MICROBIOME COMMUNITY INFERENCE IN LOW MICROBIAL BIOMASS ENVIRONMENTS

By

Vincent M. Caruso

A THESIS

Presented to the Department of Medical Informatics and Clinical Epidemiology
and the Oregon Health & Science University
School of Medicine
in partial fulfillment of the requirements for the degree of

Master of Science

March 2018

School of Medicine

Oregon Health & Science University

Certificate of Approval

This is to certify that the Master's Thesis of

Vincent M. Caruso

"Microbiome community inference in low microbial biomass environments"

Has been approved

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

First of all, I would like to thank my advisor, Lisa Karstens, for sharing her work and ideas, and then trusting me to make good judgments and do good research. I also want to thank the rest of my committee: Xubo Song, for having confidence in me when I didn't; Guanming Wu, for challenging me to be more rigorous; and Mark Asquith, for always reminding me of the biology that inspires our work. Thanks to Shannon McWeeney for brainstorming with me in the early stages and for being compassionate at the right moments, and to Diane Doctor for keeping me on track. Thanks to my friends for putting up with my long absences, and for being there when I showed up again. Thank you to my wonderful family, especially my parents, for supporting me in so many ways and mostly for just being there for me. Finally, special thanks to my girlfriend, who has stood by me, put up with my anxieties and obsessions with school, and remained supportive and loving throughout.

Abstract

Microbiome community composition is an important factor in human health, and while most microbiome research to date has focused on high microbial biomass communities, low biomass communities, such as found in the urinary tract, are also of medical importance. The relative scarcity of microbes in low biomass communities makes an accurate determination of community composition challenging, since contamination and technical noise may confound the true community signal when biomass is very low. However, the impact of these noise sources on microbial community inference is not well studied. The most common technique for estimating community composition is to sequence the 16S rRNA genes, which serve as phylogenetic barcodes for prokaryotes, in a sample. True community sequences, and hence composition, are then inferred from the noisy sequence data. OTU clustering algorithms have been the *de facto* standard for *in silico* community inference from 16S rRNA sequence data, but recently, several algorithms have emerged that use sequencer error models to improve inference. Here, we benchmarked six community inference software methods, including two OTU methods, three error model methods, and one entropy-based method, to assess their performance on samples of varying biomass. First, we compared the methods using four high microbial biomass mock communities to assess their baseline performance with "typical" microbiome data. Then, we compared the methods' performance on a dilution series made from a single mock community, to generate samples that varied only in

relative biomass. Sequences inferred by each method were classified according to their most likely source: community, technical noise, contamination, or unknown. With the high biomass data, we found that the error model methods showed both good sensitivity and specificity, whereas the other three methods showed relatively poor sensitivity, specificity, or both. Contamination was present in the inferred communities, but in very small proportions. With the dilution series data, contamination made up an increasing proportion of the inferred communities, dominating the composition of the most dilute sample, regardless of inference method. However, the error model methods showed a nearly linear association between the *number* of contaminant sequences inferred and sample concentration, whereas the entropy-based method exhibited a less linear relationship, and the OTU methods showed the least relationship between inferred contaminant sequences and sample biomass. Thus, we conclude that while no inference method on its own distinguishes true community from contaminant sequences, error model methods provide the most accurate characterization of community and contaminants, and further research into *in silico* contaminant identification using error model methods is warranted.

I. Introduction

A microbiome is a community of microbes— bacteria, archaea, viruses, and fungi— that coexist in a defined, small-scale environment. Examples of environments that host microbiomes include the human gut, skin, and mucosae, as well as soil, plant leaves, and bodies of water. The group of organisms that inhabit a given microbiome are referred to collectively as the microbiota. With the rapid growth of interest in their study, it seems that microbiomes are found almost wherever we look, including environments once thought to be sterile such as the urine,(1) the lungs,(2) and blood.(3)

Microbiomes have been closely tied to human health. Distinct microbial community compositions have been associated with a wide range of host phenotypes and diseases, from obesity to irritable bowel disease to Parkinson's disease.(4–6) In addition, there is ample evidence that the microbiota play an important role in training and regulating the immune system.(7) Although in most cases causal links have yet to be demonstrated between microbiome composition and an associated host phenotype, understanding these associations can nevertheless be of great potential benefit to the biomedical community: the human microbiome may help to diagnose and treat diseases, as well as predict response to treatments and long-term health outcomes.(8) In one example of using the microbiome for clinical application, fecal transplants have been successfully used to treat *Clostridium difficile* infections by re-establishing beneficial microbial populations.(9)

While the majority of human microbiome research to date has focused on body sites with relatively large microbial biomass, such as the gut, vagina, and oral cavity, microbial communities are also found at much lower abundance in a variety of other sites. Some examples of these low biomass microbiomes are found in urine and the urinary tract,(1) on the mucosae of the lungs,(2) and in the blood.(3) There is even evidence that the air surrounding each of us contains a cloud of microorganisms given off by our bodies.(10) Just as with the higher biomass microbiota, low biomass communities have been associated with a number of human disorders. Changes in the urinary microbiota relative to healthy population controls are seen in women who suffer from urgency urinary incontinence,(11) and distinct microbial communities are observed in patients with cystic fibrosis, asthma, and chronic obstructive pulmonary disease.(12) However, studying these low biomass communities presents special challenges to researchers. Because samples taken from urine or blood, for example, contain a relatively small amount of microbial material, accurate characterization of the community composition can be very difficult.(13,14)

Currently, the most common method for profiling a microbial community is to sequence the 16S rRNA gene. Found in all prokaryotes, the 16S rRNA gene consists of hypervariable regions, which serve as barcodes that are unique to distinct organisms, flanked by highly conserved regions that offer a target for PCR primers to isolate and amplify the region of interest in a wide range of organisms. DNA sequencing reads generated from the targeted 16S rRNA region are then processed by quality filtering and

sequence inference *in silico* to remove noise introduced by PCR and sequencing, as well as to account for intra-organism variation.

Since at least 2006,(15) clustering reads into operational taxonomic units (OTUs) has been the *de facto* standard for sequence inference with 16S rRNA data. With OTU methods, the researcher selects a radius of variability (typically 3%), within which sequence differences are assumed to be due to variation within the taxonomic group or random sequencer noise. All sequence reads within the chosen radius from each other are clustered into a single OTU, which is commonly inferred as representing one "organism" or taxonomic group. Recently, several methods have been published that take a different approach.(16–18) These algorithms, which we refer to here as error modeling methods, attempt to model the error of the sequencer, and then cluster reads so that their distribution within clusters is consistent with the error model. This approach avoids making any assumption about the variation within a taxonomic group (e.g., "intra-species 16S rRNA variation is 3%"), a weakness of OTU methods.(19) By considering both sequence similarity and abundance in the model, error modeling methods account for the broader error profiles (in terms of number of base-pair substitutions) that result from high-throughput, next-generation sequencing (NGS) experiments, which may produce many thousands of reads for a single template sequence. Hence, error modeling has the potential to simultaneously improve the sensitivity and specificity of 16S rRNA sequence inference compared to OTU methods.

As mentioned above, samples taken from an environment with low microbial biomass present distinct challenges to the researcher, and methods deemed appropriate for high biomass samples— both in the lab and in silico— may not transfer well to low biomass studies. When dealing with low biomass samples, there is less starting template DNA for the PCR reaction. Consequently, any contamination from extraction reagents or the laboratory environment makes up a larger fraction of the extracted sample when compared to high microbial biomass samples.(13) (Examples of genera commonly found as contaminants in sequencing experiments include *Bradyrhizobium*, *Rhizobium/Agrobacterium*, *Sphingomonas*, *Burkholderia*, *Microbacterium*, *Propionibacterium*, and *Pseudomonas*.(14,20)) Additionally, the greater number of PCR cycles typically required with low biomass samples may produce disproportionate quantities of chimeric and contaminant DNA sequences.(2) In other words, the sequencing of low biomass microbiome communities suffers from a low signal-to-noise ratio, a problem not encountered when sequencing high microbial biomass communities, since any contaminating sequences are overwhelmed by the community DNA in the latter case.

The low signal-to-noise ratio encountered when sequencing low microbial biomass samples can have serious implications for the biological, and even biomedical, interpretation of an experiment. On the one hand, if the strength of the noise— particularly noise from contaminating species— is too great, it may be mis-interpreted as forming an important component of the microbial community. Even noise from technical sources, such as sequencer substitution errors, can lead to inflated estimates

of sample diversity. On the other hand, the diminished strength of the community signal means that sparse but important community members may not be detected , or else may be dismissed as background noise. In the worst case, confounding noise might even lead to invalid conclusions about the biomedical role of a microbiome community, leading to a misguided understanding of the role of microbiomes in our health.

A few studies have investigated the effects of wet-lab variables on low biomass microbiome sequencing experiments, for example the impact of different quantities of starting material,(21,22) the contamination introduced by different DNA extraction and sequencing reagents,(14,23) and the effects of other potential sources of contamination.(24) Here, we instead propose to study the impact of *in silico* methods on the inference of community composition in low biomass samples. Specifically, we intend to compare the accuracy of various core algorithms in 16S rRNA read processing pipelines, namely OTU and error modeling methods, when presented with low biomass data. We only consider methods designed for or compatible with Illumina sequence data, since this has become the predominant sequencing platform,(25) and the vast majority of modern microbiome investigations currently use Illumina technology. First, we benchmark selected inference methods on mock microbial community samples with high microbial biomass to assess their baseline performance on data from high biomass communities. Then, we will compare the same methods on a mock community dilution series— a set of samples with successively decreasing DNA concentration— to observe how inference results change when sample biomass becomes small, and to evaluate how each method performs in this scenario.

II. Background

A variety of experimental techniques have been used to investigate microbiomes since at least the 1960s. The earliest studies used conventional culturing methods to grow bacteria from a sample, e.g., a swab of the oral cavity. The resulting cultures would then be examined microscopically and bacteria would be identified by their morphologic and biochemical characteristics.(26–28) However, culture methods suffer from the severe limitation that a vast majority of microorganisms will not grow in conventional culture media, a fact that was not well appreciated until culture-independent technologies were developed in the 1980s.(29) This constraint results in low sensitivity for culture experiments, and also led to the misconception that certain body sites (particularly those that are home to low biomass communities) were sterile. Another drawback to culture-based studies is their reliance on phenotype to distinguish organisms, observations that can be subjective and often ambiguous. The development of DNA probe hybridization technology provided one of the earliest culture-independent tools for querying microbiomes.(30) Although not constrained by whether or not an organism can be cultured, the sensitivity and precision of DNA probes are constrained by the set of probes used. In effect, this restriction means that only a relatively small set of known organisms can be detected by DNA hybridization.

DNA sequencing technologies represent a major advance in our ability to query microbiome communities, and the development of inexpensive, high-throughput next-generation sequencing (NGS) has fueled a huge growth in microbiome research over the

last decade. Two primary strategies exist to study microbial communities using NGS.(31) The first, shotgun whole-genome sequencing, attempts to sequence all genomic DNA present in a microbiome sample. By sequencing the entire metagenome (the genome of a community), this method affords the highest resolution of community characterization, in terms of distinguishing between similar organisms, and permits direct functional gene profiling of the microbiota. On the downside, it is expensive to achieve the sequencing depth required for shotgun metagenomic experiments. Also, when studying human microbiomes, detection of microbes is hindered by the fact that a majority of the sample DNA is contamination from the human host. In addition, analysis of the data is difficult, requiring removal of human read sequences and accurate read mapping and/or assembly of a large number of microbial genomes, all of which requires substantial computational resources.(32)

Targeted sequencing of the prokaryotic 16S rRNA gene avoids many of the difficulties encountered with shotgun sequencing, while still offering reasonable sensitivity and precision for community profiling. Found in all bacteria and archaea, the 16S rRNA gene includes regions of DNA that are highly conserved across organisms, offering an ideal target for well-designed PCR primers to isolate and amplify a gene common to community members. Other regions of the 16S rRNA gene sequence, called hypervariable regions, vary such that they can be used as barcodes to distinguish organisms from each other. Thus, by targeting and sequencing 16S rRNA genes (so-called “amplicon” sequencing) from a human host sample, a census of the bacteria living in a body habitat can be taken. Although not as precise as shotgun metagenomes, 16S

rRNA hypervariable regions are quite phylogenetically informative, typically permitting classification to genus level, with species or even strain-level distinctions sometimes possible.(16) The much smaller coverage requirement of amplicons relative to whole-genome sequencing means that 16S rRNA experiments are much less expensive to carry out. Also, the data sets produced are smaller, and don't require assembly or mapping of reads, making the analysis more tractable with only modest computational resources. Because of this accessibility and the high degree of phylogenetic information it provides, 16S rRNA sequencing is the most popular technique for studying the human microbiome, and it is this sequencing strategy that is the focus of our study.

Once a 16S rRNA sequencing experiment has been performed and the raw data obtained, computational analysis is required to extract meaningful information from the read sequences. This analysis can be separated into two stages: primary analysis, in which the raw read sequences are processed to infer the community composition; and secondary, or "downstream", analysis, which may include estimating community properties such as alpha and beta diversity, comparing samples to detect significant differences, and conducting network analysis. Here, we will address the primary processing of raw reads to infer community composition. Specifically, we will focus on data generated by the Illumina sequencing platform, which has become the predominant sequencing technology for high-throughput sequencing experiments, including microbiome experiments.(25)

The simplest approach to 16S rRNA read processing is to simply group sequences by identity, counting the number of times each unique read occurs and considering each unique read to represent a distinct organism. However, this approach is flawed because sequence data contain non-trivial errors and noise from both technical and biological sources. One source of technical error is introduced by PCR (necessary to isolate and amplify a 16S rRNA hypervariable region), and occurs when two 16S rRNA fragments from different parent sequences become fused to form a hybrid sequence, or chimera, which is not biological. Another source of technical artifacts is the sequencer, which has a finite error rate for individual base calls (on the order of 0.1% for typical Illumina reads). Some degree of biological error, either from DNA contamination of extraction reagents or from imperfectly sterile lab conditions, is also likely to be present. As a result of these various error sources, samples of a microbial community known to have only a few distinct organisms routinely produce tens of thousands or more unique read sequences.

To reduce the complexity of noisy data and allow for inference of sample composition, the most common strategy is to cluster sequences above a certain threshold of similarity into operational taxonomic units, or OTUs. **Figure 1** shows an example of OTU clustering. Typically, a threshold of 97% similarity is used, which is presumed to correspond to the natural variation within bacterial species. Thus, OTUs reduce spurious sequence identification by clustering reads presumed to differ only due to intra-species variation or technical artifacts, i.e., sequencer error. Representative sequences from each OTU cluster can then be mapped to a 16S rRNA reference database for taxonomic

identification. Such databases have been compiled over several decades and represent a compilation of millions of 16S rRNA sequences that have been deposited by researchers from the microbiology community.(33)

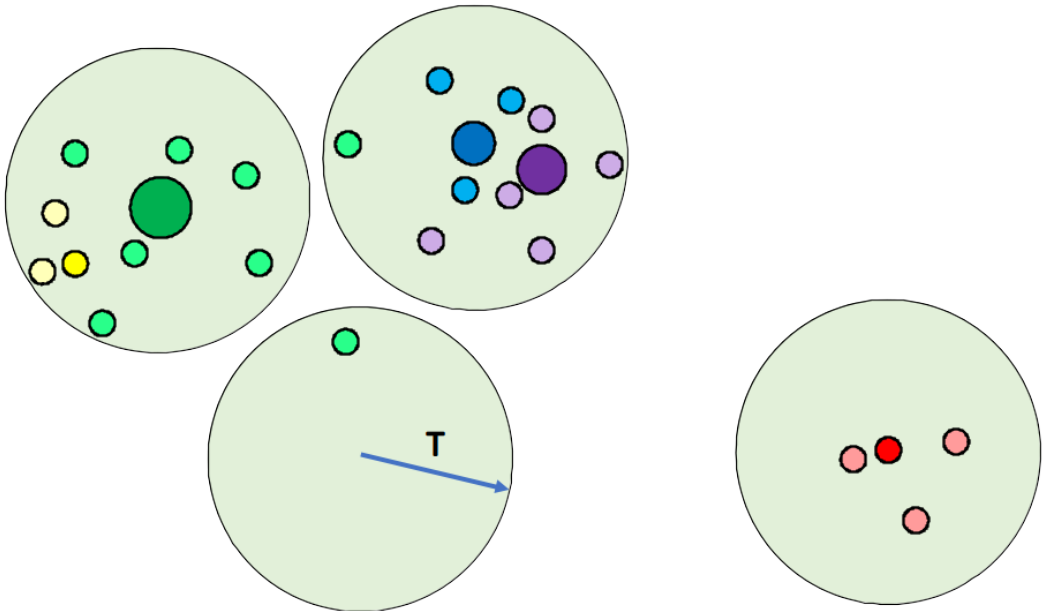

Figure 1: Example of OTU clustering. OTU clustering methods group read sequences using a fixed similarity threshold. Each dark shaded dot represents the true 16S rRNA sequence of an organism, lighter shaded dots represent read sequences with errors (derived from the true sequence of the same color). Distance between dots shows degree of difference between sequences, size of dots represents relative abundance of reads in dataset. Light grey circles are OTU clusters, with fixed similarity threshold shown by **T**. OTUs may lump organisms with similar sequences into a single cluster (as shows for blue and violet sequences above, as well as green and yellow sequences), thereby detecting only one "organism" when multiple are present. OTUs may also split reads with sequencing errors into distinct clusters (as shown by the light green sequences above), potentially resulting in false positive detections.

One issue with OTU clustering is that the fixed similarity threshold corresponds poorly to biological taxa. For example, some genera exhibit wide intra-genus variation in their 16S rRNA sequences, whereas others vary only slightly,(19) so clustering with a fixed

similarity threshold arbitrarily splits some taxa while lumping others with very similar sequences. This lumping of taxa also means that OTU clustering limits the ability to resolve closely related sequences that may represent meaningful phylogenetic differences. Similarly, the fixed threshold of OTUs does a poor job of removing sequencer error, as it does not account for the possibility that an erroneous sequence could differ from its true sequence by more than the threshold, an event that is much more likely as sequencing depth and throughput have increased with improved technologies. As a consequence, OTU clustering tends to greatly inflate the number of distinct organisms present in a sample, with estimates in the thousands commonly given even for samples likely to be of much lower diversity.(34)

Recently, a number of new methods have emerged to overcome the limitations of OTU clustering.(16–18,35) Rather than simply considering similarity to cluster reads, these algorithms use sequencer error models to distinguish erroneous reads from true biological reads. **Figure 2** shows an example of error model inference. Some of these error modeling methods employ *a priori* models of typical sequencer error profiles, while another approach is to use an adaptive model estimated from the data themselves. In addition to read similarity, error modeling methods take into account the frequency distribution of sequences, and may also consider quality scores reported by the sequencer. Once the set of biological read sequences has been inferred, all variation is considered to be biologically meaningful. Further clustering of sequences into OTUs is then possible, but may be undesirable, as it only reduces the information content and hence the resolution for organism identification. With both simulated and mock

community data, error modeling methods have been shown to increase the ability to resolve closely related organisms while decreasing the number of spurious identifications relative to OTU methods.(16,35)

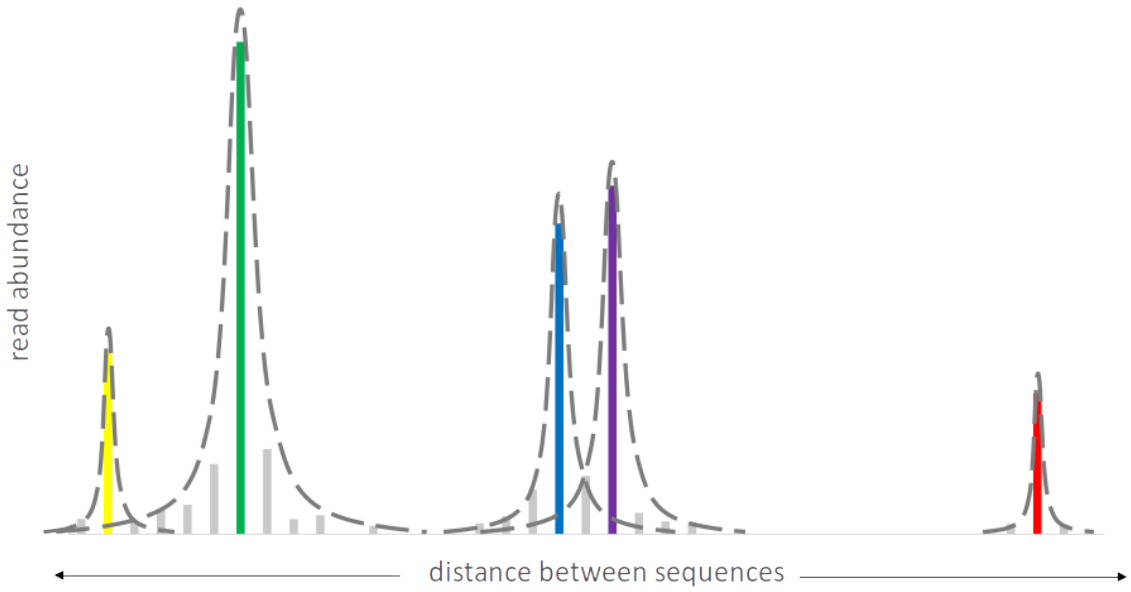

Figure 2: Example of error model sequence inference. Error modeling methods use an explicit model of sequencer error to infer community sequences. Each colored bar represents the true 16S rRNA sequence of an organism, light grey bars represent read sequences with errors. Distance between bars (x-axis) shows degree of difference between sequences, height of bars represents relative abundance of reads in dataset. Dark grey lines show a hypothetical error model, representing the maximum expected frequency of sequencer error as distance from the central sequence varies. The error model correctly infers the community composition, since each true sequence is too abundant to fall under the error profile of another sequence, while all sequences with errors are consistent with the error model for one of the true sequences.

For this study, we have selected six sequence inference methods for comparison, including two representative OTU methods, an OTU alternative that uses an

information-theoretic entropy-based approach, and three error modeling methods. The two OTU methods are called UCLUST (36) and UPARSE,(37) the entropy-based method is named Minimum Entropy Decomposition (MED),(35) and the error modeling methods are UNOISE,(17) Deblur,(18) and the Divisive Amplicon Denoising Algorithm 2 (DADA2).(16) Each is described in more detail below.

UCLUST is an OTU clustering method that we chose because it has been widely used and cited in microbiome research. Its widespread use is largely due to its inclusion, until recently, as the default inference method in the Quantitative Insights Into Microbial Ecology (QIIME) software suite,(38) a popular microbiome analysis tool. The UCLUST algorithm in QIIME begins with an abundance-sorted list of sequences. It then aligns those sequences against a database of cluster seeds, which is initially empty. Sequences are greedily clustered if they are within the radius of variation (typically 3%, or 97% similarity) of existing seeds, otherwise they become new seeds.

The second OTU method chosen, UPARSE is a newer algorithm published in 2013 by the same author as UCLUST.(36,37) UPARSE uses the same greedy clustering strategy, but precedes it with a stringent quality filtering step, which excludes lower-quality read sequences from the set used to form cluster seeds. It also removes chimeric sequences concurrently with the clustering stage (rather than prior to clustering as with UCLUST) if a query sequence is best explained as a chimera of existing cluster seeds. We selected UPARSE because it has been shown to greatly reduce the inflation of community diversity estimates that result from most OTU clustering algorithms.(34)

One of the first published 16S rRNA sequence processing methods to use a different strategy from the OTU approach was MED.(35) MED begins by placing all sequences in a single cluster and aligning them. The Shannon entropy of each alignment column is computed, and if any column has entropy higher than a threshold (computed dynamically for each cluster), the cluster is divided so as to make the entropy of the offending column zero in each new cluster. New entropy thresholds are computed, and cluster division repeats until all clusters have entropy below the threshold. MED then removes clusters whose abundance is below a minimum threshold (0.02% of all dataset reads by default), considering these to be noise. Included in the entropy calculation is a consideration of read and nucleotide frequency profiles, and thus the entropy threshold serves as an implicit model of sequencer error. **Figure 3** shows an example of entropy-based clustering.

MED was evaluated on two biological datasets,(35) and compared to UCLUST and a method that performs direct taxonomic classification of unique reads (GAST).(39) In each case, MED resolved single OTUs (from UCLUST) or single taxa (from GAST) into multiple nodes that corresponded to distinct species or host sites, demonstrating that MED has higher sensitivity for resolving ecologically meaningful information from 16S rRNA data when compared to OTU clustering. MED is included in this study for its potential to give distinct results from either OTU or error modeling methods.

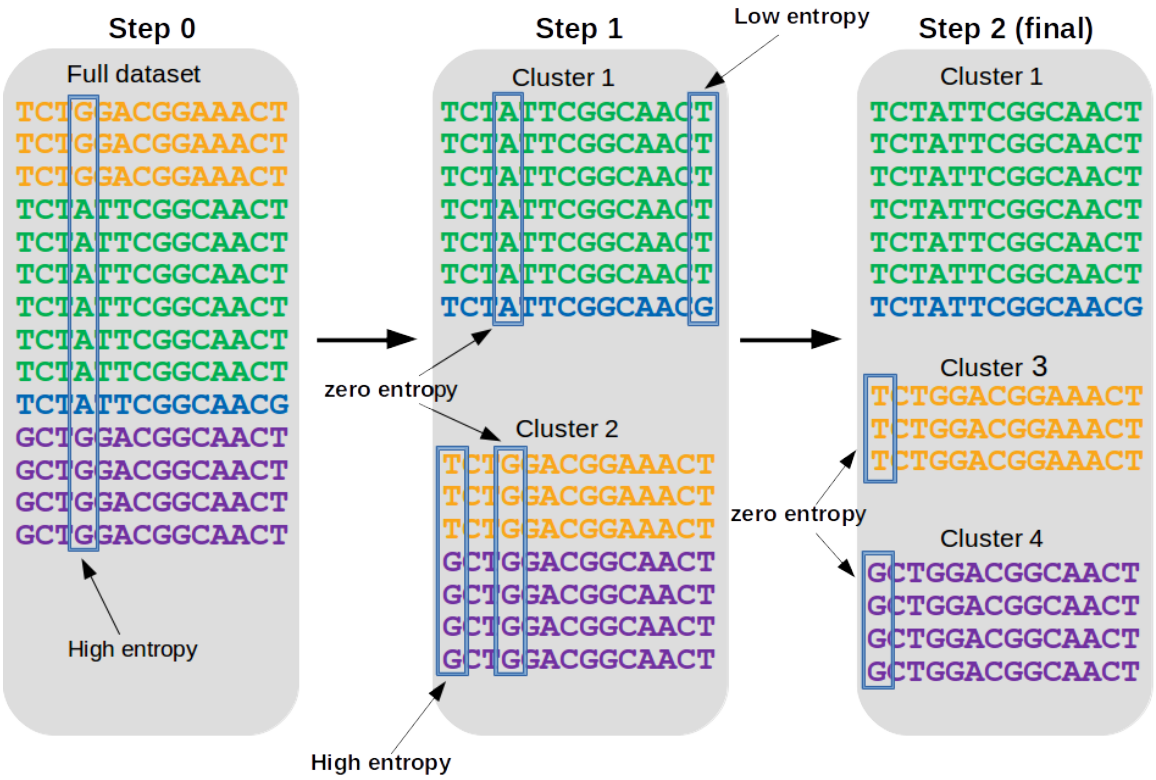

Figure 3: Example of entropy-based clustering. Entropy-based clustering considers the Shannon entropy (amount of variation) of individual columns of the aligned sequences. Sequences with the same color above are identical. If the entropy of a column is above a designated threshold, the cluster is split so as to make the entropy of the column zero in the new clusters. Columns with entropy below the threshold do not lead to cluster splitting. Cluster splitting stops when all columns in all clusters have entropy below the threshold.

Published in 2016, the first stand-alone method developed for Illumina sequencing data to use an explicit error model for sequence inference was DADA2.(16) DADA2 uses a multivariate Poisson distribution to model errors in sequencing reads, and the model parameters are estimated empirically from the data by considering quality scores reported by the sequencer. Beginning with a worst-case assumption for the error model, the algorithm alternates between clustering sequences given the error model

and estimating the error model given the clustering, until convergence. During clustering, all sequences begin in a single cluster whose centroid is the most abundant sequence, inferred to be the correct template sequence. The probabilities that all other sequences were derived from the centroid, given the error model, is calculated. If the least probable sequence is below a *p*-value threshold ($1x10^{-40}$ by default), it forms a new cluster centroid, and sequences are re-assigned to their most likely cluster. Cluster division repeats until all clusters are consistent with the error model.

The authors compared DADA2 to MED and three OTU clustering methods. On three mock community datasets, DADA2 outperformed the other methods in terms of both sensitivity and specificity: it resolved finer variation than any of the OTU methods, detected rare variants that were discarded by MED's noise-filtering abundance threshold, and generally output fewer false positive sequences than the other methods on each of the three datasets. On a natural microbiome dataset, DADA2 identified six variants of a single species that showed evidence of being ecologically distinct, which would have been lumped into a single OTU owing to their highly similar 16S rRNA sequences. This showed that using an error model, higher resolution of 16S rRNA data is possible without reducing specificity.

Another denoising algorithm, UNOISE, was originally published in 2015 (40) and later updated to its current version (UNOISE2) in 2016.(17) Similar to DADA2, UNOISE uses a parametric error model, but rather than a multivariate Poisson model, UNOISE employs

a simpler power-law distribution to approximate sequencer error, with a single user-specified tuning parameter.

UNOISE was benchmarked on three mock community datasets and three biological datasets, and results compared to those of DADA2. In general, the performance of UNOISE was very similar to that of DADA2, and in some cases UNOISE produced apparently better results. The author's evaluation showed that a simplified error model that does not consider sequence quality scores can perform comparably to one that uses quality scores to estimate many parameters (DADA2), thus requiring additional intensive computation.

Like DADA2 and UNOISE2, Deblur, published in 2017,(18) considers read frequency and inter-sequence distance to predict erroneous sequences. Instead of a smooth parametric model, Deblur employs stepwise, empirical thresholds that specify the maximum number of erroneous sequences expected at each Hamming distance (number of nucleotide substitutions) from a given query sequence, up to a maximum distance of 11. In decreasing order of abundance, each sequence's expected error abundances are computed and subtracted from the abundances of neighboring sequences. Sequences whose abundance remains above zero after all subtractions have been done are inferred as the correct template sequences.

Deblur's performance was compared to DADA2 and UNOISE on simulated, mock community, and real biological datasets. All three methods achieved high accuracy on the simulated data, with Deblur and UNOISE showing somewhat better sensitivity, and

all methods gave results close to the ground truth on the mock community data, although DADA2 predicted the fewest, and Deblur the most, false positives. Evaluated for stability on technical replicates, Deblur achieved a higher overlap of inferred sequences between replicates than DADA2 (UNOISE was not tested). Deblur thus represents another strong alternative to OTU clustering, and it was additionally shown to run an order of magnitude faster than DADA2 owing to its simple, *a priori* model.

Each of these six methods will be benchmarked according to the following two specific aims:

Aim 1

For the first aim, two methods representative of the OTU clustering algorithm, UCLUST and UPARSE, as well as MED and three error modeling methods, will be evaluated on several mock communities with high microbial biomass. The goal of this aim is twofold: first, to provide a benchmark of these methods independent of the authors who published them; and second, to establish the baseline performance of each method on "typical", high microbial biomass samples. We hypothesize that error-modeling methods will be superior to clustering methods for this purpose.

Aim 2

For the second aim, I will evaluate the same methods on a dilution series, which mimics successively lower sample biomass. There are also two goals for this stage. The first is to identify which, if any, methods perform consistently better, or better at low

concentrations. The second is to observe the extent to which low microbial biomass affects community inference, regardless of method.

The overarching goal is not only to observe the extent to which sample biomass alone can alter the community composition inferred by a given method, but also to assess the potential of various inference strategies for identifying and removing contaminating sequences computationally. Our hypothesis is that error modeling methods will be both more sensitive and specific than OTU methods, regardless of starting biomass. We also anticipate that decreasing the starting DNA concentration will lead to an increase in the inference of spurious and contaminant sequences due to the lower signal-to-noise ratio, but that error modeling will more accurately identify the true contamination present.

This study is motivated by the desire to characterize the community of low biomass microbiomes as accurately as possible using existing 16S rRNA data processing methods, in order to strengthen the validity of association between distinct microbiota profiles and phenotypes (including diseases) of interest. While a number of studies have compared the performance of these methods on typical data sets, to our knowledge this is the first to systematically examine their ability to process low biomass sample data.

III. Methods

Experimental design

To assess the performance of the six selected methods, we benchmarked each method on a number of mock community datasets. Mock communities are used instead of "natural" biological samples (such as stool or urine) because the community composition of a mock community is known in advance, which allows a more objective assessment of the accuracy of the community inferred by each method. We first compared the six methods on four high biomass (undiluted) mock community datasets, which served two purposes: first, to give an independent assessment of the merits of these methods; and second, to show the baseline performance of each method on samples representative of high microbial biomass communities (such as the gut microbiota). Three of these datasets, referred to here as "Kozich", "Schirmer", and "D'Amore" are from previously published studies.(41–43) The fourth dataset, which we call "Zymo", was generated for this study by Mark Asquith's lab at the Oregon Health & Science University.

We next benchmarked each method on a mock community dilution series, prepared from the Zymo mock community by performing eight serial dilutions (each having 1/3 the DNA concentration of the previous dilution), for a total of nine samples. The dilution series mimics samples of successively lower biomass, and allowed us to observe how each method's inference results change as biomass decreases.

Datasets

The Kozich mock community (41) comprises equal concentrations of 21 different bacterial strains, and the dataset is available as run 130403 at https://www.mothur.org/MiSeqDevelopmentData.html. The Schirmer mock community (42) was made from equal proportions of 57 prokaryotic strains (both archaea and bacteria). The dataset was obtained from the European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena), project accession PRJEB6244, run accession ERR777695 (sample metaID-35). The D'Amore mock community (43) contains the same 57 prokaryotic strains as the Schirmer community, but the D'Amore sample has DNA quantities that vary according to a logarithmic distribution. The data are also available from ENA project accession PRJEB6244, run accession ERR777739 (sample metaID-88). The Zymo community includes 8 strains with equal genomic proportions. **Table 1** gives a concise summary of these high microbial biomass mock communities.

Table 1: High microbial biomass datasets

Dataset name	Source	Number of strains	Genomic proportions	Raw reads
Kozich	Kozich, et al.(41)	21	equal	269.8k
Schirmer	Schirmer, et al.(42)	57	equal	593.9k
D'Amore	D'Amore, et al.(43)	57	logarithmic	262.1k
Zymo	Zymo Research	8	equal	427.2k

The Zymo mock community, including the high biomass (undiluted) sample and each of the dilution series samples, comprises 8 bacterial strains in equal proportions (see **Table 2** for community composition), and was prepared for this study from the ZymoBIOMICS

Microbial Community Standard, available from Zymo Research (https://www.zymoresearch.com/zymobiomics-community-standard). This mock community consists of both Gram positive and Gram negative bacteria in addition to yeast species and thus is a useful tool to ensure the success of DNA extraction (and subsequent sequencing) from a broad diversity of micro-organisms which may have intrinsic biological properties (eg. cell wall thickness) that make them more or less refractory to DNA isolation. DNA was extracted from the microbial standard with the Qiagen DNeasy blood & tissue kit following the manufacturer's recommended protocol. In brief, bacterial cells were lysed by mechanical and enzymatic lysis, followed by removal of the inhibitor by precipitation. DNA was collected by passing it through a DNA binding column. For further purification, binding products were washed to remove contaminants, and purified DNA was collected by elution.

Table 2: Zymo mock community composition

Species	Average GC (%)	Gram stain	gDNA abundance (%)
Pseudomonas aeruginosa	66.2	–	12
Escherichia coli	56.8	–	12
Salmonella enterica	52.2	–	12
Lactobacillus fermentum	52.8	+	12
Enterococcus faecalis	37.5	+	12
Staphylococcus aureus	32.7	+	12
Listeria monocytogenes	38.0	+	12
Bacillus subtilis	43.8	+	12
*Saccharomyces cerevisiae**	38.4	Yeast	2
*Cryptococcus neoformans**	48.2	Yeast	2

** These species do not have 16S rRNA genes, and hence were not sequenced in this experiment*

Subsequently, eight serial dilutions were made from the extracted DNA, where each successive aliquot was diluted with molecular grade water to 1/3 of its previous concentration, resulting in nine total samples with the following concentrations relative to the original extraction: 1/1 (undiluted), 1/3, 1/9, 1/81, 1/243, 1/729, 1/2187, and 1/6561. The V4 region of the 16S rRNA gene was amplified by PCR using Golay barcodes and the 515FB/806RB primer pair.(44–48) PCR was performed in triplicate for 35 cycles [verify] with ProMega hottaq polymerase (M5005), and amplification products confirmed with gel electrophoresis. Amplified DNA was purified with the Qiagen QIAquick PCR Purification kit. Samples were normalized to a concentration of 10 ng/µL, pooled, and sequenced on an Illumina MiSeq using a Reagent Kit V2 to generate 2x251 base-pair reads.

Sequence pre-processing

Prior to clustering, reads were first trimmed, merged, and filtered to remove low quality data. In all datasets, the first 15 nucleotides from the 5' end, which often contain pathological errors, were removed, as well as the low quality 3' tails, which varied by dataset (trim positions of forward/reverse reads: Zymo 230/210, D'Amore 250/240, Kozich 240/220, and Schirmer 240/220). After trimming, forward and reverse reads were merged and then filtered to remove low quality sequences. Merging was performed with the USEARCH (36) *fastq_mergepairs* command, with a maximum of 10 differences (*fastq_maxdiffs* = 10). To ensure that only sequences from the V4 region of the 16S rRNA were retained, merged sequences were removed if their lengths were outside the expected range for the primer pair used. These ranges are 220-225 bp for

Zymo and Kozich reads, and 258-263 bp for D'Amore and Schirmer reads (as the latter used a different primer pair that targets a longer V4 sequence). Merged sequences were further filtered to remove those with more than 2 expected errors, based on the posterior Q-scores computed by USEARCH. An exception to this protocol is for the DADA2 pipeline, in which forward and reverse reads are filtered independently, and only merged *after* sequence inference. In order to retain a similar proportion of the data as for the other methods, forward and reverse reads were filtered with a higher maximum of expected errors for the DADA2 pipeline (forward/reverse maximum errors: Zymo 2.5/2.5, D'Amore 2.5/2.5, Kozich 2.5/3.0, and Schirmer 2.5/2.5).

Sequence inference

Each of the six clustering methods was run with default parameters on each preprocessed datasets. The primary commands used for each method, as well as any additional required parameters, are described here. For the QIIME-UCLUST method, chimeras were first removed with *identify_chimeric_seqs.py* using the UCHIME (49) method (option *-m usearch61*) with the *gold.fa* reference database (available at http://drive5.com/otupipe/gold.tz). Sequences were then clustered *de novo* with the *pick_de_novo_otus.py* command and the default *uclust* algorithm. The UPARSE method was executed by calling the *cluster_otus* command in USEARCH (which concurrently removes chimeras), and then mapping reads to cluster seeds with the *otutab* command. Similarly, the UNOISE method was run by calling *unoise3* in USEARCH (which also removes chimeras), and reads were mapped to centroids with the *otutab* command. The MED method was run by invoking the *decompose* command within the Oligotyping

Pipeline.(50) As MED does not include native chimera removal, chimeras were removed with *uchime2_denovo* in USEARCH after sorting clusters by size. Deblur was run by calling *workflow* within the Deblur package, with the trim option (*-t*) set to the lower bound of the merge length window mentioned above to guarantee that all sequences have the same length, a requirement of Deblur. The DADA2 method was run with a custom R script based on the dada2 library: first, error rates were estimated with the *learnErrors* command; dereplicated reads were then clustered with *dada*, merged with *mergePairs*, and tabled with *makeSequenceTable*; sequences outside the allowed merge length window (see *Sequence preprocessing* above) were removed, and chimeras were removed with *removeBimeraDenovo*.

Evaluation

To evaluate the results from each processing method, we classified sequences into five categories, using a scheme similar to that used by Edgar (17) and by Callahan, et al.(16) Sequences that exactly matched a reference sequence from the known community were classified as "Reference". Those that differed from a more abundant Reference sequence by up to 10 nucleotides were labeled "Ref Noisy", as these are likely sequences with errors derived from Reference sequences that were incorrectly inferred as distinct. (It is also possible that such sequences arose via contamination or mutation, but this is less likely and not possible to determine conclusively.) Any unclassified sequences were compared to the National Center for Biotechnology Information's Nucleotide (NT) database (51) using BLAST.(52) The NT database is a very large collection of DNA sequences from a variety of sources. Sequences that matched an NT

sequence exactly were classified as "Contaminant", as these likely represent correctly identified contaminanting DNA in the sample. Sequences that were up to 10 nucleotides off from a more abundant Contaminant sequence were dubbed "Contam Noisy". All remaining sequences were labeled "Other", and may include unaccounted-for PCR artifacts (such as chimeras) and sequencing noise.

To obtain results at the taxonomic level, taxonomy was assigned using both *assignTaxonomy* and *assignSpecies* in the dada2 R library. For each function, the appropriately formatted Silva database (53) version 132 file was used (available at https://zenodo.org/record/1172783#.Wrb4f9ZG2kA). The resulting labels were then merged to obtain the highest available resolution, replacing any missing values for genus in one set with non-missing values in the other wherever possible. Labels at each taxonomic level were concatenated to generate a single label for each sequence, and these compound labels used to determine the number of unique taxa in each sequence category.

We further summarized results by computing recall and precision at the inferred sequence level, as well as the proportion of reads mapped to each sequence category. For sequence-level statistics, we use an observation-versus-expectation criterion: sequences that are both expected and observed (i.e., observed Reference sequences) are true positives (TP), those expected but not observed (unobserved References) are false negatives (FN), and those observed but not expected (all non-References) are false positives (FP). Sequence-level statistics give a sense of the accuracy of community

diversity estimates, whereas read-level statistics measure the bulk accuracy of a clustering algorithm.

Software

QIIME-UCLUST was implemented with scripts are from QIIME v1.9.1.(38) The *identify_chimeric_seqs.py* QIIME script requires USEARCH (36) v6.1.544, and *pick_de_novo_otus.py* calls the PyNAST alignment tool (v0.1).(54) UPARSE and UNOISE were implemented with USEARCH v10.0.240.(36) For the MED pipeline, clustering was done with v2.1 of the Oligotyping Pipeline (50); however, chimera removal was done with USEARCH version 9.2.64 (due to a known bug in v10.0.240). The Deblur pipeline uses Deblur v1.0.3,(18) which depends on VSEARCH v2.5.0,(55) MAFFT v7.3.10,(56) and SortMeRNA v2.0.(57) DADA2 was implemented in R with v1.6.0 of the dada2 package.(16) All analysis of clustering results was completed in R v3.4.3.(58) The analysis and all pipeline scripts are available at https://github.com/carusov/noisy-microbes/tree/master/community-inference/scripts.

IV. Results

High microbial biomass mock communities

We firs t benchmarked each method on the four undiluted mock community datasets to establish their baseline performance characteristics on "typical" data, representative of high microbial biomass samples. The number of sequences inferred by each method for each dataset, and their classification according the scheme described above, is shown in **Table 3**.

Table 3: Sequences inferred for high biomass communities

Dataset	Method	Total	Reference	Ref Noisy	Contaminant	Contam Noisy	Other
Zymo (8 strains)	UCLUST	200	8	78	47	1	66
	UPARSE	69	8	2	35	0	24
	MED	57	8	49	0	0	0
	UNOISE	12	8	2	1	0	1
	Deblur	8	8	0	0	0	0
	DADA2	20	8	3	5	0	4
Kozich (21 strains)	UCLUST	191	20	42	102	4	23
	UPARSE	101	20	1	75	0	5
	MED	46	22	21	3	0	0
	UNOISE	40	21	1	17	0	1
	Deblur	32	20	0	11	0	1
	DADA2	56	22	1	31	0	2
Schirmer (57 strains)	UCLUST	185	46	68	28	4	39
	UPARSE	77	46	1	26	0	4
	MED	65	56	3	6	0	0
	UNOISE	78	57	0	20	1	0
	Deblur	71	54	0	16	1	0
	DADA2	88	57	2	28	0	1
D'Amore (57 strains)	UCLUST	66	42	4	16	0	4
	UPARSE	58	42	0	15	0	1
	MED	55	50	2	3	0	0
	UNOISE	59	51	0	8	0	0
	Deblur	56	48	0	8	0	0
	DADA2	66	51	0	15	0	0

A few trends are apparent in the categorized inference results. UCLUST reports the largest number of sequences in all cases, while Deblur reports the fewest (for Zymo and Kozich) or second-fewest (for Schirmer and D'Amore). MED finds the fewest sequences on the Schirmer and D'Amore datasets but ranges in the middle on the Zymo and Kozich datasets. Among the error modeling methods, DADA2 detects the most sequences. With the less diverse Zymo and Kozich datasets, all methods recover nearly all of the expected Reference sequences (8 of 8 for Zymo and at least 20 of 21 for Kozich), but on the larger Schirmer and D'Amore data, the OTU methods detect notably fewer References (46 of 57 for Schirmer and just 42 of 57 for D'Amore). DADA2 detects the most Reference sequences in all cases, matched or closely followed by UNOISE and MED. Large numbers of Ref Noisy sequences were reported by UCLUST (42-78) for three of the four mock communities, and by MED (21-49) for two datasets, whereas all other methods inferred no more than 3 sequences in this category. In general, several sequences were detected in the Contaminant class. UCLUST and UPARSE gave the highest Contaminant counts (15-102) among all methods, while DADA2 reported the most (5-31) and Deblur the fewest (0-16) among the error modeling methods; MED is the notable exception, reporting no more than 6 Contaminant sequences. In the Contam Noisy category, UCLUST gave 4 sequences for the Kozich and Schirmer data, while no method exceeded one Contam Noisy in all other cases. Similarly, the number of Other sequences inferred typically ranged from 0-5, but UCLUST found much higher totals (23-66) for three communities, as did UPARSE (24) for the Zymo data.

To get an idea of the number of taxonomic groups represented by the inferred sequences, we assigned taxonomy to each sequence using the most specific designation available from the Silva version 132 database. **Table 4** shows the number of distinct taxonomic groups in each sequence category. (Note that the total number of groups does not, in general, equal the sum of groups in each category, since some groups may be present in more than one category.) Here we find that while the number of taxonomic groups is usually smaller than the number of sequences (in some cases considerably so), the same general trends hold. UCLUST still reports the most total distinct groups, followed by UPARSE, while either Deblur or MED reports the fewest. As with sequences, UCLUST and MED stand alone in reporting several groups in the Ref Noisy category with the Zymo and Kozich data. The exception to the trends found with the sequence-level analysis is in the Reference category: whereas the numbers of Reference *sequences* reported reveal sizable differences in sensitivity between the OTU methods and the others, these differences are not apparent at the taxonomic group level, as the number of distinct groups detected by the OTU methods is equal to or only slightly smaller than the numbers found by DADA2, UNOISE, and MED, which detected the most sequences and groups.

Table 4: Taxonomic groups inferred for high biomass communities

Dataset	Method	Total	Reference	Ref Noisy	Contaminant	Contam Noisy	Other
Zymo	UCLUST	70	8	17	36	1	31
	UPARSE	49	8	2	29	0	17
	MED	15	8	13	0	0	0
	UNOISE	12	8	2	1	0	1
	Deblur	8	8	0	0	0	0
	DADA2	18	8	3	5	0	3
Kozich	UCLUST	132	19	21	84	3	20
	UPARSE	84	19	1	62	0	5
	MED	30	19	14	3	0	0
	UNOISE	35	18	1	15	0	1
	Deblur	29	18	0	10	0	1
	DADA2	51	19	1	29	0	2
Schirmer	UCLUST	97	46	31	26	1	21
	UPARSE	74	46	1	25	0	4
	MED	55	48	2	6	0	0
	UNOISE	67	49	0	17	1	0
	Deblur	63	48	0	14	1	0
	DADA2	76	49	2	24	0	1
D'Amore	UCLUST	59	42	3	16	0	4
	UPARSE	57	42	0	15	0	1
	MED	47	43	1	3	0	0
	UNOISE	51	44	0	7	0	0
	Deblur	50	43	0	7	0	0
	DADA2	58	44	0	14	0	0

For a more concise summary of sequence inference performance, we computed the precision and recall of each method for each dataset, using sequences as the unit of analysis (**Table 5** and **Table 6**). At the sequence level, recall measures the proportion of known community members that were detected by each method, while precision gives the proportion of predicted community members that actually belong to the known community. Recall is generally high on all four undiluted mock communities, ranging from 74%-100%. DADA2 and UNOISE achieve the highest sequence recall across datasets (100% on three of four, and 89% for D'Amore). MED's recall is equal or only

slightly lower (98% for Schirmer and 88% for D'Amore), followed closely by Deblur (95%, 95%, and 84% for Kozich, Schirmer, and D'Amore, respectively). UCLUST and UPARSE give the lowest recall, which is still quite high on the two smaller communities (95%-100%), but falls notably for the larger Schirmer and D'Amore communities (81% and 74%, respectively). Precision varies much more (4%-100%) across methods and datasets, owing to the wide variation in numbers of unanticipated sequences. Deblur and UNOISE give relatively high sequence precision on all datasets (63%-100% and 53%-86%, respectively), as does MED on Schirmer and D'Amore communities (86%-91%, whereas UCLUST and UPARSE rank last on all datasets (4%-64% and 11%-72%, respecitvely). MED exhibits the most variation across datasets, ranging from 14%-91% precision.

Table 5: Sequence-level recall for the high biomass communities

Method	Dataset			
	Zymo	Kozich	Schirmer	D'Amore
UCLUST	100	95	81	74
UPARSE	100	95	81	74
MED	100	100	98	88
UNOISE	100	100	100	89
Deblur	100	95	95	84
DADA2	100	100	100	89

Table 6: Sequence-level precision for the high biomass communities

Method	Dataset			
	Zymo	Kozich	Schirmer	D'Amore
UCLUST	4	10	25	64
UPARSE	12	20	60	72
MED	14	48	86	91
UNOISE	67	53	73	86
Deblur	100	63	76	86
DADA2	40	39	65	77

To measure the bulk impact of the various noise sources on target community inference by each method, we computed the percentage of output reads assigned to Reference sequences (see **Table 7**). Here we find that a large majority of reads for all samples were mapped to the target mock community regardless of inference method, with only small differences. The largest variation is for the Zymo dataset, where UNOISE and DADA2 map ~96% of reads to Reference sequences, and MED only 92%, whereas the remaining methods map nearly 100% to References. In the case of UNOISE and DADA2, this is mainly due to a single Ref Noisy sequence with high abundance (data not shown), while for MED it is due to the large number of Ref Noisy sequences.

Table 7: Percentage of reads mapped to Reference sequences

Method	Dataset			
	Zymo	Kozich	Schirmer	D'Amore
UCLUST	99.2	98.9	96.6	98.3
UPARSE	99.8	99.1	95.2	98.0
MED	92.3	97.6	96.9	98.3
UNOISE	96.4	99.3	97.0	98.3
Deblur	100.0	99.3	96.9	98.3
DADA2	96.7	99.2	96.9	98.3

In addition to estimating the accuracy a method's sequence inference, we also wanted to observe how well it discriminates the desired biological signal from the various noise sources. Biological signal refers to the true 16S rRNA sequences of the mock community members, and noise includes contaminant species (introduced during sample processing), PCR artifacts (e.g., chimeras), and sequencer error. **Figure 4** plots the log-

scaled distributions of Reference and non-Reference sequence abundances inferred by each method studied, for each of the four high microbial biomass datasets. For the Zymo and Kozich datasets, the target signal is mostly well separated from the distribution of noisy sequences, whereas there is somewhat more overlap of signal and noise distributions on the Schirmer and D'Amore communities. Among methods, the minimum signal strength is typically highest for DADA2 and UNOISE, while that of Deblur is somewhat lower on Zymo and Kozich data, and that of UCLUST and UPARSE is much lower on all but the Zymo data. The maximum noise strength is higher for DADA2, MED, and UNOISE on the Zymo data (Deblur reports no non-Reference sequences here), but is comparable for all methods on the remaining datasets. We also see that the density of MED's noise distribution is concentrated at higher levels than that of other methods, particularly for the Zymo and Kozich data.

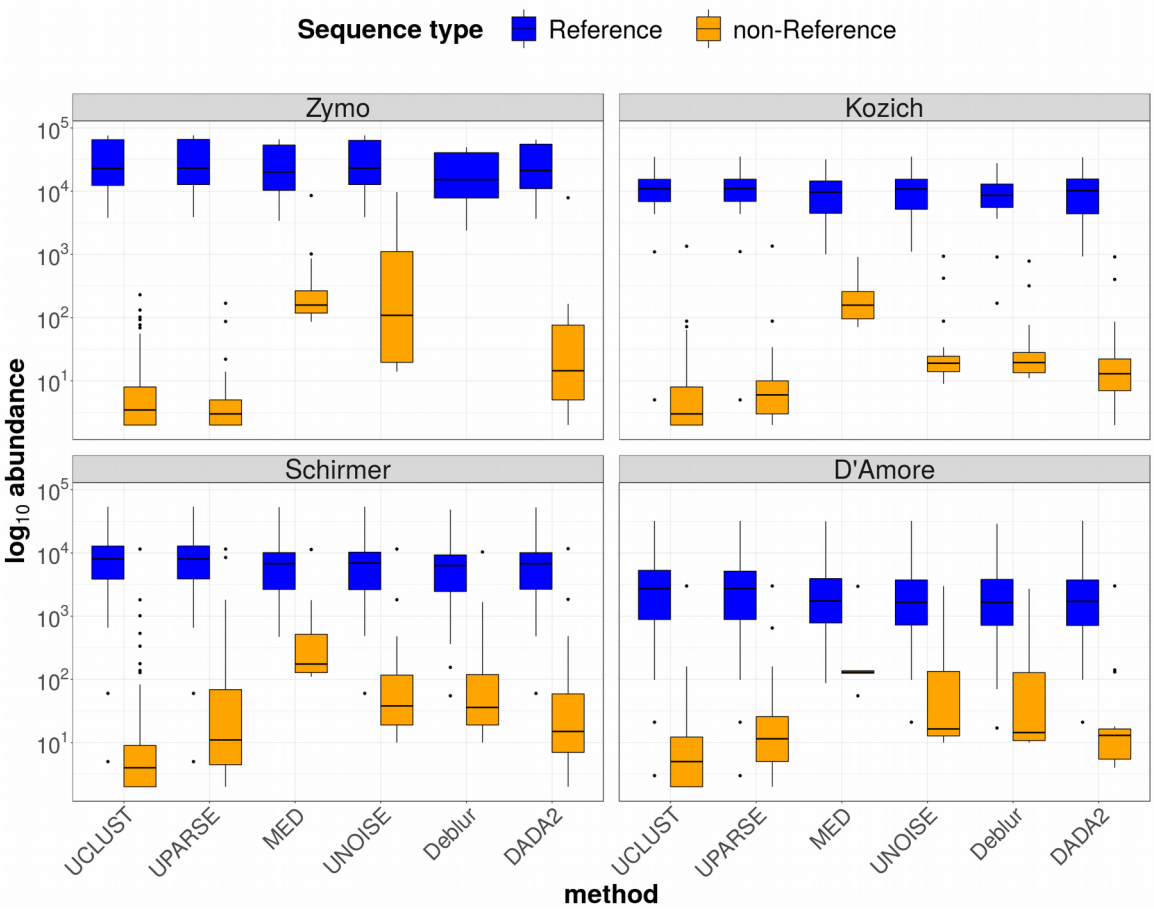

Figure 4: Abundance distributions of Reference and non-Reference sequences for high biomass communities. Log_{10}-transformed read abundance distributions of Reference sequences (those that match the 16S rRNA sequence of a known mock community member) and non-Reference sequences, as inferred by each of the six methods. Boxplots show median, IQR, 1.5 × IQR, and outliers. Each subplot shows abundance distributions for one of the four high biomass communities.

Dilution series of the Zymo mock community

To assess performance on low microbial biomass samples, we next evaluated the six inference methods on a dilution series prepared from the ZymoBIOMICS mock community. Each successive sample was diluted to 1/3 the starting DNA concentration of the previous sample in the series, for a total of 9 samples ranging from 1:1 (neat) to

1:6561 relative concentrations. A summary of sequence counts inferred by each method across the dilution series, including classification results, is shown in **Table 8**.

One trend that is immediately apparent is that as starting concentration decreases, the number of inferred sequences increases-- dramatically for some methods. (This trend appears not to hold for the two most dilute samples, but this can be explained primarily by the much lower sequencing depth obtained for these samples. See **Table 8** for the read counts of each dilution sample, after quality filtering.) At the highest concentrations (1:1 and 1:3), as observed with the high biomass communities, the error modeling methods reported the fewest sequences (8-22), but their totals increase steadily across the dilution series (again, with the exception of the two most dilute samples, as explained above) to a peak of 381-530 sequences at 1:729 dilution. The total number reported by MED is initially between those of the error model methods and those of the OTU methods for the neat sample, and remains relatively unchanging over the first four dilution samples (57-102 sequences), but then rises sharply so that it detects the highest numbers of sequences (278-570) among all methods for the three most dilute samples. In contrast, the OTU methods-- UCLUST in particular-- begin at the high end for the three highest concentration samples (69-288 for UPARSE, 202-450 for UCLUST), with a sharp spike for the 1:9 sample, but their numbers level off over the rest of the dilution series, and UPARSE actually reports the fewest sequences (142-304) for the four most dilute samples.

Table 8: Sequences inferred for the dilution series of the Zymo community

Dilution	Method	Total	Reference	Ref Noisy	Contaminant	Contam Noisy	Other
1:1 (243.5k reads)	UCLUST	200	8	78	47	0	69
	UPARSE	69	8	2	35	0	24
	MED	57	8	49	0	0	0
	UNOISE	12	8	2	1	1	0
	Deblur	8	8	0	0	0	0
	DADA2	20	8	3	5	1	3
1:3 (198.0k reads)	UCLUST	179	8	52	62	0	57
	UPARSE	77	8	0	50	0	19
	MED	86	8	78	0	0	0
	UNOISE	17	8	1	8	0	0
	Deblur	10	8	0	2	0	0
	DADA2	22	8	3	9	1	1
1:9 (282.0k reads)	UCLUST	450	8	64	218	25	135
	UPARSE	288	8	0	197	2	81
	MED	78	8	64	6	0	0
	UNOISE	119	8	1	97	3	10
	Deblur	85	8	0	75	0	2
	DADA2	114	8	3	91	0	12
1:27 (270.8k reads)	UCLUST	272	8	27	172	13	52
	UPARSE	206	8	0	164	4	30
	MED	102	8	71	23	0	0
	UNOISE	283	8	1	198	47	29
	Deblur	207	8	0	157	21	21
	DADA2	169	8	1	132	6	22
1:81 (243.5k reads)	UCLUST	336	8	23	200	14	91
	UPARSE	269	8	1	186	2	72
	MED	153	8	66	76	2	1
	UNOISE	449	8	2	277	91	71
	Deblur	339	8	0	237	38	56
	DADA2	261	8	1	195	9	48
1:243 (151.3k reads)	UCLUST	313	8	3	197	18	87
	UPARSE	259	8	0	187	3	61
	MED	398	8	50	227	63	50
	UNOISE	444	8	1	281	79	75
	Deblur	373	8	0	254	43	68
	DADA2	309	8	1	219	24	57
1:729 (144.3k reads)	UCLUST	377	8	2	239	37	91
	UPARSE	304	8	0	228	5	63
	MED	570	8	30	349	139	44
	UNOISE	530	8	1	330	123	68
	Deblur	430	8	0	293	68	61
	DADA2	381	8	2	270	49	52
1:2187 (46.1k reads)	UCLUST	153	8	1	97	2	45
	UPARSE	142	8	1	97	0	36
	MED	278	8	63	127	42	38
	UNOISE	190	8	1	119	22	40
	Deblur	168	8	1	111	11	37
	DADA2	148	8	0	100	10	30
1:6561 (49.4k reads)	UCLUST	195	8	2	127	9	49
	UPARSE	183	8	1	126	2	46
	MED	325	8	25	177	64	51
	UNOISE	267	8	4	161	39	55
	Deblur	226	8	2	152	16	48
	DADA2	193	8	1	129	11	44

Examining the sequence classifications gives insight into the source of the observed variation. Here we find that the variation in total sequences is largely driven by Contaminant sequences (illustrated in **Figure 5**), which tend to increase as samples become more dilute. However, the trend lines in **Figure 6** show that this increase in Contaminant sequences is nearly linear for the error model methods (with the exception of the two most dilute samples, as explained above), somewhat less so for MED, while the OTU methods exhibit the least association between sample dilution and number of Contaminants. We observed a smaller but roughly proportional trend for Contam Noisy sequences with error modeling methods and MED. The remaining Other sequences were somewhat less linear with dilution for the error models, and showed almost no relationship with dilution for OTU methods and MED. We also found that among error models, DADA2 generally reported the fewest Contaminant, Contam Noisy, and Other sequences, although it gave slightly higher numbers for the three most concentrated samples. All methods detected all 8 expected community members, regardless of sample dilution. Just as we observed with the high microbial biomass communities, MED and UCLUST are the only methods that infer a high number of Ref Noisy sequences, but whereas the number remains high for MED across the dilution series, it declines almost to zero for UCLUST at the lowest concentrations (see **Figure 6**).

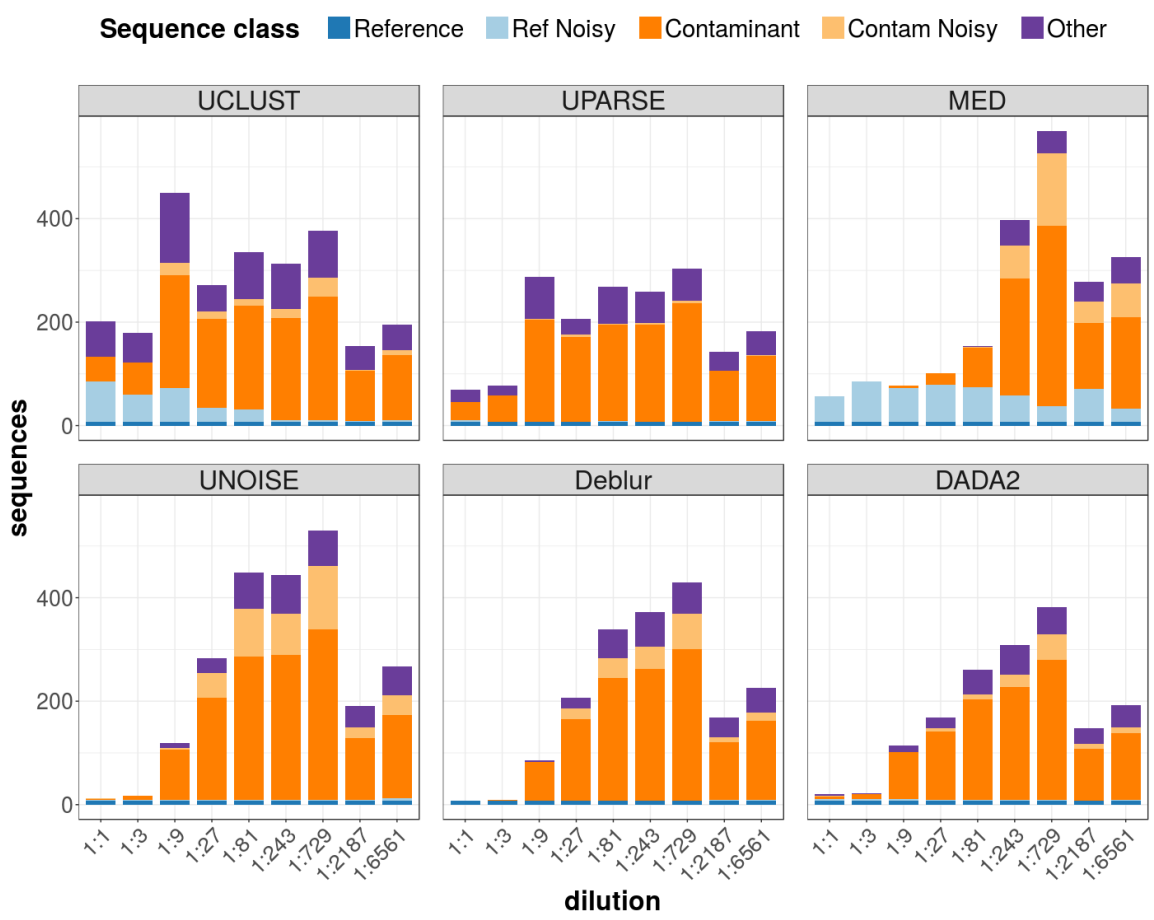

Figure 5: Sequence-level composition of dilution series samples. Composition of each sample in terms of the number of sequences in each category, as sample concentration decreases (x-axis). Categories are defined in Methods. Each subplot shows sample compositions inferred by one of the six methods benchmarked.

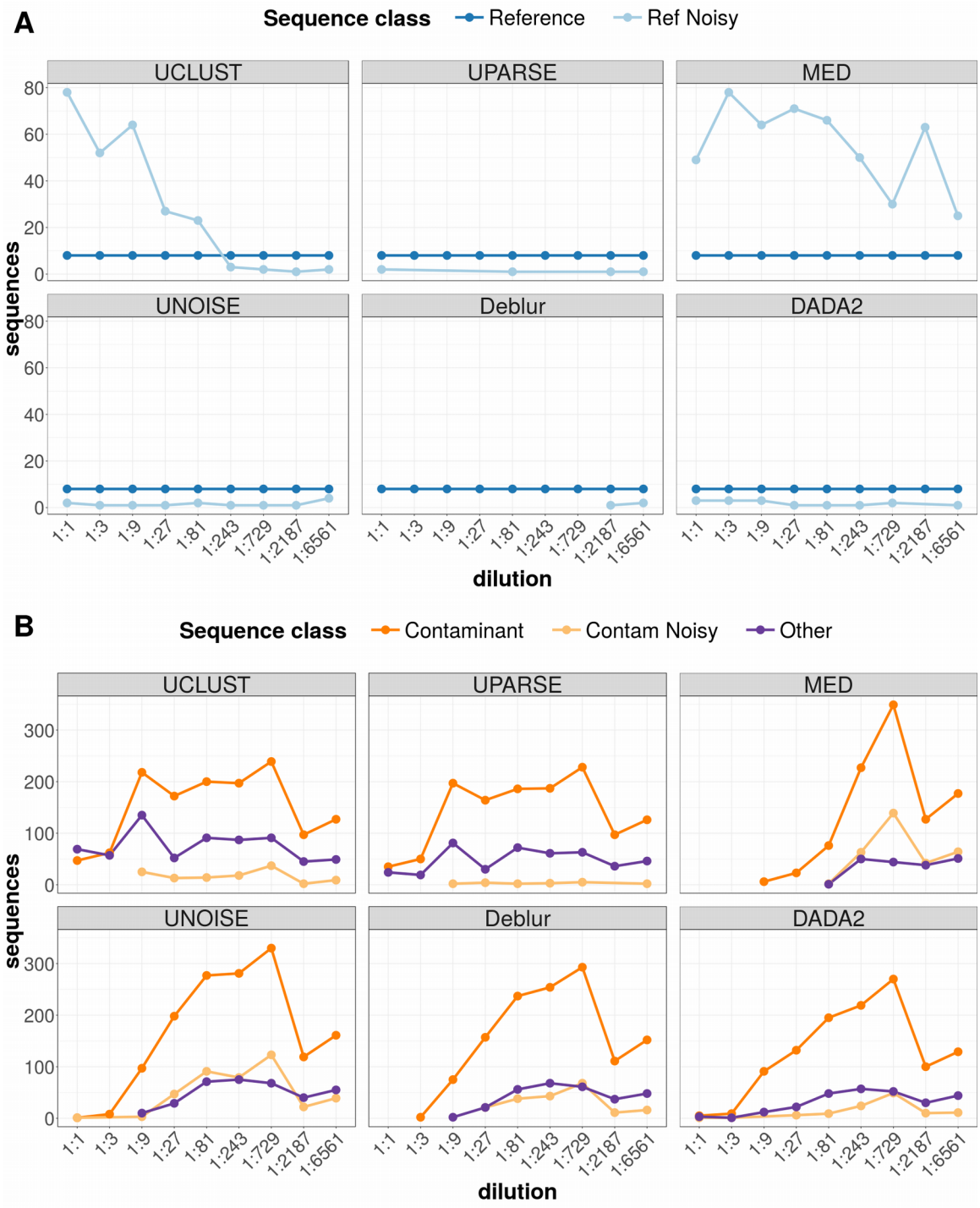

Figure 6: Trend lines of inferred sequences across the dilution series. (A) Number of Reference and Ref Noisy sequences inferred vs. decreasing sample concentration. **(B)** Number of Contaminant, Contam Noisy, and Other sequences inferred vs. decreasing sample concentration. Sequence categories are defined in Methods. Each subplot shows trend lines for sequences inferred by one of the six methods benchmarked.

Table 9 shows the results of grouping sequences according to taxonomic classification, as was done for the high biomass communities. Just as with the sequence-level results, inferred taxonomic groups tend to rise as community biomass decreases, but with distinct patterns depending on the inference algorithm. As illustrated in **Figure 7**, the increase in total inferred taxonomic groups is driven by Contaminants, and the OTU methods exhibit the most jagged trend lines, with a rapid initial increase in Contaminant taxonomic groups that then slows or flattens at higher dilutions. MED, on the other hand, shows the smallest initial increase in Contaminants, with the number increasing sharply as concentration continues to decrease. The number of Contaminant groups inferred by the error modeling methods increases more steadily across the dilution series, although not quite as linearly as observed for sequences (see **Figure 6**). The number of Ref Noisy taxonomic groups is still notably higher for UCLUST and MED compared to all other methods, but is typically less than half the number of sequences in this category. Also, DADA2 continues to report the fewest Contaminant, Contam Noisy, and Other among the error modeling methods at the taxonomic group level, with the exception of the three most concentrated samples.

Table 9: Taxonomic groups inferred for dilution series of the Zymo community

Dilution	Method	Total	Reference	Ref Noisy	Contaminant	Contam Noisy	Other
1:1	UCLUST	69	8	16	36	0	31
	UPARSE	50	8	2	29	0	17
	MED	15	8	13	0	0	0
	UNOISE	12	8	2	1	1	0
	Deblur	8	8	0	0	0	0
	DADA2	18	8	3	5	1	2
1:3	UCLUST	88	8	17	46	0	37
	UPARSE	56	8	0	38	0	15
	MED	16	8	15	0	0	0
	UNOISE	17	8	1	8	0	0
	Deblur	10	8	0	2	0	0
	DADA2	21	8	3	9	1	1
1:9	UCLUST	167	8	16	117	13	71
	UPARSE	134	8	0	107	2	45
	MED	20	8	13	5	0	0
	UNOISE	74	8	1	61	3	8
	Deblur	60	8	0	51	0	2
	DADA2	88	8	3	69	0	11
1:27	UCLUST	105	8	11	80	8	29
	UPARSE	92	8	0	78	4	18
	MED	30	8	14	15	0	0
	UNOISE	90	8	1	76	20	16
	Deblur	79	8	0	67	14	14
	DADA2	78	8	1	65	4	15
1:81	UCLUST	130	8	9	98	8	44
	UPARSE	119	8	1	95	2	42
	MED	55	8	12	41	2	1
	UNOISE	128	8	2	110	28	32
	Deblur	119	8	0	101	21	31
	DADA2	115	8	1	94	6	28
1:243	UCLUST	133	8	3	102	12	42
	UPARSE	123	8	0	101	3	35
	MED	116	8	13	92	22	28
	UNOISE	132	8	1	109	27	33
	Deblur	127	8	0	106	20	32
	DADA2	129	8	1	107	16	31
1:729	UCLUST	160	8	1	135	15	43
	UPARSE	156	8	0	132	4	39
	MED	160	8	10	133	40	26
	UNOISE	169	8	1	146	46	35
	Deblur	161	8	0	141	36	31
	DADA2	162	8	2	136	22	31
1:2187	UCLUST	90	8	1	67	2	30
	UPARSE	87	8	1	66	0	25
	MED	98	8	14	67	25	24
	UNOISE	89	8	1	69	14	24
	Deblur	86	8	1	67	9	23
	DADA2	87	8	0	67	8	23
1:6561	UCLUST	108	8	2	87	5	29
	UPARSE	105	8	1	86	2	27
	MED	113	8	13	87	21	24
	UNOISE	108	8	4	87	16	26
	Deblur	103	8	2	85	8	24
	DADA2	103	8	1	84	7	24

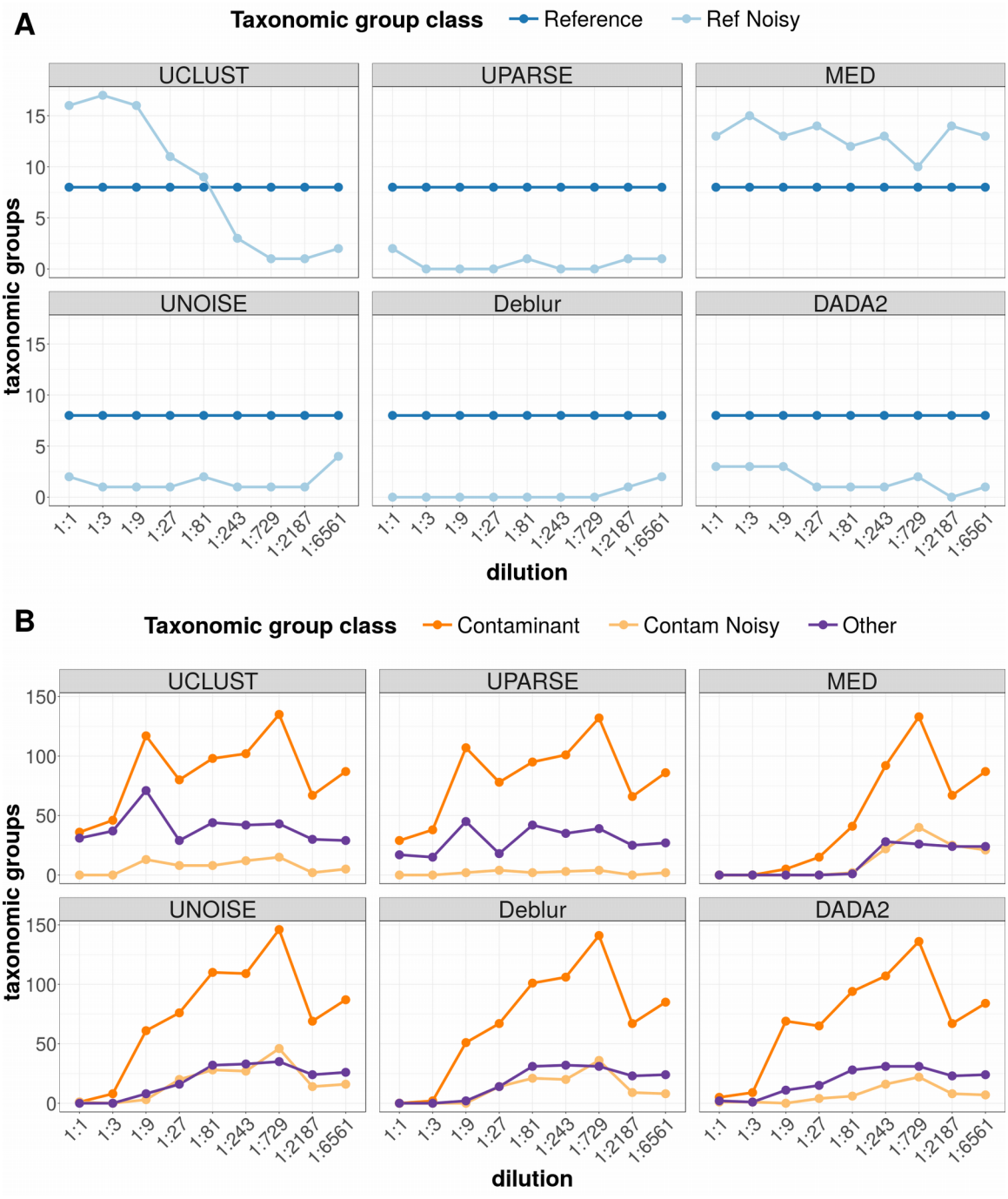

Figure 7: Trend lines of inferred taxonomic groups across the dilution series. (A) Number of taxonomic groups in the Reference and Ref Noisy categories vs. decreasing sample concentration. **(B)** Number of taxonomic groups in the Contaminant, Contam Noisy, and Other sequences inferred vs. decreasing sample concentration. Sequence categories are defined in Methods. Taxonomic groups were determined by assigning the highest possible resolution of taxonomy to each sequence using the SILVA v123 16S rRNA database, then counting the number of unique taxonomic labels. Each subplot shows trend lines for taxonomic groups inferred by one of the six methods benchmarked.

Sequence-level recall was perfect across the dilution series-- all methods detected all 8 expected strains at every sample concentration-- likely due to the low diversity and even genomic proportions of the Zymo mock community. However, sequence precision is more variable, which reflects the high numbers of false positives at lower concentrations. For the two highest concentrations, precision is best with Deblur (80%-100%) and worst for UCLUST (~4%), but MED has the best precision at middle concentrations (1:9 to 1:81), while UPARSE has the best precision at lowest concentrations (1:243 to 1:6561). However, beyond 1:9 dilution, all methods achieve well below 10% precision.

Figure 8 shows the proportion of reads in each class for each method, across the dilution series. Immediately we see that as concentration drops, the proportion of Reference reads declines considerably with all methods, reaching a minimum of less than 20% for the most dilute sample. At the highest concentrations (1:1 and 1:3), Deblur and the OTU methods assign over 99% of reads to Reference sequences. UNOISE and DADA2 have small but notable proportions of Ref Noisy reads over the first six dilutions (1%-4%, primarily due to a single Ref Noisy sequence, data not shown), and MED assigns the largest proportion to Ref Noisy across all methods and dilutions (2%-11%). Beginning with the 1:9 dilution sample, reads assigned to Contaminant sequences become apparent, increasing steadily in proportion until they dominate the inferred composition for the three lowest dilutions, making up 43%-70% of the sample. Reads from Other and Contam Noisy sequences also generally increase across the dilution series. The former typically comprise a larger fraction, reaching a maximum of ~10% at the lowest

concentration, but MED and UNOISE exhibit Contam Noisy fractions comparable to those of Other reads.

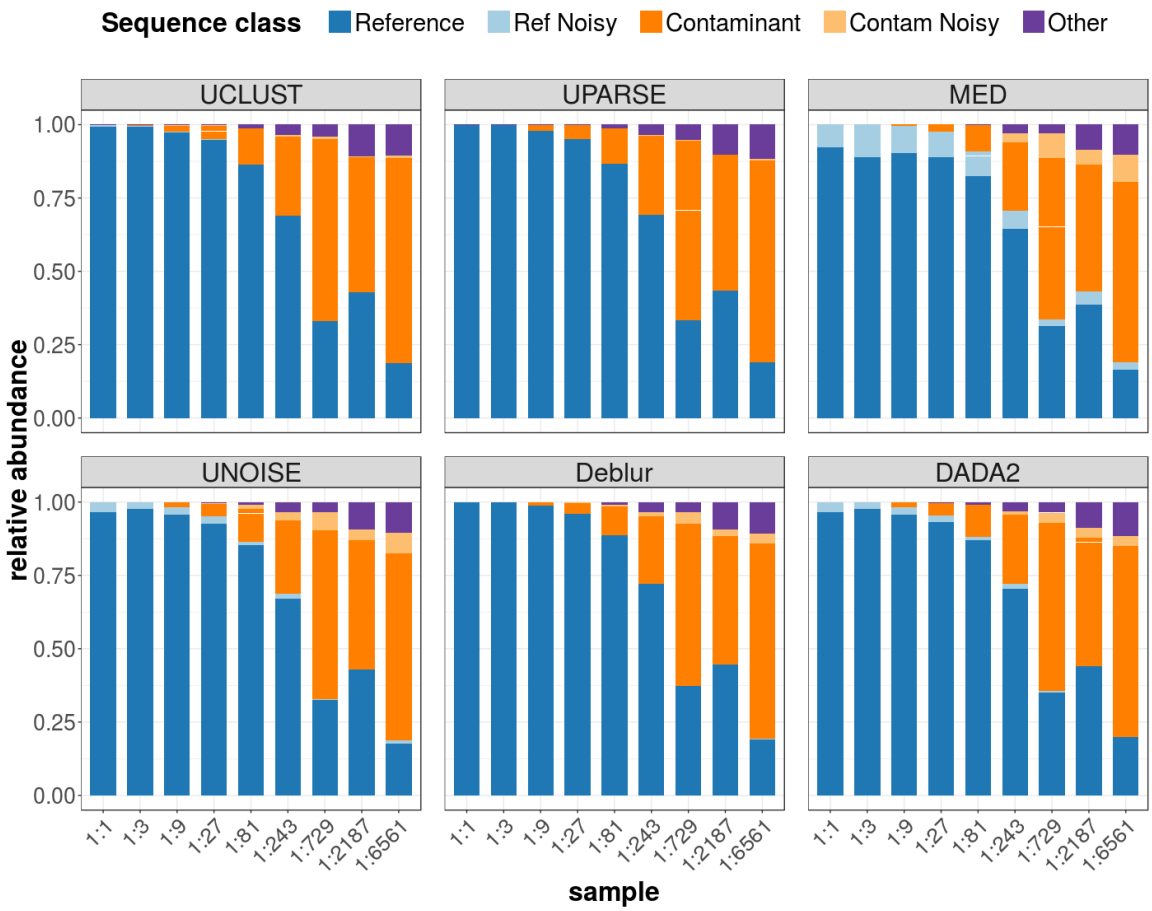

Figure 8: Read-level composition of dilution series samples. Composition of each sample, in terms of proportion of total sample reads in each sequence category, as community biomass decreases, for the Zymo mock community. Categories are defined in Methods. Each subplot shows the sample compositions inferred by one of the six methods benchmarked.

As with the undiluted mock communities, we compared target community signal distribution with that of the noise for the dilution series samples. The results are shown in **Figure 9**. Regardless of inference method, we observed that as concentration decreases, the signal and noise distributions steadily converge, as the signal weakens

and noise grows. Over the first few dilution samples (1:1 to 1:81 relative concentration), the signal distribution remains nearly constant, and even though the strength of the noise increases, signal and noise are generally well separated (a single non-Reference sequence overlaps the Reference distributions for DADA2, MED, and UNOISE). Deblur achieves the largest separation for the three highest concentrations, while MED's gap is much smaller. However, with decreasing DNA concentration, the distributions begin to overlap considerably, as the signal strength steadily declines while the noise strength continues to increase, and there is little difference between methods at lower concentrations.

One interesting phenomenon is that sequence counts for the individually processed dilution series samples changed when the samples were pooled and processed as a single dataset. In general, total inferred sequences, as well as counts in each category, increased when the samples were processed together as one dataset. The exception was for DADA2, which reported the same counts regardless of whether the neat sample was processed individually or not.

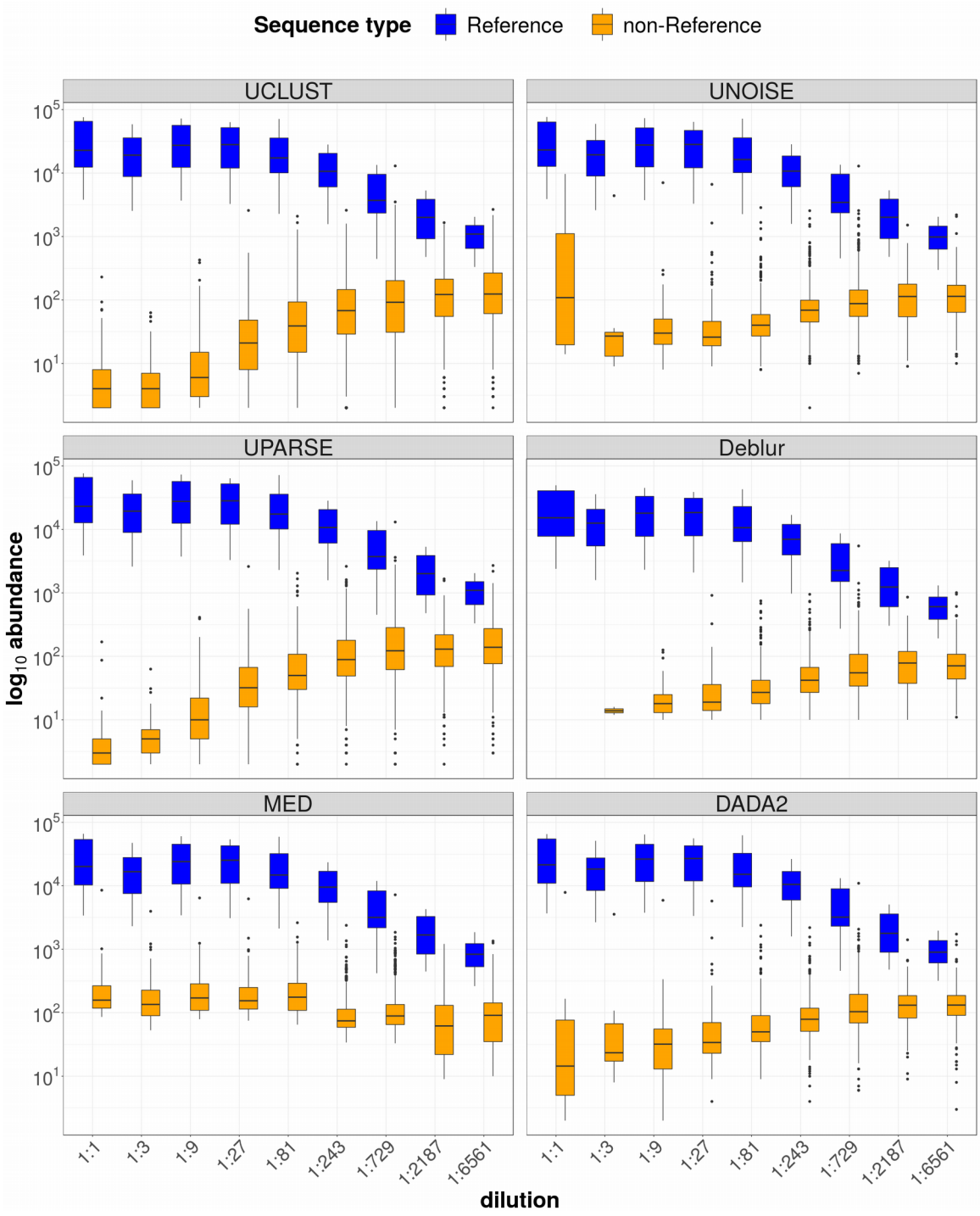

Figure 9: Abundance distributions of Reference and non-Reference sequences for the dilution series. Log_{10}-transformed read abundance distributions of Reference sequences (those that match the 16S rRNA sequence of a known mock community member) and non-Reference sequences vs. decreasing community biomass for the Zymo mock community. Boxplots show median, IQR, 1.5 × IQR, and outliers. Each subplot shows abundance distributions inferred by one of the six methods benchmarked.

V. Discussion

High microbial biomass benchmarking

On the high biomass mock communities, the numbers of sequences in each class inferred by the error modeling methods were in good agreement, and also correlated with the known diversity of each community much better than with the other three methods. Deblur exhibited the best specificity of the three in the sense that it reported the fewest unexpected sequences, but DADA2 and UNOISE had somewhat better sensitivity for expected sequences. The higher Contaminant counts detected by DADA2 and UNOISE suggest that they may be more sensitive to low-abundance contamination, but it is also possible that their models underestimate the error, resulting in over-splitting of sequence clusters, i.e., poorer specificity. The error modeling methods as a group also out-performed the other three algorithms in terms of their ability to distinguish the true signal from sources of noise. Thus, error models in general offer superior performance than OTU or entropy-based algorithms, combining good sensitivity with good specificity, and the choice of which error modeling method to use depends upon the goals of the research. If minimizing detection of spurious sequences and contaminants is most important, Deblur seems more appropriate. However, if maximizing detection of true community members and/or real biological contaminants is the priority, DADA2 or UNOISE appear to be better suited.

Results for MED were less consistent. While sequence sensitivity was quite high for all undiluted samples, specificity was highly variable. Specificity was excellent on the

Schirmer and D'Amore communities, but suffered on the Zymo and Kozich samples due to a large number of Ref Noisy sequences (49 and 21, respectively). Virtually all of these Ref Noisy were within 1 nt of a reference strain, strongly suggesting that they are indeed false positives that should have been clustered with a Reference sequence. This false detection is also reflected in the signal and noise distributions: MED's inference of several Ref Noisy sequences with relatively high abundance results in the poor separation of signal from noise observed with the Zymo and Kozich datasets (**Figure 4**). Hence, although MED was designed to distinguish biological strains with extremely similar 16S rRNA sequences, it does so unreliably, as it is prone to reporting spurious sequences that arise from sequencer errors. However, in studies for which such fine distinctions are not deemed important, MED has the potential to yield accurate results by simply collapsing inferred sequences that differ by a single nucleotide.

The OTU methods produced both the poorest recall and specificity. Since these methods rely only on a distance metric for clustering, their reduced sensitivity must result from lumping together distinct strains with high 16S rRNA sequence identity, ignoring read abundance. On the other hand, the high numbers of non-Reference sequences reported by the OTU methods relative to the other methods are almost certainly the result of splitting into distinct clusters those reads with sequencing errors that fall outside the similarity threshold of the template sequence from which they were derived. UPARSE had identical sensitivity to UCLUST, but showed considerably better specificity, especially in terms of Ref Noisy sequences. This is best explained by UPARSE's strict quality filtering step prior to cluster inference, which presumably removes the majority

of reads with several sequencing errors. Some of the many Contaminant sequences inferred by the OTU methods probably represent real biological contamination, as some Contaminants are reported for each sample by most or all methods. However, it is revealing that the communities for which UCLUST and UPARSE diverge most from the consensus are Zymo and Kozich, which have poorer read quality profiles compared to Schirmer and D'Amore, and thus are expected to have more sequencing errors. Clearly, the fixed similarity threshold of OTU methods is ill-equipped to deal with such scenarios. In addition, UCLUST and UPARSE tended to report one or more Reference sequences with much lower signal strength than the other methods studied (see **Figure 4**). These observations show that OTU methods are inferior to the other algorithms on multiple counts, and thus their continued use for research is not advised.

Grouping inferred sequences by taxonomic classification tended to mask some of the distinctions between methods. The OTU methods still tended to report higher numbers of Contaminant and Other taxonomic, and UCLUST and MED found several more Ref Noisy groups than other methods, indicating poorer specificity for the non-error modeling methods, but the differences were less dramatic than with the sequence-level analysis. This observation suggests that in some cases, the lack of specificity exhibited by OTU methods and MED may not have a large impact on community inference, since many of the spurious sequences may still fall into the same taxonomic group as the template sequence from which they were derived. However, the differences between the numbers of sequences and the numbers of taxonomic groups in the Reference category illustrates perhaps the most important utility of error model inference. The

taxonomic grouping might lead to the conclusion that all methods have very similar sensitivities, since there is little variation in the number of Reference groups detected. In contrast, at the sequence level the error model methods and MED show clearly superior sensitivity for the most diverse Schirmer and D'Amore communities, since they distinguished several strains within some taxonomic groups— distinctions which we know exist, given the mock community composition. Hence, if strain-level sensitivity is potentially important, then OTU methods are inadequate.

Dilution series benchmarking

The dilution series results clearly show that as starting DNA concentration decreases, sequences derived from noise sources comprise an increasing proportion of the inferred community, both in terms of number of distinct sequences present as well as their abundances. The dramatic rise across methods in numbers of Contaminant sequences detected at lower starting concentrations suggests that, in comparison to samples of high microbial biomass, there are in fact many more contaminating species present at detectable levels in sequencing libraries prepared from samples with low microbial biomass. This can be explained by the observation that when less sample DNA is present for PCR amplification, small levels of microbial contaminants (introduced via non-sterile lab equipment, imperfect aseptic technique, and reagent kits [citations needed]) make up a larger proportion of the total DNA. Hence, the contaminant sequences are amplified with higher probability and to a greater extent than would be the case with a high microbial biomass sample (in which the sample DNA overwhelms the contamination). Thus, the "Contaminant"-labeled sequences detected in the dilution

samples with low microbial biomass most likely reflect biological contamination introduced during sample processing, rather than indicating false positives.

Given this likelihood, the strong association we observed between sample dilution and the number of Contaminant sequences inferred by each of the error modeling methods is evidence that these three methods detect the true contamination present more accurately than the OTU methods or MED. This is consistent with the higher sensitivity and specificity exhibited by these three methods on the high microbial biomass communities. The differences between the three methods in numbers of Contaminant and Contam Noisy sequences, which maintain their order across most of the dilution series (UNOISE reports more than Deblur, which reports more than DADA2), can be attributed to differences in their error models. Either UNOISE and Deblur are more sensitive than DADA2 for this dataset, or else the former two underestimate the actual sequencing error, leading to poorer specificity. Based on the results with the undiluted mock communities, for which DADA2 showed the highest sensitivity among all methods, it seems unlikely that DADA2 should now have the lowest sensitivity of the three. Furthermore, UNOISE and Deblur use a fixed error model (with the default settings), whereas DADA2 estimates its model dynamically from the data; and since the dilution series dataset had a relatively poor read quality profile, it therefore seems plausible that DADA2's model is better adjusted to a higher degree of error, and that UNOISE and Deblur underestimate the error profile on this dataset. Changing the error model parameters from their defaults for UNOISE and Deblur may yield better results for

datasets with lower read quality, but establishing guidelines for doing so may be challenging and is beyond the scope of this study.

MED, the lone entropy-based inference method, is unique in inferring none to very few Contaminants at higher concentrations, then rising sharply to infer the most Contaminants at very low concentrations. This phenomenon is primarily explained by MED's use of a read abundance threshold (set to 0.02% of total dataset reads by default) to filter out clusters arising from noise. The filter greatly limits the number of Contaminant sequences detected at higher concentrations, when true contamination is sparse, but as the proportion of contaminant DNA rises, many more contamination-derived sequences exceed the abundance filter. Thus, the effectiveness of this type of filter for removing biological noise depends unpredictably on the sample's DNA concentration (as well as on dataset size), and it comes with the risk that lower-abundance species present in the target community will also be removed, reducing sensitivity.

Additionally, the relatively high number of Ref Noisy sequences inferred by MED at all sample dilutions, as well as the high number of Contaminant and Contam Noisy inferred at low concentrations, show that the entropy criterion used to divide sequence clusters is too sensitive, i.e., it underestimates the sequencer error, resulting in many false positives. As with the error model parameters of UNOISE and Deblur, MED's default entropy criterion can be adjusted to better reflect the error rate for a given dataset, but choosing an appropriate value would require validation by the user.

To some extent, inference results for the OTU methods followed the trend of increasing sequence counts as concentration decreases, driven by a rise in Contaminant sequences. However, the number of inferred Contaminant sequences was much less associated with sample dilution than we observed for error model-based inference. This makes sense in light of the reduced sensitivity observed previously on the high microbial biomass communities. As the samples become more dilute and noise (i.e., contamination) grows relative to the target signal, OTU methods initially generate many new clusters. But as relative contamination continues to increase and more sequences are amplified, some distinct contaminant sequences may be lumped together by the similarity threshold, leading to the plateau in Contaminant sequences observed with OTU methods. With UCLUST, the drop in the number of Ref Noisy sequences over the dilution series reflects the OTU clustering strategy. For the highest concentration samples, large quantities of community template DNA lead to high sequencing depth and a corresponding wide range of sequencer errors for these sequences, producing high numbers of Ref Noisy sequences due to reads with errors that fall outside the OTU similarity threshold. Decreasing concentration brings decreased sequencing depth and a concomitant reduction in the range of errors, so fewer community-derived sequences have errors that exceed the cluster radius. This phenomenon clearly illustrates that OTU methods in general are not well suited to modern high-throughput sequencing data: the typical similarity radius does not account for the wide error profiles that occur with deep 16S rRNA sequencing, and increasing the radius would only further degrade the already reduced sensitivity.

The significant overlap of the signal and noise abundance distributions at low DNA concentrations illustrates the difficulty of separating the target community from noise sources for low microbial biomass samples. Clearly, an abundance filter (as is commonly used with high microbial biomass samples to remove contamination) would be ineffective in this scenario, as any choice of threshold risks either removing several community species or retaining several non-community species. Since this overlap was observed for all methods at low concentrations, we may conclude that none of the inference methods studied here is sufficient on its own to adequately distinguish community signal from noise. However, the clear positive correlation of starting DNA concentration with signal strength, and the negative correlation with noise strength, suggests a possible strategy for detecting contaminant sequences. Namely, a dilution series could be prepared from a mock community or sample aliquot and processed in parallel with the samples under study. Sequences inferred for the dilution series whose abundance increases with decreasing concentration could then be labeled as contaminants, and removed from the inferred communities of study samples. Indeed, such an approach has already been implemented in one form by Callahan, et al. [citation needed], and other variations on this approach are possible. Furthermore, based on the results of this study, it is our hypothesis that such a strategy would work best when an error modeling method is used for sequence inference, since these algorithms produced the best correlation between concentration and both number and abundance of noise sequences (see **Figure 6** and **Figure 9**), likely due to superior sensitivity and specificity as evidenced for the high biomass communities. MED shows some relationship between

concentration and non-Reference sequences detected, but abundances appear much less correlated; and while the OTU methods exhibit stronger abundance correlation, their demonstrated lack of sensitivity and specificity for Reference and apparently for non-Reference sequences would likely limit their accuracy when paired with such a contaminant-detection strategy.

Although not strictly related to microbial biomass, the observation that sequence counts differed for most methods depending on whether samples were pooled or processed independently gives further insight into the performance of these inference methods. UCLUST, UPARSE, UNOISE, and MED all infer clusters based on the entire dataset, regardless of the number of samples; reads from each sample are then mapped to clusters to obtain sample-wise abundances. Thus, a larger dataset (e.g., the full dilution series) will tend to produce more clusters than a smaller dataset. As demonstrated by the inconsistent results for the dilution samples, employing such a strategy may mean that community inference is difficult to reproduce, even for technical replicates. In contrast, Deblur and DADA2 perform inference on each sample independently. After the initial inference step, UNOISE, MED, and Deblur all attempt to control false positives with a minimum abundance filter, using abundance across samples. UNOISE and Deblur employ absolute minima (8 and 10 reads per inferred sequence by default, respectively), whereas MED uses a relative minimum (0.02% of total datasets reads, per sequence, by default). In either case, these abundance filters may complicate accurate sample inference by creating a dependence on dataset size, and MED's proportional threshold may remove many rare species, regardless of origin, which may be particularly

problematic in the case of low microbial biomass community. In the instance of Deblur and UNOISE, the use of such a filter may also have given an overly optimistic estimate of specificity, since pooled sample processing showed that Deblur's error model actually inferred several more sequences that were filtered out when samples were processed independently.

The generalizability of our findings may be limited by certain characteristics of this study. The ZymoBIOMICS Microbial Standard has low diversity and uniform genomic proportions, and as such is not representative of a typical microbiome sample. This mock community clearly exhibits the impact of contaminant noise at low microbial biomass, but the high proportion of each species did not challenge the sensitivity of inference at lower DNA concentrations. Dilution series data from a more varied community structure might permit further distinctions between inference methods, particularly among the error modeling algorithms. Another limitation is that a dilution series cannot perfectly mimic low microbial biomass samples taken from a study subject, as the act of diluting itself may introduce noise in the form of contamination. However, similar to dilution series preparation, low microbial biomass samples typically require additional processing to produce enough DNA for sequencing, which can also introduce contaminants. Thus, a dilution series provides a good approximation of the outsized impact that even small levels of contamination can have when starting DNA concentration is low. Finally, the classification scheme we used to analyze inferred sequences is inexact— there will inevitably be some overlap between categories that represent technical noise (Ref Noisy and Contam Noisy) and those representing other

noise sources (Contaminant and Other). Nevertheless, this scheme provides a logical framework within which to compare and assess inference methods, and demonstrates clear differences between the methods studied.

VI. Summary and Conclusions

Despite its limitations, this study revealed some clear distinctions not only between methods, but also between community inference for high and low microbial biomass communities. From the high biomass results, we found that both of the OTU inference methods studied suffered from a lack of sensitivity relative to newer inference algorithms, and UCLUST in particular— a method which has been used in a large number of published microbiome studies— also showed very poor specificity in the face of technical sequencer noise. MED, the lone entropy-based method, showed generally good sensitivity, but similar to UCLUST, it tended to overestimate community diversity by mis-inferring community-derived sequences with errors as distinct organisms. The three error modeling methods, on the other hand, all exhibited good sensitivity and specificity. Among the three, Deblur appeared to be the least sensitive and/or most specific, while DADA2 showed evidence of being the most sensitive and/or least specific, but we believe further research is needed to confidently characterize the differences between these three inference methods.

Aside from the differences in method performance, we found that with high microbial biomass communities, biological noise, or contamination, had a small impact on community inference. In general, every method detected several sequences that were likely introduced contaminants, but their abundances and proportion of the overall inferred community was very small relative to the known community members.

Results from the dilution series benchmarking tended to support the differences in method performance found on the high biomass communities. The nearly linear association between sample dilution and purported contaminant sequences detected by each of the error modeling methods showed further evidence of good accuracy for these methods, in this case representing accurate detection of biological noise. In contrast, the OTU methods displayed much less correlation between community biomass and detection of new unexpected sequences. MED detected the fewest purported contaminants at high DNA concentration, and the greatest number at low concentration, exhibiting the effect of MED's abundance filter in the former case, while supporting a lack of specificity in the latter case.

However, in stark contrast to the high microbial biomass scenario, our observations of the dilution series inference make it clear that for low microbial biomass samples, biological noise can be a considerable confounding factor. In sequencing the 16S rRNA of low biomass communities, very small levels of DNA contamination can be amplified such that their abundance becomes comparable to that of the targeted community sequences. We also found that none of the inference methods studied here is sufficient on its own to distinguish the target community from biological contamination in this scenario. Clearly, more research is needed to develop reliable techniques for removing contamination, and since perfectly aseptic sample processing may be a practical impossibility, contaminant removal by *in silico* methods may also be required. Owing to their demonstrated superior sensitivity and specificity when challenged with technical noise, as well as their apparently better accuracy for detecting biological noise, we

believe that error modeling methods are the best choice currently available for microbial community inference from 16S rRNA sequence data. In addition, we hypothesize that an error modeling method combined with a dilution series as a positive control may provide a viable tool for detecting biological noise, and propose this for future study.

VII. References

1. Siddiqui H, Nederbragt AJ, Lagesen K, Jeansson SL, Jakobsen KS. Assessing diversity of the female urine microbiota by high throughput sequencing of 16S rDNA amplicons. BMC Microbiol [Internet]. 2011;11(1):244. Available from: http://bmcmicrobiol.biomedcentral.com/articles/10.1186/1471-2180-11-244

2. Aho VTE, Pereira P a B, Haahtela T, Pawankar R, Auvinen P, Koskinen K. The microbiome of the human lower airways: a next generation sequencing perspective. World Allergy Organ J [Internet]. 2015;1–13. Available from: http://dx.doi.org/10.1186/s40413-015-0074-z

3. Païssé S, Valle C, Servant F, Courtney M, Burcelin R, Amar J, et al. Comprehensive description of blood microbiome from healthy donors assessed by 16S targeted metagenomic sequencing. Transfusion. 2016;56(5):1138–47.

4. Delzenne NM, Cani PD. Interaction between obesity and the gut microbiota: relevance in nutrition. Annu Rev Nutr [Internet]. 2011;31(1):15–31. Available from: http://www.annualreviews.org/doi/10.1146/annurev-nutr-072610-145146

5. Kassinen A, Krogius-Kurikka L, Mäkivuokko H, Rinttilä T, Paulin L, Corander J, et al. The Fecal Microbiota of Irritable Bowel Syndrome Patients Differs Significantly From That of Healthy Subjects. Gastroenterology. 2007;133(1):24–33.

6. Mulak A, Bonaz B. Brain-gut-microbiota axis in Parkinson's disease. World J Gastroenterol [Internet]. 2015 [cited 2017 May 18];21(37):10609–20. Available from: http://www.wjgnet.com/1007-9327/full/v21/i37/10609.htm

7. Chassaing B, Kumar M, Baker MT, Singh V, Vijay-Kumar M. Mammalian gut immunity. Biomed J [Internet]. 2014;37(5):246–58. Available from: http://www.biomedj.org/article.asp?issn=2319-4170;year=2014;volume=37;issue=5;spage=246;epage=258;aulast=Chassaing

8. Nayak RR, Turnbaugh PJ. Mirror, mirror on the wall: which microbiomes will help heal them all? BMC Med [Internet]. 2016;14(1):72. Available from: http://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-016-0622-6

9. Youngster I, Sauk J, Pindar C, Wilson RG, Kaplan JL, Smith MB, et al. Fecal microbiota transplant for relapsing clostridium difficile infection using a frozen inoculum from unrelated donors: a randomized, open-label, controlled pilot study. Clin Infect Dis. 2014;58(11):1515–22.

10. Meadow JF, Altrichter AE, Bateman AC, Stenson J, Brown G, Green JL, et al. Humans differ in their personal microbial cloud. PeerJ [Internet]. 2015;3(1):e1258. Available from: https://peerj.com/articles/1258

11. Karstens L, Asquith M, Davin S, Stauffer P, Fair D, Gregory WT, et al. Does the urinary microbiome play a role in urgency urinary incontinence and its severity? Front Cell Infect Microbiol [Internet]. 2016 Jul 27 [cited 2017 Jan 25];6(78):1–13. Available from: http://journal.frontiersin.org/Article/10.3389/fcimb.2016.00078/abstract

12. Dickson RP, Erb-Downward JR, Huffnagle GB. The role of the bacterial microbiome in lung disease. Expert Rev Respir Med [Internet]. 2013 Jun 9 [cited 2017 May 1];7(3):245–57. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4007100&tool=pmcentrez&rendertype=abstract%5Cnhttp://www.tandfonline.com/doi/full/10.1586/ers.13.24

13. Lazarevic V, Gaïa N, Emonet S, Girard M, Renzi G, Despres L, et al. Challenges in the culture-independent analysis of oral and respiratory samples from intubated patients. Front Cell Infect Microbiol [Internet]. 2014 [cited 2017 May 18];4:65. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24904840

14. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol [Internet]. 2014;12(1):87. Available from: http://www.biomedcentral.com/1741-7007/12/87

15. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics [Internet]. 2006 Jul 1 [cited 2016 Nov 11];22(13):1658–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16731699

16. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: high-resolution sample inference from Illumina amplicon data. Nat Methods [Internet]. 2016 Jul [cited 2016 Sep 18];13(7):581–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27214047

17. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv [Internet]. 2016; Available from: http://biorxiv.org/content/early/2016/10/15/081257.abstract

18. Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. Gilbert JA, editor. mSystems [Internet]. 2017 Apr 25 [cited 2017 Mar 8];2(e00191-16):1–7. Available from: http://msystems.asm.org/lookup/doi/10.1128/mSystems.00191-16

19. Schloss PD, Westcott SL. Assessing and improving methods used in operational taxonomic unit-based approaches for 16S rRNA gene sequence analysis. Appl Environ Microbiol [Internet]. 2011 May [cited 2016 Oct 31];77(10):3219–26. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21421784

20. Laurence M, Hatzis C, Brash DE. Common contaminants in next-generation sequencing that hinder discovery of low-abundance microbes. Gilbert T, editor. PLoS One [Internet]. 2014 May 16 [cited 2018 Mar 8];9(5):e97876. Available from: http://dx.plos.org/10.1371/journal.pone.0097876

21. Chafee M, Maignien L, Simmons SL. The effects of variable sample biomass on comparative metagenomics. Environ Microbiol [Internet]. 2015 Jul [cited 2017 Apr 8];17(7):2239–53. Available from: http://doi.wiley.com/10.1111/1462-2920.12668

22. Bowers RM, Clum A, Tice H, Lim J, Singh K, Ciobanu D, et al. Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. BMC Genomics [Internet]. 2015;16(856):1–12. Available from: http://www.biomedcentral.com/1471-2164/16/856

23. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog [Internet]. 2016;8(1):24. Available from: http://gutpathogens.biomedcentral.com/articles/10.1186/s13099-016-0103-7

24. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol [Internet]. 2014;15(12):564. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4311479&tool=pmcentrez&rendertype=abstract

25. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature [Internet]. 2017 Nov 1 [cited 2018 Mar 26];551(7681):457. Available from: http://www.nature.com/doifinder/10.1038/nature24621

26. Eutick ML, O'Brien RW, Slaytor M. Bacteria from the gut of Australian termites. Appl Environ Microbiol. 1978;35(5):823–8.

27. Socransky SS, Gibbons RJ, Dale AC, Bortnick L, Rosenthal E, Macdonald JB. The microbiota of the gingival crevice area of man—I. Arch Oral Biol [Internet]. 1963 May [cited 2017 May 18];8(3):275–80. Available from: http://linkinghub.elsevier.com/retrieve/pii/0003996963900190

28. O'Rourke J, Lee A, McNeill J. Differences in the gastrointestinal microbiota of specific pathogen free mice: an often unknown variable in biomedical research. Lab Anim [Internet]. 1988;22(4):297–303. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2976437

29. Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. Microbial ecology and evolution: a ribosomal RNA approach. Ann Rev Microbiol [Internet]. 1986 [cited 2018 Mar 26];40:337–65. Available from: https://www-annualreviews-org.liboff.ohsu.edu/doi/pdf/10.1146/annurev.mi.40.100186.002005

30. Smith GL, Socransky SS, Smith CM. Non-isotopic DNA probes for the identification of subgingival microorganisms. Oral Microbiol Immunol [Internet]. 1989 Mar [cited 2017 May 18];4(1):41–6. Available from: http://www.ncbi.nlm.nih.gov/pubmed/2628867

31. Weinstock GM. Genomic approaches to studying the human microbiota. Nature [Internet]. 2012 Sep 13 [cited 2018 Mar 25];489(7415):250–6. Available from: http://www.nature.com/articles/nature11553

32. Ferretti P, Farina S, Cristofolini M, Girolomoni G, Tett A, Segata N. Experimental metagenomics and ribosomal profiling of the human skin microbiome. Exp Dermatol [Internet]. 2017;26:211–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27623553%5Cnhttp://doi.wiley.com/10.1111/exd.13210

33. Glöckner FO, Yilmaz P, Quast C, Gerken J, Beccati A, Ciuprina A, et al. 25 years of serving the community with ribosomal RNA gene reference databases and tools. J Biotechnol [Internet]. 2017 Nov 10 [cited 2018 Mar 25];261:169–76. Available from: https://www.sciencedirect.com/science/article/pii/S0168165617314943

34. Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, et al. Open-source sequence clustering methods improve the state of the art. mSystems. 2016;1(1):1–16.

35. Eren AM, Morrison HG, Lescault PJ, Reveillaud J, Vineis JH, Sogin ML. Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. ISME J [Internet]. 2014 Oct 17 [cited 2016 Nov 13];9(4):968–79. Available from: http://www.nature.com/doifinder/10.1038/ismej.2014.195

36. Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics [Internet]. 2010 Oct 1 [cited 2016 Nov 11];26(19):2460–1. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20709691

37. Edgar RC. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nat Methods [Internet]. 2013 Aug 18 [cited 2017 Feb 11];10(10):996–8. Available from: http://www.nature.com/doifinder/10.1038/nmeth.2604

38. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods [Internet]. 2010 May 11 [cited 2018 Jan 7];7(5):335–6. Available from: http://www.nature.com/articles/nmeth.f.303

39. Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. Eisen JA, editor. PLoS Genet [Internet]. 2008 Nov 21 [cited 2017 Apr 23];4(11):e1000255. Available from: http://dx.plos.org/10.1371/journal.pgen.1000255

40. Edgar RC, Flyvbjerg H. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics [Internet]. 2015 Nov 1 [cited 2017 Apr 26];31(21):3476–82. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv401

41. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. Appl Environ Microbiol [Internet]. 2013 Sep 1 [cited 2017 Aug 15];79(17):5112–20. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23793624

42. Schirmer M, Ijaz UZ, D'Amore R, Hall N, Sloan WT, Quince C. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res [Internet]. 2015 Mar 31 [cited 2017 Sep 11];43(6):e37–e37. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gku1341

43. D'Amore R, Ijaz UZ, Schirmer M, Kenny JG, Gregory R, Darby AC, et al. A comprehensive benchmarking study of protocols and sequencing platforms for 16S rRNA community profiling. BMC Genomics [Internet]. 2016 Dec 14 [cited 2017 Sep 16];17(1):55. Available from: http://www.biomedcentral.com/1471-2164/17/55

44. Apprill A, McNally S, Parsons R, Weber L. Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. Aquat Microb Ecol [Internet]. 2015 Jun 4 [cited 2018 Mar 25];75(2):129–37. Available from: http://www.int-res.com/abstracts/ame/v75/n2/p129-137/

45. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, et al. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A [Internet]. 2011;108 Suppl(Supplement_1):4516–22. Available from: http://www.pnas.org/content/108/Supplement_1/4516.short

46. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J [Internet]. 2012 Aug 8 [cited 2018 Mar 25];6(8):1621–4. Available from: http://www.nature.com/articles/ismej20128

47. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. Environ Microbiol [Internet]. 2016 May 1 [cited 2018 Mar 25];18(5):1403–14. Available from: http://doi.wiley.com/10.1111/1462-2920.13023

48. Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, et al. Improved bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. mSystems [Internet]. 2016 Feb 25 [cited 2018 Mar 25];1(1):e00009-15. Available from: http://www.ncbi.nlm.nih.gov/pubmed/27822518

49. Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R. UCHIME improves sensitivity and speed of chimera detection. Bioinformatics [Internet]. 2011 Aug 15 [cited 2018 Jan 7];27(16):2194–200. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btr381

50. Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, et al. Oligotyping: differentiating between closely related microbial taxa using 16S rRNA gene data. Freckleton R, editor. Methods Ecol Evol [Internet]. 2013 Dec [cited 2016 Oct 19];4(12):1111–9. Available from: http://doi.wiley.com/10.1111/2041-210X.12114

51. Nucleotide [Internet]. Bethesda, MD: National Center for Biotechnology Information; 1988. Available from: https://www.ncbi.nlm.nih.gov/nucleotide

52. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol [Internet]. 1990 Oct 5 [cited 2018 Jan 9];215(3):403–10. Available from: http://www.sciencedirect.com.liboff.ohsu.edu/science/article/pii/S0022283605803602?via%3Dihub

53. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res [Internet]. 2012 Nov 27 [cited 2018 Mar 26];41(D1):D590–6. Available from: http://academic.oup.com/nar/article/41/D1/D590/1069277/The-SILVA-ribosomal-RNA-gene-database-project

54. Caporaso JG, Bittinger K, Bushman FD, DeSantis TZ, Andersen GL, Knight R. PyNAST: a flexible tool for aligning sequences to a template alignment. Bioinformatics [Internet]. 2010 Jan 15 [cited 2018 Jan 7];26(2):266–7. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btp636

55. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. VSEARCH: a versatile open source tool for metagenomics. PeerJ [Internet]. 2016 Oct 18 [cited 2017 Feb 22];4(e2584):1–22. Available from: https://peerj.com/articles/2584

56. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res [Internet]. 2002 Jul 15 [cited 2018 Jan 7];30(14):3059–66. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkf436

57. Kopylova E, Noe L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics [Internet]. 2012 Dec 1 [cited 2017 Feb 17];28(24):3211–7. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts611

58. Team RC. R: A language and environment for statistical computing. [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2017. Available from: https://www.r-project.org