Microbiome differential abundance methods produce disturbingly different results across 38 datasets

Jacob T. Nearing[1,*], Gavin M. Douglas[1,*], Molly Hayes[2], Jocelyn MacDonald[3], Dhwani Desai[4], Nicole Allward[5], Casey M. A. Jones[6], Robyn Wright[6], Akhilesh Dhanani[4], André M. Comeau[4], Morgan G. I. Langille[4,6]

[1]Department of Microbiology and Immunology, Dalhousie University
[2]Department of Mathematics and Statistics, Dalhousie University
[3]Department of Computer Science, Dalhousie University
[4]Integrated Microbiome Resource, Dalhousie University
[5]Department of Civil and Resource Engineering, Dalhousie University
[6]Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

Email for correspondence: jacob.nearing@dal.ca

[*]These authors contributed equally.

Abstract

Identifying differentially abundant microbes is a common goal of microbiome studies. Multiple methods have been applied for this purpose, which are largely used interchangeably in the literature. Although it has been observed that these tools can produce different results, there have been very few large-scale comparisons to describe the scale and significance of these differences. In addition, it is challenging for microbiome researchers to know which differential abundance tools are appropriate for their study and how these tools compare to one another. Here, we have investigated these questions by analyzing 38 16S rRNA gene datasets with two sample groups for differential abundance testing. We tested for differences in amplicon sequence variants and operational taxonomic units (referred to as ASVs for simplicity) between these groups with 14 commonly used differential abundance tools. Our findings confirmed that these tools identified drastically different numbers and sets of significant ASVs, however, for many tools the number of features identified correlated with aspects of the tested study data, such as sample size, sequencing depth, and effect size of community differences. We also found that the ASVs identified by each method were dependent on whether the abundance tables were prevalence-filtered before testing. ALDEx2 and ANCOM produced the most consistent results across studies and agreed best with the intersect of results from different approaches. In contrast, several methods, such as LEfSe, limma voom, and edgeR, produced inconsistent results and in some cases were unable to control the false discovery rate. In addition to these observations, we were unable to find supporting evidence for a recent recommendation that limma voom, corncob, and DESeq2 are more reliable overall compared with other methods. Although ALDEx2 and

ANCOM are two promising conservative methods, we argue that those researchers requiring more sensitive methods should use a consensus approach based on multiple differential abundance methods to help ensure robust biological interpretations.

Introduction

Microbial communities are frequently characterized with DNA sequencing. Marker gene sequencing, such as 16S rRNA gene sequencing, is the most common form of microbiome profiling and enables the relative abundances of taxa to be compared across different samples. A frequent and seemingly basic question to investigate with this type of data is: which taxa significantly differ in abundance between sample groupings? Newcomers to the microbiome field may be surprised to learn that there is little consensus on how best to approach this question. Indeed, there are numerous ongoing debates regarding the best practices for differential abundance (DA) testing with microbiome data (Allaband et al., 2019; Pollock et al., 2018).

One area of disagreement is whether read count tables should be rarefied (i.e., subsampled) to correct for differing read depths across samples (Weiss et al., 2017). This approach has been heavily criticized because excluding data could reduce statistical power and introduce biases. In particular, using rarefied count tables for standard tests, such as the t-test and Wilcoxon test, can result in unacceptably high false positive rates (McMurdie and Holmes, 2014). Nonetheless, microbiome data is still frequently rarefied because it can simplify analyses, particularly for methods that do not control for variation in read depth across samples. For example, LEfSe (Segata et al., 2011) is a popular method for identifying differentially abundant taxa that first converts read counts to percentages. Accordingly, read count tables are often rarefied before being input into this tool so that variation in sample read depth does not bias analyses. Without addressing the variation in depth across samples by some approach, the richness can drastically differ between samples due to read depth alone.

A related question to whether data should be rarefied is whether rare taxa should be filtered out. This question arises in many high-throughput datasets, where the burden of correcting for many tests can greatly reduce statistical power. Filtering out potentially uninformative features before running statistical tests can help address this problem, although this can also have unexpected effects (Bourgon et al., 2010). Importantly, this filtering must be independent of the test statistic evaluated (referred to as Independent Filtering). For instance, hard cut-offs for the prevalence and abundance of taxa across samples, and not within one group compared with another, are commonly used to exclude rare taxa (Schloss, 2020). This data filtering could be especially important for microbiome datasets because they are often extremely sparse. Nonetheless, it remains unclear whether filtering rare taxa has much effect on DA results in practice.

Another contentious area is regarding which statistical distributions are most appropriate for analyzing microbiome data. Statistical frameworks based on a range of distributions have been developed for modelling read count data. For example, DESeq2 (Love et al., 2014) and edgeR (Robinson and Oshlack, 2010) are both tools that assume normalized read counts follow a negative binomial distribution. To identify differentially abundant taxa, a null and alternative hypothesis are compared for each taxon. The null hypothesis is that the same parameters for the negative binomial solution explain the distribution of taxa across all sample groupings. The alternative hypothesis is that different parameters are needed to account for differences between sample groupings. If the null hypothesis can be rejected for a specific taxon then it is considered differentially abundant. This idea is the foundation of distribution-based DA tests, including other methods such as corncob (Martin et al., 2020) and metagenomeSeq (Paulson et al., 2013), which model microbiome data with the beta-binomial and zero-inflated Gaussian distributions, respectively.

Compositional data analysis (CoDa) methods represent an alternative to these approaches. It has recently become more widely appreciated that sequencing data are compositional (Gloor et al., 2017) meaning that sequencing only provides information on the relative abundance of features and that each feature is dependent on the relative abundance of all other features. This characteristic means that false inferences are commonly made when standard methods, intended for absolute abundances, are used with taxa relative abundances. CoDa methods circumvent this issue by reframing the focus to ratios of taxa relative abundances (Aitchison, 1982; Morton et al., 2019). The difference between CoDa methods considered in this paper is what quantity is used as the denominator, or the reference, for the transformation. The centred log-ratio (CLR) transformation is a CoDa approach that uses the geometric mean of the relative abundance of all taxa within a sample as the reference for that sample. An extension of this approach is implemented in the tool ALDEx2 (Fernandes et al., 2014) . The additive log-ratio transformation is an alternative approach where the reference is the relative abundance of a single taxon, which should be present with low variance in read counts across samples. ANCOM is one tool that implements this additive log-ratio approach (Mandal et al., 2015).

Evaluating the numerous options for analyzing microbiome data outlined above has proven difficult. This is largely because there are no gold standards to compare with DA tool results. Simulating datasets with specific taxa that are differentially abundant is a partial solution to this problem, but it is imperfect. For example, it has been noted that parametric simulations can result in circular arguments for specific tools (Hawinkel et al., 2019). It is unsurprising that distribution-based methods perform best when applied to simulated data based on that distribution. Nonetheless, simulated data with no expected differences has been valuable for evaluating the false discovery rate (FDR) of these methods. Based on this approach it has become clear that many of the methods output unacceptably high numbers of false positives (Calgaro et al., 2020; Feijen et al., 2016; Thorsen et al., 2016; Weiss et al., 2017). Similarly, based on simulated datasets with spiked taxa it has been shown that these methods can drastically vary in statistical power as well (Hawinkel et al., 2019; Thorsen et al., 2016).

Although these general observations have been well substantiated, there is less agreement regarding the performance of tools across evaluation studies. Certain observations have been reproducible, such as the higher FDR of edgeR and metagenomeSeq. Similarly, ALDEx2 has been repeatedly shown to have low power to detect differences (Calgaro et al., 2020; Hawinkel et al., 2019). In contrast, both ANCOM and limma voom (Law et al., 2014; Ritchie et al., 2015) have been implicated as both accurately and poorly controlling the FDR, depending on the study (Calgaro et al., 2020; Hawinkel et al., 2019; Weiss et al., 2017). To further complicate comparisons, different sets of tools and dataset types have been analyzed across evaluation studies. This means that, on some occasions, the best performing method in one evaluation is missing from another. In addition, certain popular microbiome-specific methods, such as MaAsLin2 (Mallick et al., 2021), have been missing from past evaluations. Finally, many evaluations limit their analysis to a small number of datasets that do not represent the breadth of datasets found in 16S rRNA gene sequencing studies.

Given the inconsistencies across these studies it is important that additional, independent evaluations be performed to elucidate the performance of current DA methods. Accordingly, herein we have conducted additional evaluations of common DA tools across 38 two-group 16S rRNA gene datasets. We first present the concordance of the methods on these datasets to investigate how consistently the methods cluster and perform in general, with and without the removal of rare taxa. Next, based on artificially subsampling the datasets into two groups where no differences are expected, we present the observed FDR for each DA tool. Lastly, we present an evaluation of how consistent biological interpretations would be across diarrheal datasets depending on which tool was applied. Our work enables improved assessment of these DA tools and highlights which key recommendations made by previous studies hold in an independent evaluation. Furthermore, our analysis shows various characteristics of DA tools that authors can use to evaluate published literature within the field.

Methods

Code and data availability

All code used for processing and analyzing the data is available in this GitHub repository: https://github.com/nearinj/Comparison_of_DA_microbiome_methods. The processed datasets and metadata files analyzed in this study are available on figshare: https://figshare.com/articles/dataset/16S_rRNA_Microbiome_Datasets/14531724. The accessions and/or locations of the raw data for each tested dataset are listed in **Supplementary Table 1**.

Dataset processing

Thirty-eight different datasets were included in our main analyses for assessing the characteristics of microbiome differential abundance tools. Two additional datasets were also included for a comparison of differential abundance consistency across diarrhea-related microbiome datasets. All datasets presented herein have been previously published or are publicly available (Alkanani et al., 2015; Baxter et al., 2016; Chase et al., 2016; De Tender et al., 2015; Dinh et al., 2015; Douglas et al., 2018; Dranse et al., 2018; Duvallet et al., 2017; Frère et al., 2018; Gonzalez et al., 2018; Goodrich et al., 2014; Hoellein et al., 2017; Ji et al., 2015; Kesy et al., 2019; Lamoureux et al., 2017; Lozupone et al., 2013; McCormick et al., 2016; Mejía-León et al., 2014; Nearing et al., 2019; Noguera-Julian et al., 2016; Oberbeckmann et al., 2016; Oliveira et al., 2018; Papa et al., 2012; Pop et al., 2014; Rosato et al., 2020; Ross et al., 2015; Scheperjans et al., 2015; Scher et al., 2013; Schneider et al., 2017; Schubert et al., 2014; Singh et al., 2015; Son et al., 2015; Turnbaugh et al., 2009; Vincent et al., 2013; Wu et al., 2019; Yurgel et al., 2017; Zeller et al., 2014; Zhu et al., 2013) (**Supp. Table 1**). Most datasets were already available in table format with ASV or operational taxonomic unit abundances while a minority needed to be processed from raw sequences. These raw sequences were processed with QIIME 2 version 2019.7 (Bolyen et al., 2018) based on the Microbiome Helper standard operating procedure (Comeau et al., 2017). Primers were removed using cutadapt (Martin, 2011) and stitched together using the QIIME 2 VSEARCH (Rognes et al., 2016) *join-pairs* plugin. Stitched reads were then quality filtered using the *quality-filter* plugin and reads were denoised using Deblur (Amir et al., 2017) to produce amplicon sequence variants (ASVs). Abundance tables of ASVs for each sample were then output into tab-delimited files. Rarefied tables were also produced for each dataset, where the rarefied read depth was taken to be the lowest read depth of any sample in the dataset over 2000 reads (with samples below this threshold discarded).

Differential abundance testing

We created a custom shell script (run_all_tools.sh) that ran each differential abundance tool on each dataset within this study. As input the script took a tab-delimited ASV abundance table, a rarefied version of that same table, and a metadata file that contained a column that split the samples into two groups for testing. This script also accepted a prevalence cut-off filter to remove ASVs below a minimum cut-off, which was set to 10% (i.e., ASVs found in fewer than 10% of samples were removed) for the filtered data analyses we present. Note that in a minority of cases a genus abundance table was input instead, in which case all options were kept the same. When the prevalence filter option was set, the script also generated new filtered rarefied tables based on an input rarefaction depth.

Following these steps, each individual differential abundance method was run on the input data using either the rarefied or non-rarefied table, depending on which is recommended for that tool. The workflow used to run each differential abundance tool (with run_all_tools.sh) is described below. The first step in each of these workflows was to read the dataset tables into R (version 3.6.3) with a custom script and then ensure that samples within the metadata and feature abundance tables were in the same order. An alpha-value of 0.05 was chosen as our significance cutoff and FDR-corrected p-values were used for methods that output p-values.

ALDEx2

We passed the non-rarefied feature table and the corresponding sample metadata to the *aldex* function from the ALDEx2 R package (Fernandes et al., 2014) which generated Monte Carlo samples of Dirichlet distribution for each sample, using a uniform prior, performed CLR transformation of each realization, and performed Wilcoxon tests on the transformed realizations. The function then returned the expected Benjamini-Hochberg (BH) FDR-corrected P-value was then returned for each feature based on the results across Monte Carlo samples.

ANCOM-II

We ran the non-rarefied feature table through the R ANCOM-II (Kaul et al., 2017; Mandal et al., 2015) (https://github.com/FrederickHuangLin/ANCOM) function *feature_table_pre_process*, which first examined the abundance table to identify outlier zeros and structural zeros (Kaul et al., 2017). Outlier zeros, identified by finding outliers in the distribution of taxon counts within each sample grouping, were ignored during differential abundance analysis and replaced with NA. Structural zeros, taxa that were absent in one grouping but present in the other, were ignored during data analysis and automatically called as differentially abundant. A pseudo count of 1 was then applied across the dataset to allow for log transformation. Using the main function *ANCOM,* all additive log-ratios for each taxon were then tested for significance using Wilcoxon rank-sum tests, and p-values were FDR-corrected using the BH method. ANCOM then applied a detection threshold as described in the original paper (Mandal et al., 2015), whereby a taxon was called as DA if the number of corrected p-values reaching nominal significance for that taxon was greater than 90% of the maximum possible number of significant comparisons.

corncob

We converted the metadata and non-rarefied feature tables into a phyloseq object, which we input to corncob's *differentialTest* function (Martin et al., 2020). This function first converted the data into relative abundances and then fit each taxon abundance to a beta-binomial model, using logit link functions for both the mean and overdispersion. Because corncob models each of these simultaneously and performs both differential abundance and differential variability testing (Martin et al., 2020), we set the null overdispersion model to be the same as the non-null model so that only taxa having differential abundances were identified. Finally, the function performed significance testing, for which we chose Wald tests (with the default non-bootstrap setting), and we obtained BH FDR-corrected p-values as output.

DESeq2

We first passed the non-rarefied feature tables to the *DESeq* function (Love et al., 2014) with default settings, except that instead of the default relative log expression (also known as the median-of-ratios method) the estimation of size factors was set to use "poscounts", which calculates a modified relative log expression that helps account for features missing in at least one sample. The function performed three steps: (1) estimation of size factors, which are used to normalize library sizes in a model-based fashion; (2) estimation of dispersions from the negative binomial likelihood for each feature, and subsequent shrinkage of each dispersion estimate towards the parametric (default) trendline by empirical Bayes; (3) fitting each feature to the specified class groupings with negative binomial generalized linear models and performing hypothesis testing, for which we chose the default Wald test. Finally, using the *results* function, we obtained the resulting BH FDR-corrected p-values.

edgeR

Using the phyloseq_to_edgeR function (https://joey711.github.io/phyloseq-extensions/edgeR.html), we added a pseudocount of 1 to the non-rarefied feature table and used the function *calcNormFactors* from the edgeR R package (Robinson and Oshlack, 2010) to compute relative log expression normalization factors. Negative binomial dispersion parameters were then estimated using the functions *estimateCommonDisp* followed by *estimateTagwiseDisp* to shrink feature-wise dispersion estimates through an empirical Bayes approach. We then used the *exactTest* for negative binomial data (Robinson and Oshlack, 2010) to identify features that differ between the specified groups. The resulting p-values were then corrected for multiple testing with the BH method with the function *topTags*.

LEfSe

The rarefied feature table was first converted into LEfSe format using the LEfSe script *format_input.py* (Segata et al., 2011). We then ran LEfSe on the formatted table using the *run_lefse.py* script with default settings and no subclass specifications. Briefly, this command first normalized the data using total sum scaling, which divides each feature count by the total library size. Then it performed a Kruskal-Wallis (which in our two-group case reduces to the Wilcoxon rank-sum) hypothesis test to identify potential differentially abundant features, followed by linear discriminant analysis of class labels on abundances to estimate the effect sizes for significant features. From these, only those features with scaled logarithmic linear discriminant analysis scores above the threshold score of 2.0 (default) were called as differentially abundant.

limma voom

We first normalized the non-rarefied feature table using the edgeR *calcNormFactors* function, with either the trimmed mean of M-values (TMM) or TMM with singleton pairing (TMMwsp) option. During this step (for both options), a single sample was chosen to be a reference sample using upper-quartile normalization. This step failed in some highly sparse abundance tables, in which cases, we instead chose the sample with the largest sum of square-root transformed feature abundances to be the reference sample. After normalization, we used the limma R package (Ritchie et al., 2015) function *voom* to convert normalized counts to $\log_2$-counts-per-million and assign precision weights to each observation based on the mean-variance trend. We then used the functions *lmFit, eBayes*, and *topTable* to fit weighted linear regression models, perform tests based on an empirical Bayes moderated t-statistic (Phipson et al., 2016) and obtain BH FDR-corrected p-values.

MaAsLin2

We entered either a rarefied or non-rarefied feature table into the main *Maaslin2* function within the Maaslin2 R package(Mallick et al., 2021). We specified arcsine square-root transformation as in the package vignette (instead of the default log) and total sum scaling normalization. For consistency with other tools, we specified no random effects and turned off default standardization. The function fit a linear model to each feature's transformed abundance on the specified sample grouping, tested significance using a Wald test, and output BH FDR-corrected p-values.

metagenomeSeq

We first entered the counts and sample information to the function *newMRexperiment* from the metagenomeSeq R package (Paulson et al., 2013). Next, we used *cumNormStat* and *cumNorm* to apply cumulative sum-scaling normalization, which attempts to normalize sequence counts based on the lower-quartile abundance of features. We then used *fitFeatureModel* to fit normalized feature counts with zero-inflated log-normal models (with pseudo-counts of 1 added prior to $\log_2$ transformation) and perform empirical Bayes moderated t-tests, and *MRfulltable* to obtain BH FDR-corrected p-values.

t-test

We applied total sum scaling normalization to the rarefied feature table, and performed an unpaired Welch's t-test for each feature to compare the specified groups. We corrected the resulting p-values for multiple testing with the BH method.

Wilcoxon test

Using raw feature abundances in the rarefied case, and CLR-transformed abundances (after applying a pseudocount of 1) in the non-rarefied case, we performed Wilcoxon rank-sum tests for each feature to compare the specified sample groupings. We corrected the resulting p-values with the BH method.

Comparing numbers of significant hits between tools

We compared the number of significant ASVs each tool identified in 38 different datasets. Each tool was run as described above using default settings with some modifications suggested by the tool authors, as noted above. A heatmap representing the number of significant hits found by each tool was constructed using the pheatmap R package (Kolde, 2012). Spearman correlations between the number of significant ASVs identified by a tool and the following dataset characteristics were computed using the cor.test function in R: sample size, Aitchison's distance effect size as computed using a PERMANOVA test (adonis; vegan) (Dixon, 2003), sparsity, mean sample ASV richness, median sample read depth, read depth range between samples and the coefficient of variation for read depth within a dataset.

Cross-tool, within-study differential abundance consistency analysis

We compared the consistency between different tools within all datasets by pooling all ASVs identified as being significant by at least one tool in the 38 different datasets. The number of methods that identified each ASV as differentially abundant were then tallied.

False positive analysis

To estimate the false positives a method might produce during data analysis, eight datasets were selected for analysis. These datasets were chosen based on having the largest sample sizes, while also being from diverse environment types. In each dataset, the most frequent sample group was chosen for analysis to help ensure similar composition among samples tested. Within this grouping, random labels of either case or control were assigned to samples and the various differential abundance methods were tested on them. This was repeated 10 times for each dataset and each tool tested with an additional 90 replications for all tools except for ALDEx2, ANCOM-II and corncob due to high resource requirements. After analysis was completed, the number of differentially abundant ASVs identified by each tool was assessed at an alpha value of 0.05.

Cross-study differential abundance consistency analysis

Two additional datasets were acquired to bring the number of diarrhea-related datasets to five. The ASVs in each of these datasets were previously taxonomically classified and so we used these classifications to collapse all feature abundances to the genus level. Note that taxonomic classification was performed using several different methods, which represents another source of technical variation. We excluded unclassified and *sensu stricto*-labelled genus levels. We then ran all differential abundance tools on these datasets at the genus level. These comparisons were between the diarrhea and non-diarrhea sample groups. The same processing workflow was used for the supplementary obesity dataset comparison as well.

For each tool and study combination, we determined which genera were significantly different at an alpha of 0.05 (where relevant). For each tool we then tallied up the number of times each genus was significant, i.e., how many datasets each genus was significant in based on a given tool. The null expectation distributions of these counts per tool were generated by randomly sampling genera from each dataset for 100,000 replicates. The probability of sampling a genus (i.e., calling it significant) was set to be equal to the proportion of actual significant genera. For each replicate we tallied up the number of times each genus was sampled across datasets. We then compared the observed and expected distributions of the number of studies each genus was found to be significant in. Note that to simplify this analysis we ignored the directionality of the significance (e.g., whether it was higher in case or control samples). We excluded genera never found to be significant. We performed bootstrap Kolmogorov-Smirnov tests (10,000 replicates) using the *ks.boot* function from the Matching R package (Sekhon, 2011) to compare the expected and observed distributions for each tool.

Results

Microbiome differential abundance methods produce a highly variable number of significant ASVs within the same microbiome datasets

To investigate how different DA tools impact biological interpretations across microbiome datasets, we tested 14 different differential abundance testing approaches (**Table 1**) on 38 different microbiome datasets. These datasets corresponded to a range of environments, including the human gut, plastisphere, freshwater lakes, and urban environments (**Supp. Table 1**). The features in these datasets corresponded to both ASVs and clustered operational taxonomic units, but we refer to them all as ASVs below for simplicity.

We also investigated how prevalence filtering each dataset prior to analysis impacted the observed results. We chose to either use no prevalence filtering (**Fig. 1A**) or a 10% prevalence filter that removed any ASVs found in fewer than 10% of samples within each dataset (**Fig. 1B**).

We found that in both the filtered and unfiltered analyses the percentage of significant ASVs identified by each DA method varied widely across datasets, with means ranging from 3.8-32.5% and 0.8-40.5%, respectively. Interestingly, we found that many tools behaved differently between datasets. Specifically, some tools identified the most features in one dataset while identifying only an intermediate number in other datasets. This was especially evident in the unfiltered datasets (**Fig. 1A**).

To investigate possible factors driving this variation we examined how the number of ASVs identified by each tool correlated with several variables. These variables included dataset richness, variation in sequencing depth between samples, dataset sparsity, and Aitchison's distance effect size (based on PERMANOVA tests). As expected, we found that all tools positively correlated with the effect size between test groups with rho values ranging between 0.35-0.72 with unfiltered data (**Fig. 2A**) and 0.31-0.52 for filtered data (**Fig. 2B**). We also found in the filtered datasets that the number of features found by all tools significantly correlated with the median read depth, range in read depth, and sample size. There was much less consistency in these correlations across the unfiltered data. For instance, only the t-test, both Wilcoxon methods, and both limma voom methods correlated significantly with the range in read depth **(Fig. 2B**). We also found that edgeR was negatively correlated with mean sample richness in the unfiltered analysis.

Table 1: Differential abundance tools compared in this study

Tool	Input	Norm.	Trans.	Distribution	Covariates	Random Effects	Hypothesis test	FDR Corr.	CoDa	Dev. For
ALDEx2	Counts	None	CLR	Dirichlet-multinomial	Yes*	No	Wilcoxon rank-sum	Yes	Yes	RNA-seq, 16S, MGS
ANCOM-II	Counts	None	ALR	Non-parametric	Yes	Yes	Wilcoxon rank-sum	Yes	Yes	MGS
corncob	Counts	None	None	Beta-binomial	Yes	No	Wald (default)	Yes	No	16S, MGS
DESeq2	Counts	Modified RLE (default is RLE)	None	Negative binomial	Yes	No	Wald (default)	Yes	No	RNA-seq, 16S, MGS
edgeR	Counts	RLE (default is TMM)	None	Negative binomial	Yes*	No	Exact	Yes	No	RNA-seq
LEFse	Rarefied relative abundance	TSS	None	Non-parametric	Subclass factor only	No	Kruskal-Wallis	No	No	16S, MGS
MaAsLin2	Counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
MaAsLin2 (rare)	Rarefied counts	TSS	AST (default is log)	Normal (default)	Yes	Yes	Wald	Yes	No	MGS
metagenomeSeq	Counts	CSS	Log	Zero-inflated log-Normal	Yes	No	Moderated t	Yes	No	16S. MGS
limma voom (TMM)	Counts	TMM	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
limma voom (TMMwsp)	Counts	TMMwsp	Log; Precision weighting	Normal (default)	Yes	Yes	Moderated t	Yes	No	RNA-seq
t-test (rare)	Rarefied Counts	None	None	Normal	No	No	Welch's t-test	Yes	No	N/A
Wilcoxon (CLR)	CLR abundances	None	CLR	Non-parametric	No	No	Wilcoxon rank-sum	Yes	Yes	N/A
Wilcoxon (rare)	Rarefied counts	None	None	Non-parametric	No	No	Wilcoxon rank-sum	Yes	No	N/A

** The tool supports additional covariates if they are provided. ANCOM-II automatically performs ANOVA in this case, ALDEx2 requires that users select the test, and edgeR requires use of a different function (glmFit or glmQLFit instead of exactTest).*

Abbreviations: ALR, additive log-ratio; AST, arcsine square-root transformation; CLR, centered log-ratio; CoDa, compositional data analysis; CSS, cumulative sum scaling; FDR Corr., false-discovery rate correction; MGS, metagenomic sequencing; RLE, relative log expression; TMM, trimmed mean of M-values; Trans., transformation; TSS, total sum scaling

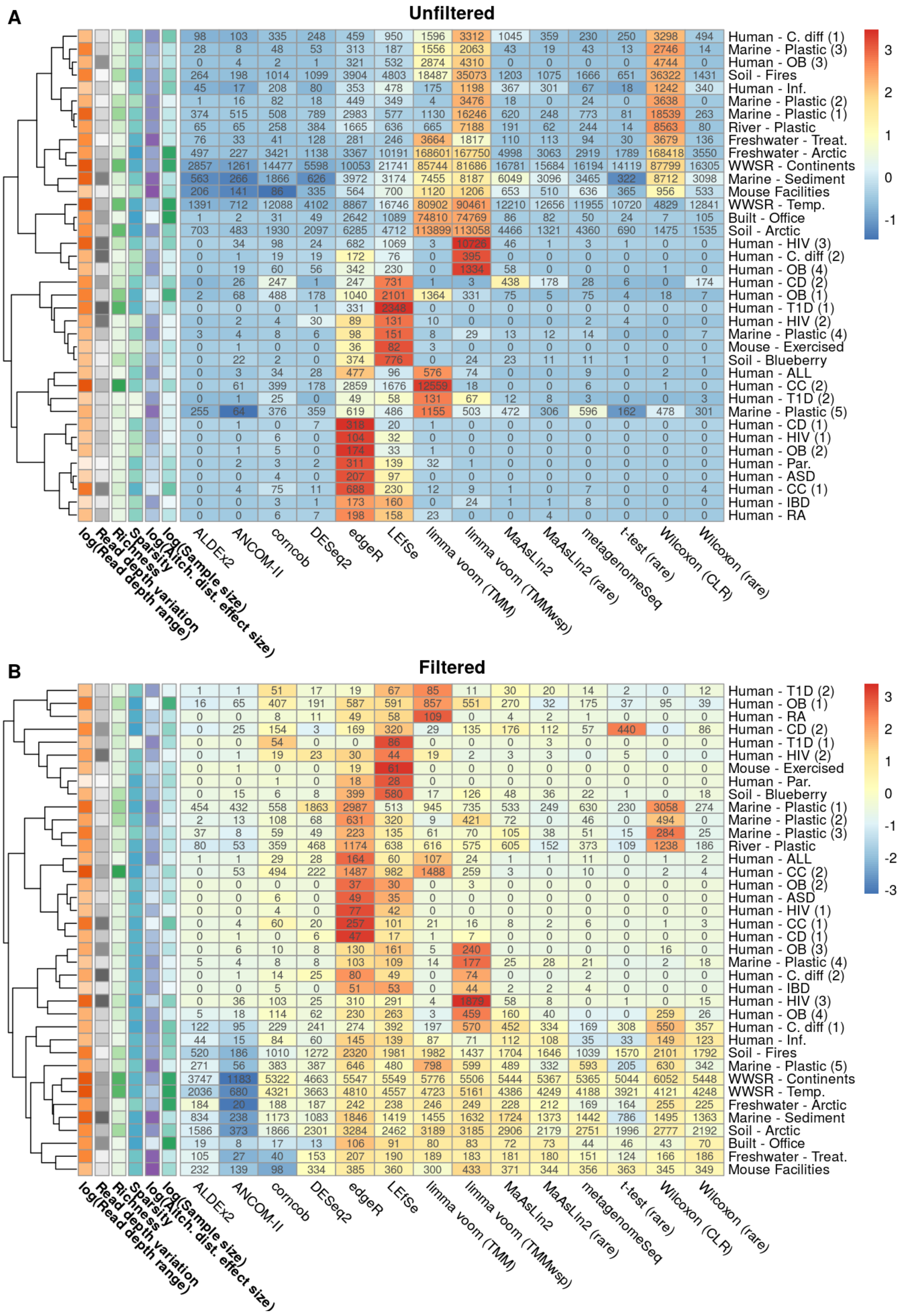

A
Unfiltered
98 103 335 248 459 950 1596 3312 1045 359 230 250 3298 494 Human - C. diff (1)
28 8 48 53 313 187 1556 2063 43 19 43 13 2746 14 Marine - Plastic (3)
0 4 2 1 321 532 2874 4310 0 0 0 0 4744 0 Human - OB (3)
264 198 1014 1099 3904 4803 18487 35073 1203 1075 1666 651 36322 1431 Soil - Fires
45 17 208 80 353 478 175 1198 367 301 67 18 1242 340 Human - Inf.
1 16 82 18 449 349 4 3476 18 0 24 0 3638 0 Marine - Plastic (2)
374 515 508 789 2983 577 1130 16246 620 248 773 81 18539 263 Marine - Plastic (1)
65 65 258 384 1665 636 665 7188 191 62 244 14 8563 80 River - Plastic
76 33 41 128 281 246 3664 1817 110 113 94 30 3679 136 Freshwater - Treat.
497 227 3421 1138 3367 10191 168601 167750 4998 3063 2919 1789 168418 3550 Freshwater - Arctic
2857 1261 14477 5598 10053 21741 85744 81686 16781 15684 16194 14119 87799 16305 WWSR - Continents
563 266 1866 626 3972 3174 7455 8187 6049 3096 3465 322 8712 3098 Marine - Sediment
206 141 86 335 564 700 1120 1206 653 510 636 365 956 533 Mouse Facilities
1391 712 12088 4102 8867 16746 80902 90461 12210 12656 11955 10720 4829 12841 WWSR - Temp.
1 2 31 49 2642 1089 74810 74769 86 82 50 24 7 105 Built - Office
703 483 1930 2097 6285 4712 113899 113058 4466 1321 4360 690 1475 1535 Soil - Arctic
0 34 98 24 682 1069 3 10726 46 1 3 1 0 0 Human - HIV (3)
0 1 19 19 172 76 0 395 0 0 0 0 0 0 Human - C. diff (2)
0 19 60 56 342 230 0 1334 58 0 0 0 1 0 Human - OB (4)
0 26 247 1 247 731 1 3 438 178 28 6 0 174 Human - CD (2)
2 68 488 178 1040 2101 1364 331 75 5 75 4 18 7 Human - OB (1)
0 0 0 1 331 2348 0 0 0 0 0 0 0 0 Human - T1D (1)
0 2 4 30 89 131 10 0 0 0 2 4 0 0 Human - HIV (2)
3 4 8 6 98 151 8 29 13 12 14 0 0 7 Marine - Plastic (4)
0 1 0 0 36 82 3 0 0 0 0 0 0 0 Mouse - Exercised
0 22 2 0 374 776 0 24 23 11 11 1 0 1 Soil - Blueberry
0 3 34 28 477 96 576 74 0 0 9 0 2 0 Human - ALL
0 61 399 178 2859 1676 12559 18 0 0 6 0 1 0 Human - CC (2)
0 1 25 0 49 58 131 67 12 8 3 0 0 3 Human - T1D (2)
255 64 376 359 619 486 1155 503 472 306 596 162 478 301 Marine - Plastic (5)
0 1 0 7 318 20 1 0 0 0 0 0 0 0 Human - CD (1)
0 0 6 0 104 32 0 0 0 0 0 0 0 0 Human - HIV (1)
0 1 5 0 174 33 1 0 0 0 0 0 0 0 Human - OB (2)
0 2 3 2 311 139 32 1 0 0 0 0 0 0 Human - Par.
0 0 4 0 207 97 0 0 0 0 0 0 0 0 Human - ASD
0 4 75 11 688 230 12 9 1 0 7 0 0 4 Human - CC (1)
0 0 3 1 173 160 0 24 1 1 8 0 0 0 Human - IBD
0 0 6 7 198 158 23 0 0 4 0 0 0 0 Human - RA
log(Read depth range)
Read depth variation
Richness
Sparsity
log(Aitch. dist. effect size)
log(Sample size)
ALDEx2
ANCOM-II
corncob
DESeq2
edgeR
LEfSe
limma voom (TMM)
limma voom (TMMwsp)
MaAsLin2
MaAsLin2 (rare)
metagenomeSeq
t-test (rare)
Wilcoxon (CLR)
Wilcoxon (rare)
3
2
1
0
-1
B
Filtered
1 1 51 17 19 67 85 11 30 20 14 2 0 12 Human - T1D (2)
16 65 407 191 587 591 857 551 270 32 175 37 95 39 Human - OB (1)
0 0 8 11 49 58 109 0 4 2 1 0 0 0 Human - RA
0 25 154 3 169 320 29 135 176 112 57 440 0 86 Human - CD (2)
0 0 54 0 0 86 0 0 0 3 0 0 0 0 Human - T1D (1)
0 1 19 23 30 44 19 2 3 3 0 5 0 0 Human - HIV (2)
0 1 0 0 19 61 0 0 0 0 0 0 0 0 Mouse - Exercised
0 0 1 0 18 28 0 0 0 0 0 0 0 0 Human - Par.
0 15 6 8 399 580 17 126 48 36 22 1 0 18 Soil - Blueberry
454 432 558 1863 2987 513 945 735 533 249 630 230 3058 274 Marine - Plastic (1)
2 13 108 68 631 320 9 421 72 0 46 0 494 0 Marine - Plastic (2)
37 8 59 49 223 135 61 70 105 38 51 15 284 25 Marine - Plastic (3)
80 53 359 468 1174 638 616 575 605 152 373 109 1238 186 River - Plastic
1 1 29 28 164 60 107 24 1 1 11 0 1 2 Human - ALL
0 53 494 222 1487 982 1488 259 3 0 10 0 2 4 Human - CC (2)
0 0 0 0 37 30 0 3 0 0 0 0 0 0 Human - OB (2)
0 0 6 0 49 35 0 0 0 0 0 0 0 0 Human - ASD
0 0 4 0 77 42 0 0 0 0 0 0 0 0 Human - HIV (1)
0 4 60 20 257 101 21 16 8 2 6 0 1 3 Human - CC (1)
0 1 0 6 47 17 1 7 0 0 0 0 0 0 Human - CD (1)
0 6 10 8 130 161 5 240 0 0 0 0 16 0 Human - OB (3)
5 4 8 8 103 109 14 177 25 28 21 0 2 18 Marine - Plastic (4)
0 1 14 25 80 49 0 74 0 0 2 0 0 0 Human - C. diff (2)
0 0 5 0 51 53 0 44 2 2 4 0 0 0 Human - IBD
0 36 103 25 310 291 4 1879 58 8 0 1 0 15 Human - HIV (3)
5 18 114 62 230 263 3 459 160 40 0 0 259 26 Human - OB (4)
122 95 229 241 274 392 197 570 452 334 169 308 550 357 Human - C. diff (1)
44 15 84 60 145 139 87 71 112 108 35 33 149 123 Human - Inf.
520 186 1010 1272 2320 1981 1982 1437 1704 1646 1039 1570 2101 1792 Soil - Fires
271 56 383 387 646 480 798 599 489 332 593 205 630 342 Marine - Plastic (5)
3747 1183 5322 4663 5547 5549 5776 5506 5444 5367 5365 5044 6052 5448 WWSR - Continents
2036 680 4321 3663 4810 4557 4723 5161 4386 4249 4188 3921 4121 4248 WWSR - Temp.
184 20 188 187 242 238 246 249 228 212 169 164 255 225 Freshwater - Arctic
834 238 1173 1083 1846 1419 1455 1632 1724 1373 1442 786 1495 1363 Marine - Sediment
1586 373 1866 2301 3284 2462 3189 3185 2906 2179 2751 1996 2777 2192 Soil - Arctic
19 8 17 13 106 91 80 83 72 73 44 46 43 70 Built - Office
105 27 40 153 207 190 189 183 181 180 151 124 166 186 Freshwater - Treat.
232 139 98 334 385 360 300 433 371 344 356 363 345 349 Mouse Facilities
log(Read depth range)
Read depth variation
Richness
Sparsity
log(Aitch. dist. effect size)
log(Sample size)
ALDEx2
ANCOM-II
corncob
DESeq2
edgeR
LEfSe
limma voom (TMM)
limma voom (TMMwsp)
MaAsLin2
MaAsLin2 (rare)
metagenomeSeq
t-test (rare)
Wilcoxon (CLR)
Wilcoxon (rare)
3
2
1
0
-1
-2
-3

Figure 1: Variation in the proportion of significant features depending on the differential abundance method and dataset. Heatmaps indicate the numbers of significant amplicon sequence variants (ASVs) identified in each dataset by the corresponding tool based on (A) unfiltered data and (B) 10% prevalence-filtered data. Cells are coloured based on the standardized (scaled and mean centred) percentage of significant ASVs for each dataset. Additional coloured cells in the left-most six columns indicate the dataset characteristics we hypothesized could be driving variation in these results. Darker colours indicate higher values in these six columns. Datasets were hierarchically clustered based on Euclidean distances using the complete method.

Despite the variability of tool performance between datasets, we did find that several tools tended to identify more significant hits (**Supp. Fig 1C-D**). In the unfiltered datasets, we found that limma voom (TMMwsp; mean: 40.5% / TMM; mean: 29.7%), Wilcoxon (CLR; mean: 30.7%), LEfSe (mean: 12.6%), and edgeR (mean: 12.4%) tended to find the largest number of significant ASVs compared with other methods. Interestingly, in a few datasets, such as the Human-ASD and Human-OB (2) datasets, edgeR found a significantly higher proportion of significant ASVs than any other tool. In addition, we found that limma voom (TMMwsp) found the majority of ASVs to be significant (73.5%) in the Human-HIV (3) dataset while the other tools found 0-11% ASVs to be significant (**Fig. 1A**). Similarly, we found that both limma voom methods identified over 99% of ASVs to be significant in several cases such as the Built-Office and Freshwater-Arctic datasets. We found similar, although not as extreme, trends with LEfSe where in some datasets, such as the Human-T1D (1) dataset, the tool found a much higher percentage of significant hits (3.5%) compared with all other tools (0-0.4%). This observation is most likely a result of LEfSe filtering significant features by effect size rather than using FDR correction to reduce the number of false positives. We found that two of the three compositionally aware methods we tested identified fewer significant ASVs than the other tools tested. Specifically, ALDEx2 (mean: 1.4%) and ANCOM-II (mean: 0.8%) identified the fewest significant ASVs. We found the conservative behavior of these tools to be consistent across all 38 datasets we tested.

Overall, the results based on the filtered tables were similar, although there was a smaller range in the number of significant features identified by each tool. All tools except for ALDEx2 found a lower number of total significant features when compared with the unfiltered dataset (**Supp. Fig 1C-D**). As with the unfiltered data, ANCOM-II was the most stringent method (mean: 3.8%), while edgeR (mean: 32.5%), LEfSe (mean: 27.6%), limma voom (TMMwsp; mean: 27.3% / TMM; mean: 23.5%), and Wilcoxon (CLR; mean: 25.4%) tended to output the highest numbers of significant ASVs (**Fig. 1B**).

Finally, we examined the mean relative abundance of the features identified by each tool to determine whether tools may be biased toward the identification of highly abundant features. We found that both ALDEx2 (median: 0.013%), ANCOM-II (median: 0.024%) and to a lesser degree DESeq2 (median: 0.006%) tended to find significant features that were higher in relative abundance in the unfiltered datasets. A similar trend for ALDEx2 (median: 0.011%) and ANCOM-II (median: 0.029%) was also found in the filtered datasets (**Supp. Fig 1A-B**).

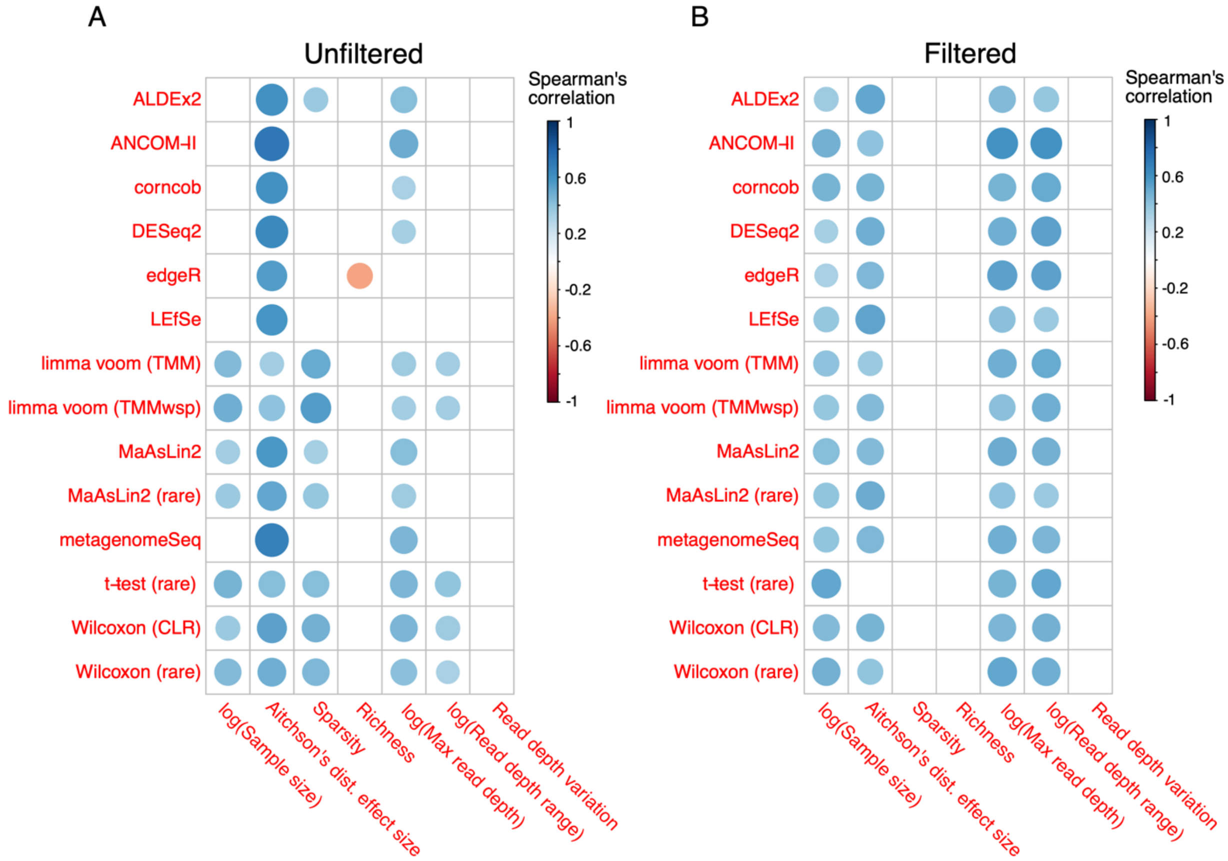


Figure 2: Dataset characteristics associated with percentage of significant amplicon sequence variants. The correlation coefficients (Spearman's rho) are displayed by size and color. These correspond to the dataset characteristics correlated with the percentage of significant amplicon sequence variants identified by that tool per dataset. Only significant correlations ($p < 0.05$) are displayed.

High variability of overlapping significant ASVs

We next investigated the overlap in significant ASVs across tools within each dataset. Based on the unfiltered data, we found that both limma voom methods identified similar sets of significant ASVs that were different from those of most other tools (**Fig. 3A**). However, we also found that many of the ASVs identified by the limma voom methods were also identified as significant based on the Wilcoxon (CLR) approach, despite these being highly methodologically distinct tools. Furthermore, the two Wilcoxon test approaches had different consistency profiles despite using the same hypothesis test. In contrast, we found that both MaAsLin2 approaches had similar consistency profiles, although the non-rarefied method found slightly lower-ranked features. We also found that the most conservative tools, ALDEx2 and ANCOM-II, primarily identified features that were also identified by almost all other methods. In contrast, edgeR and LEfSe, two tools that often identified the most significant ASVs, output the highest percentage of ASVs that were not identified by any other tool: 12.1% and 11.1%, respectively. Corncob, metagenomeSeq, and DESeq2 identified ASVs at more intermediate consistency profiles.

The overlap in significant ASVs based on the prevalence-filtered data was similar overall to the unfiltered data results (**Fig. 3B**). One important exception was that the limma voom approaches identified a much higher proportion of ASVs that were also identified by most other tools, compared with the unfiltered data. Nonetheless, similar to the unfiltered data results, the Wilcoxon (CLR) significant ASVs displayed a bimodal distribution and a strong overlap with limma voom methods. We also found that overall, the proportion of ASVs consistently identified as significant by more than 12 tools was much higher in the filtered data (mean: 38.5; SD: 15.8) compared with the unfiltered data (mean: 17.3; SD: 22.1). In contrast with the unfiltered results, corncob, metagenomeSeq, and DESeq2 had lower proportions of ASVs at intermediate consistency ranks. However, ALDEx2 and ANCOM-II once again produced significant ASVs that largely overlapped with other tools.

The above analyses summarized the variation in tool performance across datasets, but it is difficult to discern which tools performed most similarly from these results alone. To identify overall similarly performing tools we conducted principal coordinates analysis based on the Jaccard distance between significant sets of ASVs (**Fig. 3C, 3D**). One clear trend for both unfiltered and filtered data is that edgeR and LEfSe cluster together and are separated from other methods on the first principal coordinate. Interestingly, corncob, which is a methodologically distinct approach, also clusters relatively close to these two methods on the first PC. The major outliers on the second principal coordinate differ depending on whether the data was prevalence-filtered. For the unfiltered data, the main outliers are the limma voom methods, followed by Wilcoxon (CLR; **Fig. 3C**). In contrast, ANCOM-II is the sole major outlier on the second principal component based on filtered data (**Fig. 3D**). These visualizations highlight the major tool clusters based on the resulting sets of significant ASVs. However, the percentage of variation explained by the top two components is relatively low in each case, which means that substantial information regarding potential tool clustering is missing from these panels (**Supp. Fig. 2** and **Supp. Fig. 3**). For instance, ANCOM-II and corncob are major outliers on the third and fourth principal coordinates, respectively, of the unfiltered data analysis, which highlights the uniqueness of these methods.

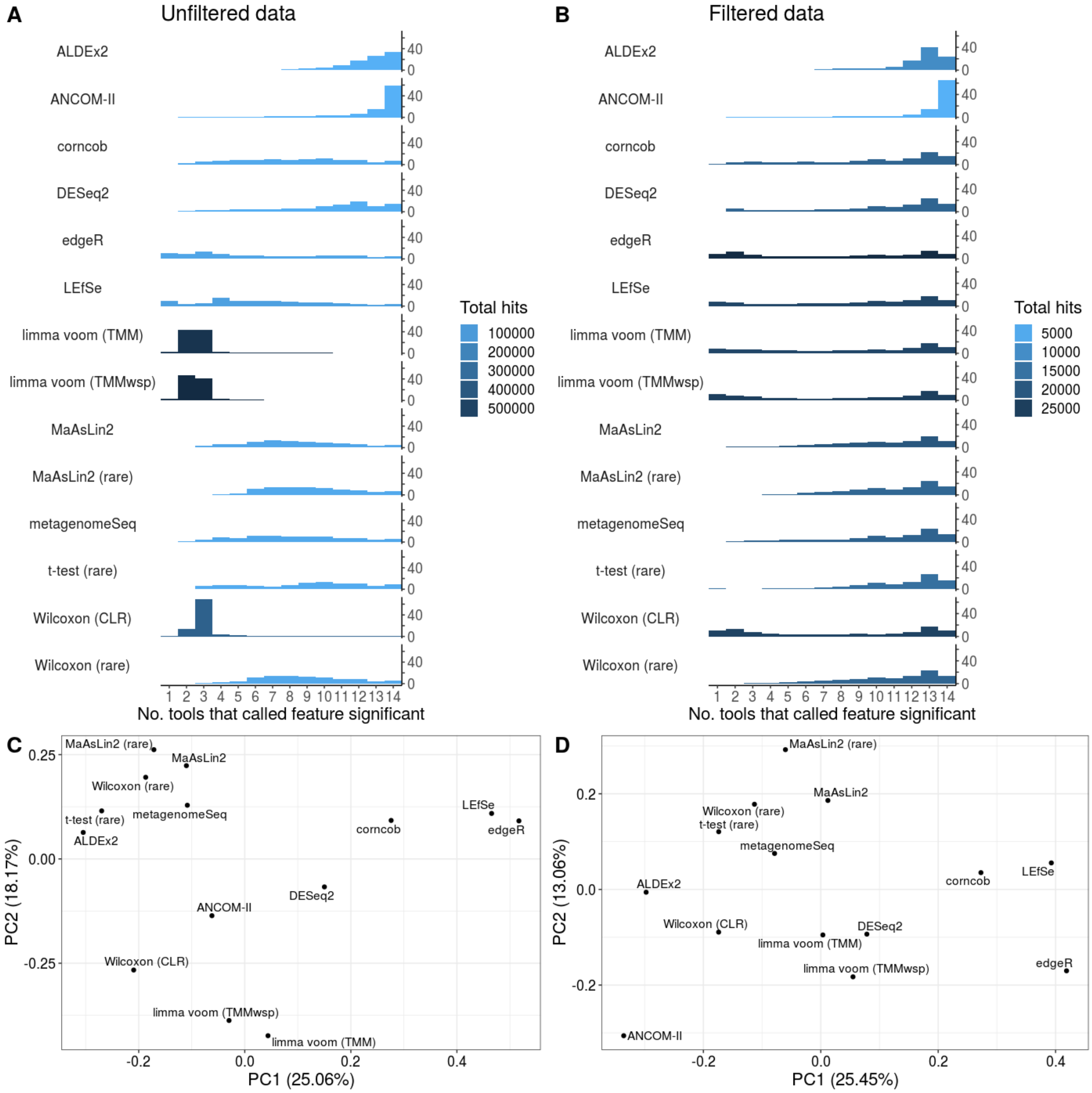


Figure 3: Overlap of significant features across tools and tool clustering. (A and B) The number of tools that called each feature significant, stratified by features called by each individual tool for the (A) unfiltered and (B) 10% prevalence-filtered data. The features correspond to the amplicon sequence variants (and operational taxonomic units) from all 38 tested datasets. Results are shown as a percentage of all ASVs identified by each tool. The total number of significant features identified by each tool is indicated by the bar colors. (C and D) Plots are displayed for the first two principal coordinates (PCs) for both (C) non-prevalence-filtered and (D) 10% prevalence-filtered data. These plots are based on the mean inter-tool Jaccard distance across the 38 main datasets that we analyzed, computed by averaging over the inter-tool distance matrices for all individual datasets in order to weight each dataset equally.

False discovery rate of microbiome differential abundance tools depends on the dataset

We next investigated the FDR of each DA tool across eight datasets. For each dataset we selected the most frequently sampled group and randomly reassigned them as case or control samples. Each DA tool was then run on those samples and results were compared. This was repeated a total of 10 times for each filtered dataset and 100 times for each unfiltered dataset (except for corncob, ANCOM-II and ALDEx2). We used a higher number of replicates for the unfiltered datasets because they were less stable across replicates. The percentage of significant ASVs after BH multiple-test correction was relatively low for most tested tools on the filtered data (**Fig. 4B**). Two outliers were edgeR (mean: 10.3%; SD: 9.0%) and LEfSe (mean: 4.4%; SD: 1.2%), which consistently identified more significant hits compared with other tools (range of other tool means: 0% - 2.2%). Both limma voom methods output highly variable percentages of significant ASVs, especially based on the unfiltered data (**Fig. 4A**). In particular, in 5/8 of the unfiltered datasets, the limma voom methods identified more than 5% of ASVs as significant on average. Interestingly, while these two methods exhibited similar performance overall, the performance within the unfiltered Freshwater-Treatment dataset was highly different between the methods with the TMMwsp method identifying 0.001% and the TMM method identifying 9.0%. Only ALDEx2 and the t-test (rare) approach consistently identified no ASVs as significantly different in this analysis.

Overall, we found that the raw numbers of significant ASVs were lower in the filtered dataset than in the unfiltered data (as expected due to many ASVs being filtered out), and that most tools identified only a small percentage of significant ASVs, regardless of filtering procedure. The exceptions were the two limma voom methods, which had high FDRs with unfiltered data, and edgeR and LEfSe, which had high FDRs on the filtered data. Although these tools stand out on average, we also observed that in several replicates on the unfiltered datasets, the Wilcoxon (CLR) approach identified almost all features as statistically significant (**Supp. Fig. 4).** This was also true for both limma voom methods, which highlights that a minority of replicates are driving up the average FDR of these methods. Such extreme values were not observed for the filtered data (**Supp. Fig 5**).

Investigation into these outlier replicates for the Wilcoxon (CLR) approach revealed that the mean differences in read depth between the two tested groups were consistently higher in replicates in which 30% or more of ASVs were significant. Interestingly, this pattern was absent when examining replicates for the limma voom methods (**Supp. Fig 6**).

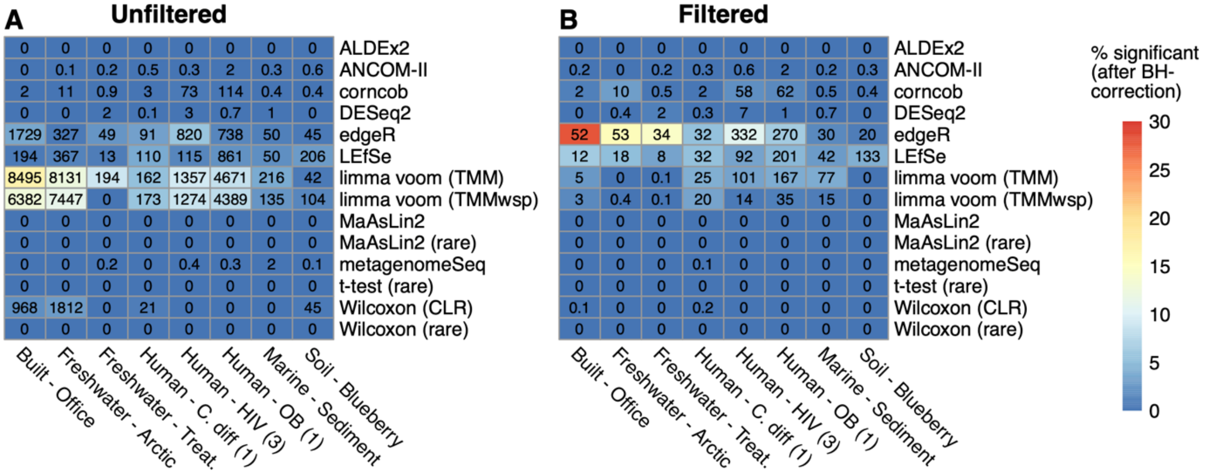


Figure 4: Variation in false discovery rate across the tested differential abundance tools in the context of simple simulations. Heatmap of the percentage and number of significant amplicon sequence variants (ASVs) identified by each tool in eight simulation datasets based on applying (A) no prevalence filter and (B) a 10% prevalence filter. Cell colours indicate the percentage and the number of significant ASVs is written in each cell. Mean numbers higher than one were rounded to the nearest integer for visualization. Significant ASVs were identified after applying the Benjamini-Hochberg (BH) false discovery rate procedure and then using a cut-off of 0.05.

Tools vary in how consistently they identify the same significant genera within diarrhea case-control datasets

Separate from the above analysis comparing consistency between tools on the same dataset, we next investigated whether certain tools provide more consistent signals across datasets of the same disease. This analysis focused on the genus-level across tools to help limit inter-study variation. We specifically focused on diarrhea as a phenotype, which has been shown to exhibit a strong effect on the microbiome and to be relatively reproducible across studies (Duvallet et al., 2017).

We acquired five datasets for this analysis representing the microbiome of individuals with diarrhea compared with individuals without diarrhea (see Methods). We ran all DA tools on each individual filtered dataset. Similar to our ASV-level analyses, the tools substantially varied in terms of the number of significant genera identified. For instance, ALDEx2 identified a mean of 17.6 genera as significant in each dataset (SD: 17.4), while edgeR identified a mean of 46.0 significant genera (SD: 12.9). Tools that identify more genera as significant in general are accordingly more likely to identify genera as consistently significant compared with tools with fewer significant hits. Accordingly, inter-tool comparisons of the number of times each genus was identified as significant would not be informative.

Instead, we analyzed the observed distribution of the number of studies that each genus was identified as significant in compared with the expected distribution given random data. This approach enabled us to compare the tools based on how much more consistently each tool performed relative to its own random expectation. For instance, on average edgeR identified significant genera more consistently across studies compared with ALDEx2 (mean numbers of datasets that genera were found in across studies were 1.67 and 1.54 for edgeR and ALDEx2, respectively). However, this observation was simply driven by the increased number of significant genera identified by edgeR. Indeed, when compared with the random expectation, ALDEx2 displayed a 1.35-fold increase ($p < 0.0001$) of consistency in calling significant genera in the observed data. In contrast, edgeR produced results that were only 1.10-fold more consistent compared with the random expectation ($p = 0.02$).

ALDEx2 and edgeR represent the extremes of how consistently tools identify the same genera as significant across studies, but there is a large range (**Fig. 5**). Notably, all tools were significantly more consistent than the random expectation across these datasets ($p < 0.05$) (**Table 2**). In addition to ALDEx2, the other top performing approaches based on this evaluation included both MaAsLin2 workflows, limma voom (TMM), and ANCOM-II.

We conducted a similar investigation across five obesity 16S datasets, which was more challenging to interpret due to the lower consistency in general (**Supp. Table 2**). Specifically, most significant genera were called in only a single study and only MaAslin2 (both with non-rarefied and rarefied data) and the t-test (rare) approaches performed significantly better than expected by chance ($p < 0.05$). The MaAsLin2 (rare) approach produced by far the most consistent results based on these datasets (fold difference: 1.23; $p = 0.006$).

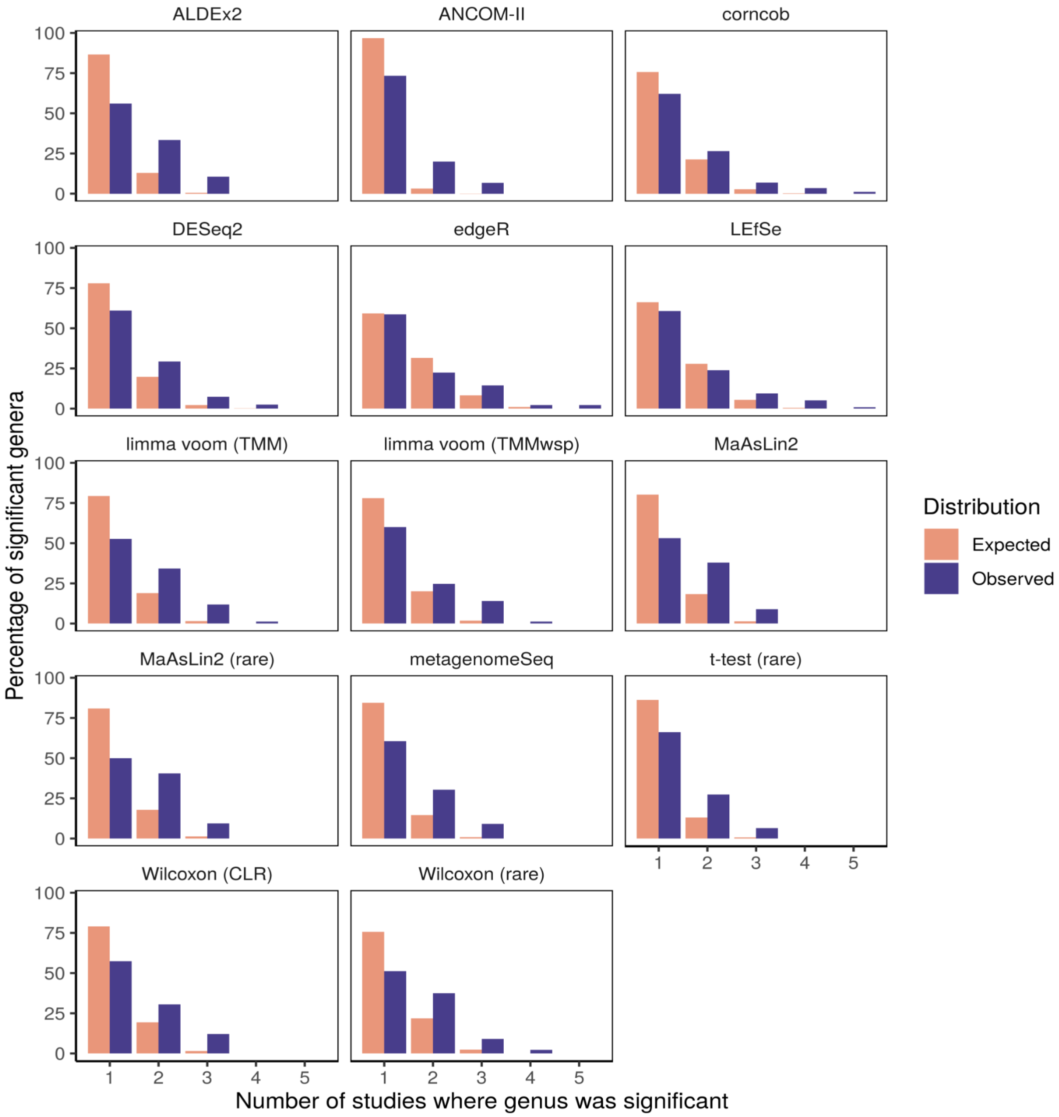
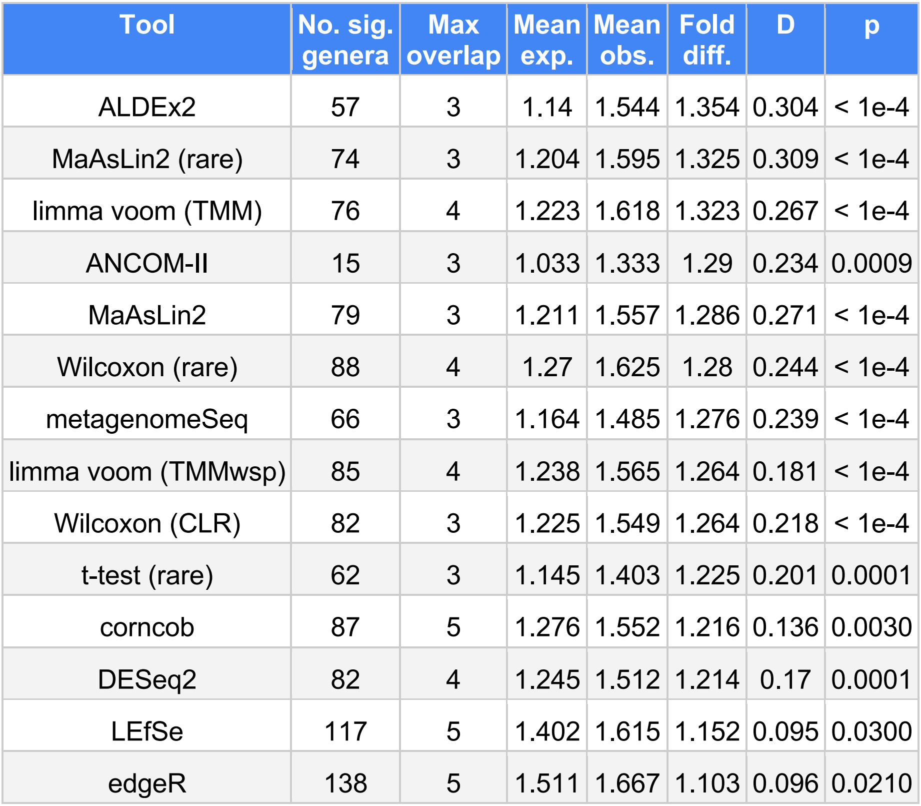


Figure 5: Observed consistency of significant genera across diarrhea datasets is higher than the random expectation overall. These barplots illustrate the distributions of the number of studies for which each genus was identified as significant (excluding genera never found to be significant). The random expectation distribution is based on replicates of randomly selecting genera as significant and then computing the consistency across studies.

Table 2: Results of Kolmogorov-Smirnov tests comparing observed and expected consistency in differentially abundant genera across five diarrhea datasets



Tool	No. sig. genera	Max overlap	Mean exp.	Mean obs.	Fold diff.	D	p
ALDEx2	57	3	1.14	1.544	1.354	0.304	< 1e-4
MaAsLin2 (rare)	74	3	1.204	1.595	1.325	0.309	< 1e-4
limma voom (TMM)	76	4	1.223	1.618	1.323	0.267	< 1e-4
ANCOM-II	15	3	1.033	1.333	1.29	0.234	0.0009
MaAsLin2	79	3	1.211	1.557	1.286	0.271	< 1e-4
Wilcoxon (rare)	88	4	1.27	1.625	1.28	0.244	< 1e-4
metagenomeSeq	66	3	1.164	1.485	1.276	0.239	< 1e-4
limma voom (TMMwsp)	85	4	1.238	1.565	1.264	0.181	< 1e-4
Wilcoxon (CLR)	82	3	1.225	1.549	1.264	0.218	< 1e-4
t-test (rare)	62	3	1.145	1.403	1.225	0.201	0.0001
corncob	87	5	1.276	1.552	1.216	0.136	0.0030
DESeq2	82	4	1.245	1.512	1.214	0.17	0.0001
LEfSe	117	5	1.402	1.615	1.152	0.095	0.0300
edgeR	138	5	1.511	1.667	1.103	0.096	0.0210

Column descriptions:

No. sig. genera: Number of genera significant in at least one dataset

Max overlap: Max number of datasets where a genus was called significant by this tool

Mean exp.: Mean number of datasets that each genera is expected to be significant in (of the genera that are significant at least once)

Mean obs.: Mean number of datasets that each genera was observed to be significant in (of the genera that are significant at least once)

Fold diff.: Fold difference of mean observed over mean expected number of times significant genera are found across multiple datasets

D: Kolmogorov-Smirnov test statistic

Discussion

Herein we have compared the performance of commonly used DA tools, primarily on actual 16S datasets. While it might be argued that differences in tool outputs are expected given that they test different hypotheses, we believe this perspective ignores how these tools are used in practice. In particular, these tools are frequently used interchangeably in the microbiome literature. Accordingly, an improved understanding of the variation in DA method performance is crucial to properly interpret microbiome studies. We have illustrated here that these tools can produce substantially different results, which highlights that many biological interpretations based on microbiome data analysis are likely not robust to DA tool choice. Our findings should serve as a cautionary tale for researchers conducting their own microbiome data analysis and reinforce the need to honestly report the findings of a representative set of different analysis options to ensure robust results are reported. Despite the high variation across DA tool results, we were able to characterize several consistent patterns produced by various tools that researchers should keep in mind when assessing both their own results and results from published work.

Two major groups of DA tools could be distinguished by how many significant ASVs they tended to identify. We found that limma voom, edgeR, Wilcoxon (CLR), and LEfSe output a high number of significant ASVs on average. In contrast, ALDEx2 and ANCOM-II tended to identify only a relatively small number of ASVs as significant. We hypothesize that these latter tools are more conservative and have higher precision, but with a concomitant probable loss in sensitivity. This hypothesis is related to our observation that significant ASVs identified by these two tools tended to also be identified by almost all other differential abundance methods, which we interpret to be ASVs that are more likely to be true positives.

Given that ASVs commonly identified as significant are likely more reliable, it is noteworthy that significant ASVs in the unfiltered data tended to be called by fewer tools. This was particularly true for both limma voom approaches and the Wilcoxon (CLR) approach. Although it is possible that many of these significant ASVs are incorrectly missed by other tools, it is more likely that these tools are simply performing especially poorly on unfiltered data due to several reasons, such as data sparsity.

This issue with the limma voom approaches was also highlighted by high false positive rates on several unfiltered randomized datasets, which agrees with a past FDR assessment of this approach (Hawinkel et al., 2019). It is also important to acknowledge that our randomized approach for estimating FDR is not a perfect representation of real data; that is, real sample groupings will likely contain some systematic differences in microbial abundances—although the effect size may be very small—whereas our randomized datasets should have none. Accordingly, identifying only a few significant ASVs under this approach is not necessarily proof that a tool has a low FDR in practice. However, tools that identified many significant ASVs in the absence of distinguishing signals likely also have high FDR on real data.

Two additional particularly problematic tools based on this analysis were edgeR and LEfSe. The edgeR method has been previously found to exhibit a high FDR on several occasions (Hawinkel et al., 2019; Thorsen et al., 2016) Although metagenomeSeq also has been flagged as such (Thorsen et al., 2016), that was not the case in our analysis. This agrees with a recent report that metagenomeSeq (using the zero-inflated log-normal approach, as we did) appropriately controlled the FDR, but exhibited low power (Lin and Peddada, 2020). There have been mixed results previously regarding whether ANCOM appropriately controls the FDR (Hawinkel et al., 2019; Weiss et al., 2017), but the results from our limited analysis suggest that this method is conservative and controls the FDR while potentially missing true positives.

Related to this point, we found that ANCOM-II performed better than average at identifying the same genera as significantly DA across five diarrhea-related datasets despite only identifying a mean of four genera as significant per dataset. Nonetheless, the ANCOM-II results were less consistent than ALDEx2, both MaAsLin2 workflows, and limma voom (TMM). The tools that produced the least consistent results across datasets (relative to the random expectation) included the t-test (rare) approach, LEfSe, and edgeR. The random expectation in this case was quite simplistic; it was generated based on the assumption that all genera were equally likely to be significant by chance. This assumption must be invalid to some degree simply because some genera are more prevalent than others across samples. Accordingly, it is surprising that the tools produced only marginally more consistent results than expected.

Although this cross-data consistency analysis was informative, it was interesting to note that not all environments and datasets are appropriate for this comparison. Specifically, we found that the consistency of significant genera across five datasets comparing obese and control individuals was no higher than expected by chance for most tools. This observation does not necessarily reflect that there are few consistent genera that differ between obese and non-obese individuals; it could instead simply reflect technical and/or biological factors that differ between the particular datasets we analyzed (Pollock et al., 2018). Despite these complicating factors, it is noteworthy that the MaAsLin2 workflows produced more consistent results than expected based on these datasets.

We believe the above observations regarding DA tools are valuable, but many readers are likely primarily interested in hearing specific recommendations. Indeed, the need for standardized practices in microbiome analysis have recently become better appreciated(Hill, 2020). One goal of our work was to validate the recommendations of another recent DA method evaluation paper, which found that limma voom, corncob, and DESeq2 performed best overall (Calgaro et al., 2020). Based on our results we do not recommend these tools as the sole methods used for data analysis, and instead would suggest using more conservative methods such as ALDEx2 and ANCOM-II. Although these methods have lower statistical power (Calgaro et al., 2020; Hawinkel et al., 2019), we believe this an acceptable trade-off given the higher cost of identifying false positives as differentially abundant. However, MaAsLin2 (particularly with rarefied data) could also be a reasonable choice for users looking for increased statistical power at the potential cost of more false positives. We can clearly recommend that users avoid using edgeR (a tool primarily intended for RNA-seq data) as well as LEfSe for conducting DA testing with 16S data. Users should also be aware that limma voom and the Wilcoxon (CLR) approaches may perform especially poorly on unfiltered data. This is especially true for the Wilcoxon (CLR) approach when read depths greatly differ between groups of interest.

More generally, we recommend that users employ several methods and focus on significant features identified by most tools, while keeping in mind the characteristics of the tools presented within this manuscript. For example, authors may want to present identified taxonomic markers in categories based on the tool characteristics presented within this paper or the number of tools that agree upon its identification. Importantly, applying multiple DA tools to the same dataset should be reported explicitly. Clearly this approach would make results more difficult to biologically interpret, but it would provide a clearer perspective on which differentially abundant features are robust to reasonable changes in the analysis.

A common counterargument to using consensus approaches with DA tools is that there is no assurance that the intersection of the tool outputs is more reliable; it is possible that the tools are simply picking up the same noise as significant. Although we think this is unlikely, in any case running multiple DA tools is still important to give context to reporting significant features. For example, researchers might be using a tool that produces highly non-overlapping sets of significant features compared with other DA approaches. Even if the researchers are confident in their approach, these discrepancies should be made clear when the results are summarized. This is crucial for providing honest insight into how robust specific findings are expected to be across independent studies, which often use different DA approaches.

How and whether to conduct independent filtering of data prior to conducting DA tests are other important open questions regarding microbiome data analysis (Schloss, 2020). Although statistical arguments regarding the validity of independent filtering are beyond the scope of this work, intuitively it is reasonable to exclude features found in only a small number of samples (regardless of which groups those samples are in). The basic reason for this is that otherwise the burden of multiple-test correction becomes so great as to nearly prohibit identifying any differentially abundant features. Despite this drawback, many tools identified large numbers of significant ASVs in the unfiltered data. However, these significant ASVs tended to be more tool-specific in the unfiltered data and there was much more variation in the percentage of significant ASVs across tools. Accordingly, we would suggest performing prevalence filtering (e.g., at 10%) of features prior to DA testing, although we acknowledge that more work is needed to estimate an optimal cut-off rather than just arbitrarily selecting one (McMurdie and Holmes, 2014).

Another common question is whether microbiome data should be rarefied prior to DA testing. It is possible that the question of whether to rarefy data has received disproportionate attention in the microbiome field: there are numerous other factors affecting an analysis pipeline that likely affect results more. Indeed, tests based on rarefied data in our analyses did not perform substantially worse than other methods on average. More specifically, the most consistent inter-tool methods, ANCOM-II and ALDEx2, are based on non-rarefied data, but MaAsLin2 based on rarefied data produced the most consistent results across datasets of the same phenotype. Accordingly, we cannot definitively conclude that rarefying data prior to DA testing is always inadvisable. It should be noted that we are referring only to rarefying in the context of DA testing: whether rarefying is advisable for other analyses, such as prior to computing diversity metrics, is beyond the scope of this work (McMurdie and Holmes, 2014; Weiss et al., 2017).

In conclusion, the high variation in the output of DA tools across numerous 16S rRNA gene sequencing datasets highlights an alarming reproducibility crisis facing microbiome researchers. Unfortunately, this high variation across tools implies that biological interpretations will often drastically differ depending on which DA tool is used. One incomplete solution to this problem would be to normalize the practice of reporting results based on a range of DA tools, which would help ensure that any key conclusions were robust to the researchers' analysis choices.

Acknowledgements

We would like to thank everyone who responded to MGIL's queries on Twitter regarding which differential abundance tools to evaluate. We would also like to thank the authors of all DA tools and datasets used in this study for making their code and data freely available. JTN is funded by both a Nova Scotia Graduate Scholarship and a ResearchNS Scotia Scholars award. GMD was funded by a Canadian Graduate Scholarship (Doctoral) from NSERC. MGIL is funded through a National Sciences and Engineering Research Council (NSERC) Discovery Grant and the Canada Research Chairs program.

Competing interests

The authors declare that they have no competing interests.

References

Aitchison J. 1982. The Statistical Analysis of Compositional Data. *J R Stat Soc Ser B* **44**:139–160. doi:https://doi.org/10.1111/j.2517-6161.1982.tb01195.x

Alkanani AK, Hara N, Gottlieb PA, Ir D, Robertson CE, Wagner BD, Frank DN, Zipris D. 2015. Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes. *Diabetes* **64**:3510 LP – 3520. doi:10.2337/db14-1847

Allaband C, McDonald D, Vázquez-Baeza Y, Minich JJ, Tripathi A, Brenner DA, Loomba R, Smarr L, Sandborn WJ, Schnabl B, Dorrestein P, Zarrinpar A, Knight R. 2019. Microbiome 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clin Gastroenterol Hepatol* **17**:218–230. doi:10.1016/j.cgh.2018.09.017

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP, Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. *mSystems* **2**.

Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med* **8**:37. doi:10.1186/s13073-016-0290-3

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang K Bin, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik A V, Metcalf JL, Morgan SC, Morton J, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson Michael S II, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso JG. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science. *PeerJ Prepr* **6**:e27295v2. doi:10.7287/peerj.preprints.27295v2

Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for high-throughput experiments. *Proc Natl Acad Sci* **107**:9546 LP – 9551. doi:10.1073/pnas.0914005107

Calgaro M, Romualdi C, Waldron L, Risso D, Vitulo N. 2020. Assessment of statistical methods from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome Biol* **21**:191. doi:10.1186/s13059-020-02104-1

Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, Siegel J, Caporaso JG. 2016. Geography and Location Are the Primary Drivers of Office Microbiome Composition. *mSystems* **1**:e00022-16. doi:10.1128/mSystems.00022-16

Comeau AM, Douglas GM, Langille MGI. 2017. Microbiome Helper: a Custom and Streamlined Workflow for Microbiome Research. *mSystems* **2**.

De Tender CA, Devriese LI, Haegeman A, Maes S, Ruttink T, Dawyndt P. 2015. Bacterial Community Profiling of Plastic Litter in the Belgian Part of the North Sea. *Environ Sci Technol* **49**:9629–9638. doi:10.1021/acs.est.5b01093

Dinh DM, Volpe GE, Duffalo C, Bhalchandra S, Tai AK, Kane A V, Wanke CA, Ward HD. 2015. Intestinal Microbiota, Microbial Translocation, and Systemic Inflammation in Chronic HIV Infection. *J Infect Dis* **211**:19–27. doi:10.1093/infdis/jiu409

Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J Veg Sci* **14**. doi:10.1111/j.1654-1103.2003.tb02228.x

Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, Tayler R, El-Omar EM, Russell RK, Hold GL, Langille MGI, Van Limbergen J. 2018. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* **6**:13. doi:10.1186/s40168-018-0398-3

Dranse HJ, Zheng A, Comeau AM, Langille MGI, Zabel BA, Sinal CJ. 2018. The impact of chemerin or chemokine-like receptor 1 loss on the mouse gut microbiome. *PeerJ* **6**:e5494. doi:10.7717/peerj.5494

Duvallet C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome studies identifies disease-specific and shared responses. *Nat Commun* **8**:1784. doi:10.1038/s41467-017-01973-8

Feijen EAM, Font-Gonzalez A, van Dalen EC, van der Pal HJH, Reulen RC, Winter DL, Kuehni CE, Haupt R, Alessi D, Byrne J, Bardi E, Jakab Z, Grabow D, Garwicz S, Jankovic M, Levitt GA, Skinner R, Zadravec Zaletel L, Hjorth L, Tissing WJE, de Vathaire F, Hawkins MM, Kremer LCM, consortium P. 2016. Late Cardiac Events after Childhood Cancer: Methodological Aspects of the Pan-European Study PanCareSurFup. *PLoS One* **11**:e0162778.

Fernandes AD, Reid JNS, Macklaim JM, McMurrough TA, Edgell DR, Gloor GB. 2014. 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. doi:10.1186/2049-2618-2-15

Frère L, Maignien L, Chalopin M, Huvet A, Rinnert E, Morrison H, Kerninon S, Cassone A-L, Lambert C, Reveillaud J, Paul-Pont I. 2018. Microplastic bacterial communities in the Bay of Brest: Influence of polymer type and size. *Environ Pollut* **242**:614–625. doi:https://doi.org/10.1016/j.envpol.2018.07.023

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* **8**:2224. doi:10.3389/fmicb.2017.02224

Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H, Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M, Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-analysis. *Nat Methods* **15**:796–798. doi:10.1038/s41592-018-0141-9

Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M, Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human Genetics Shape the Gut Microbiome. *Cell* **159**:789–799. doi:10.1016/j.cell.2014.09.053

Hawinkel S, Mattiello F, Bijnens L, Thas O. 2019. A broken promise: microbiome differential abundance methods do not control the false discovery rate. *Brief Bioinform* **20**:210–221. doi:10.1093/bib/bbx104

Hill C. 2020. You have the microbiome you deserve. *Gut Microbiome* **1**:e3. doi:DOI: 10.1017/gmb.2020.3

Hoellein TJ, McCormick AR, Hittie J, London MG, Scott JW, Kelly JJ. 2017. Longitudinal patterns of microplastic concentration and bacterial assemblages in surface and benthic habitats of an urban river. *Freshw Sci* **36**:491–507. doi:10.1086/693012

Ji P, Parks J, Edwards MA, Pruden A. 2015. Impact of Water Chemistry, Pipe Material and Stagnation on the Building Plumbing Microbiome. *PLoS One* **10**:e0141087.

Kaul A, Mandal S, Davidov O, Peddada SD. 2017. Analysis of Microbiome Data in the Presence of Excess Zeros. *Front Microbiol* **8**:2114. doi:10.3389/fmicb.2017.02114

Kesy K, Oberbeckmann S, Kreikemeyer B, Labrenz M. 2019. Spatial Environmental Heterogeneity Determines Young Biofilm Assemblages on Microplastics in Baltic Sea Mesocosms . *Front Microbiol* .

Kolde R. 2012. Pheatmap: pretty heatmaps. *R Packag version* **1**.

Lamoureux E V, Grandy SA, Langille MGI. 2017. Moderate Exercise Has Limited but Distinguishable Effects on the Mouse Microbiome. *mSystems* **2**:e00006-17. doi:10.1128/mSystems.00006-17

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol* **15**:R29. doi:10.1186/gb-2014-15-2-r29

Lin H, Peddada S Das. 2020. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ biofilms microbiomes* **6**:60. doi:10.1038/s41522-020-00160-w

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**:550. doi:10.1186/s13059-014-0550-8

Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, Gebert MJ, Knight R, Fontenot AP, Palmer BE. 2013. Alterations in the Gut Microbiota Associated with HIV-1 Infection. *Cell Host Microbe* **14**:329–339. doi:10.1016/j.chom.2013.08.006

Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren B, Schwager EH, Chatterjee S, Thompson KN, Wilkinson JE, Subramanian A, Lu Y, Waldron L, Paulson JN, Franzosa EA, Bravo HC, Huttenhower C. 2021. Multivariable Association Discovery in Population-scale Meta-omics Studies. *bioRxiv* 2021.01.20.427420. doi:10.1101/2021.01.20.427420

Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. 2015. Analysis of composition of microbiomes: a novel method for studying microbial composition. *Microb Ecol Health Dis* **26**:27663. doi:10.3402/mehd.v26.27663

Martin BD, Witten D, Willis AD. 2020. Modeling microbial abundances and dysbiosis with beta-binomial regression. *Ann Appl Stat* **14**:94–115. doi:10.1214/19-AOAS1283

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1 Next Gener Seq Data Anal.*

McCormick AR, Hoellein TJ, London MG, Hittie J, Scott JW, Kelly JJ. 2016. Microplastic in surface waters of urban rivers: concentration, sources, and associated bacterial assemblages. *Ecosphere* **7**:e01556. doi:https://doi.org/10.1002/ecs2.1556

McMurdie PJ, Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Comput Biol* **10**:e1003531.

Mejía-León ME, Petrosino JF, Ajami NJ, Domínguez-Bello MG, de la Barca AMC. 2014. Fecal microbiota imbalance in Mexican children with type 1 diabetes. *Sci Rep* **4**:3814. doi:10.1038/srep03814

Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R. 2019. Establishing microbial composition measurement standards with reference frames. *Nat Commun* **10**:2719. doi:10.1038/s41467-019-10656-5

Nearing JT, Connors J, Whitehouse S, Van Limbergen J, Macdonald T, Kulkarni K, Langille MGI. 2019. Infectious Complications Are Associated With Alterations in the Gut Microbiome in Pediatric Patients With Acute Lymphoblastic Leukemia. *Front Cell Infect Microbiol* **9**:28. doi:10.3389/fcimb.2019.00028

Noguera-Julian M, Rocafort M, Guillén Y, Rivera J, Casadellà M, Nowak P, Hildebrand F, Zeller G, Parera M, Bellido R, Rodríguez C, Carrillo J, Mothe B, Coll J, Bravo I, Estany C, Herrero C, Saz J, Sirera G, Torrela A, Navarro J, Crespo M, Brander C, Negredo E, Blanco J, Guarner F, Calle ML, Bork P, Sönnerborg A, Clotet B, Paredes R. 2016. Gut Microbiota Linked to Sexual Preference and HIV Infection. *EBioMedicine* **5**:135–146. doi:10.1016/j.ebiom.2016.01.032

Oberbeckmann S, Osborn AM, Duhaime MB. 2016. Microbes on a Bottle: Substrate, Season and Geography Influence Community Composition of Microbes Colonizing Marine Plastic Debris. *PLoS One* **11**:e0159289.

Oliveira FS, Brestelli J, Cade S, Zheng J, Iodice J, Fischer S, Aurrecoechea C, Kissinger JC, Brunk BP, Stoeckert Jr CJ, Fernandes GR, Roos DS, Beiting DP. 2018. MicrobiomeDB: a systems biology platform for integrating, mining and analyzing microbiome experiments.

Nucleic Acids Res **46**:D684–D691. doi:10.1093/nar/gkx1027

Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, Giannoukos G, Ciulla D, Tabbaa D, Ingram J, Schauer DB, Ward D V, Korzenik JR, Xavier RJ, Bousvaros A, Alm EJ. 2012. Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with Inflammatory Bowel Disease. *PLoS One* **7**:e39242.

Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial marker-gene surveys. *Nat Methods* **10**:1200–1202. doi:10.1038/nmeth.2658

Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. 2016. ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *Ann Appl Stat* **10**:946–963. doi:10.1214/16-AOAS920

Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The Madness of Microbiome: Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies. *Appl Environ Microbiol* **84**:e02627-17. doi:10.1128/AEM.02627-17

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. 2014. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol* **15**:R76. doi:10.1186/gb-2014-15-6-r76

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**:e47–e47. doi:10.1093/nar/gkv007

Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* **11**:R25. doi:10.1186/gb-2010-11-3-r25

Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**:e2584. doi:10.7717/peerj.2584

Rosato A, Barone M, Negroni A, Brigidi P, Fava F, Xu P, Candela M, Zanaroli G. 2020. Microbial colonization of different microplastic types and biotransformation of sorbed PCBs by a marine anaerobic bacterial community. *Sci Total Environ* **705**:135790. doi:10.1016/j.scitotenv.2019.135790

Ross MC, Muzny DM, McCormick JB, Gibbs RA, Fisher-Hoch SP, Petrosino JF. 2015. 16S gut community of the Cameron County Hispanic Cohort. *Microbiome* **3**:7. doi:10.1186/s40168-015-0072-y

Scheperjans F, Aho V, Pereira PAB, Koskinen K, Paulin L, Pekkonen E, Haapaniemi E, Kaakkola S, Eerola-Rautio J, Pohja M, Kinnunen E, Murros K, Auvinen P. 2015. Gut microbiota are related to Parkinson’s disease and clinical phenotype. *Mov Disord* **30**:350–358. doi:https://doi.org/10.1002/mds.26069

Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, Rostron T, Cerundolo V, Pamer EG, Abramson SB, Huttenhower C, Littman DR. 2013. Expansion of intestinal Prevotella copri correlates with enhanced susceptibility to arthritis. *Elife* **2**:e01202. doi:10.7554/eLife.01202

Schloss PD. 2020. Removal of rare amplicon sequence variants from 16S rRNA gene sequence surveys biases the interpretation of community structure data. *bioRxiv* 2020.12.11.422279. doi:10.1101/2020.12.11.422279

Schneider D, Thürmer A, Gollnow K, Lugert R, Gunka K, Groß U, Daniel R. 2017. Gut bacterial communities of diarrheic patients with indications of Clostridioides difficile infection. *Sci Data* **4**:170152. doi:10.1038/sdata.2017.152

Schubert AM, Rogers MAM, Ring C, Mogle J, Petrosino JP, Young VB, Aronoff DM, Schloss PD. 2014. Microbiome Data Distinguish Patients with &lt;span class=&quot;named-content genus-species&quot; id=&quot;named-content-1&quot;&gt;Clostridium difficile&lt;/span&gt; Infection and Non-&lt;span class=&quot;named-content genus-species&quot; id=&quot;named-content-2&quot;&gt;C. difficile&lt;/span&gt;-Associated Diarrhea from Healthy Controls. *MBio* **5**:e01021-14. doi:10.1128/mBio.01021-14

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011. Metagenomic biomarker discovery and explanation. *Genome Biol* **12**:R60. doi:10.1186/gb-2011-12-6-r60

Sekhon JS. 2011. Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *J Stat Software; Vol 1, Issue 7* .

Singh P, Teal TK, Marsh TL, Tiedje JM, Mosci R, Jernigan K, Zell A, Newton DW, Salimnia H, Lephart P, Sundin D, Khalife W, Britton RA, Rudrik JT, Manning SD. 2015. Intestinal microbial communities associated with acute enteric infections and disease recovery. *Microbiome* **3**:45. doi:10.1186/s40168-015-0109-2

Son JS, Zheng LJ, Rowehl LM, Tian X, Zhang Y, Zhu W, Litcher-Kelly L, Gadow KD, Gathungu G, Robertson CE, Ir D, Frank DN, Li E. 2015. Comparison of Fecal Microbiota in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons Simplex Collection. *PLoS One* **10**:e0137725.

Thorsen J, Brejnrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, Sørensen S, Bisgaard H, Waage J. 2016. 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. doi:10.1186/s40168-016-0208-8

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE. 2009. A core gut microbiome in obese and lean twins. *Nature* **457**. doi:10.1038/nature07540

Vincent C, Stephens DA, Loo VG, Edens TJ, Behr MA, Dewar K, Manges AR. 2013. Reductions in intestinal Clostridiales precede the development of nosocomial Clostridium difficile infection. *Microbiome* **1**:18. doi:10.1186/2049-2618-1-18

Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR, Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* **5**:27. doi:10.1186/s40168-017-0237-y

Wu Linwei, Ning D, Zhang B, Li Y, Zhang P, Shan X, Zhang Qiuting, Brown MR, Li Z, Van Nostrand JD, Ling F, Xiao N, Zhang Ya, Vierheilig J, Wells GF, Yang Y, Deng Y, Tu Q, Wang A, Acevedo D, Agullo-Barcelo M, Alvarez PJJ, Alvarez-Cohen L, Andersen GL, de Araujo JC, Boehnke KF, Bond P, Bott CB, Bovio P, Brewster RK, Bux F, Cabezas A, Cabrol L, Chen S, Criddle CS, Deng Y, Etchebehere C, Ford A, Frigon D, Sanabria J, Griffin JS, Gu AZ, Habagil M, Hale L, Hardeman SD, Harmon M, Horn H, Hu Z, Jauffur S, Johnson DR, Keller J, Keucken A, Kumari S, Leal CD, Lebrun LA, Lee J, Lee M, Lee ZMP, Li Y, Li Z, Li M, Li X, Ling F, Liu Y, Luthy RG, Mendonça-Hagler LC, de Menezes FGR, Meyers AJ, Mohebbi A, Nielsen PH, Ning D, Oehmen A, Palmer A, Parameswaran P, Park J, Patsch D, Reginatto V, de los Reyes FL, Rittmann BE, Noyola A, Rossetti S, Shan X, Sidhu J, Sloan WT, Smith K, de Sousa OV, Stahl DA, Stephens K, Tian R, Tiedje JM, Tooker NB, Tu Q, Van Nostrand JD, De los Cobos Vasconcelos D, Vierheilig J, Wagner M, Wakelin S, Wang A, Wang B, Weaver JE, Wells GF, West S, Wilmes P, Woo S-G, Wu Linwei, Wu J-H, Wu Liyou, Xi C, Xiao N, Xu M, Yan T, Yang Y, Yang M, Young M, Yue H, Zhang B, Zhang P, Zhang Qiuting, Zhang Ya, Zhang T, Zhang Qian, Zhang W, Zhang Yu, Zhou H, Zhou J, Wen X, Curtis TP, He Q, He Z, Brown MR, Zhang T, He Z, Keller J, Nielsen PH, Alvarez PJJ, Criddle CS, Wagner M, Tiedje JM, He Q, Curtis TP, Stahl DA, Alvarez-Cohen L, Rittmann BE, Wen X, Zhou J, Consortium GWM. 2019. Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol* **4**:1183–1195. doi:10.1038/s41564-019-0426-5

Yurgel SN, Douglas GM, Comeau AM, Mammoliti M, Dusault A, Percival D, Langille MGI. 2017. Variation in Bacterial and Eukaryotic Communities Associated with Natural and Managed Wild Blueberry Habitats. *Phytobiomes J* **1**:102–113. doi:10.1094/PBIOMES-03-17-0012-R

Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F, Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P, Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol Syst Biol* **10**:766. doi:https://doi.org/10.15252/msb.20145645

Zhu L, Baker SS, Gill C, Liu W, Alkhouri R, Baker RD, Gill SR. 2013. Characterization of gut microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between endogenous alcohol and NASH. *Hepatology* **57**:601–609. doi:https://doi.org/10.1002/hep.26093