

1 Microbiome differential abundance methods produce disturbingly different results 2 across 38 datasets

3
4

5 Jacob T. Nearing^{1,*}, Gavin M. Douglas^{1,*}, Molly Hayes², Jocelyn MacDonald³, Dhwani Desai⁴,
6 Nicole Allward⁵, Casey M. A. Jones⁶, Robyn Wright⁶, Akhilesh Dhanani⁴, André M. Comeau⁴,
7 Morgan G. I. Langille^{4,6}

8

9 ¹Department of Microbiology and Immunology, Dalhousie University

10 ²Department of Mathematics and Statistics, Dalhousie University

11 ³Department of Computer Science, Dalhousie University

12 ⁴Integrated Microbiome Resource, Dalhousie University

13 ⁵Department of Civil and Resource Engineering, Dalhousie University

14 ⁶Department of Pharmacology, Dalhousie University, Halifax, Nova Scotia, Canada

15

16 Email for correspondence: jacob.nearing@dal.ca

17

18 *These authors contributed equally.

19

20 **Abstract**

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

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

44

45 **Introduction**

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

54 One area of disagreement is whether read count tables should be rarefied (i.e.,
55 subsampled) to correct for differing read depths across samples (Weiss et al., 2017). This
56 approach has been heavily criticized because excluding data could reduce statistical power and
57 introduce biases. In particular, using rarefied count tables for standard tests, such as the t-test and
58 Wilcoxon test, can result in unacceptably high false positive rates (McMurdie and Holmes,
59 2014). Nonetheless, microbiome data is still frequently rarefied because it can simplify analyses,
60 particularly for methods that do not control for variation in read depth across samples. For
61 example, LefSe (Segata et al., 2011) is a popular method for identifying differentially abundant
62 taxa that first converts read counts to percentages. Accordingly, read count tables are often
63 rarefied before being input into this tool so that variation in sample read depth does not bias
64 analyses. Without addressing the variation in depth across samples by some approach, the
65 richness can drastically differ between samples due to read depth alone.

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

77 Another contentious area is regarding which statistical distributions are most appropriate
78 for analyzing microbiome data. Statistical frameworks based on a range of distributions have
79 been developed for modelling read count data. For example, DESeq2 (Love et al., 2014) and
80 edgeR (Robinson and Oshlack, 2010) are both tools that assume normalized read counts follow a

81 negative binomial distribution. To identify differentially abundant taxa, a null and alternative
82 hypothesis are compared for each taxon. The null hypothesis is that the same parameters for the
83 negative binomial solution explain the distribution of taxa across all sample groupings. The
84 alternative hypothesis is that different parameters are needed to account for differences between
85 sample groupings. If the null hypothesis can be rejected for a specific taxon then it is considered
86 differentially abundant. This idea is the foundation of distribution-based DA tests, including
87 other methods such as corncob (Martin et al., 2020) and metagenomeSeq (Paulson et al., 2013),
88 which model microbiome data with the beta-binomial and zero-inflated Gaussian distributions,
89 respectively.

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

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

117 Although these general observations have been well substantiated, there is less agreement
118 regarding the performance of tools across evaluation studies. Certain observations have been
119 reproducible, such as the higher FDR of edgeR and metagenomeSeq. Similarly, ALDEx2 has
120 been repeatedly shown to have low power to detect differences (Calgaro et al., 2020; Hawinkel

121 et al., 2019). In contrast, both ANCOM and limma voom (Law et al., 2014; Ritchie et al., 2015)
122 have been implicated as both accurately and poorly controlling the FDR, depending on the study
123 (Calgaro et al., 2020; Hawinkel et al., 2019; Weiss et al., 2017). To further complicate
124 comparisons, different sets of tools and dataset types have been analyzed across evaluation
125 studies. This means that, on some occasions, the best performing method in one evaluation is
126 missing from another. In addition, certain popular microbiome-specific methods, such as
127 MaAsLin2 (Mallick et al., 2021), have been missing from past evaluations. Finally, many
128 evaluations limit their analysis to a small number of datasets that do not represent the breadth of
129 datasets found in 16S rRNA gene sequencing studies.

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

142

143 **Methods**

144

145 Code and data availability

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

152

153 Dataset processing

154 Thirty-eight different datasets were included in our main analyses for assessing the
155 characteristics of microbiome differential abundance tools. Two additional datasets were also
156 included for a comparison of differential abundance consistency across diarrhea-related
157 microbiome datasets. All datasets presented herein have been previously published or are
158 publicly available (Alkanani et al., 2015; Baxter et al., 2016; Chase et al., 2016; De Tender et al.,
159 2015; Dinh et al., 2015; Douglas et al., 2018; Dranse et al., 2018; Duvallet et al., 2017; Frère et
160 al., 2018; Gonzalez et al., 2018; Goodrich et al., 2014; Hoeslein et al., 2017; Ji et al., 2015; Kesy

161 et al., 2019; Lamoureux et al., 2017; Lozupone et al., 2013; McCormick et al., 2016; Mejía-León
162 et al., 2014; Nearing et al., 2019; Noguera-Julian et al., 2016; Oberbeckmann et al., 2016;
163 Oliveira et al., 2018; Papa et al., 2012; Pop et al., 2014; Rosato et al., 2020; Ross et al., 2015;
164 Scheperjans et al., 2015; Scher et al., 2013; Schneider et al., 2017; Schubert et al., 2014; Singh et
165 al., 2015; Son et al., 2015; Turnbaugh et al., 2009; Vincent et al., 2013; Wu et al., 2019; Yurgel
166 et al., 2017; Zeller et al., 2014; Zhu et al., 2013) (**Supp. Table 1**). Most datasets were already
167 available in table format with ASV or operational taxonomic unit abundances while a minority
168 needed to be processed from raw sequences. These raw sequences were processed with QIIME 2
169 version 2019.7 (Bolyen et al., 2018) based on the Microbiome Helper standard operating
170 procedure (Comeau et al., 2017). Primers were removed using cutadapt (Martin, 2011) and
171 stitched together using the QIIME 2 VSEARCH (Rognes et al., 2016) *join-pairs* plugin. Stitched
172 reads were then quality filtered using the *quality-filter* plugin and reads were denoised using
173 Deblur (Amir et al., 2017) to produce amplicon sequence variants (ASVs). Abundance tables of
174 ASVs for each sample were then output into tab-delimited files. Rarefied tables were also
175 produced for each dataset, where the rarefied read depth was taken to be the lowest read depth of
176 any sample in the dataset over 2000 reads (with samples below this threshold discarded).

177

178 Differential abundance testing

179 We created a custom shell script (`run_all_tools.sh`) that ran each differential abundance tool on
180 each dataset within this study. As input the script took a tab-delimited ASV abundance table, a
181 rarefied version of that same table, and a metadata file that contained a column that split the
182 samples into two groups for testing. This script also accepted a prevalence cut-off filter to
183 remove ASVs below a minimum cut-off, which was set to 10% (i.e., ASVs found in fewer than
184 10% of samples were removed) for the filtered data analyses we present. Note that in a minority
185 of cases a genus abundance table was input instead, in which case all options were kept the same.
186 When the prevalence filter option was set, the script also generated new filtered rarefied tables
187 based on an input rarefaction depth.

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

195

196 *ALDEx2*

197 We passed the non-rarefied feature table and the corresponding sample metadata to the *aldex*
198 function from the ALDEx2 R package (Fernandes et al., 2014) which generated Monte Carlo
199 samples of Dirichlet distribution for each sample, using a uniform prior, performed CLR
200 transformation of each realization, and performed Wilcoxon tests on the transformed

201 realizations. The function then returned the expected Benjamini-Hochberg (BH) FDR-corrected
202 P-value was then returned for each feature based on the results across Monte Carlo samples.

203

204 *ANCOM-II*

205 We ran the non-rarefied feature table through the R ANCOM-II (Kaul et al., 2017; Mandal et al.,
206 2015) (<https://github.com/FrederickHuangLin/ANCOM>) function *feature_table_pre_process*,
207 which first examined the abundance table to identify outlier zeros and structural zeros (Kaul et
208 al., 2017). Outlier zeros, identified by finding outliers in the distribution of taxon counts within
209 each sample grouping, were ignored during differential abundance analysis and replaced with
210 NA. Structural zeros, taxa that were absent in one grouping but present in the other, were ignored
211 during data analysis and automatically called as differentially abundant. A pseudo count of 1 was
212 then applied across the dataset to allow for log transformation. Using the main function *ANCOM*,
213 all additive log-ratios for each taxon were then tested for significance using Wilcoxon rank-sum
214 tests, and p-values were FDR-corrected using the BH method. ANCOM then applied a detection
215 threshold as described in the original paper (Mandal et al., 2015), whereby a taxon was called as
216 DA if the number of corrected p-values reaching nominal significance for that taxon was greater
217 than 90% of the maximum possible number of significant comparisons.

218

219 *corncob*

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

229

230 *DESeq2*

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

240 hypothesis testing, for which we chose the default Wald test. Finally, using the *results* function,
241 we obtained the resulting BH FDR-corrected p-values.

242

243 *edgeR*

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

253

254 *LEfSe*

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

265

266 *limma voom*

267 We first normalized the non-rarefied feature table using the edgeR *calcNormFactors* function,
268 with either the trimmed mean of M-values (TMM) or TMM with singleton pairing (TMMwsp)
269 option. During this step (for both options), a single sample was chosen to be a reference sample
270 using upper-quartile normalization. This step failed in some highly sparse abundance tables, in
271 which cases, we instead chose the sample with the largest sum of square-root transformed feature
272 abundances to be the reference sample. After normalization, we used the limma R package
273 (Ritchie et al., 2015) function *voom* to convert normalized counts to log₂-counts-per-million and
274 assign precision weights to each observation based on the mean-variance trend. We then used the
275 functions *lmFit*, *eBayes*, and *topTable* to fit weighted linear regression models, perform tests
276 based on an empirical Bayes moderated t-statistic (Phipson et al., 2016) and obtain BH FDR-
277 corrected p-values.

278

279 *MaAsLin2*

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

287

288 *metagenomeSeq*

289 We first entered the counts and sample information to the function *newMRexperiment* from the
290 *metagenomeSeq* R package (Paulson et al., 2013). Next, we used *cumNormStat* and *cumNorm* to
291 apply cumulative sum-scaling normalization, which attempts to normalize sequence counts based
292 on the lower-quartile abundance of features. We then used *fitFeatureModel* to fit normalized
293 feature counts with zero-inflated log-normal models (with pseudo-counts of 1 added prior to log₂
294 transformation) and perform empirical Bayes moderated t-tests, and *MRfulltable* to obtain BH
295 FDR-corrected p-values.

296

297 *t-test*

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

301

302 *Wilcoxon test*

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

307

308 Comparing numbers of significant hits between tools

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

318

319 Cross-tool, within-study differential abundance consistency analysis

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

323

324 False positive analysis

325 To estimate the false positives a method might produce during data analysis, eight datasets were
326 selected for analysis. These datasets were chosen based on having the largest sample sizes, while
327 also being from diverse environment types. In each dataset, the most frequent sample group was
328 chosen for analysis to help ensure similar composition among samples tested. Within this
329 grouping, random labels of either case or control were assigned to samples and the various
330 differential abundance methods were tested on them. This was repeated 10 times for each dataset
331 and each tool tested with an additional 90 replications for all tools except for ALDEx2,
332 ANCOM-II and corncob due to high resource requirements. After analysis was completed, the
333 number of differentially abundant ASVs identified by each tool was assessed at an alpha value of
334 0.05.

335

336 Cross-study differential abundance consistency analysis

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

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

358

359

360 **Results**

361

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

364

365 To investigate how different DA tools impact biological interpretations across microbiome
366 datasets, we tested 14 different differential abundance testing approaches (**Table 1**) on 38
367 different microbiome datasets. These datasets corresponded to a range of environments,
368 including the human gut, plastisphere, freshwater lakes, and urban environments (**Supp. Table**
369 **1**). The features in these datasets corresponded to both ASVs and clustered operational
370 taxonomic units, but we refer to them all as ASVs below for simplicity.

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

374 We found that in both the filtered and unfiltered analyses the percentage of significant
375 ASVs identified by each DA method varied widely across datasets, with means ranging from
376 3.8-32.5% and 0.8-40.5%, respectively. Interestingly, we found that many tools behaved
377 differently between datasets. Specifically, some tools identified the most features in one dataset
378 while identifying only an intermediate number in other datasets. This was especially evident in
379 the unfiltered datasets (**Fig. 1A**).

380 To investigate possible factors driving this variation we examined how the number of
381 ASVs identified by each tool correlated with several variables. These variables included dataset
382 richness, variation in sequencing depth between samples, dataset sparsity, and Aitchison's
383 distance effect size (based on PERMANOVA tests). As expected, we found that all tools
384 positively correlated with the effect size between test groups with rho values ranging between
385 0.35-0.72 with unfiltered data (**Fig. 2A**) and 0.31-0.52 for filtered data (**Fig. 2B**). We also found
386 in the filtered datasets that the number of features found by all tools significantly correlated with
387 the median read depth, range in read depth, and sample size. There was much less consistency in
388 these correlations across the unfiltered data. For instance, only the t-test, both Wilcoxon
389 methods, and both limma voom methods correlated significantly with the range in read depth
390 (**Fig. 2B**). We also found that edgeR was negatively correlated with mean sample richness in the
391 unfiltered analysis.

392

393

394

395

396

397

398

399

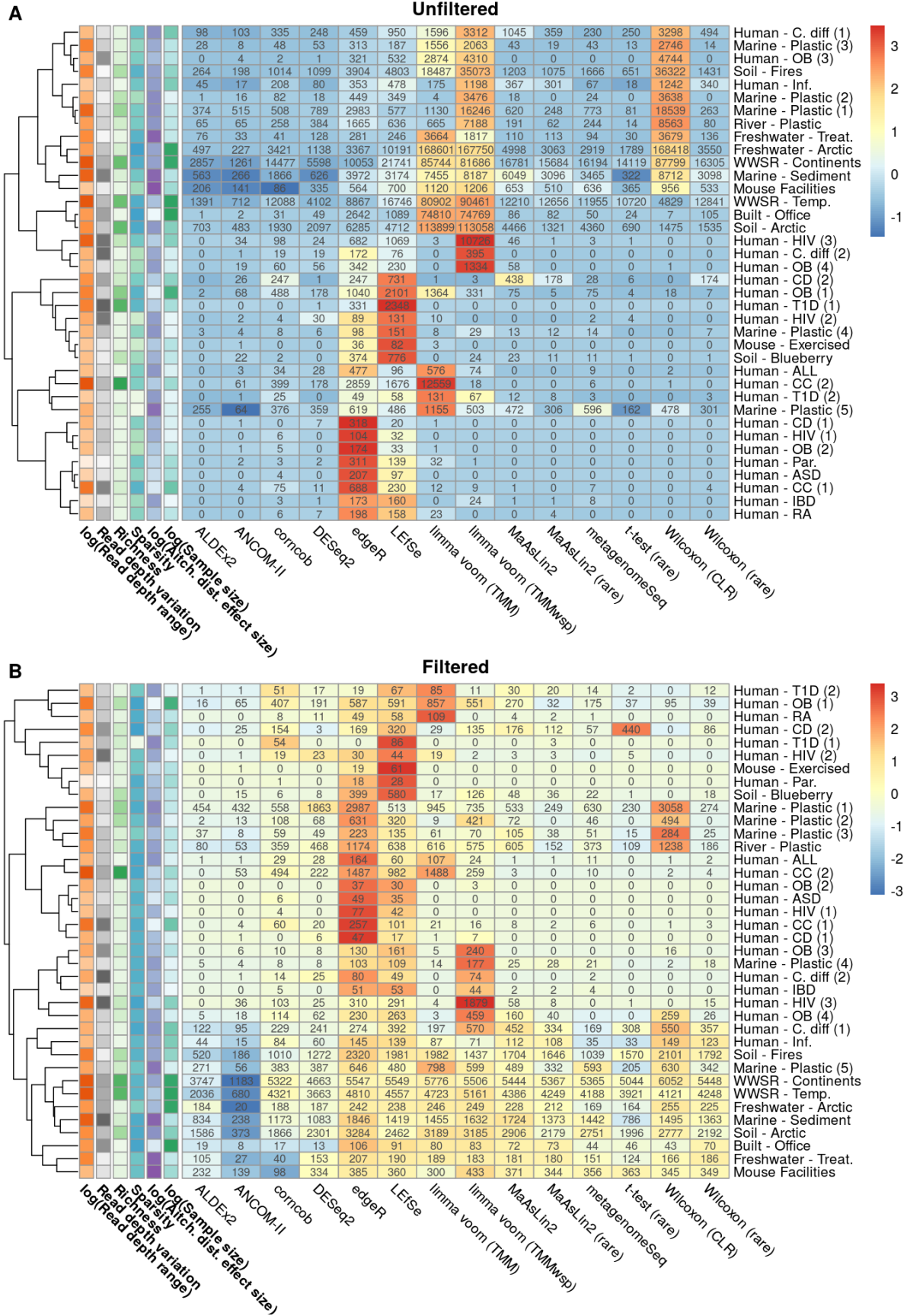
400 **Table 1: Differential abundance tools compared in this study**

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

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

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

408
 409



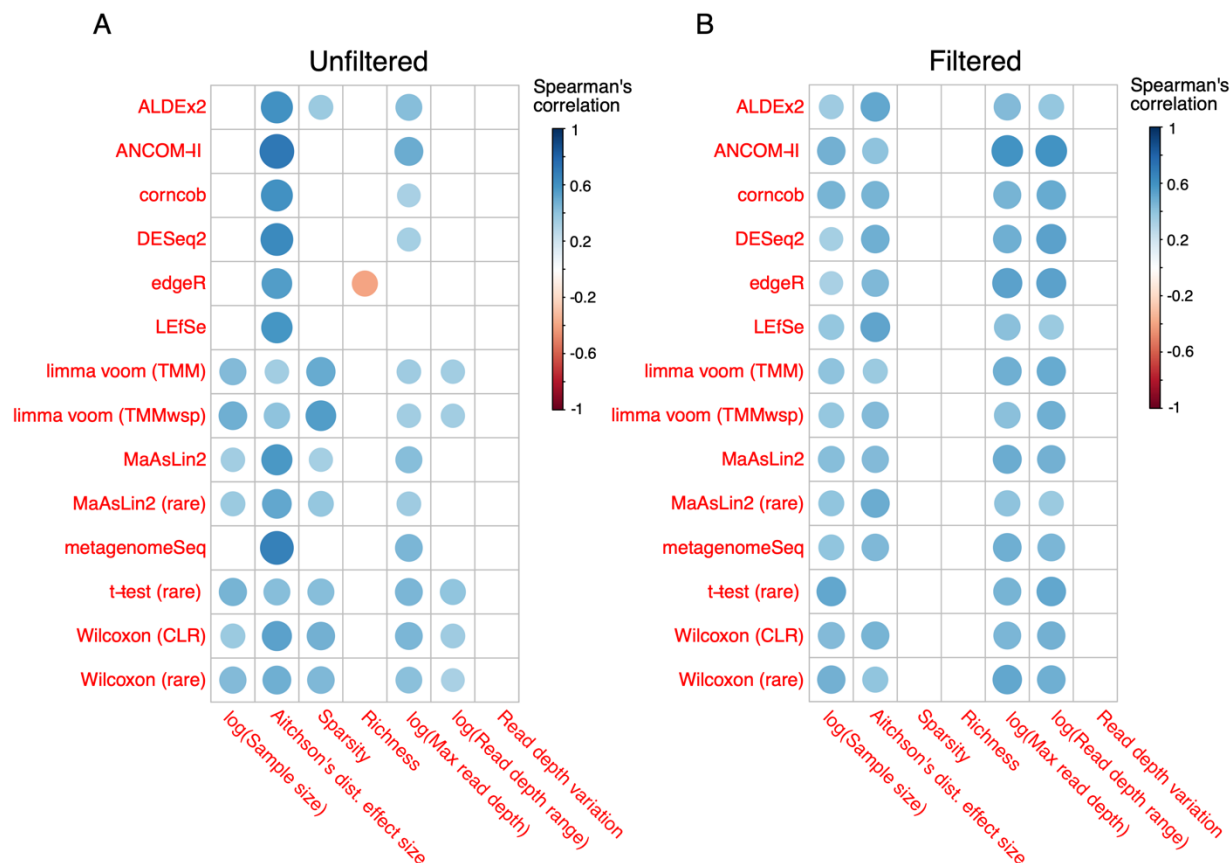
411 **Figure 1: Variation in the proportion of significant features depending on the differential**
412 **abundance method and dataset.** Heatmaps indicate the numbers of significant amplicon
413 sequence variants (ASVs) identified in each dataset by the corresponding tool based on (A)
414 unfiltered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

420
421 Despite the variability of tool performance between datasets, we did find that several
422 tools tended to identify more significant hits (**Supp. Fig 1C-D**). In the unfiltered datasets, we
423 found that limma voom (TMMwsp; mean: 40.5% / TMM; mean: 29.7%), Wilcoxon (CLR;
424 mean: 30.7%), LefSe (mean: 12.6%), and edgeR (mean: 12.4%) tended to find the largest
425 number of significant ASVs compared with other methods. Interestingly, in a few datasets, such
426 as the Human-ASD and Human-OB (2) datasets, edgeR found a significantly higher proportion
427 of significant ASVs than any other tool. In addition, we found that limma voom (TMMwsp)
428 found the majority of ASVs to be significant (73.5%) in the Human-HIV (3) dataset while the
429 other tools found 0-11% ASVs to be significant (**Fig. 1A**). Similarly, we found that both limma
430 voom methods identified over 99% of ASVs to be significant in several cases such as the Built-
431 Office and Freshwater-Arctic datasets. We found similar, although not as extreme, trends with
432 LefSe where in some datasets, such as the Human-T1D (1) dataset, the tool found a much higher
433 percentage of significant hits (3.5%) compared with all other tools (0-0.4%). This observation is
434 most likely a result of LefSe filtering significant features by effect size rather than using FDR
435 correction to reduce the number of false positives. We found that two of the three
436 compositionally aware methods we tested identified fewer significant ASVs than the other tools
437 tested. Specifically, ALDEx2 (mean: 1.4%) and ANCOM-II (mean: 0.8%) identified the fewest
438 significant ASVs. We found the conservative behavior of these tools to be consistent across all
439 38 datasets we tested.

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

447 Finally, we examined the mean relative abundance of the features identified by each tool
448 to determine whether tools may be biased toward the identification of highly abundant features.
449 We found that both ALDEx2 (median: 0.013%), ANCOM-II (median: 0.024%) and to a lesser
450 degree DESeq2 (median: 0.006%) tended to find significant features that were higher in relative

451 abundance in the unfiltered datasets. A similar trend for ALDEx2 (median: 0.011%) and
 452 ANCOM-II (median: 0.029%) was also found in the filtered datasets (**Supp. Fig 1A-B**).



453
 454 **Figure 2: Dataset characteristics associated with percentage of significant amplicon**
 455 **sequence variants.** The correlation coefficients (Spearman's rho) are displayed by size and
 456 color. These correspond to the dataset characteristics correlated with the percentage of
 457 significant amplicon sequence variants identified by that tool per dataset. Only significant
 458 correlations ($p < 0.05$) are displayed.

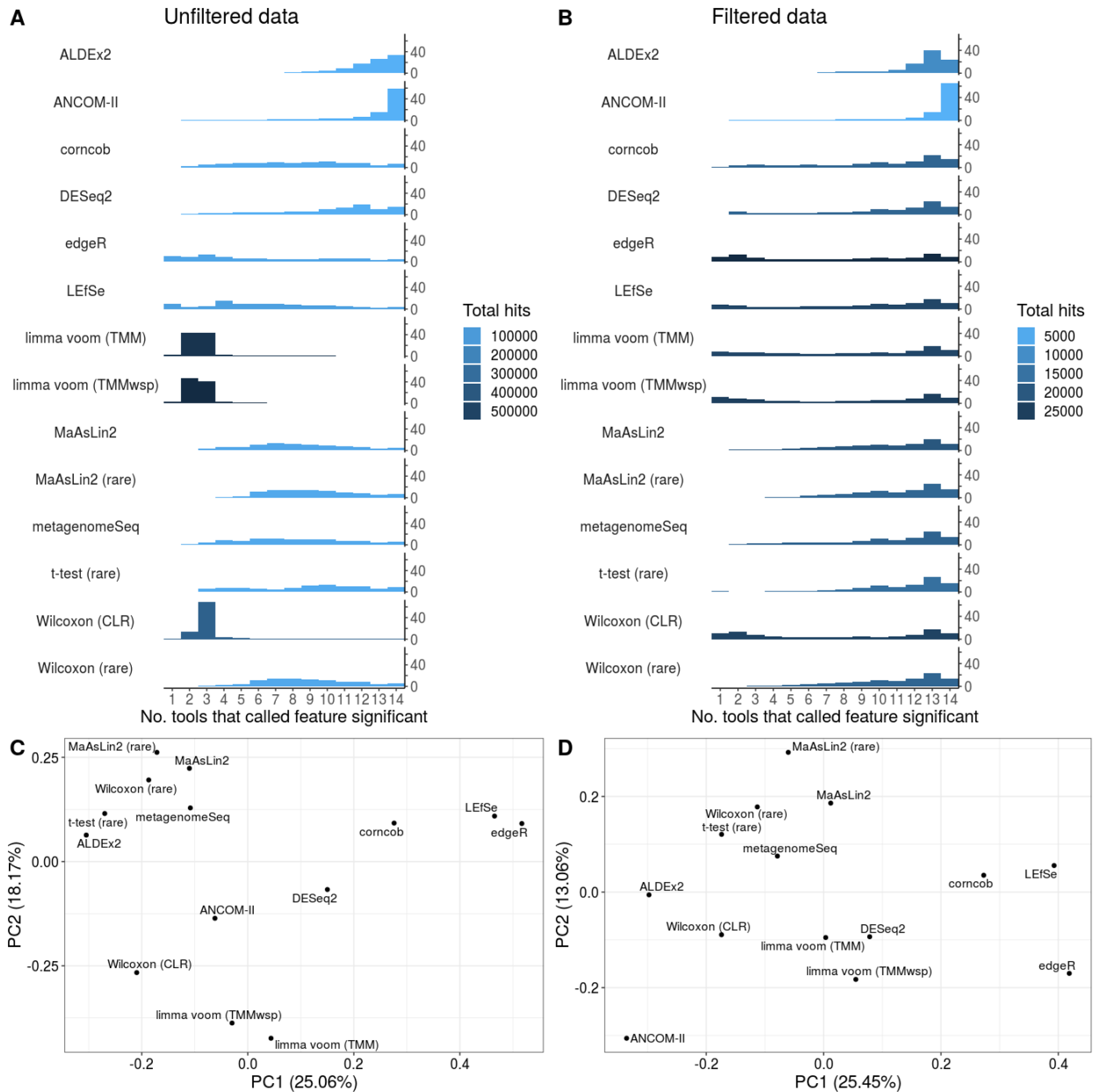
459
 460 **High variability of overlapping significant ASVs**

461 We next investigated the overlap in significant ASVs across tools within each dataset. Based on
 462 the unfiltered data, we found that both limma voom methods identified similar sets of significant
 463 ASVs that were different from those of most other tools (**Fig. 3A**). However, we also found that
 464 many of the ASVs identified by the limma voom methods were also identified as significant
 465 based on the Wilcoxon (CLR) approach, despite these being highly methodologically distinct
 466 tools. Furthermore, the two Wilcoxon test approaches had different consistency profiles despite
 467 using the same hypothesis test. In contrast, we found that both MaAsLin2 approaches had similar
 468 consistency profiles, although the non-rarefied method found slightly lower-ranked features. We
 469 also found that the most conservative tools, ALDEx2 and ANCOM-II, primarily identified
 470 features that were also identified by almost all other methods. In contrast, edgeR and LefSe, two

471 tools that often identified the most significant ASVs, output the highest percentage of ASVs that
472 were not identified by any other tool: 12.1% and 11.1%, respectively. Corncob, metagenomeSeq,
473 and DESeq2 identified ASVs at more intermediate consistency profiles.

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

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



502
503
504
505
506
507
508
509
510
511
512
513

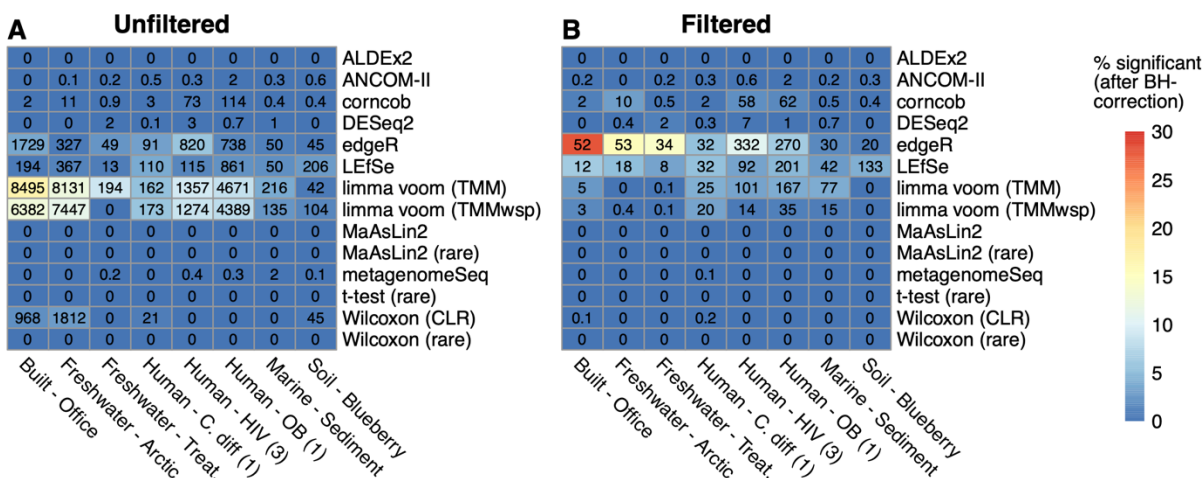
Figure 3: Overlap of significant features across tools and tool clustering. (A and B) The number of tools that called each feature significant, stratified by features called by each individual tool for the (A) unfiltered and (B) 10% prevalence-filtered data. The features correspond to the amplicon sequence 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.

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

515 We next investigated the FDR of each DA tool across eight datasets. For each dataset we
516 selected the most frequently sampled group and randomly reassigned them as case or control
517 samples. Each DA tool was then run on those samples and results were compared. This was
518 repeated a total of 10 times for each filtered dataset and 100 times for each unfiltered dataset
519 (except for corncob, ANCOM-II and ALDEx2). We used a higher number of replicates for the
520 unfiltered datasets because they were less stable across replicates. The percentage of significant
521 ASVs after BH multiple-test correction was relatively low for most tested tools on the filtered
522 data (**Fig. 4B**). Two outliers were edgeR (mean: 10.3%; SD: 9.0%) and LEfSe (mean: 4.4%; SD:
523 1.2%), which consistently identified more significant hits compared with other tools (range of
524 other tool means: 0% - 2.2%). Both limma voom methods output highly variable percentages of
525 significant ASVs, especially based on the unfiltered data (**Fig. 4A**). In particular, in 5/8 of the
526 unfiltered datasets, the limma voom methods identified more than 5% of ASVs as significant on
527 average. Interestingly, while these two methods exhibited similar performance overall, the
528 performance within the unfiltered Freshwater-Treatment dataset was highly different between the
529 methods with the TMMwsp method identifying 0.001% and the TMM method identifying 9.0%.
530 Only ALDEx2 and the t-test (rare) approach consistently identified no ASVs as significantly
531 different in this analysis.

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

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



546
547 **Figure 4: Variation in false discovery rate across the tested differential abundance tools in**
548 **the context of simple simulations.** Heatmap of the percentage and number of significant
549 amplicon sequence variants (ASVs) identified by each tool in eight simulation datasets based on
550 applying (A) no prevalence filter and (B) a 10% prevalence filter. Cell colours indicate the
551 percentage and the number of significant ASVs is written in each cell. Mean numbers higher
552 than one were rounded to the nearest integer for visualization. Significant ASVs were identified
553 after applying the Benjamini-Hochberg (BH) false discovery rate procedure and then using a cut-
554 off of 0.05.

555

556

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

559 Separate from the above analysis comparing consistency between tools on the same dataset, we
560 next investigated whether certain tools provide more consistent signals across datasets of the
561 same disease. This analysis focused on the genus-level across tools to help limit inter-study
562 variation. We specifically focused on diarrhea as a phenotype, which has been shown to exhibit a
563 strong effect on the microbiome and to be relatively reproducible across studies (Duvall et al.,
564 2017).

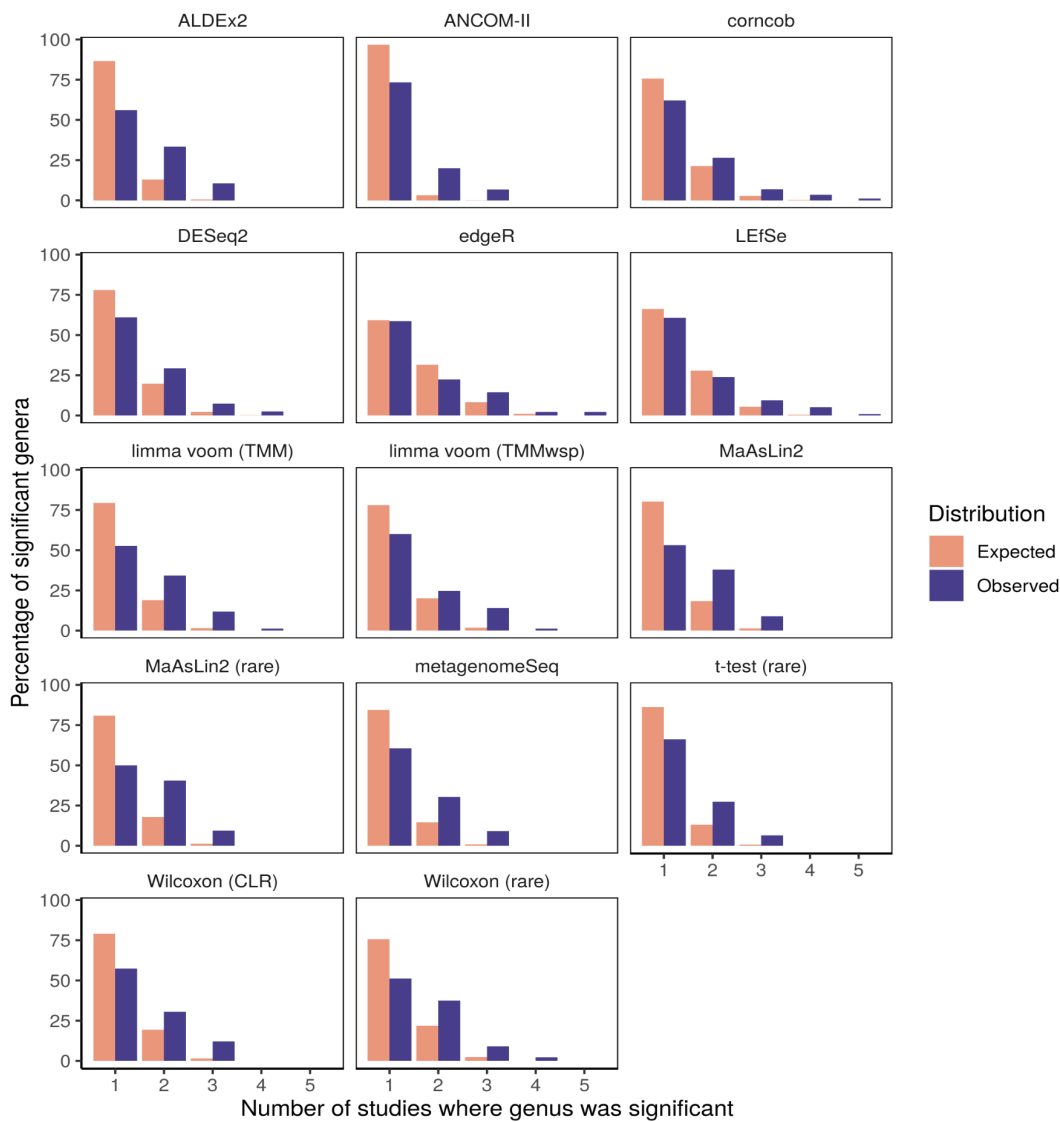
565 We acquired five datasets for this analysis representing the microbiome of individuals
566 with diarrhea compared with individuals without diarrhea (see Methods). We ran all DA tools on
567 each individual filtered dataset. Similar to our ASV-level analyses, the tools substantially varied
568 in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

574 Instead, we analyzed the observed distribution of the number of studies that each genus
575 was identified as significant in compared with the expected distribution given random data. This

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

585 ALDEx2 and edgeR represent the extremes of how consistently tools identify the same
586 genera as significant across studies, but there is a large range (**Fig. 5**). Notably, all tools were
587 significantly more consistent than the random expectation across these datasets ($p < 0.05$) (**Table**
588 **2**). In addition to ALDEx2, the other top performing approaches based on this evaluation
589 included both MaAsLin2 workflows, limma voom (TMM), and ANCOM-II.

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



596
597

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

603
604
605
606
607
608

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

611

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

612

613 **Column descriptions:**

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

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

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

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

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

619 **D:** Kolmogorov-Smirnov test statistic

620

621

622

623

624

625

626

627

628

629

630

631

632 **Discussion**

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

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

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

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

669 Two additional particularly problematic tools based on this analysis were edgeR and
670 LefSe. The edgeR method has been previously found to exhibit a high FDR on several occasions
671 (Hawinkel et al., 2019; Thorsen et al., 2016) Although metagenomeSeq also has been flagged as

672 such (Thorsen et al., 2016), that was not the case in our analysis. This agrees with a recent report
673 that metagenomeSeq (using the zero-inflated log-normal approach, as we did) appropriately
674 controlled the FDR, but exhibited low power (Lin and Peddada, 2020). There have been mixed
675 results previously regarding whether ANCOM appropriately controls the FDR (Hawinkel et al.,
676 2019; Weiss et al., 2017), but the results from our limited analysis suggest that this method is
677 conservative and controls the FDR while potentially missing true positives.

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

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

697 We believe the above observations regarding DA tools are valuable, but many readers are
698 likely primarily interested in hearing specific recommendations. Indeed, the need for
699 standardized practices in microbiome analysis have recently become better appreciated (Hill,
700 2020). One goal of our work was to validate the recommendations of another recent DA method
701 evaluation paper, which found that limma voom, corncob, and DESeq2 performed best overall
702 (Calgaro et al., 2020). Based on our results we do not recommend these tools as the sole methods
703 used for data analysis, and instead would suggest using more conservative methods such as
704 ALDEx2 and ANCOM-II. Although these methods have lower statistical power (Calgaro et al.,
705 2020; Hawinkel et al., 2019), we believe this an acceptable trade-off given the higher cost of
706 identifying false positives as differentially abundant. However, MaAsLin2 (particularly with
707 rarefied data) could also be a reasonable choice for users looking for increased statistical power
708 at the potential cost of more false positives. We can clearly recommend that users avoid using
709 edgeR (a tool primarily intended for RNA-seq data) as well as LefSe for conducting DA testing
710 with 16S data. Users should also be aware that limma voom and the Wilcoxon (CLR) approaches

711 may perform especially poorly on unfiltered data. This is especially true for the Wilcoxon (CLR)
712 approach when read depths greatly differ between groups of interest.

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

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

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

743 Another common question is whether microbiome data should be rarefied prior to DA
744 testing. It is possible that the question of whether to rarefy data has received disproportionate
745 attention in the microbiome field: there are numerous other factors affecting an analysis pipeline
746 that likely affect results more. Indeed, tests based on rarefied data in our analyses did not
747 perform substantially worse than other methods on average. More specifically, the most
748 consistent inter-tool methods, ANCOM-II and ALDEx2, are based on non-rarefied data, but
749 MaAsLin2 based on rarefied data produced the most consistent results across datasets of the
750 same phenotype. Accordingly, we cannot definitively conclude that rarefying data prior to DA

751 testing is always inadvisable. It should be noted that we are referring only to rarefying in the
752 context of DA testing: whether rarefying is advisable for other analyses, such as prior to
753 computing diversity metrics, is beyond the scope of this work (McMurdie and Holmes, 2014;
754 Weiss et al., 2017).

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

762

763 **Acknowledgements**

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

771

772 **Competing interests**

773 The authors declare that they have no competing interests.

774

775 **References**

776

- 777 Aitchison J. 1982. The Statistical Analysis of Compositional Data. *J R Stat Soc Ser B* **44**:139–
778 160. doi:<https://doi.org/10.1111/j.2517-6161.1982.tb01195.x>
- 779 Alkanani AK, Hara N, Gottlieb PA, Ir D, Robertson CE, Wagner BD, Frank DN, Zipris D. 2015.
780 Alterations in Intestinal Microbiota Correlate With Susceptibility to Type 1 Diabetes.
781 *Diabetes* **64**:3510 LP – 3520. doi:10.2337/db14-1847
- 782 Allaband C, McDonald D, Vázquez-Baeza Y, Minich JJ, Tripathi A, Brenner DA, Loomba R,
783 Smarr L, Sandborn WJ, Schnabl B, Dorrestein P, Zarrinpar A, Knight R. 2019. Microbiome
784 101: Studying, Analyzing, and Interpreting Gut Microbiome Data for Clinicians. *Clin*
785 *Gastroenterol Hepatol* **17**:218–230. doi:10.1016/j.cgh.2018.09.017
- 786 Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, Kightley EP,
787 Thompson LR, Hyde ER, Gonzalez A, Knight R. 2017. Deblur Rapidly Resolves Single-
788 Nucleotide Community Sequence Patterns. *mSystems* **2**.
- 789 Baxter NT, Ruffin MT, Rogers MAM, Schloss PD. 2016. Microbiota-based model improves the
790 sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med* **8**:37.
791 doi:10.1186/s13073-016-0290-3
- 792 Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm
793 EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ,

- 794 Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope E, Da Silva R,
795 Dorrestein PC, Douglas GM, Durall DM, Duvall C, Edwardson CF, Ernst M, Estaki M,
796 Fouquier J, Gauglitz JM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes
797 S, Holste H, Huttenhower C, Huttley G, Janssen S, Jarmusch AK, Jiang L, Kaehler B, Kang
798 K Bin, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolk T, Kreps J, Langille
799 MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD,
800 McDonald D, McIver LJ, Melnik A V, Metcalf JL, Morgan SC, Morton J, Naimey AT,
801 Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML,
802 Pruesse E, Rasmussen LB, Rivers A, Robeson Michael S II, Rosenthal P, Segata N, Shaffer
803 M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P,
804 Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y,
805 Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC,
806 Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R, Caporaso
807 JG. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data
808 science. *PeerJ Prepr* 6:e27295v2. doi:10.7287/peerj.preprints.27295v2
- 809 Bourgon R, Gentleman R, Huber W. 2010. Independent filtering increases detection power for
810 high-throughput experiments. *Proc Natl Acad Sci* 107:9546 LP – 9551.
811 doi:10.1073/pnas.0914005107
- 812 Calgario M, Romualdi C, Waldron L, Risso D, Vitulo N. 2020. Assessment of statistical methods
813 from single cell, bulk RNA-seq, and metagenomics applied to microbiome data. *Genome*
814 *Biol* 21:191. doi:10.1186/s13059-020-02104-1
- 815 Chase J, Fouquier J, Zare M, Sonderegger DL, Knight R, Kelley ST, Siegel J, Caporaso JG.
816 2016. Geography and Location Are the Primary Drivers of Office Microbiome
817 Composition. *mSystems* 1:e00022-16. doi:10.1128/mSystems.00022-16
- 818 Comeau AM, Douglas GM, Langille MGI. 2017. Microbiome Helper: a Custom and Streamlined
819 Workflow for Microbiome Research. *mSystems* 2.
- 820 De Tender CA, Devriese LI, Haegeman A, Maes S, Ruttink T, Dawyndt P. 2015. Bacterial
821 Community Profiling of Plastic Litter in the Belgian Part of the North Sea. *Environ Sci*
822 *Technol* 49:9629–9638. doi:10.1021/acs.est.5b01093
- 823 Dinh DM, Volpe GE, Duffalo C, Bhalchandra S, Tai AK, Kane A V, Wanke CA, Ward HD.
824 2015. Intestinal Microbiota, Microbial Translocation, and Systemic Inflammation in
825 Chronic HIV Infection. *J Infect Dis* 211:19–27. doi:10.1093/infdis/jiu409
- 826 Dixon P. 2003. VEGAN, a package of R functions for community ecology. *J Veg Sci* 14.
827 doi:10.1111/j.1654-1103.2003.tb02228.x
- 828 Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, Tayler R, El-Omar
829 EM, Russell RK, Hold GL, Langille MGI, Van Limbergen J. 2018. Multi-omics
830 differentially classify disease state and treatment outcome in pediatric Crohn’s disease.
831 *Microbiome* 6:13. doi:10.1186/s40168-018-0398-3
- 832 Dranse HJ, Zheng A, Comeau AM, Langille MGI, Zabel BA, Sinal CJ. 2018. The impact of
833 chemerin or chemokine-like receptor 1 loss on the mouse gut microbiome. *PeerJ* 6:e5494.
834 doi:10.7717/peerj.5494
- 835 Duvall C, Gibbons SM, Gurry T, Irizarry RA, Alm EJ. 2017. Meta-analysis of gut microbiome
836 studies identifies disease-specific and shared responses. *Nat Commun* 8:1784.
837 doi:10.1038/s41467-017-01973-8
- 838 Feijen EAM, Font-Gonzalez A, van Dalen EC, van der Pal HJH, Reulen RC, Winter DL, Kuehni
839 CE, Haupt R, Alessi D, Byrne J, Bardi E, Jakab Z, Grabow D, Garwicz S, Jankovic M,

- 840 Levitt GA, Skinner R, Zdravec Zaletel L, Hjorth L, Tissing WJE, de Vathaire F, Hawkins
841 MM, Kremer LCM, consortium P. 2016. Late Cardiac Events after Childhood Cancer:
842 Methodological Aspects of the Pan-European Study PanCareSurFup. *PLoS One*
843 **11**:e0162778.
- 844 Fernandes AD, Reid JNS, Macklaim JM, McMurrugh TA, Edgell DR, Gloor GB. 2014.
845 Unifying the analysis of high-throughput sequencing datasets: characterizing RNA-seq, 16S
846 rRNA gene sequencing and selective growth experiments by compositional data analysis.
847 *Microbiome* **2**:15. doi:10.1186/2049-2618-2-15
- 848 Frère L, Maignien L, Chalopin M, Huvet A, Rinnert E, Morrison H, Kerninon S, Cassone A-L,
849 Lambert C, Reveillaud J, Paul-Pont I. 2018. Microplastic bacterial communities in the Bay
850 of Brest: Influence of polymer type and size. *Environ Pollut* **242**:614–625.
851 doi:<https://doi.org/10.1016/j.envpol.2018.07.023>
- 852 Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome Datasets Are
853 Compositional: And This Is Not Optional. *Front Microbiol* **8**:2224.
854 doi:10.3389/fmicb.2017.02224
- 855 Gonzalez A, Navas-Molina JA, Kosciolk T, McDonald D, Vázquez-Baeza Y, Ackermann G,
856 DeReus J, Janssen S, Swafford AD, Orchanian SB, Sanders JG, Shorenstein J, Holste H,
857 Petrus S, Robbins-Pianka A, Brislawn CJ, Wang M, Rideout JR, Bolyen E, Dillon M,
858 Caporaso JG, Dorrestein PC, Knight R. 2018. Qiita: rapid, web-enabled microbiome meta-
859 analysis. *Nat Methods* **15**:796–798. doi:10.1038/s41592-018-0141-9
- 860 Goodrich JK, Waters JL, Poole AC, Sutter JL, Koren O, Blekhman R, Beaumont M,
861 Van Treuren W, Knight R, Bell JT, Spector TD, Clark AG, Ley RE. 2014. Human Genetics
862 Shape the Gut Microbiome. *Cell* **159**:789–799. doi:10.1016/j.cell.2014.09.053
- 863 Hawinkel S, Mattiello F, Bijmans L, Thas O. 2019. A broken promise: microbiome differential
864 abundance methods do not control the false discovery rate. *Brief Bioinform* **20**:210–221.
865 doi:10.1093/bib/bbx104
- 866 Hill C. 2020. You have the microbiome you deserve. *Gut Microbiome* **1**:e3. doi:DOI:
867 10.1017/gmb.2020.3
- 868 Hoellein TJ, McCormick AR, Hittie J, London MG, Scott JW, Kelly JJ. 2017. Longitudinal
869 patterns of microplastic concentration and bacterial assemblages in surface and benthic
870 habitats of an urban river. *Freshw Sci* **36**:491–507. doi:10.1086/693012
- 871 Ji P, Parks J, Edwards MA, Pruden A. 2015. Impact of Water Chemistry, Pipe Material and
872 Stagnation on the Building Plumbing Microbiome. *PLoS One* **10**:e0141087.
- 873 Kaul A, Mandal S, Davidov O, Peddada SD. 2017. Analysis of Microbiome Data in the Presence
874 of Excess Zeros. *Front Microbiol* **8**:2114. doi:10.3389/fmicb.2017.02114
- 875 Kesy K, Oberbeckmann S, Kreikemeyer B, Labrenz M. 2019. Spatial Environmental
876 Heterogeneity Determines Young Biofilm Assemblages on Microplastics in Baltic Sea
877 Mesocosms . *Front Microbiol* .
- 878 Kolde R. 2012. Pheatmap: pretty heatmaps. *R Packag version* **1**.
- 879 Lamoureux E V, Grandy SA, Langille MGI. 2017. Moderate Exercise Has Limited but
880 Distinguishable Effects on the Mouse Microbiome. *mSystems* **2**:e00006-17.
881 doi:10.1128/mSystems.00006-17
- 882 Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model
883 analysis tools for RNA-seq read counts. *Genome Biol* **15**:R29. doi:10.1186/gb-2014-15-2-
884 r29
- 885 Lin H, Peddada S Das. 2020. Analysis of microbial compositions: a review of normalization and

- 886 differential abundance analysis. *NPJ biofilms microbiomes* **6**:60. doi:10.1038/s41522-020-
887 00160-w
- 888 Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for
889 RNA-seq data with DESeq2. *Genome Biol* **15**:550. doi:10.1186/s13059-014-0550-8
- 890 Lozupone CA, Li M, Campbell TB, Flores SC, Linderman D, Gebert MJ, Knight R, Fontenot
891 AP, Palmer BE. 2013. Alterations in the Gut Microbiota Associated with HIV-1 Infection.
892 *Cell Host Microbe* **14**:329–339. doi:10.1016/j.chom.2013.08.006
- 893 Mallick H, Rahnavard A, McIver LJ, Ma S, Zhang Y, Nguyen LH, Tickle TL, Weingart G, Ren
894 B, Schwager EH, Chatterjee S, Thompson KN, Wilkinson JE, Subramanian A, Lu Y,
895 Waldron L, Paulson JN, Franzosa EA, Bravo HC, Huttenhower C. 2021. Multivariable
896 Association Discovery in Population-scale Meta-omics Studies. *bioRxiv*
897 2021.01.20.427420. doi:10.1101/2021.01.20.427420
- 898 Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. 2015. Analysis of
899 composition of microbiomes: a novel method for studying microbial composition. *Microb*
900 *Ecol Health Dis* **26**:27663. doi:10.3402/mehd.v26.27663
- 901 Martin BD, Witten D, Willis AD. 2020. Modeling microbial abundances and dysbiosis with
902 beta-binomial regression. *Ann Appl Stat* **14**:94–115. doi:10.1214/19-AOAS1283
- 903 Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads.
904 *EMBnet.journal; Vol 17, No 1 Next Gener Seq Data Anal.*
- 905 McCormick AR, Hoellein TJ, London MG, Hittie J, Scott JW, Kelly JJ. 2016. Microplastic in
906 surface waters of urban rivers: concentration, sources, and associated bacterial assemblages.
907 *Ecosphere* **7**:e01556. doi:https://doi.org/10.1002/ecs2.1556
- 908 McMurdie PJ, Holmes S. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is
909 Inadmissible. *PLOS Comput Biol* **10**:e1003531.
- 910 Mejía-León ME, Petrosino JF, Ajami NJ, Domínguez-Bello MG, de la Barca AMC. 2014. Fecal
911 microbiota imbalance in Mexican children with type 1 diabetes. *Sci Rep* **4**:3814.
912 doi:10.1038/srep03814
- 913 Morton JT, Marotz C, Washburne A, Silverman J, Zaramela LS, Edlund A, Zengler K, Knight R.
914 2019. Establishing microbial composition measurement standards with reference frames.
915 *Nat Commun* **10**:2719. doi:10.1038/s41467-019-10656-5
- 916 Nearing JT, Connors J, Whitehouse S, Van Limbergen J, Macdonald T, Kulkarni K, Langille
917 MGI. 2019. Infectious Complications Are Associated With Alterations in the Gut
918 Microbiome in Pediatric Patients With Acute Lymphoblastic Leukemia. *Front Cell Infect*
919 *Microbiol* **9**:28. doi:10.3389/fcimb.2019.00028
- 920 Noguera-Julian M, Rocafort M, Guillén Y, Rivera J, Casadellà M, Nowak P, Hildebrand F,
921 Zeller G, Parera M, Bellido R, Rodríguez C, Carrillo J, Mothe B, Coll J, Bravo I, Estany C,
922 Herrero C, Saz J, Sirera G, Torrela A, Navarro J, Crespo M, Brander C, Negro E, Blanco
923 J, Guarner F, Calle ML, Bork P, Sönnnerborg A, Clotet B, Paredes R. 2016. Gut Microbiota
924 Linked to Sexual Preference and HIV Infection. *EBioMedicine* **5**:135–146.
925 doi:10.1016/j.ebiom.2016.01.032
- 926 Oberbeckmann S, Osborn AM, Duhaim MB. 2016. Microbes on a Bottle: Substrate, Season and
927 Geography Influence Community Composition of Microbes Colonizing Marine Plastic
928 Debris. *PLoS One* **11**:e0159289.
- 929 Oliveira FS, Brestelli J, Cade S, Zheng J, Iodice J, Fischer S, Aurrecochea C, Kissinger JC,
930 Brunk BP, Stoeckert Jr CJ, Fernandes GR, Roos DS, Beiting DP. 2018. MicrobiomeDB: a
931 systems biology platform for integrating, mining and analyzing microbiome experiments.

- 932 *Nucleic Acids Res* **46**:D684–D691. doi:10.1093/nar/gkx1027
- 933 Papa E, Docktor M, Smillie C, Weber S, Preheim SP, Gevers D, Giannoukos G, Ciulla D,
934 Tabbaa D, Ingram J, Schauer DB, Ward D V, Korzenik JR, Xavier RJ, Bousvaros A, Alm
935 EJ. 2012. Non-Invasive Mapping of the Gastrointestinal Microbiota Identifies Children with
936 Inflammatory Bowel Disease. *PLoS One* **7**:e39242.
- 937 Paulson JN, Stine OC, Bravo HC, Pop M. 2013. Differential abundance analysis for microbial
938 marker-gene surveys. *Nat Methods* **10**:1200–1202. doi:10.1038/nmeth.2658
- 939 Phipson B, Lee S, Majewski IJ, Alexander WS, Smyth GK. 2016. ROBUST
940 HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE
941 GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION. *Ann*
942 *Appl Stat* **10**:946–963. doi:10.1214/16-AOAS920
- 943 Pollock J, Glendinning L, Wisedchanwet T, Watson M. 2018. The Madness of Microbiome:
944 Attempting To Find Consensus “Best Practice” for 16S Microbiome Studies. *Appl Environ*
945 *Microbiol* **84**:e02627-17. doi:10.1128/AEM.02627-17
- 946 Pop M, Walker AW, Paulson J, Lindsay B, Antonio M, Hossain MA, Oundo J, Tamboura B, Mai
947 V, Astrovskaya I, Bravo HC, Rance R, Stares M, Levine MM, Panchalingam S, Kotloff K,
948 Ikumapayi UN, Ebruke C, Adeyemi M, Ahmed D, Ahmed F, Alam MT, Amin R, Siddiqui
949 S, Ochieng JB, Ouma E, Juma J, Mailu E, Omore R, Morris JG, Breiman RF, Saha D,
950 Parkhill J, Nataro JP, Stine OC. 2014. Diarrhea in young children from low-income
951 countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol*
952 **15**:R76. doi:10.1186/gb-2014-15-6-r76
- 953 Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers
954 differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids*
955 *Res* **43**:e47–e47. doi:10.1093/nar/gkv007
- 956 Robinson MD, Oshlack A. 2010. A scaling normalization method for differential expression
957 analysis of RNA-seq data. *Genome Biol* **11**:R25. doi:10.1186/gb-2010-11-3-r25
- 958 Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source
959 tool for metagenomics. *PeerJ* **4**:e2584. doi:10.7717/peerj.2584
- 960 Rosato A, Barone M, Negroni A, Brigidi P, Fava F, Xu P, Candela M, Zanolli G. 2020.
961 Microbial colonization of different microplastic types and biotransformation of sorbed
962 PCBs by a marine anaerobic bacterial community. *Sci Total Environ* **705**:135790.
963 doi:10.1016/j.scitotenv.2019.135790
- 964 Ross MC, Muzny DM, McCormick JB, Gibbs RA, Fisher-Hoch SP, Petrosino JF. 2015. 16S gut
965 community of the Cameron County Hispanic Cohort. *Microbiome* **3**:7. doi:10.1186/s40168-
966 015-0072-y
- 967 Scheperjans F, Aho V, Pereira PAB, Koskinen K, Paulin L, Pekkonen E, Haapaniemi E,
968 Kaakkola S, Eerola-Rautio J, Pohja M, Kinnunen E, Murros K, Auvinen P. 2015. Gut
969 microbiota are related to Parkinson’s disease and clinical phenotype. *Mov Disord* **30**:350–
970 358. doi:https://doi.org/10.1002/mds.26069
- 971 Scher JU, Sczesnak A, Longman RS, Segata N, Ubeda C, Bielski C, Rostron T, Cerundolo V,
972 Pamer EG, Abramson SB, Huttenhower C, Littman DR. 2013. Expansion of intestinal
973 *Prevotella copri* correlates with enhanced susceptibility to arthritis. *Elife* **2**:e01202.
974 doi:10.7554/eLife.01202
- 975 Schloss PD. 2020. Removal of rare amplicon sequence variants from 16S rRNA gene sequence
976 surveys biases the interpretation of community structure data. *bioRxiv* 2020.12.11.422279.
977 doi:10.1101/2020.12.11.422279

- 978 Schneider D, Thürmer A, Gollnow K, Lugert R, Gunka K, Groß U, Daniel R. 2017. Gut bacterial
979 communities of diarrheic patients with indications of *Clostridioides difficile* infection. *Sci*
980 *Data* **4**:170152. doi:10.1038/sdata.2017.152
- 981 Schubert AM, Rogers MAM, Ring C, Mogle J, Petrosino JP, Young VB, Aronoff DM, Schloss
982 PD. 2014. Microbiome Data Distinguish Patients with *Clostridium*
983 *difficile* Infection and Non-*Clostridium*
984 *difficile*-Associated Diarrhea from Healthy Controls. *MBio* **5**:e01021-14. doi:10.1128/mBio.01021-14
- 985 Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS, Huttenhower C. 2011.
986 Metagenomic biomarker discovery and explanation. *Genome Biol* **12**:R60. doi:10.1186/gb-
987 2011-12-6-r60
- 988 Sekhon JS. 2011. Multivariate and Propensity Score Matching Software with Automated
989 Balance Optimization: The Matching package for R. *J Stat Software; Vol 1, Issue 7*.
- 990 Singh P, Teal TK, Marsh TL, Tiedje JM, Mosci R, Jernigan K, Zell A, Newton DW, Salimnia H,
991 Lephart P, Sundin D, Khalife W, Britton RA, Rudrik JT, Manning SD. 2015. Intestinal
992 microbial communities associated with acute enteric infections and disease recovery.
993 *Microbiome* **3**:45. doi:10.1186/s40168-015-0109-2
- 994 Son JS, Zheng LJ, Rowehl LM, Tian X, Zhang Y, Zhu W, Litcher-Kelly L, Gadow KD,
995 Gathungu G, Robertson CE, Ir D, Frank DN, Li E. 2015. Comparison of Fecal Microbiota
996 in Children with Autism Spectrum Disorders and Neurotypical Siblings in the Simons
997 Simplex Collection. *PLoS One* **10**:e0137725.
- 1000 Thorsen J, Breenrod A, Mortensen M, Rasmussen MA, Stokholm J, Al-Soud WA, Sørensen S,
1001 Bisgaard H, Waage J. 2016. Large-scale benchmarking reveals false discoveries and count
1002 transformation sensitivity in 16S rRNA gene amplicon data analysis methods used in
1003 microbiome studies. *Microbiome* **4**:62. doi:10.1186/s40168-016-0208-8
- 1004 Turnbaugh PJ, Hamady M, Yatsunencko T, Cantarel BL, Duncan A, Ley RE. 2009. A core gut
1005 microbiome in obese and lean twins. *Nature* **457**. doi:10.1038/nature07540
- 1006 Vincent C, Stephens DA, Loo VG, Edens TJ, Behr MA, Dewar K, Manges AR. 2013.
1007 Reductions in intestinal Clostridiales precede the development of nosocomial *Clostridium*
1008 *difficile* infection. *Microbiome* **1**:18. doi:10.1186/2049-2618-1-18
- 1009 Weiss S, Xu ZZ, Peddada S, Amir A, Bittinger K, Gonzalez A, Lozupone C, Zaneveld JR,
1010 Vázquez-Baeza Y, Birmingham A, Hyde ER, Knight R. 2017. Normalization and microbial
1011 differential abundance strategies depend upon data characteristics. *Microbiome* **5**:27.
1012 doi:10.1186/s40168-017-0237-y
- 1013 Wu Linwei, Ning D, Zhang B, Li Y, Zhang P, Shan X, Zhang Qiuting, Brown MR, Li Z, Van
1014 Nostrand JD, Ling F, Xiao N, Zhang Ya, Vierheilig J, Wells GF, Yang Y, Deng Y, Tu Q,
1015 Wang A, Acevedo D, Agullo-Barcelo M, Alvarez PJJ, Alvarez-Cohen L, Andersen GL, de
1016 Araujo JC, Boehnke KF, Bond P, Bott CB, Bovio P, Brewster RK, Bux F, Cabezas A,
1017 Cabrol L, Chen S, Criddle CS, Deng Y, Etchebere C, Ford A, Frigon D, Sanabria J,
1018 Griffin JS, Gu AZ, Habagil M, Hale L, Hardeman SD, Harmon M, Horn H, Hu Z, Jauffur S,
1019 Johnson DR, Keller J, Keucken A, Kumari S, Leal CD, Lebrun LA, Lee J, Lee M, Lee
1020 ZMP, Li Y, Li Z, Li M, Li X, Ling F, Liu Y, Luthy RG, Mendonça-Hagler LC, de Menezes
1021 FGR, Meyers AJ, Mohebbi A, Nielsen PH, Ning D, Oehmen A, Palmer A, Parameswaran P,
1022 Park J, Patsch D, Reginatto V, de los Reyes FL, Rittmann BE, Noyola A, Rossetti S, Shan
1023 X, Sidhu J, Sloan WT, Smith K, de Sousa OV, Stahl DA, Stephens K, Tian R, Tiedje JM,

1024 Tooker NB, Tu Q, Van Nostrand JD, De los Cobos Vasconcelos D, Vierheilig J, Wagner M,
1025 Wakelin S, Wang A, Wang B, Weaver JE, Wells GF, West S, Wilmes P, Woo S-G, Wu
1026 Linwei, Wu J-H, Wu Liyou, Xi C, Xiao N, Xu M, Yan T, Yang Y, Yang M, Young M, Yue
1027 H, Zhang B, Zhang P, Zhang Qiuting, Zhang Ya, Zhang T, Zhang Qian, Zhang W, Zhang
1028 Yu, Zhou H, Zhou J, Wen X, Curtis TP, He Q, He Z, Brown MR, Zhang T, He Z, Keller J,
1029 Nielsen PH, Alvarez PJJ, Criddle CS, Wagner M, Tiedje JM, He Q, Curtis TP, Stahl DA,
1030 Alvarez-Cohen L, Rittmann BE, Wen X, Zhou J, Consortium GWM. 2019. Global diversity
1031 and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol*
1032 **4**:1183–1195. doi:10.1038/s41564-019-0426-5

1033 Yurgel SN, Douglas GM, Comeau AM, Mammoliti M, Dusault A, Percival D, Langille MGI.
1034 2017. Variation in Bacterial and Eukaryotic Communities Associated with Natural and
1035 Managed Wild Blueberry Habitats. *Phyobiomes J* **1**:102–113. doi:10.1094/PBIOMES-03-
1036 17-0012-R

1037 Zeller G, Tap J, Voigt AY, Sunagawa S, Kultima JR, Costea PI, Amiot A, Böhm J, Brunetti F,
1038 Habermann N, Hercog R, Koch M, Luciani A, Mende DR, Schneider MA, Schrotz-King P,
1039 Tournigand C, Tran Van Nhieu J, Yamada T, Zimmermann J, Benes V, Kloor M, Ulrich
1040 CM, von Knebel Doeberitz M, Sobhani I, Bork P. 2014. Potential of fecal microbiota for
1041 early-stage detection of colorectal cancer. *Mol Syst Biol* **10**:766.
1042 doi:<https://doi.org/10.15252/msb.20145645>

1043 Zhu L, Baker SS, Gill C, Liu W, Alkhoury R, Baker RD, Gill SR. 2013. Characterization of gut
1044 microbiomes in nonalcoholic steatohepatitis (NASH) patients: A connection between
1045 endogenous alcohol and NASH. *Hepatology* **57**:601–609.
1046 doi:<https://doi.org/10.1002/hep.26093>

1047
1048