E. & Satija, R. Single-cell RNA sequencing to explore immune cell heterogeneity. *Nat Rev Immunol* **18**, 35–45 (2018).

6. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat Rev Genet* **16**, 133–145 (2015).
7. Finak, G. *et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data*. vol. 16 (2015).
8. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat Methods* **11**, 740–742 (2014).
9. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. *Microbiome Datasets Are Compositional: And This Is Not Optional*. vol. 8 (2017).
10. Hawinkel, S., Mattiello, F., Bijnens, L. & Thas, O. *A broken promise: microbiome differential abundance methods do not control the false discovery rate*. vol. 20 (2019).
11. Weiss, S. *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**, 27 (2017).
12. Thorsen, J. *et al.* Large-scale benchmarking reveals false discoveries and count transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in microbiome studies. *Microbiome* **4**, 62 (2016).
13. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**, 1200–1202 (2013).
14. Fernandes, A. D. *et al.* Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S rRNA gene sequencing and selective growth experiments by compositional data analysis. *Microbiome* **2**, 15 (2014).
15. Russel, J. *et al. DAtest: a framework for choosing differential abundance or expression method*.
16. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

17. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).

18. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**, R29 (2014).

19. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).

20. Risso, D., Perraudeau, F., Gribkova, S., Dudoit, S. & Vert, J.-P. A general and flexible method for signal extraction from single-cell RNA-seq data. *Nat Commun* **9**, 284 (2018).

21. Van den Berge, K. *et al.* Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. *Genome Biol* **19**, 24 (2018).

22. Schiffer, L. *et al.* HMP16SData: Efficient Access to the Human Microbiome Project Through Bioconductor. *Am J Epidemiol* **188**, 1023–1026 (2019).

23. Pasolli, E. *et al.* Accessible, curated metagenomic data through ExperimentHub. *Nat Methods* **14**, 1023–1024 (2017).

24. Quinn, T. P., Erb, I., Richardson, M. F. & Crowley, T. M. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics* **34**, 2870–2878 (2018).

25. Murdoch, D. J., Tsai, Y.-L. & Adcock, J. P -Values are Random Variables. *Am Stat* **62**, 242–245 (2008).

26. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**, 766 (2014).

27. Castro-Nallar, E. *et al.* Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls. *PeerJ* **3**, e1140 (2015).

28. Consortium, T. H. M. P. & The Human Microbiome Project Consortium. *Structure, function and diversity of the healthy human microbiome*. vol. 486 (2012).

29. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat Methods* **2**, 345–350 (2005).
30. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).
31. Townes, F. W., William Townes, F., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature Selection and Dimension Reduction for Single Cell RNA-Seq based on a Multinomial Model. *biorXiv* (2019).
32. Svensson, V. Droplet scRNA-seq is not zero-inflated. *biorXiv* (2019).
33. McMurdie, P. J. & Holmes, S. phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One* **8**, e61217 (2013).
34. Quinn, T. P., Crowley, T. M. & Richardson, M. F. Benchmarking differential expression analysis tools for RNA-Seq: normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics* **19**, 274 (2018).
35. Calle, M. L. Statistical Analysis of Metagenomics Data. *Genomics Inf.* **17**, e6 (2019).
36. Aitchison, J. *The Statistical Analysis of Compositional Data*. vol. 44 (1982).
37. Kumar, M. S. *et al.* Analysis and correction of compositional bias in sparse sequencing count data. *BMC Genomics* **19**, 799 (2018).
38. Mandal, S. *et al.* Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb. Ecol. Health Dis.* **26**, (2015).
39. Mao, J., Chen, Y. & Ma, L. *Bayesian Graphical Compositional Regression for Microbiome Data*. (2019).
40. Bogomolov, M., Peterson, C. B., Benjamini, Y. & Sabatti, C. Testing hypotheses on a tree: new error rates and controlling strategies. *arXiv* (2017).
41. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**, R25 (2010).