Differential richness inference for 16S rRNA marker gene surveys

M. Senthil Kumar[1,2,*], Eric V. Slud[3,4], Christine Hehnly[5,6,9], Lijun Zhang[5,9], James Broach[5,6], Rafael P. Irizarry[1,2], Steven J. Schiff[7,+], Joseph N. Paulson[8,+,*]

[1]Department of Data Science, The Dana-Farber Cancer Institute, Boston, MA
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA
[3]Department of Mathematics, University of Maryland, College Park, MD
[4]Center for Statistical Research and Methodology U.S. Census Bureau, Suitland, MD
[5]Penn State Institute for Personalized Medicine, The Pennsylvania State University College of Medicine, Hershey, PA
[6]Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA
[7]Department of Engineering Science and Mechanics, Pennsylvania State University, State College, PA
[8]Department of Biostatistics, Product Development, Genentech, San Francisco, CA.
[9]Contributed equally.
[+]Co-senior.
*Corresponding: MSK: senthil@ds.dfci.harvard.edu; JNP: paulson.joseph@gene.com

Preprint server: Biorxiv, link: https://www.biorxiv.org/content/10.1101/2021.11.07.467583v1
Classification. : Physical sciences > Biophysics and Computational Biology, Biological sciences > Microbiology
Keywords. : richness, species misclassification, 16S microbiome, mathematical model, amplification, sequencing

Abstract

Individual and environmental health outcomes are frequently linked to changes in the diversity of associated microbial communities. This makes deriving health indicators based on microbiome diversity measures essential.

While microbiome data generated using high throughput 16S rRNA marker gene surveys are appealing for this purpose, 16S surveys also generate a plethora of spurious microbial taxa. When this artificial inflation in the observed number of taxa (i.e., richness, a diversity measure) is ignored, we find that changes in the abundance of detected taxa confound current methods for inferring differences in richness.

Here we argue that the evidence of our own experiments, theory guided exploratory data analyses and existing literature, support the conclusion that most sub-genus discoveries are spurious artifacts of clustering 16S sequencing reads. We proceed based on this finding to model a 16S survey's systematic patterns of sub-genus taxa generation as a function of genus abundance to derive a robust control for false taxa accumulation.

Such controls unlock classical regression approaches for highly flexible differential richness inference at various levels of the surveyed microbial assemblage: from sample groups to specific taxa collections. The proposed methodology for differential richness inference is available through an R package, *Prokounter*.

Package availability: https://github.com/mskb01/prokounter

1. Introduction

Clinically relevant health outcomes are often accompanied by changes in the diversity of associated microbial communities. For instance, decreased gut microbiome diversity accompanies childhood diarrhea[1], enteric infections[2], and has been shown to predict the onset of infant type I diabetes[3]. Distinct intra-tumoral microbial diversity levels are associated with cancer sub-types[4–6]. Thus, inferring disease associated changes in microbiome diversity metrics is useful for characterizing disease pathology and progression.

Among the various diversity measures, *richness* quantifies the number of taxonomic groups in a community[7,8]. Changes in species richness of biological communities have informed key environmental management practices that are relevant to public health and well-being[7–23]. Of the technologies available for characterizing microbial communities, 16S rRNA gene surveys are widely adopted for their high throughput and low cost. As a broad screening tool, they largely avoid the need for laborious culturing of microbes. This makes them especially attractive for deriving health metrics based on the microbiome.

In this work, we focus on inferring changes in richness of microbial communities between sample groups (i.e., differential richness) with 16S survey data.

To infer differential richness, one first estimates richness for the specific communities of interest in each survey sample. The estimated values are then compared between sample groups with either fixed or mixed effects models, or with non-parametric statistical tests, possibly adjusting for sampling effort[24–27]. There are two types of sample-level estimates of richness. *Observed* richness refers to the number of taxa observed in a sample. *Asymptotic* richness is obtained by adding an estimate of the number of unobserved taxa to the number of observed taxa. Approaches to estimate asymptotic richness vary, but often assume that relatively uncommon taxa are the most informative[28]. Both types of richness estimates enable valid comparisons among *macro*-ecological communities[24,25,28–30].

However, direct application of the aforementioned richness estimates and comparisons to 16S microbiome data would ignore the plethora of uncommon and spurious taxa that inflate observed richness estimates in 16S survey data[27,31–34]. When this artificial inflation in observed richness is ignored, we find that differential abundance of detected taxa confounds current methods for differential richness inference. The problem is severe when between-sample richness comparisons are made at lower taxonomic levels, e.g., genus. Thus, direct application of classical methods to microbiome differential richness inference is unreliable.

Attempts to overcome sequencing noise have been made. Chiu & Chao[35], noting that singleton taxa are highly susceptible to sequencing noise, establish an improved estimator for undetected richness by relying on more abundant taxa (also see Willis[36]). However, the estimator is often numerically undefined at lower levels of the taxonomy, and still takes observed richness at face value.

Our results indicate that the observed frequencies of spurious taxa are determined by the output abundances of input sequences, and thus need not be restricted to singleton frequencies alone.

We therefore aimed to develop a flexible differential richness inference procedure for 16S microbiome surveys — one that would not only allow investigators to seek sample-wide richness changes across experimental groups (as is commonly done in modern metagenomics), but also within genera or taxa collections of any particular interest, while accounting for false taxa accumulations.

The paper is divided into several sections. Section 2.1, based on our own experiments and exploratory data analyses guided by theory, presents multiple lines of evidence supporting the view that most sub-genus taxa currently identified in 16S surveys are spurious. This allows us to exploit within-genus taxa accumulation data to derive a robust control for false taxa accumulations (Methods section). Section 2.2 illustrates the confounded differential richness inferences arising from current methods, when detected taxa exhibit a net non-zero relative abundance fold change between sample-groups. Section 2.3 applies the proposed procedure (*Prokounter*) to a variety of datasets and illustrates the value that differential richness inferences at lower taxonomic levels add to clinical and public health related microbiome data analyses. For example, application of *Prokounter* to a gut microbiome survey of a traveling individual[2] identifies invading genera with increased richness in member taxa, during and after an enteric infection.

2. Results

2.1 Most sub-genus taxa in 16S surveys are likely technical artifacts

16S surveys reconstruct target microbial populations by clustering sequencing reads. Spurious microbial taxa occur when the clustering procedure's error model fails to capture the entirety of sequence variation induced by the technical steps in 16S sequencing. These steps include, but are not limited to, PCR amplification of 16S material and sequencing (Fig. 1A).

To identify the major parameters underlying false taxa accumulations, we mathematically model the nucleotide substitution errors introduced by a chain of PCR amplification and sequencing processes allowing for back mutations (Supplementary Note 1). Under reasonable assumptions, we find that the rate of falsely classifying an error variant of a source sequence (type I error) using a priori fixed sequence similarity thresholds, strongly increases with the source sequence's recovered (i.e., output) abundance. The average recovered abundance is multiplicative in the source sequence's apparent input abundance and the total sampling depth (Supplementary Note 1). Thus, false sequence clustering decisions, and hence the resulting false clusters, increasingly accumulate with the true source sequence's recovered abundance, and not necessarily sample depth. We therefore identify a mechanism through which spurious clusters of sequences are increasingly identified as microbial taxa, regardless of the underlying biological reality.

Given the empirical observation that 16S genetic segments are mostly limited in resolution to prokaryotic genera[37–44], we explored within-genus taxa accumulations (i.e., the number of detected sub-genus taxa as a function of recovered genus abundances), in several publicly available 16S surveys. In general, we expect genera to vary in their true richness and the relative abundances of member taxa. This must accordingly induce biological variation in the genus-specific taxa accumulation patterns. However, this expectation did not broadly hold in the several microbiome surveys analyzed here. Within-genus taxa accumulation patterns were highly concordant for several genera within study (Fig. 2A, Fig. S1-S3). Relative to the number of detected genera, which ranged from 60-400 across studies, a clustering analysis indicates that within-genus taxa accumulation data supports only 2-8 distinct accumulation patterns in each study (Table. S1). Multiple dominant genera can be clustered to the same accumulation pattern. In addition, relative to study specific covariates, a robust trend estimate of the within-genus taxa accumulation data explains the bulk of the variation in genus-specific and sample-wide taxa accumulations (Tables 1-2) in each study. Similar qualitative conclusions follow when genus recovered abundance is used as a predictor, instead of an estimated trend (Tables S2-S3). Finally, these qualitative and quantitative attributes of the accumulation patterns were obtained regardless of the 16S clustering approach used (Tables 1-2). These results indicate a strong within-study regularity in observed taxa accumulations across genera and sample groups - as if most genera have similar taxa richness and evenness - suggesting a likely technical origin.

Single colony experiment To further verify these conclusions, we conducted a 16S sequencing experiment on a target *Pseudomonas aeruginosa* population. The experimental sample was by itself overnight derived from a single *P. aeruginosa* colony (Supplementary Note 1). In a series of experimental samples, we varied both the input abundance of Pseudomonas cells and the PCR

142 amplification cycles. Our mathematical model (Supplementary Note 1), which tracked the
143 probability distribution of cell division induced nucleotide substitutions over generations, indicates
144 that under no selection pressure, we can expect one biological 16S genotype in our input. An
145 upper bound on the number of our input taxa is given by the number of 16S genes generally found
146 within the *Pseudomonas* genus (~ 4), times two for taxa clusters corresponding to forward and
147 reverse complement strands. What we found was a rather different representation, rich with low
148 abundant and poorly replicating taxa: the total numbers of observed *Pseudomonas* taxa were
149 1050 and 300 for clustering methods based on sequence similarity with respective thresholds of
150 99% and 97%. The bulk of the newly identified *Pseudomonas* taxa preferentially contributed to
151 the low frequency regime of the taxa abundance histogram (Fig. S4), suggesting that they are
152 likely clusters of rare, erroneous 16S sequencing reads generated during amplification and
153 sequencing. Notably, taxa within-Pseudomonas, despite having a noisy occurrence with respect
154 to amplification cycles and input cells (Fig. 3A), accumulated along the Pseudomonas genus
155 recovered abundance axis in a clear, robust fashion (Fig. 3B). As expected, the stricter the
156 sequence similarity threshold, the stronger the rate of taxa accumulations along the recovered
157 abundance axis (Fig. 3B). Furthermore, taxa accumulations from several detected genera
158 followed quantitatively similar patterns (Fig. 3C, Tables 1-2). From prior experiments in our
159 laboratory and from control samples, we know *Pseudomonas* lab contaminants have very weak
160 relative abundances. Restricting the above analysis to only those Pseudomonas taxa that track
161 input cells, does not change the aforementioned conclusions qualitatively (Fig. S5).

162 Similar results on taxa accumulation patterns were also obtained for the *multiple*-genera Oral and
163 Gut mock communities of the microbiome quality control project, handling lab B (MBQC[27]).

164 Because true taxa are expected to replicate across study samples, we next explored sub-genus
165 taxa occurrence rates (Fig. S19). In all studies, we find that over 50% of sub-genus taxa in over
166 50% of the detected genera did not replicate in more than 10% of the samples. Mock experimental
167 communities are expected to represent a greater degree of homogeneity than real world
168 communities as the latter may contain rare variants. Restricting analysis to experimental
169 communities with single- and multiple- mock genera, we find that in eight out of nine datasets,
170 over 50% of sub-genus taxa in over 50% of the mock genera replicated in less than 50% of the
171 samples (Fig. S19). These results indicate poor within-study replicability of most sub-genus taxa.

172 Finally, because we expect true taxa richness and evenness to vary along the taxonomic tree, we
173 explored taxa accumulations for the various taxonomic levels (i.e., family, order, class and
174 phylum) in each study. Remarkably, the total number of observed taxa at any level of the
175 taxonomic tree, was strongly predicted by recovered abundance alone and was not dependent
176 on the taxonomic level considered (Fig. 2A, Fig. S20). These results indicate a strong regularity
177 in taxa accumulations across taxonomic levels.

178 Taken together, our results indicate that most sub-genus taxa in 16S surveys are likely spurious.

179

2.2 Spurious taxa confound differential richness inference

That observed spurious sub-genus taxa increasingly accumulate with genus recovered abundances leaves us with two expectations.

First, without appropriate corrections, inferring differences in a genus' number of associated taxa (i.e., genus-wise differential richness) are highly likely to be confounded by the genus's respective difference in the recovered abundances (differential abundance). We observe that estimated genus-wise richness values from asymptotic estimators grew systematically with the genus-specific recovered abundances (Fig. S6). In addition to observed richness, estimates of unobserved richness can exhibit similar behavior (Fig. S7). This in turn induces an artifactual positive correlation between the resulting genus-wise differential richness fold changes and the genus-wise differential abundance fold changes (Fig. 2B, S8).

Second, inferring differential richness between sample-groups (i.e., sample-wide differential richness) are highly likely to be confounded by a net non-zero relative abundance fold change of detected genera. Straightforward simulations where spurious taxa are generated in an abundance dependent fashion illustrate this behavior (Fig. S6). Interestingly, illustrative examples of the same were rare in several 16S surveys, suggesting that spurious taxa accumulations are comparable at the sample-level. Indeed, in many datasets, the relative abundance log fold changes of member genera were symmetric and concentrated around zero (Figs. S9-S10). Nevertheless, exceptions with asymmetric relative abundance log fold change distributions exist and a case in point is offered by the long-term time series study discussed below (Fig. S11).

In Supplementary note 2, we model the abundance dependent generation of spurious taxa in 16S surveys within the sample theoretic framework of Chao[29] and Harris[45] and find that the above observations agree with theory.

2.3 Prokounter enables flexible differential richness inference

To overcome the aforementioned biases when applying current richness estimators to 16S surveys and to establish a flexible differential richness inference approach, we developed *Prokounter* and applied it to several microbial communities including those from a long-term time series study, hydrocephalus cohort, waste-water treatment plant and our pseudomonas dilution experiment.

While zero-truncated statistical models offer one route to modeling member inclusions in a population survey, the same can be achieved by incorporating appropriate predictors in a regression context[46]. The former is the approach taken by some classical richness estimators to model species abundance[28,47]. We take the latter view and proceed as follows. Based on the results from section 2.1, we assume that most sub-genus taxa in 16S surveys are false. This allows us to exploit a 16S survey's overall sub-genus taxa accumulation trend, along with any systematic genus-specific effects, as a sampling effort dependent control for false taxa accumulation (Methods). This control is exploited within standard regression methods for differential richness inference.

With a few 16S surveys, we illustrate the insights offered by the proposed procedure, *Prokounter*, in achieving genus-specific and sample-wide differential richness inferences.

Unlike other estimators analyzed here (Chao1[29], ACE[48] and Breakaway[49]), the uncorrelatedness of Prokounter's richness statistics with genus-wide differential abundance statistics is clear in each dataset (Fig. 2B,C, Fig. S12). Breakaway's estimates were the most variable, often accompanied by wide confidence intervals. On several occasions, genus-specific differential richness estimates were not well defined in numerical value when using current richness estimators for numerical, and not necessarily statistical identifiability reasons. Sample-wide inferences agreed among all methods in most cases, except when detected genera exhibited a net non-zero relative abundance fold change distribution.

In all surveys below, asymptotic genus-wise and sample-wide richness estimates heavily tracked their respective observed richness values (97-100% Pearson correlations, Figs. S13-S17).

<u>*Long-term time series study*</u> Based on a clustering analysis of abundance profiles, David et al.,[2] identified that a distinct sub-group of the phyla *Firmicutes* replaced another *Firmicutes* sub-group, post-enteric infection, in the gut microbiome of an individual relocating to a different country. Prokounter refines this result further by identifying several Firmicutes genera (*Faecalibacterium*, *[Ruminococcus]*, *Oscillospora*) that are less rich post-infection. On the other hand, *Dorea* and *Coprobacillus*, members of Firmicutes, were found to have significantly increased richness in infection and post-infection samples. The genus *Acinetobacter* from the phylum *Tenericutes* was found to have significantly higher richness in samples collected during infection, while this was not the case post-infection. Thus, differential richness adds another state variable to the microbiome state specifications of the original study.

In David et al's dataset, sample-wide inferences disagreed among the methods compared. Prokounter produced negative richness inferences for both infection and post-infection samples consistent with antibiotic exposure. Chao1/Betta and ACE/Betta indicated reduced richness post-infection with a relatively weak significance for reduced richness in infection samples. Breakaway/Betta failed to reject any of the corresponding null hypotheses (p-value=0.99 infection and p-value=0.74 post-infection), potentially owing to the very high variability of Breakaway estimates. As established in the previous subsection, these differences in inferences likely stem from the asymmetric differential abundance of detected genera in the samples collected during infection.

<u>*Pathogenesis*</u> We applied Prokounter to a 16S survey of the cerebrospinal fluid from hydrocephalus children hypothesized to have infectious (PIH) and non-infectious (NPIH) origins[50]. We intuitively expected, and observed, that the cerebrospinal fluid enveloping the central nervous system to register lower richness compared to laboratory controls. PIH samples had relatively lower richness compared to clinical control samples.

A genus that is positively differentially abundant, along with a negative differential richness estimate might indicate invasion of a sub-species. Genus-specific differential richness inference with Prokounter yields two genera as having lowered richness in the PIH samples: *Paenibacillus*

257 and *Streptococcus*. *Paenibacillus* was the dominant pathogenic genus identified with the PIH
258 phenotype using 16S data[50].

259 *Waste-water treatment* To demonstrate an ecological monitoring application, we applied
260 Prokounter to 16S data arising from a waste-water treatment plant[51]. The method indicates that
261 relative to the effluent, sample groups from each of the post-treatment stages have significant
262 negative microbial richnesses. These results readily agree with our expectation of a publicly
263 implemented waste water treatment protocol. Chao1/Betta, ACE/Betta produced similar results.
264 Breakaway/Betta failed to reject the null for sample groups corresponding to effluent (p=.065) and
265 inlet to pumphouse (p=.692).

266 Using differential abundance analysis, the original study highlighted the persistence of *Legionella*
267 and *Mycobacterium* in post-treatment samples calling into question the efficacy of the treatment
268 process. Performing genus-specific differential richness analysis with Prokounter indicates that
269 the treatment plant reduces the richness associated with several types including *Mycobacterium*.
270 We did not detect *Legionella* as reduced in richness in the effluent. These results indicate that
271 waste-water treatment has been effective with removing *Mycobacterium* sub-types.

272 *Pseudomonas dilution study* The Pseudomonas dilution experiment varied two parameters of a
273 16S experimental pipeline: amplification cycles, and input cells of a single colony derived
274 microbial population.

275 Increased amplification cycles can allow increased sampling of both contaminant and input
276 genera. Thus, within further sampling constraints imposed by the multiplexed nature of the
277 experiment, we expect sample-wide richness to grow with amplification cycles. Sample-wide
278 differential richness inference from all methods matched this expectation.

279 It is well known that the abundance of lab contaminants falls with input loads [34]. If the dynamic
280 range in input loads is sufficiently high, we can expect inferred sample-wide richness to fall with
281 input Pseudomonas cells. Results from Prokounter, Chao1/Betta and ACE/Betta matched this
282 expectation. Breakaway/Betta failed to reject the corresponding null hypotheses (p=0.4).

283 The genus of principal interest in this experiment is Pseudomonas. The genus-wide differential
284 richness results from Prokounter indicated a decrease in richness with respect to input cells and
285 an increase with respect to amplification cycles. This is in line with our expectations as we expect
286 the detection rate of lab contaminant Pseudomonas species to grow with amplification cycles,
287 and fall with input Pseudomonas loads. In direct contrast, Chao1/Betta and ACE/Betta,
288 confounded by input Pseudomonas's increasing abundance, indicated a Pseudomonas richness
289 *increase* with input cells (p=0 for both), and Breakaway/Betta failed to reject (p=0.251).

3. Discussion

Summary: 16S microbiome surveys reconstruct target microbial populations by clustering sequencing reads. Spurious microbial taxa occur when the clustering procedure's error model fails to capture the entirety of sequence variation induced by the technical steps in 16S sequencing (Supplementary Note 1, Fig. 1A). We have shown that the false taxa thus generated not only inflates the estimate of a (microbial) community's richness (Supplementary Note 2, Fig. S6), but they also cause taxa differential abundance to confound differential richness inferences (Fig. 2B, S8). This occurs because every false taxon is generated through errors from one or a few true (i.e., input) taxa, and hence, their rates of production increase with the output abundance of the corresponding source taxa (Supplementary Note 1). Based on our result that most sub-genus discoveries are likely false (Section 2.1), we have established abundance dependent controls for false taxa accumulations using a given survey's within-genus taxa accumulation data (Methods, Fig.2C, S2, S18). We have shown that our strategy overcomes the confounding problem (Fig. 2C, S12). And we have illustrated the utility of differential richness inferences in individual and public health related microbiome data analyses (Section 2.3).

Assumption: Our approach assumes that most sub-genus taxa in 16S surveys are spurious and are poor representatives of the underlying microbial community. We have provided several lines of evidence to support this conclusion: First, our mock experiment of an overnight derived microbial population indicated that observed richness can be severely inflated (Fig. 3, S5). Our expectation was set in part by a mathematical model of cellular reproduction, where we tracked the probability distribution over substitutions, over generations (Supplementary Note 1). Second, in a manner similar to what we would expect of low probability errors, most sub-genus taxa in both controlled mock and real world datasets are rare and show poor replicability across samples (Fig. S19). Third, within-genus taxa accumulation patterns in several publicly available datasets, including those from single- and multi-genera mock experiments, appear remarkably regular as if most genera in 16S surveys have similar richness and taxa evenness (Fig. 2-3, S1-S3, Tables 1-2, S2-S3). Fourth, the total number of taxa observed for any taxonomic level was strongly determined by the category's recovered abundance alone and was not dependent on the level itself (Fig. 2, S20). Finally, the literature offers abundant support for abundance dependent false taxa generation in 16S surveys, of which we note a closely related few. Kunin et al.,[32] demonstrate the large number of false *Escherichia* taxa that arise in a 16S survey of a target *E.coli* population (also see Degnan and Ochman[52], Pinto and Raskin[53]). Based on the empirical observation that the number of false taxa generated are sampling effort dependent, Schloss et al.,[54] recommend that community-level comparisons be made at comparable sampling depths. Haas et al.,[55] illustrate the predictable, abundance dependent generation of false chimeric taxa within genera in mock communities.

Implications for richness theory and automated ecological surveys: False microbial taxa in 16S surveys arise because automated procedures to reconstruct taxa misclassify sequencing reads from their true types. In Supplementary Note 1, we analyzed the influence of amplification and sequencing induced substitutions in causing misclassifications (also see Schloss[56], and Sze and Schloss[57]). In Supplementary Note 2, we mathematically modeled the false taxa that arise

through misclassification and showed in part that a traditional asymptotic richness estimator (Chao1[29]) is biased under this more general sampling scenario. The severity of bias is determined by sampling parameters. Together with the results mentioned in the previous paragraphs, we conclude that classical richness theory, which predominantly focuses on estimating undetected richness while assuming observed richness at face value, should be generalized for observed species misclassifications in modern high throughput and highly automated surveys.

Asymptotic richness estimators track observed richness values in 16S surveys: In the several 16S surveys considered here, asymptotic richness estimates tracked observed richness values both sample-wide and at within-genera levels (Fig. S13-S17). Our mathematical models and simulations that incorporate false taxa accumulations within the sampling theoretic framework of Chao[29] and Harris[45] indicate that such tracking can arise when the apparent richness (i.e, the true plus false richness) and not necessarily true richness, is undersampled in a survey (Supplementary Note 2). This explains the observed tracking in the *Pseudomonas* genus in the *Pseudomonas* dilution experiment, where we do not expect undersampling of the true *Pseudomonas* community (Fig. S13).

False discovery control in differential richness analysis, confounding with differential abundance: Hughes et al.,[58] argue that traditional macroecological richness estimators continue to enable robust sample-wide richness comparisons in 16S surveys. Our analysis identifies exceptions (Section 2.3, Long-term time series study) and clarifies the practical conditions under which controlling for spurious discoveries become important. In particular, we find that false taxa accumulations cause abundance dependent inflation in observed taxa numbers and their frequencies (Supplementary Note 1, 2), causing differential (relative) abundances of detected taxa to confound differential richness inference with traditional methods (Fig. 2B, S6-S8, S11). When spurious taxa accumulations are comparable across contrasted experimental groups, no such confounding arises (Fig. S9-S10). Our empirical analyses indicate that such an assumption is too strong for making differential richness inferences at lower taxonomic levels (e.g., genus-specific) of a microbial assemblage (Fig. 2B).

Relaxing microbiome richness comparisons to taxonomic groups: Microbiome analyses frequently restrict richness comparisons to the entire microbial assemblage obtained in study samples (sample-wide richness inference). From the perspective of deriving health and ecological indicators based on community assemblages, analysis of a community's finer organization levels is equally interesting[2,8,10–12,17]. Our genus-wise differential richness results (Section 2.3) indicate that contrasting richness for taxonomic sub-groups can enable practically useful inferences and add interesting dimensions to microbiome state space descriptions.

Within-genus taxa accumulation structure and the trend estimator: Our results document reliable across-genera regularity in the patterns of within-genus taxa accumulations, across many studies and genus-specific experiments (Fig. 2-3, S1-S3, Tables 1-2, S2-S3). We speculate that genus abundances, in contrast to sampling depth, more accurately track the sampling rate of false sequence variation in 16S surveys for at least two reasons. First, commonly exploited 16S rRNA target segments are limited in resolution beyond genus level[37–44]. Second, genus recovered

abundances, unlike total sampling depth, normalize for the sampling rates of distinct genera. This restricts us from mixing taxa accumulation statistics over truly disparate input biological sequences from distinct genera, while allowing us to preserve any systematic genus specific effects. We used a robust trend estimate of the within-genus taxa accumulation data to model spurious taxa accumulation (Methods, Fig. 2-3, S1-S3). The coherent accumulation of a large number of detected taxa translated to low estimation uncertainties. These curves were not necessarily linear in the recovered genus abundances (Fig. S1-S3). The systematic genus-specific contributions to this trend can arise due to between-genera variation in both detectable true input sequence diversity (copy number[43] or number of distinct cell types) and 16S sequencing noise[56,57].

Abundance dependent control in bioinformatic sequence analysis: Beyond differential richness inference, there is a need for recovered abundance dependent control in other (meta)genomic sequence analyses e.g., sequencing read mapping and taxonomic annotation, which exploit fixed sequence similarity thresholds. Probabilistic methods have a natural incorporation of abundance in clustering/mapping decisions. In all cases however, poor error models would continue to drive false taxa accumulations. It must be noted that we have not analyzed false negative rates in this study[59,60].

Limitations of differential richness inference Observed (and reportedly, asymptotic[61]) richness estimates cannot forecast crossing over of species accumulation curves that can in principle occur with additional sampling effort. However, differential analysis of both these estimates over realized sampling effort is still useful for detecting perturbations to the evenness of a biological community[58,62], and is thus effective for deriving predictors of individual and environmental health.

Future work. There are several avenues for future research. First, an integrated estimation procedure of false taxa accumulation rates and differential richness fold changes would lead to more appropriate p-values under the assumed statistical models. Second, development of ecological richness estimators in the presence of species misclassifications would be a valuable addition to the literature. Supplementary Note 2 considers a simple but a useful special case. Third, 16S surveys on mixtures of microbial species with varied relatedness and controlled input richness levels, would enable a joint characterization of detectable 16S resolution, taxa reconstruction algorithms and richness estimators. Fourth, control for multiple testing over tree structured hypotheses can be incorporated if one wishes to automate hypothesis testing over taxa collections defined by subtrees of a taxonomic tree[63,64]. Finally, all our empirical observations were based on a set of 16S surveys that operate over partial 16S gene targets. Because full length 16S surveys also involve amplification, and sequencing protocols[41], we expect the qualitative nature of our results to generalize to such surveys, perhaps at a lower taxonomic level (e.g., species), and this can be explored.

Taken together, this paper significantly clarifies the dynamics of spurious discovery accumulation in 16S surveys, presents strategies for modeling their generation, demonstrates the need to control for the observed false discoveries in microecological surveys while deriving differential richness inferences, and offers a flexible practical solution to achieve the same.

4. Methods

Prokounter

Our proposed procedure for differential richness inference works in two steps. A control for false taxa accumulation is established first. The estimated control is subsequently exploited within standard generalized linear models for differential richness inference.

Let n_{gj} denote the reconstructed number of taxa for genus g in sample j, y_{gj} denote the corresponding recovered abundance (i.e., genus's total count in the sample), and τ represent the sample depth.

Let $f_t(log\ y_{gj})$ indicate the logged technical contribution to taxa accumulation for a given genus and its recovered abundance level. This function is used to model the log of the expected false taxa accumulation. Its estimate $\hat{f}_t\ (log\ \ y_{gj})$ is obtained using within-genus taxa accumulation data as follows.

<u>Estimating the technical contribution $\hat{f}_t$</u> We explored two strategies to estimate a robust within-genus accumulation trend.

A semi-parameteric smoothing spline model is assumed on $z_{gj} = log\ n_{gj}$,

$$z_{gj}|g, y_{gj} = \eta(g, y_{gj}) + \varepsilon_{gj} = \kappa + f_R(log\ y_{gj}) + f_G(g) + f_{GR}(g, log\ y_{gj}) + \varepsilon_{gj} \quad (1)$$

with $\varepsilon_{gj} \sim N(0, \sigma^2)$, and appropriate side conditions are placed on $f_{\cdot}(\cdot)$ (Chapters 2-3[65]). Here κ and $f_R(\cdot)$ denote the intercept and recovered abundance dependent components; f_G and f_{GR} indicate the genus and its respective interaction functions with the recovered genus abundance.

Briefly, η is estimated as a unique solution to the penalized optimization problem: $\hat{\eta} = arg\ min_{h \in H}\ l(\ h\ |\ \tilde{y}_{\cdot}, x_{\cdot}) + \lambda J(h)$, where $l(\cdot\ |\tilde{y}_{\cdot}, x_{\cdot})$ is the negative log likelihood, λ is a regularization parameter and $J(\cdot)$is a roughness penalty that penalizes overfitting of h to the data. The specification of $J(\cdot)$ involves, in part, integrals of squared second order derivatives of the estimand over the range of $log\ y_{gj}$, thereby enforcing smoothness. Supplementary Note 4 offers more details on the model construction and an exact correspondence to example 2.7 in Gu[81]. Numerical optimization is performed using the R package *gss*[65]. Supplementary figures S2 and S18 offer examples of the fits that result.

The technical contribution to taxa growth is estimated as $\hat{f}_t(g, log\ y_{gj}) = \kappa + \hat{f}_R(log\ y_{gj}) + \hat{f}_G(g)$. Only the significant genus effects are retained after multiple testing correction with the Benjamini-Hochberg procedure. When the genera contributions are null or similar, as we observed empirically in several datasets (e.g., Fig. S9, S10), $\hat{f}_t(g, log\ y_{gj}) \propto \hat{f}_R(log\ y_{gj})$.

The latter observation inspires the following alternative strategy: estimate $\hat{f}_t(\cdot)$ as a net average

within-genus accumulation curve using the *loess* smoother. Both options are made available in our software. As expected, inferences arising and the results in tables 1-2 are similar with both approaches. Fig. S3 offers examples of the fitted trends. The spline strategy does offer better control in the presence of systematic genus effects (Fig. S18).

For consistency, in this paper, we have chosen the spline strategy.

The fitted $\widehat{\boldsymbol{f}}_{\boldsymbol{t}}$ can be used to control for false taxa accumulation in standard differential richness inference procedures. In *Prokounter*, we incorporate it through the models presented below.

Differential richness inference We use Greek letters to indicate regression parameters. A ·in the subscript indicates vectorizing over the subscript. *X* denotes the experimental design matrix. Genus-specific, sample-wide and taxa collection models are presented in equations (2)-(4) below. In each case, given the quantity modeled, reasonable transformations of the estimated logged technical contribution, $\widehat{\boldsymbol{f}}_{\boldsymbol{t}}$**,** based on eqn. (1), are used. Terms involving X below can be viewed to approximate the effects arising from genus-recovered abundance interaction terms from eqn. (1).

Genus-specific differential richness inference the conditional mean of the observed richness is modeled through the link:

$$log\ E[n_{gj} \mid y_{g\cdot}, X, f_t(\cdot)] = X_j{}^T \mu_g\ +\ \nu_g f_t(log\ y_{gj}) \tag{2}$$

where the right hand side is an approximate form for the log of the conditional expectation of the right hand side of eqn. (1).

Sample-wide differential richness inference For inference across sample groups, we posit:

$$log\ E[n_{+j} \mid y_{g\cdot}, X, f_t(\cdot)] = X_j{}^T \zeta +\ \gamma\ log \sum_{g:y_{gj}>0} e^{\ f_t(log y_{gj})} \tag{3}$$

where the $+$ indicates summation over a subscript. As in eqn.(2) the right hand side of eqn.(3) is an approximate form for the log of the conditional expectation of the right hand side of eqn.(1), now summed over g. The net sample-wise technical contribution is modeled as a simple sum of the technical contributions from the genera detected in the sample. Although eqn.(3) does not immediately arise from eqn.(2), we find the simplicity and emphasis on dominant contributors to the sum, the more abundant genera, appealing. In addition, we often find that $\nu_g \approx 1$ and $\gamma \approx 1$in applications.

Differential richness inference for arbitrary collections of genera For an arbitrary taxonomic group k (e.g., phyla), with a set of member genera G_k, we assume :

$$log\ E[n_{kj} \mid y_{g\cdot}, X, f_t(\cdot)] = X_j{}^T \psi_k\ +\ \gamma_k\ log \sum_{g \in G_k \cap\ y_{gj}>0} e^{\ f_t(log y_{gj})} \tag{4}$$

As with the sample-wide model, here too we have modeled the sample-wise technical contribution for each collection k based on the sum of genus-level technical contributions, but now restricted

only to those genera considered within the collection.

Keeping to the traditional theme of continuous Poisson mixtures driving sample-wide species accumulations, we chose Negative Binomial variance functions when performing sample-wide inferences, and Poisson variance functions for genus-specific richness inferences. For the several studies considered here, the estimated overdispersion coefficients for sample-wide Negative Binomial models were in the range of 10^{-3} to 10^{-1}. For well expressed genera, inferences and model diagnostics were not sensitive to the two distribution assumptions. Parameter estimation and inference on the regression parameters μ_g, ζ and ψ_k were performed using R's *glm* function. Maximum likelihood estimation with iteratively reweighted least squares converges rapidly in about ten iterations or less. Speaking to the explanatory power of $\widehat{f_t}$, as implied by tables 1-2, the residual deviance is often small, on the order of the residual degrees of freedom. To gauge reproducibility of inferences over fitted $\hat{f}_t(\cdot)$, confidence intervals based on the bootstrap *t*[66] are also available for the regression coefficients of the sample-wide differential richness inference model.

The above models, which were used to generate the results in the applications section, exploit observed richness as response variables and are therefore non-asymptotic in nature. In the several 16S surveys considered here, asymptotic genus-wise and sample-wide richness estimates heavily tracked their respective observed richness values (97-100% Pearson correlations, Figs. S13-S17). We therefore propose the same regression models above for standard inverse variance weighted regression analyses of asymptotic richness estimates. As expected, results from such a procedure were similar to those obtained with observed richness as the response variable. Also see reference[26] for a heterogeneity test of potential interest.

We implement these procedures in an R package *Prokounter*. Supplementary Note 2 presents further discussions on the regression models above.

Package and code availability:

The R package *Prokounter* is available from the link: https://github.com/mskb01/prokounter

Code for the paper is available from the link : https://github.com/mskb01/prokounterPaper

Richness estimators and differential analyses: Estimates and standard errors for Chao1 and ACE estimators were calculated using the R package *vegan*[67]. Breakaway estimates and standard errors were obtained using the R package Breakaway. Differential richness inferences corresponding to the three estimators were obtained with the R package Betta[26]. Rarefaction based interpolated and extrapolated richness estimates and standard errors were obtained using the package *iNext*[68]. The R package doParallel[69] was used for several parallel computing tasks.

The following datasets and study design variables were used to construct design matrices for sample-wide and genus-specific differential analyses reported in the applications section.

1. Hydrocephalus[50] (PIH100 FST97) - Control and Case.

2. Wastewater[51] (WW FST99) - Influent, Effluent, Before UV treatment, After UV treatment, Pond storage, and Inlet to pumphouse for subsequent spray irrigation.
3. MBQC, Handling lab B (MBQC-HLB) - Gut mock, Oral mock, the rest of the stool samples were typed as Other.
4. Time series study[2] (TS FST97, Donor B) - based on the original study, three time windows were established to define sample groups: days up to to 150 were categorized as *pre-infection*, days from 151 upto 159 as *infection*, and days post 159 were typed as *post-infection*.
5. Pseudomonas dilution study (Pseudomonas FST97) - number of cycles and logged number of input Pseudomonas cells.

Dilution experiment: A monoisolate was prepared overnight from a Luria-Bertani (LB) agar plate into a 5 mL LB liquid, which grew to 10^9 cells. A ten fold serial dilution of cells from 10^5 to 10 cells in phosphate buffer saline (PBS) was generated. DNA was isolated, 16S amplified and sequencing libraries were prepared as previously described[50]. Briefly, DNA was isolated using the Zymobiomics DNA miniprep kit following manufacturers protocol with bead beating and proteinase K treatment. For 16S amplification, primer-extension polymerase chain reaction (PE-PCR) of the V1-V2 region was performed using an M13 tagged 336R universal primer as previously described[70] and amplification cycles were varied. Briefly, target DNA was mixed with a 10 µl of 10X buffer, and annealed with M13 tagged 336R by first heating to 95°C and then cooling to 40°C slowly. The annealed product was extended using Klenow polymerase (5U/µl and primers digested with 20U/µl Exo I (NEB, USA), then amplified with 500 nM primers (805R and M13) using the MolTaq 16S Mastermix (Molzym GmbH & Co Kg, Germany). Library preparation was done with the Hyper Prep Kit (KAPA Biosystems, USA) following the manufacturer's protocol and libraries were sequenced on MiSeq using the 600 cycle v3 kit.

16S datasets and taxa reconstruction pipelines: The mouse microbiome 16S data was obtained from the R/Bioconductor package *metagenomeSeq*[71]. The moderate to severe diarrheal 16S survey was obtained from the R/Bioconductor package *msd16S*[72]. The long-term time series 16S survey[2] was obtained from the supplementary data of the corresponding paper. The wastewater 16S survey[51] was obtained on request from the authors of the original study. MBQC handling laboratory B's (HL-B) sequencing reads was obtained from the Microbiome Quality Control (MBQC) project[27].

We generated three varieties of taxa count data from each of the *Pseudomonas*, *PIH100* 16S and MBQC *HL-B (handling lab B)* sequencing data. These include sequence similarity threshold based taxa clustering methods for 99% and 97% sequence similarities (*Qiime1*), and a probabilistic taxa clustering method (*Dada2*) as follows.

Quality filtering of sequencing reads: Paired-end reads were processed with *Trimmomatic*[73] (v0.38) to remove universal adapters and low-quality reads. Reads with ambiguous bases were removed or truncated using *Dada2*'s *filterAndTrim*[74] function. The 16S V1-V2 regions in both our Pseudomonas and PIH100 data were sequenced using 2x300bp paired-end reads. Based on sequencing read quality score profiles, we retained the first 240bp and 210bp in the forward and

568 reverse reads for the *Pseudomonas* dataset. These numbers were respectively 200bp and 190bp
569 for *PIH100*. For HL-B, we removed the first 2bp following the primers in the forward and reverse
570 reads. This allowed us to neglect the trailing low quality bases adversely affecting the taxa
571 reconstructions, while still allowing for sufficient overlap to merge paired-end reads.
572 Reads with either the designed primers or their reverse complements were filtered using
573 *cutadapt*[75]. The quality filtered reads were then clustered with Qiime1[76] and Dada2[74] as below.
574
575 **Qiime 1 :** Quality filtered forward and reverse reads were merged using *Pear*[77], and then clustered
576 using *pick_open_reference_otus.py (Qiime1 version 1.9.1),* which implements the Qiime1 open
577 reference OTU clustering algorithm. Briefly, closed reference clustering of merged reads were
578 performed against the *Silva132* database at 97% and 99% sequence similarity thresholds, using
579 Uclust[78] v.1.2.22q . Reads that did not map to the database were subsampled and used as new
580 centroids for a *de novo* OTU clustering step at the respective sequence similarity thresholds.
581 Remaining unmapped reads were subsequently close clustered against the *de novo* OTUs.
582 Finally, another step of *de novo* clustering was performed on the remaining unmapped reads.
583 Taxonomy was assigned to taxa representative sequences with Uclust based on the *Silva132*[79]
584 database . These sequences were filtered with *Pynast*[80], and OTU tables generated.
585
586 **Dada2**: Dada2 allows denoising forward and the reverse reads independently. Error rates were
587 estimated separately for the quality filtered forward and reverse reads for each sample. This
588 estimation step is based on a sample of reads for computational tractability. Reads were
589 deduplicated and sequence clusters inferred based on the estimated error rates. Taxa from
590 forward and reverse reads were merged at the end of the workflow. Chimeric taxa were removed
591 with the function *removeChimeraDenovo*. The resulting taxa were assigned taxonomic labels
592 based on the *Silva132* database, using their naïve Bayes classifier.

Acknowledgements

This project was supported by an NIH Director's Transformative Award 1R01AI145057. We thank the Genome Science Facility at the Penn State University College of Medicine for performing the sequencing for the *Pseudomonas* dilution study.

Figures

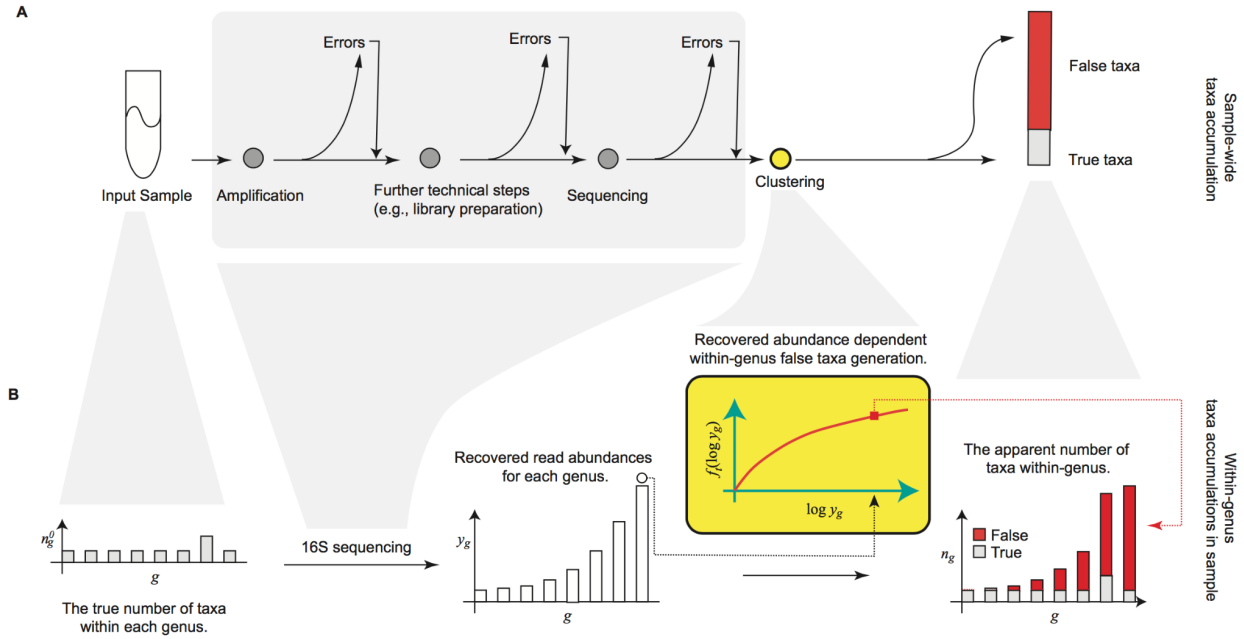

Figure 1. Within-genus false taxa accumulation structure. (A) Sequences in input samples are subjected to various technical steps during 16S sequencing (gray shade). The output reads from 16S sequencing are clustered for sequence similarity using a methodology of choice. Of the number of taxa (clusters) thus reconstructed, some are true, i.e., equal in sequence to those in the input sample, the rest are spurious i.e., false (red). (B) For every genus, the accumulation is determined as a function of its recovered abundances. Notation: ${n_g}^0$ the respective true number of taxa associated (true richness), y_g the genus recovered abundance, $f_t(\cdot)$ the abundance dependent technical component driving false taxa accumulations within-genus.

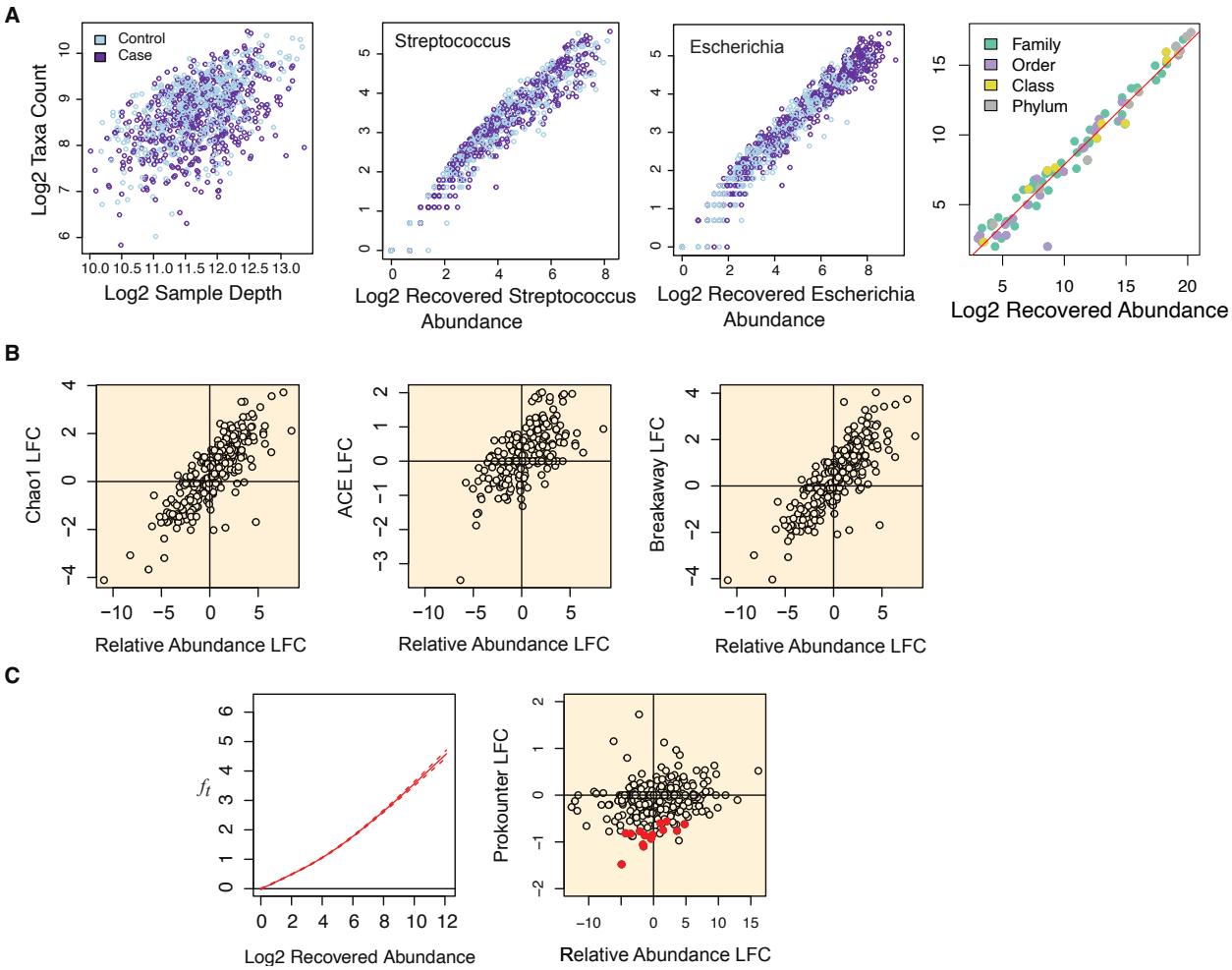

Figure 2. Concordant taxa accumulations across genera, confounded differential richness inference and the Prokounter strategy. (A) Sample-wide taxa accumulations are visualized with respect to sample depth (left). Within-genus taxa accumulations are visualized with respect to the total recovered genus abundances for two genera, i.e., the sum of the abundances of all taxa within the genus (center). Dataset-wide taxa accumulations for any taxonomic level is strongly predicted by recovered abundance alone (right). Red line illustrates a linear fit. (B) Differential richness log-fold changes (LFC, y-axis) track differential relative abundance fold changes (LFC, x-axis) in the waste-water treatment survey. (C) Prokounter exploits within-genus accumulation data to model false taxa accumulation rates. When exploited in a standard Poisson regression setting, the resulting differential richness fold changes are uncorrelated with genus-wide differential abundance statistics (right). Dashed lines represent confidence intervals. Points colored in red are the genus-specific differential richness inferences for the waste-water treatment survey.

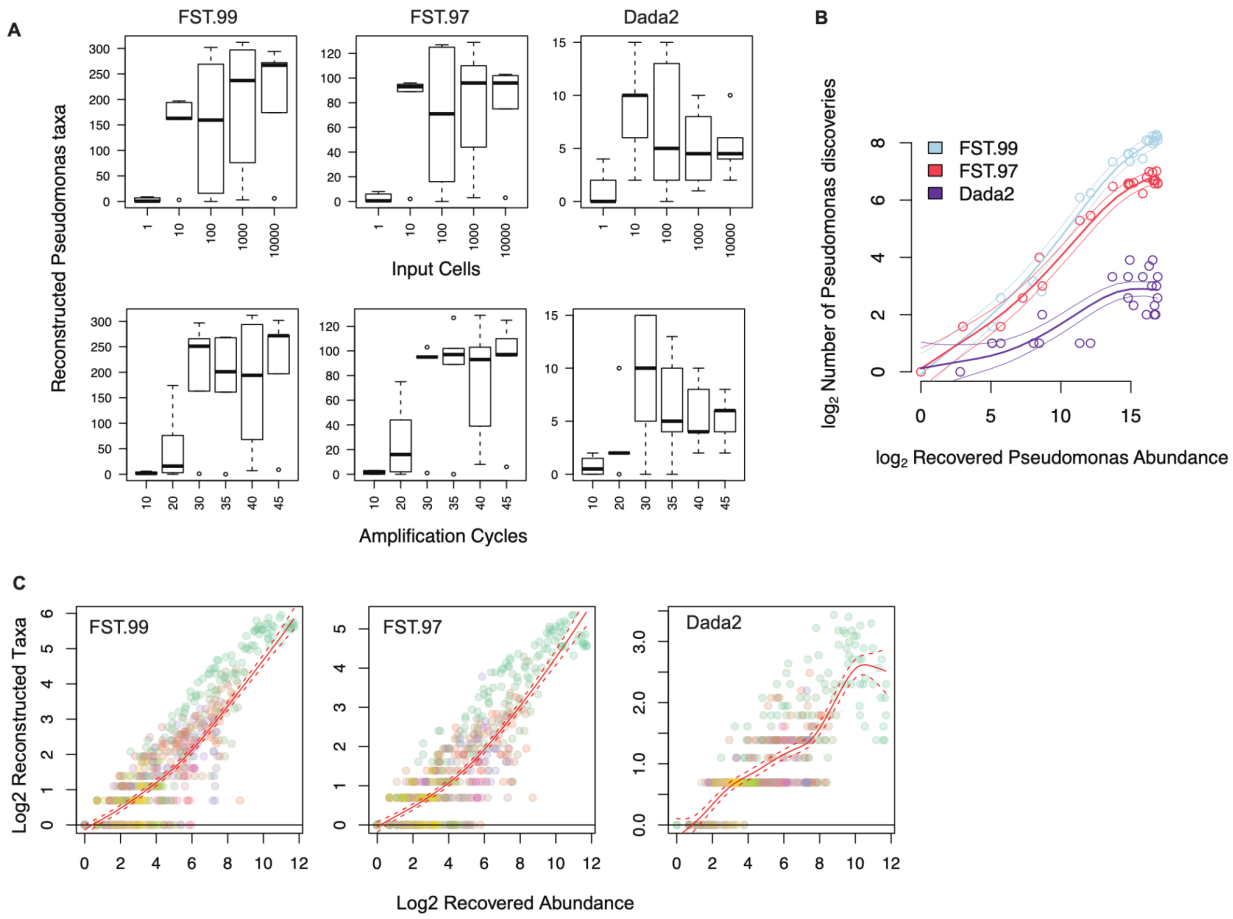

Figure 3. False microbial discoveries accumulate along the recovered abundance axis in the Pseudomonas dilution study. (A) For each taxa clustering method, the observed variation in within-genus Pseudomonas taxa accumulations are driven by experimental and technical parameters. Contaminant Pseudomonas are expected to fall with input loads, indicating false discovery accumulations at higher recovered Pseudomonas abundances. (B) The genus recovered abundance axis offers a succinct representation for taxa accumulations. Average and the 95% point-wise confidence intervals for the logged within-Pseudomonas taxa accumulation trends are shown with colored lines for each method, with colored circles indicating the respective observations. (C) An overlay of taxa accumulations across multiple detected genera in the study. Colors indicate genera.

Tables

Dataset	Year	16S segment, Sequencing& Clustering	Pseudo R^2_{trend}	Pseudo $R^2_{trend+design}$	AIC_{trend}	$AIC_{trend+design}$
Mouse [59, 60]	2009	V2, 454, FST.97	96.82%	98.68%	1.250×10^4	1.0273×10^4
Diarrhea [1]	2014	V12, 454, FST.99	98.52%	98.96%	1.2323×10^5	1.1455×10^5
Time series [2]	2014	V4, GAIIx, FST.97	94.43%	98.46%	3.9401×10^4	3.0813×10^4
Wastewater [48]	2018	V34, MiSeq, FST.99	91.70%	95.51%	3.7400×10^4	2.3682×10^4
MBQC-HLB$^{(97)}$ [27]	2017	V4, MiSeq, FST.97	97.15%	98.66%	1.0710×10^5	9.2275×10^4
MBQC-HLB$^{(99)}$ [27]	2017	V4, MiSeq, FST.99	98.67%	99.29%	1.089×10^5	9.6674×10^4
MBQC-HLB$^{(D)}$ [27]	2017	V4, MiSeq, Dada2	61.03%	81.72%	3.82×10^4	3.3877×10^4
PIH100$^{(97)}$ [47]	2020	V12, MiSeq, FST.97	94.04%	97.32%	1.5740×10^4	1.3858×10^4
PIH100$^{(99)}$ [47]	2020	V12, MiSeq, FST.99	97.66%	98.90%	1.7351×10^4	1.5739×10^4
PIH100$^{(D)}$ [47]	2020	V12, MiSeq, Dada2	88.37%	91.01%	8.8630×10^3	9.0263×10^3
Pseudomonas$^{(97)}$	2021	V12, MiSeq, FST.97	97.49%	98.92%	3.0762×10^3	2.8235×10^3
Pseudomonas$^{(99)}$	2021	V12, MiSeq, FST.99	98.71%	99.46%	3.3810×10^3	3.0717×10^3
Pseudomonas$^{(D)}$	2021	V12, MiSeq, Dada2	89.26%	93.77%	1881.62	1889.10

Table 1: Relative to study variables, within-genus taxa accumulation trends capture bulk of the systematic variation in 16S surveys' *genus-specific* taxa accumulations. For each 16S survey dataset mentioned in column 1, the year of publication is listed in column 2, the partial 16S segment targeted, machine technology and sequence clustering approach used are specified in column 3. FST.x refers to sequence clustering at an a priori fixed sequence similarity threshold of x%. McFadden's pseudo-R^2 for explaining genus-specific taxa accumulations with two negative binomial regressions (NB) are listed in columns 4 and 5. The fourth column is obtained when the NB regression includes within-genus taxa accumulation trend ($\hat{f}_R(\cdot)$, Methods) alone as predictor. The fifth column additionally includes the genus identifier, total sample depth, and experimental design matrix for each dataset as predictors (methods). Corresponding Akaike Information Criteria (AIC) are listed in columns 6 and 7.

Dataset	Year	16S segment, Sequencing& Clustering	Pseudo R^2_{trend}	Pseudo $R^2_{trend+design}$	AIC_{trend}	$AIC_{trend+design}$
Mouse [59, 60]	2009	V2, 454, FST.97	99.91%	99.92%	1.2971×10^3	1.2853×10^3
Diarrhea [1]	2014	V12, 454, FST.99	99.94%	99.94%	1.2508×10^4	1.2406×10^4
Time series [2]	2014	V4, GAIIx, FST.97	99.95%	99.95%	2.0972×10^3	1.9238×10^3
Wastewater [48]	2018	V34, MiSeq, FST.99	99.96%	99.97%	6.7384×10^2	6.3222×10^2
MBQC-HLB$^{(97)}$ [27]	2017	V4, MiSeq, FST.97	99.98%	99.99%	2.5052×10^3	2.4587×10^3
MBQC-HLB$^{(99)}$ [27]	2017	V4, MiSeq, FST.99	99.992%	99.993%	2.7456×10^3	2.6955×10^3
MBQC-HLB$^{(D)}$ [27]	2017	V4, MiSeq, Dada2	99.86%	99.03%	1.6748×10^3	1.5003×10^3
PIH100$^{(97)}$ [47]	2020	V12, MiSeq, FST.97	98.96%	99.07%	1.2430×10^3	1.1989×10^3
PIH100$^{(99))}$ [47]	2020	V12, MiSeq, FST.99	99.94%	99.95%	1.4471×10^3	1.3988×10^3
PIH100$^{(D)}$ [47]	2020	V12, MiSeq, Dada2	99.73%	99.73%	9.8011×10^2	9.8415×10^2
Pseudomonas$^{(97)}$	2021	V12, MiSeq, FST.97	99.94%	99.95%	3.0641×10^2	2.8237×10^2
Pseudomonas$^{(99)}$	2021	V12, MiSeq, FST.99	99.97%	99.98%	3.1263×10^2	2.9741×10^2
Pseudomonas$^{(D)}$	2021	V12, MiSeq, Dada2	99.83%	99.84%	1.8723×10^2	1.9202×10^2

Table 2: Relative to study variables, within-genus taxa accumulation trends capture bulk of the systematic variation in 16S surveys' *sample-wide* taxa accumulations. For each 16S survey dataset mentioned in column 1, the year of publication is listed in column 2, the partial 16S segment targeted, machine technology and sequence clustering approach used are specified in column 3. FST.x refers to sequence clustering at an a priori fixed sequence similarity threshold of x%. McFadden's pseudo-R^2 for explaining sample-wide taxa accumulations with two negative binomial regressions (NB) are listed in columns 4 and 5. The fourth column is obtained when the NB regression includes within-genus taxa accumulation trend ($\hat{f}_R(\cdot)$, Methods) alone as predictor. The fifth column additionally includes the total sample depth, and experimental design matrix for each dataset as predictors (methods). Corresponding Akaike Information Criteria (AIC) are listed in columns 6 and 7.

657 # References

658 1. Pop, M. *et al.* Diarrhea in young children from low-income countries leads to large-scale
659 alterations in intestinal microbiota composition. *Genome Biol.* **15**, R76 (2014).
660 2. David, L. A. *et al.* Host lifestyle affects human microbiota on daily timescales. *Genome Biol.*
661 15, 1–15 (2014).
662 3. Kostic, A. D. *et al.* The Dynamics of the Human Infant Gut Microbiome in Development and
663 in Progression toward Type 1 Diabetes. *Cell Host Microbe* 17, 260–273 (2015).
664 4. Riquelme, E. *et al.* Tumor Microbiome Diversity and Composition Influence Pancreatic
665 Cancer Outcomes. *Cell* 178, 795-806.e12 (2019).
666 5. Nejman, D. *et al.* The human tumor microbiome is composed of tumor type–specific
667 intracellular bacteria. *Science* 368, 973–980 (2020).
668 6. Poore, G. D. *et al.* Microbiome analyses of blood and tissues suggest cancer diagnostic
669 approach. *Nature* 579, 567–574 (2020).
670 7. Magurran, A. E. *Ecological Diversity and Its Measurement*. (Princeton University Press,
671 1988).
672 8. Magurran, A. E. & McGill, B. J. *Biological diversity: frontiers in measurement and*
673 *assessment*. (Oxford University Press, 2011).
674 9. Hooper, D. U. *et al.* A global synthesis reveals biodiversity loss as a major driver of
675 ecosystem change. *Nature* 486, 105–108 (2012).
676 10. Purvis, A. & Hector, A. Getting the measure of biodiversity. *Nature* 405, 212–219 (2000).
677 11. Fleishman, E., Noss, R. F. & Noon, B. R. Utility and limitations of species richness metrics
678 for conservation planning. *Ecol. Indic.* 6, 543–553 (2006).
679 12. Adams, W. M., Small, R. D. S. & Vickery, J. A. The impact of land use change on migrant
680 birds in the Sahel. *Biodiversity* 15, 101–108 (2014).
681 13. Hallmann, C. A., Foppen, R. P. B., van Turnhout, C. A. M., de Kroon, H. & Jongejans, E.
682 Declines in insectivorous birds are associated with high neonicotinoid concentrations.

683 *Nature* 511, 341–343 (2014).
684 14. Stanton, R. L., Morrissey, C. A. & Clark, R. G. Analysis of trends and agricultural drivers of
685 farmland bird declines in North America: A review. *Agric. Ecosyst. Environ.* 254, 244–254
686 (2018).
687 15. Inger, R. *et al.* Common European birds are declining rapidly while less abundant species'
688 numbers are rising. *Ecol. Lett.* 18, 28–36 (2015).
689 16. Sambell, C. E., Holland, G. J., Haslem, A. & Bennett, A. F. Diverse land-uses shape new
690 bird communities in a changing rural region. *Biodivers. Conserv.* 28, 3479–3496 (2019).
691 17. Spellerberg, I. F. *Monitoring Ecological Change*. (Cambridge University Press, 2005).
692 doi:10.1017/CBO9780511614699.
693 18. Adams, J. *Species richness: patterns in the diversity of life*. (Springer Science & Business
694 Media, 2010).
695 19. Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D. & Van Oudenaarden, A.
696 Regulation of noise in the expression of a single gene. *Nat. Genet.* 31, 69–73 (2002).
697 20. McFall-Ngai, M. *et al.* Animals in a bacterial world, a new imperative for the life sciences.
698 *Proc. Natl. Acad. Sci.* 110, 3229–3236 (2013).
699 21. Redford, K. H., Segre, J. A., Salafsky, N., Rio, C. M. del & McAloose, D. Conservation and
700 the Microbiome. *Conserv. Biol.* 26, 195–197 (2012).
701 22. Jiménez, R. R. & Sommer, S. The amphibian microbiome: natural range of variation,
702 pathogenic dysbiosis, and role in conservation. *Biodivers. Conserv.* 26, 763–786 (2017).
703 23. West, A. G. *et al.* The microbiome in threatened species conservation. *Biol. Conserv.* 229,
704 85–98 (2019).
705 24. Gotelli, N. J. & Colwell, R. K. Quantifying biodiversity: procedures and pitfalls in the
706 measurement and comparison of species richness. *Ecol. Lett.* 4, 379–391 (2001).
707 25. Gotelli, N. J. & Colwell, R. K. Estimating species richness. 16.
708 26. Willis, A., Bunge, J. & Whitman, T. Improved detection of changes in species richness in

709 high diversity microbial communities. *J. R. Stat. Soc. Ser. C Appl. Stat.* 66, 963–977 (2017).
710 27. Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the
711 Microbiome Quality Control (MBQC) project consortium. *Nat. Biotechnol.* 35, 1077–1086
712 (2017).
713 28. Chao, A. & Chiu, C.-H. Species Richness: Estimation and Comparison. in *Wiley StatsRef:*
714 *Statistics Reference Online* 1–26 (American Cancer Society, 2016).
715 doi:10.1002/9781118445112.stat03432.pub2.
716 29. Chao, A. Nonparametric Estimation of the Number of Classes in a Population. *Scand. J.*
717 *Stat.* 11, 265–270 (1984).
718 30. Bunge, J. & Fitzpatrick, M. Estimating the Number of Species: A Review. *J. Am. Stat.*
719 *Assoc.* 88, 364–373 (1993).
720 31. Bent, S. J. & Forney, L. J. The tragedy of the uncommon: understanding limitations in the
721 analysis of microbial diversity. *ISME J.* 2, 689–695 (2008).
722 32. Kunin, V., Engelbrektson, A., Ochman, H. & Hugenholtz, P. Wrinkles in the rare biosphere:
723 pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environ.*
724 *Microbiol.* 12, 118–123 (2010).
725 33. Huse, S. M., Welch, D. M., Morrison, H. G. & Sogin, M. L. Ironing out the wrinkles in the rare
726 biosphere through improved OTU clustering. *Environ. Microbiol.* 12, 1889–1898 (2010).
727 34. Schloss, P. D. Reintroducing mothur: 10 Years Later. *Appl. Environ. Microbiol.* 86, (2020).
728 35. Chiu, C.-H. & Chao, A. Estimating and comparing microbial diversity in the presence of
729 sequencing errors. *PeerJ* 4, e1634 (2016).
730 36. Willis, A. Species richness estimation with high diversity but spurious singletons. (2016).
731 37. Fox, G. E., Wisotzkey, J. D. & Jurtshuk JR, P. How close is close: 16S rRNA sequence
732 identity may not be sufficient to guarantee species identity. *Int. J. Syst. Evol. Microbiol.* 42,
733 166–170 (1992).
734 38. Janda, J. M. & Abbott, S. L. 16S rRNA Gene Sequencing for Bacterial Identification in the

735 Diagnostic Laboratory: Pluses, Perils, and Pitfalls. *J. Clin. Microbiol.* 45, 2761–2764 (2007).
736 39. Knight, R. *et al.* Best practices for analysing microbiomes. *Nat. Rev. Microbiol.* 16, 410–422
737 (2018).
738 40. Hillmann, B. *et al.* Evaluating the Information Content of Shallow Shotgun Metagenomics.
739 *mSystems* 3, (2018).
740 41. Johnson, J. S. *et al.* Evaluation of 16S rRNA gene sequencing for species and strain-level
741 microbiome analysis. *Nat. Commun.* 10, 5029 (2019).
742 42. Yarza, P. *et al.* Uniting the classification of cultured and uncultured bacteria and archaea
743 using 16S rRNA gene sequences. *Nat. Rev. Microbiol.* 12, 635–645 (2014).
744 43. Větrovskỳ, T. & Baldrian, P. The variability of the 16S rRNA gene in bacterial genomes and
745 its consequences for bacterial community analyses. *PloS One* 8, e57923 (2013).
746 44. Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D. & Konstantinidis, K. T. Strengths
747 and Limitations of 16S rRNA Gene Amplicon Sequencing in Revealing Temporal Microbial
748 Community Dynamics. *PLOS ONE* 9, e93827 (2014).
749 45. Harris, B. Determining bounds on integrals with applications to cataloging problems. *Ann.*
750 *Math. Stat.* 521–548 (1959).
751 46. Gelman, A. Struggles with Survey Weighting and Regression Modeling. *Stat. Sci.* 22,
752 (2007).
753 47. Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species
754 and the Number of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.*
755 12, 42–58 (1943).
756 48. Chao, A. & Lee, S.-M. Estimating the Number of Classes via Sample Coverage. *J. Am. Stat.*
757 *Assoc.* 87, 210–217 (1992).
758 49. Willis, A. & Bunge, J. Estimating diversity via frequency ratios. *Biometrics* 71, 1042–1049
759 (2015).
760 50. Paulson, J. N. *et al.* Paenibacillus infection with frequent viral coinfection contributes to

761 postinfectious hydrocephalus in Ugandan infants. *Sci. Transl. Med.* 12, (2020).
762 51. Kulkarni, P. *et al.* Conventional wastewater treatment and reuse site practices modify
763 bacterial community structure but do not eliminate some opportunistic pathogens in
764 reclaimed water. *Sci. Total Environ.* 639, 1126–1137 (2018).
765 52. Degnan, P. H. & Ochman, H. Illumina-based analysis of microbial community diversity.
766 *ISME J.* 6, 183–194 (2012).
767 53. Pinto, A. J. & Raskin, L. PCR biases distort bacterial and archaeal community structure in
768 pyrosequencing datasets. (2012).
769 54. Schloss, P. D., Gevers, D. & Westcott, S. L. Reducing the effects of PCR amplification and
770 sequencing artifacts on 16S rRNA-based studies. *PloS One* 6, e27310 (2011).
771 55. Haas, B. J. *et al.* Chimeric 16S rRNA sequence formation and detection in Sanger and 454-
772 pyrosequenced PCR amplicons. *Genome Res.* 21, 494–504 (2011).
773 56. Schloss, P. D. & Westcott, S. L. Assessing and Improving Methods Used in Operational
774 Taxonomic Unit-Based Approaches for 16S rRNA Gene Sequence Analysis. *Appl. Environ.*
775 *Microbiol.* 77, 3219–3226 (2011).
776 57. Sze, M. A. & Schloss, P. D. The Impact of DNA Polymerase and Number of Rounds of
777 Amplification in PCR on 16S rRNA Gene Sequence Data. *mSphere* 4, (2019).
778 58. Hughes, J. B., Hellmann, J. J., Ricketts, T. H. & Bohannan, B. J. M. Counting the
779 Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Appl. Environ.*
780 *Microbiol.* 67, 4399–4406 (2001).
781 59. Olson, N. D. *et al.* A framework for assessing 16S rRNA marker-gene survey data analysis
782 methods using mixtures. *Microbiome* 8, 1–18 (2020).
783 60. Prodan, A. *et al.* Comparing bioinformatic pipelines for microbial 16S rRNA amplicon
784 sequencing. *PLOS ONE* 15, e0227434 (2020).
785 61. Haegeman, B. *et al.* Robust estimation of microbial diversity in theory and in practice. *ISME*
786 *J.* 7, 1092–1101 (2013).

787 62. Flather, C. Fitting species–accumulation functions and assessing regional land use impacts
788 on avian diversity. *J. Biogeogr.* 23, 155–168 (1996).
789 63. Goeman, J. J. & Finos, L. The inheritance procedure: multiple testing of tree-structured
790 hypotheses. *Stat. Appl. Genet. Mol. Biol.* 11, (2012).
791 64. Meijer, R. J. & Goeman, J. J. A multiple testing method for hypotheses structured in a
792 directed acyclic graph. *Biom. J.* 57, 123–143 (2015).
793 65. Gu, C. *Smoothing spline ANOVA models*. (Springer, 2002).
794 66. Gu, C. *Smoothing spline ANOVA models*. vol. 297 (Springer Science & Business Media,
795 2013).
796 67. DiCiccio, T. J. & Efron, B. Bootstrap confidence intervals. *Stat. Sci.* 11, 189–228 (1996).
797 68. Oksanen, J. *et al.* The vegan package. *Community Ecol. Package* 10, 719 (2007).
798 69. Hsieh, T. C., Ma, K. H. & Chao, A. iNEXT: an R package for rarefaction and extrapolation of
799 species diversity (H ill numbers). *Methods Ecol. Evol.* 7, 1451–1456 (2016).
800 70. Weston, S. & Calaway, R. Getting Started with doParallel and foreach. *Vignette CRAN URL*
801 957, (2019).
802 71. Chang, S.-S., Hsu, H.-L., Cheng, J.-C. & Tseng, C.-P. An efficient strategy for broad-range
803 detection of low abundance bacteria without DNA decontamination of PCR reagents. *PloS*
804 *One* 6, e20303 (2011).
805 72. Paulson, J. N. *et al. metagenomeSeq: Statistical analysis for sparse high-throughput*
806 *sequencing*. (Bioconductor version: Release (3.13), 2021).
807 doi:10.18129/B9.bioc.metagenomeSeq.
808 73. Paulson, J. N., Bravo, H. C., Pop, M. & biocViews ExperimentData, S. Package 'msd16s'.
809 (2015).
810 74. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
811 data. *Bioinformatics* 30, 2114–2120 (2014).
812 75. Callahan, B. J. *et al.* DADA2: high-resolution sample inference from Illumina amplicon data.

Nat. Methods (2016).

76. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12 (2011).
77. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336 (2010).
78. Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. PEAR: a fast and accurate Illumina Paired-End reAd mergeR. *Bioinformatics* 30, 614–620 (2014).
79. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinforma. Oxf. Engl.* 26, 2460–2461 (2010).
80. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 41, D590–D596 (2013).
81. Caporaso, J. G. *et al.* PyNAST: a flexible tool for aligning sequences to a template alignment. *Bioinformatics* 26, 266–267 (2010).