Received: 25 May 2021 | Revised: 10 September 2021 | Accepted: 14 September 2021

DOI: 10.1002/edn3.255

Environmental DNA
Open Access
Dedicated to the study and use of environmental DNA for basic and applied sciences

WILEY

ORIGINAL ARTICLE

Dataset complexity impacts both MOTU delimitation and biodiversity estimates in eukaryotic 18S rRNA metabarcoding studies

Alejandro De Santiago[1,2] | Tiago José Pereira[1,2] | Sarah L. Mincks[3] | Holly M. Bik[1,2]

[1]Department of Marine Sciences, University of Georgia, Athens, GA, USA

[2]Institute of Bioinformatics, University of Georgia, Athens, GA, USA

[3]College of Fisheries and Ocean Sciences, University of Alaska, Fairbanks, AK, USA

Correspondence
Holly M. Bik, Department of Marine Sciences, University of Georgia, 325 Sanford Drive, Athens, GA 30602, USA.
Email: hbik@uga.edu

Funding information
Institutional startup funding from the University of Georgia; North Pacific Research Board, Grant/Award Number: NPRB project 1303; Gulf of Mexico Research Initiative

Abstract

How does the evolution of bioinformatics tools impact the biological interpretation of high-throughput sequencing datasets? For eukaryotic metabarcoding studies, in particular, researchers often rely on tools originally developed for the analysis of 16S ribosomal RNA (rRNA) datasets. Such tools do not adequately account for the complexity of eukaryotic genomes, the ubiquity of intragenomic variation in eukaryotic metabarcoding loci, or the differential evolutionary rates observed across eukaryotic genes and taxa. Recently, metabarcoding workflows have shifted away from the use of operational taxonomic units (OTUs) toward delimitation of amplicon sequence variants (ASVs). We assessed how the choice of bioinformatics algorithm impacts the downstream biological conclusions that are drawn from eukaryotic 18S rRNA metabarcoding studies. We focused on four workflows including UCLUST and VSearch algorithms for OTU clustering, and DADA2 and Deblur algorithms for ASV delimitation. We used two 18S rRNA datasets to further evaluate whether dataset complexity had a major impact on the statistical trends and ecological metrics: a "high complexity" (HC) environmental dataset generated from community DNA in Arctic marine sediments, and a "low complexity" (LC) dataset representing individually barcoded nematodes. Our results indicate that ASV algorithms produce more biologically realistic metabarcoding outputs, with DADA2 being the most consistent and accurate pipeline regardless of dataset complexity. In contrast, OTU clustering algorithms inflate the metabarcoding-derived estimates of biodiversity, consistently returning a high proportion of "rare" molecular operational taxonomic units (MOTUs) that appear to represent computational artifacts and sequencing errors. However, species-specific MOTUs with high relative abundance are often recovered regardless of the bioinformatics approach. We also found high concordance across pipelines for downstream ecological analysis based on beta-diversity and alpha-diversity comparisons that

Alejandro De Santiago and Tiago José Pereira are joint first authors.

Alejandro De Santiago and Tiago José Pereira contributed equally to this work.

utilize taxonomic assignment information. Analyses of LC datasets and rare MOTUs are especially sensitive to the choice of algorithms and better software tools may be needed to address these scenarios.

KEYWORDS
18S rRNA gene, ASVs, biodiversity, metabarcoding, microbial eukaryotes, nematodes, OTUs

1 | INTRODUCTION

Over the last ten years, there has been a rapid expansion of metabarcoding methods applied toward the study of eukaryotic taxa. These studies typically utilize loci such as the cytochrome c oxidase subunit I gene (COI) or other mitochondrial genes (e.g., 12S) for large vertebrates and macroinvertebrates (Arribas et al., 2016; Elbrecht & Leese, 2017; Hebert et al., 2003; Leray & Knowlton, 2016; Machida et al., 2012), the 18S or 28S ribosomal RNA (rRNA) nuclear genes for microbial invertebrates and single-celled eukaryotes (Creer et al., 2010; Pawlowski et al., 2012), ITS rRNA loci for fungi (Lindner et al., 2013; Toju & Baba, 2018), or the *rbcL* and *matK* plastid genes for plants and algae (Akita et al., 2019; Bell et al., 2017; Group et al., 2009; Wolf & Vis, 2020). Since metabarcoding studies utilize high-throughput sequencing (HTS) technologies (e.g., Illumina sequencing) and PCR primers that amplify a broad range of taxa from numerous samples at once, the resulting datasets have provided unprecedented biodiversity insights that have quickly expanded our understanding of spatio-temporal patterns and ecological interactions in diverse environments (Bik, Porazinska et al., 2012; Deiner et al., 2017; Zinger et al., 2019).

The massive amount of data produced by HTS platforms, however, requires stringent quality control steps to differentiate real biological reads from erroneous sequences. Proper study design is paramount for drawing robust biological and ecological conclusions from metabarcoding datasets (Abellan-Schneyder et al., 2021; Alberdi et al., 2018; Murray et al., 2015). However, even the most well-designed studies can contain artifacts stemming from PCR, sequencing errors (e.g., Illumina "tag bleeding"), run batch effects (Fidler et al., 2020; Leek et al., 2012; Schnell et al., 2015), and microbial contamination (e.g., introduced via kit reagents, [Salter et al., 2014]). While some of these artifacts can be detected and bioinformatically eliminated via the incorporation of rigorous control samples (e.g., DNA extraction blanks, PCR negative controls, and mock community standards, [Alberdi et al., 2018; Deiner et al., 2017; Hornung et al., 2019]) and data cleaning software (e.g., Decontam, microDecon, [Davis et al., 2018; McKnight et al., 2019]), other "artifacts" appear to result from the biological nuances of the chosen metabarcoding locus (i.e., differential patterns of gene evolution across taxa). For example, intragenomic variation persists within nuclear rRNA loci, even though rRNA tandem repeat arrays are subjected to concerted evolution within eukaryotic genomes (Bik et al., 2013; Kumar et al., 2017; Pawłowska et al., 2020; Zhao et al., 2019). As a result, an individual eukaryote will typically be represented by a "Head–Tail" pattern in 18S rRNA metabarcoding datasets, exhibiting a dominant "Head" sequence with high relative abundance that represents the species-specific DNA barcode, as well as a "Tail" of rarer low-abundance sequences that exhibit high pairwise similarity to the diagnostic reference barcode (Pereira et al., 2020; Porazinska et al., 2010, 2010). This phenomenon has also been reported for other metabarcoding makers (e.g., ITS; Anslan et al., 2018), despite less prevalent in the COI gene (Macheriotou et al., 2019).

The clustering of raw metabarcoding reads into molecular operational taxonomic units (MOTUs) is one ubiquitous approach for minimizing artifacts and reducing confounding biological variation such as intragenomic rRNA variation. MOTU is a general term for a DNA-based "species approximation" (Blaxter, 2016; Blaxter et al., 2005; Creer et al., 2010) that can be used to calculate traditional biodiversity indices and perform ecological analyses, either using taxa presence/absence or using relative abundance. Two distinct classes of MOTUs seen in the scientific literature are associated with specific bioinformatics algorithms for processing raw HTS data: operational taxonomic units (OTUs) and amplicon sequence variants (ASVs), the latter also known as exact sequence variation (ESVs) or zero-radius OTUs (ZOTUs) (Callahan et al., 2017; Edgar, 2017, 2018) In fact, the terminology for denoising methods such as ASVs has been used interchangeably (Antich et al., 2021; Terrat et al., 2020). OTUs emerged as an early approach for analyzing data from diverse HTS platforms, relying on the distance-based clustering of raw reads according to a set pairwise similarity cutoff (e.g., 97% and 99% for prokaryotes and eukaryotes, respectively [Bik, Porazinska et al., 2012; Deiner et al., 2017; He et al., 2015]). A plethora of OTU clustering algorithms now exist, which can cluster reads in myriad ways (Jackson et al., 2016; Prodan et al., 2020), with endless parameter choices for the strictness (or inclusivity) of the OTU cluster generation steps. Two highly popular OTU clustering algorithms include UCLUST (Edgar, 2010) and VSearch (Rognes et al., 2016), respectively implemented in the QIIME1 (Caporaso et al., 2010) and QIIME2 (Bolyen et al., 2019) software suites for microbial ecology. Newer ASV algorithms utilize Illumina error-correction approaches to perform quality checks and processing on raw sequence reads. Currently, the two most popular tools for ASV generation are DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017), although other ASV algorithms are rapidly emerging (e.g., UNOISE2 for ZOTUs, [Edgar, 2017]). The most important distinction between these two classes of MOTUs is the increased sophistication of ASV algorithms, which include

complex mathematical modelling approaches aimed at eliminating false positives, chimeras, and sequencing artifacts (although note that more simplistic chimera checking steps are also ingrained in many OTU clustering pipelines, [Edgar et al., 2011; Mysara et al., 2017]). Furthermore, ASVs are thought to represent the true DNA sequences present in the target community and are thus reusable and reproducible across metabarcoding studies (in contrast to *de novo* OTUs which are emergent features of a dataset and must be treated as study-specific clusters [Callahan et al., 2017]). The impact of different MOTU generation workflows, and the biological implications for the move toward ASVs, has not been extensively documented for eukaryotes. Worth mentioning, BIOCOM-PIPE (Djemiel et al., 2020) and PEMA (Zafeiropoulos et al., 2020) are flexible pipelines allowing the analysis of multiple gene markers (e.g., 16S, 18S, ITS, and COI) at once, and offering different clustering methods in addition to ASVs. Similarly, Brandt et al. (2021) have combined denoising and clustering approaches to evaluate both prokaryote and eukaryote diversity in metabarcoding datasets. For eukaryotes specifically, Giebner et al. (2020) have recently compared the diversity level of COI and 18S rRNA markers as well as metabarcoding and sequence capture approaches. Still, the vast majority of benchmarking studies and algorithm comparisons have been performed on 16S rRNA studies of bacterial and archaeal communities (Caruso et al., 2019; Escudié et al., 2018; Ghodsi et al., 2011; Jackson et al., 2016; Nearing et al., 2018; Prodan et al., 2020). Although these prokaryotic-focused studies have provided important insights on the performance and ecological relevance of different computational workflows, their relevance for eukaryotic taxa is not always clear. Eukaryotic genome complexity and evolution may pose challenges for bioinformatics algorithms originally optimized for use with bacterial/archaeal 16S rRNA datasets. Additionally, many eukaryotic metabarcoding studies include complementary visual surveys of taxa (e.g., microscopy, trawls, kick sampling, and video transects), which can provide critical independent observations for validating DNA-based inferences derived from HTS datasets (Cahill et al., 2018; Dell'Anno et al., 2015; Djurhuus et al., 2018; Geisen et al., 2018; Schuelke et al., 2018).

The current eukaryotic metabarcoding literature is heavily focused on methods comparisons for field sampling (Beentjes et al., 2019; Koziol et al., 2019; Turner et al., 2015) and wet laboratory protocols (e.g., DNA extraction comparisons, design of improved primer sets, [Alberdi et al., 2018; Bradley et al., 2016; Brannock & Halanych, 2015]), and comparison of intraspecific vs. interspecific genetic diversity for the COI gene (Elbrecht et al., 2018; Leray et al., 2013; Macheriotou et al., 2019). Yet, method comparisons of eukaryotic metabarcoding workflows and benchmarking of other loci (i.e., 18S rRNA and ITS rRNA datasets) are still limited, particularly when assessing the performance of algorithms producing distinct classes of MOTUs (i.e., OTUs vs. ASVs, [Bálint et al., 2014; Macheriotou et al., 2019; Pauvert et al., 2019]). Thus, there is a pressing need for downstream comparisons of bioinformatics tools in eukaryotic metabarcoding studies especially those that evaluate the biological implications of different MOTU generation approaches and how the choice of software tools impacts downstream ecological analyses (but see above; [Brandt et al., 2021; Djemiel et al., 2020; Giebner et al., 2020; Zafeiropoulos et al., 2020]).

In the present study, we aimed to evaluate how distinct bioinformatics pipelines may impact the biological inferences drawn from 18S rRNA metabarcoding datasets. We focused on four computational workflows representing two distinct classes of MOTU generation (Figure 1): OTU clustering using VSearch and UCLUST, and ASV generation using DADA2 and Deblur. Our study compared results from two 18S rRNA metabarcoding datasets generated as part of other ongoing projects: a "high complexity" (HC) environmental dataset generated from bulk community DNA in Arctic marine sediments, and a "low complexity" (LC) dataset representing metabarcoding profiles from individually barcoded nematodes, where morphological identification of each specimen was verified under light microscopy (Schuelke et al., 2018). Both datasets represent the same 18S rRNA gene region (V1–V2 hypervariable regions) and were generated using the same PCR primer set and molecular wet laboratory protocols, thus facilitating direct comparison of downstream bioinformatics outputs. We hypothesized that (1) ASVs would represent a more biologically relevant approach for eukaryotic metabarcoding data (i.e., a more accurate representation of the biodiversity), (2) species-specific DNA barcodes (Head MOTUs exhibiting high relative abundance) would be consistently recovered across all bioinformatics pipelines, (3) computational algorithms and parameters would strongly influence downstream estimates of alpha- and beta diversity, and (4) dataset complexity (i.e., the level of biodiversity contained in each metabarcoding sample) would impact the performance of bioinformatics pipelines in unanticipated ways.

2 | MATERIALS AND METHODS

2.1 | Sample collection and data generation

In this study, we used a "high complexity" (HC) environmental dataset generated from community DNA in Arctic marine sediments, and a "low complexity" (LC) dataset representing individually barcoded nematodes. We defined the datasets as HC and LC based on the expected diversity of each sample. Single-nematode metabarcoding samples are expected to display a lower diversity, which is represented by the nematode-specific DNA barcode and other additional sequences of associated taxa (e.g., gut content and parasites). On the contrary, marine sediment samples are likely to be represented by highly diverse benthic communities, including multiple phyla.

The Arctic metabarcoding dataset (HC) was generated using 127 marine sediment samples collected from the continental shelf/slope in the Northeast Chukchi and Beaufort Seas (Figure S1, Table 1). This dataset was generated as part of a broad study focused on evaluating benthic meiofauna community structure through a combination of morphological taxonomy and omics

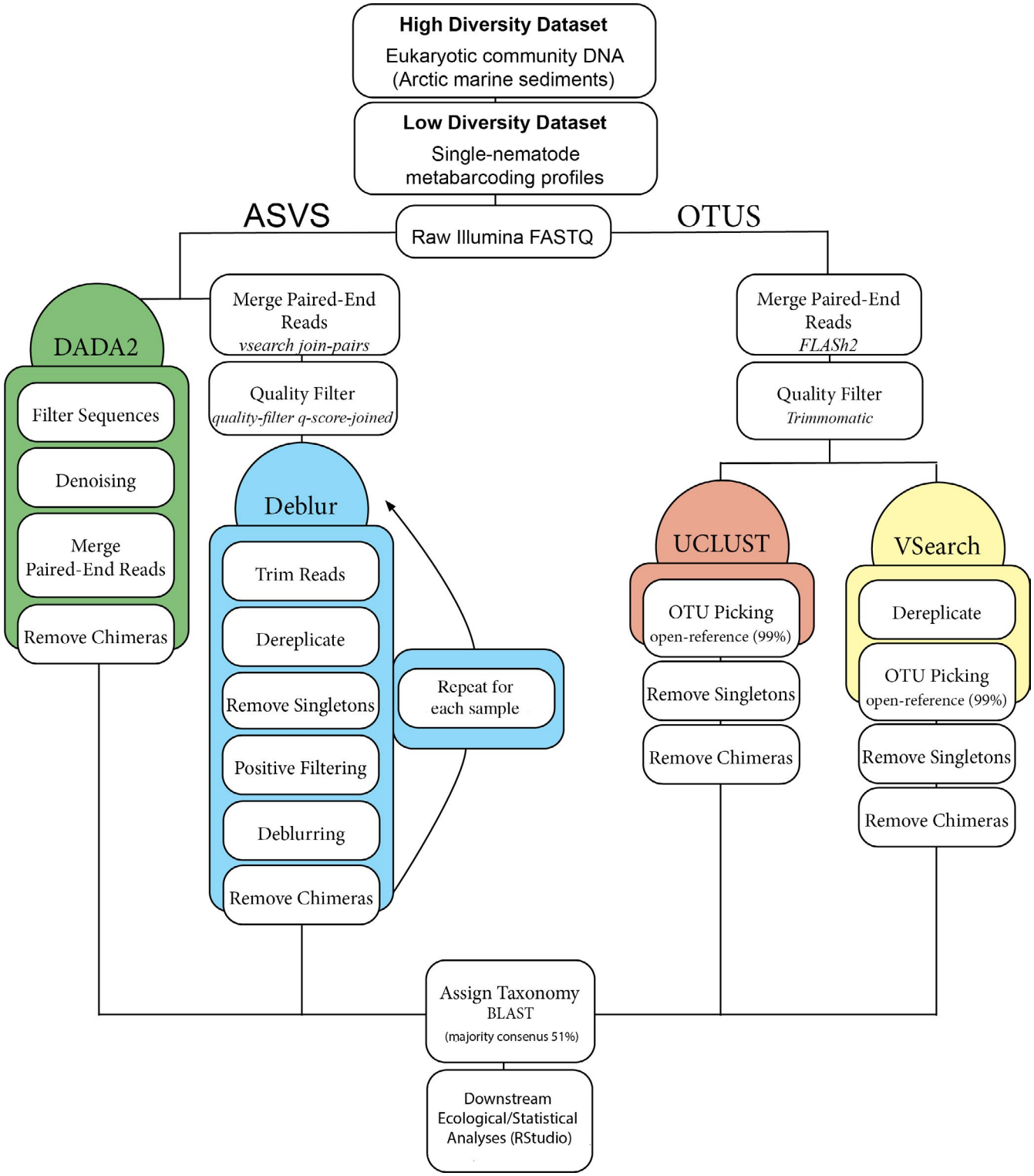


FIGURE 1 Workflow diagram of the four different bioinformatics pipelines used in this study. For the UCLUST and VSearch pipelines, reads were first quality controlled and merged using FLASH2 and Trimmomatic. Reads were clustered into OTUs using a 99% similarity threshold. Singletons and chimeric sequences were removed. For DADA2, the demultiplexed FASTQ files were used to estimate ASVs. For Deblur, denoising was done on merged quality-filtered reads. Unlike the DADA2 algorithm, Deblur processes each sample independently. Four matrices were produced for each dataset. Taxonomy was assigned using BLAST+ (BLAST majority consensus) in QIIME2 v2019.4

approaches. Additional details on sediment sample collection and sample processing are provided in Mincks et al. (2021). Briefly, frozen sediment samples were thawed and meiofaunal organisms were isolated via decantation over a 63-μm mesh sieve. Total genomic DNA was extracted from material retained on the sieve using MoBio PowerSoil® kits (MoBio Laboratories, Inc.). A fragment of

TABLE 1 Description of geographic locations and number of samples included in this study

Alaskan continental shelf High complexity (HC) Dataset			
Geographic Subregion	**Number of samples**	**Region**	**Depth range (m)**
Alaskan Beaufort Shelf	8	Arctic	50–500
Amundsen Gulf	30	Arctic	32–352
Banks Island	7	Arctic	51–379
Camden Bay	13	Arctic	20–350
Chukchi Sea	32	Arctic	25.4–51.1
Mackenzie River Plume	37	Arctic	17–1200
Nematode microbiomes Low complexity (LC) dataset			
Nematode family[a]	**Number of samples**[b]	**Region**	**Depth range (m)**
Chromadoridae	31	Arctic, Gulf of Mexico, Southern California	0–1000
Comesomatidae	46	Arctic, Gulf of Mexico	20–1000
Desmoscolecidae	29	Arctic, Gulf of Mexico	200–1000
Oxystominidae	27	Arctic, Gulf of Mexico	20–1000
Other	94	Arctic, Gulf of Mexico, Southern California	0–2239

[a]The category "Other" includes less common nematode families (a total of 20).

[b]For the Low Complexity (LC) dataset, number of samples corresponds to the number of nematode specimens belonging to each respective family.

the 18S rRNA gene (~400-bp, V1–V2 hypervariable regions) was amplified from environmental DNA extracts using the F04/R22 eukaryotic primers (Blaxter et al., 1998). For marine meiofauna, including nematodes, the V1–V2 or V9 regions of 18S rRNA gene provide the best binding and taxonomic coverage across meiofaunal phyla. However, the former region is more variable, providing higher resolution in separating taxa (Creer et al., 2010), as well as better represented in molecular databases. PCR products from each sediment sample were tagged with a unique nucleotide barcode and pooled before sequencing.

The single-specimen metabarcoding dataset (LC) was generated from individual marine nematodes collected in the Pacific, Arctic, and Gulf of Mexico (Figure S1, Table 1). This dataset was originally generated by Schuelke et al. (2018) who evaluated microbiome patterns (archaea and bacteria) associated with diverse marine nematode genera. Further exploration of intragenomic rRNA patterns in marine nematodes was carried out by Pereira et al. (2020) in a refined version of the original dataset (i.e., 227 samples). Both studies provide detailed methods regarding nematode sorting, taxonomic identification, and genus-level diversity patterns. Wet laboratory protocols for single-nematode metabarcoding included an extensive suite of blank/control samples as a checkpoint for contamination in these low biomass samples. After morphological vouchering of each nematode specimen, samples were submitted to molecular procedures described in Pereira et al. (2020). Briefly, DNA was extracted from individual nematodes using a Proteinase K "Worm Lysis Buffer" protocol. PCR amplification of the 18S rRNA gene was carried out using the same eukaryotic primer set (F04/R22; [Blaxter et al., 1998]) and multiplexing/pooling procedures as described for the HC dataset above.

All PCR products were purified using magnetic beads following the manufacturer's protocol (Agencourt AMPure XP beads; Beckman Coulter). Sample concentrations were subsequently measured using a Qubit® 3.0 Fluorometer and a Qubit® dsDNA HS (High Sensitivity) Assay Kit (Thermo Fisher Scientific). Normalization values were calculated to ensure that approximately equivalent DNA concentrations were pooled across all samples, including controls and blank samples. The final pooled libraries were subjected to an additional magnetic bead cleanup step, followed by size selection on a BluePippin (Sage Science) to remove any remaining primer dimer and isolate target PCR amplicons within the range of 300–700 bp. A Bioanalyzer trace was run on each size-selected pool as a quality control measure, and the pooled 18S rRNA amplicon libraries were sequenced in two separate runs on the Illumina MiSeq platform (2 × 300-bp paired-end runs) at the UC Davis Genomics Core Facility. All wet laboratory protocols, sample mapping files, and downstream bioinformatics scripts used in this study have been deposited on GitHub (https://github.com/BikLab/OTU-ASV-euk-benchmarking).

2.2 | OTU and ASV pipeline designs

OTU clustering was carried out using the UCLUST and VSearch algorithms, while ASVs were generated using the DADA2 and Deblur pipelines (Figure 1). For OTU picking workflows, raw Illumina reads were first demultiplexed, merged using FLASH2 (Magoč & Salzberg, 2011), and quality filtered using Trimmomatic (Bolger et al., 2014) in conjunction with a custom script described in Schuelke et al. (2018). UCLUST (Edgar, 2010) was implemented within QIIME1 v1.9.1 (Caporaso et al., 2010), whereby quality-filtered reads were clustered

at 99% using the subsampled open-reference protocol (*pick_open_reference_otus.py*), with 10% subsampling of failed reads, and discarding of singletons from the final OTU table (Rideout et al., 2014). For metazoans, and nematodes in particular, clustering at 99% has been widely accepted as a proxy for species delimitation (Fonseca et al., 2017; Macheriotou et al., 2019). Chimeric sequences were identified and removed from the OTU table using USEARCH61, and the resulting OTUs were aligned to the QIIME-formatted SILVA132 database using the Pynast aligner and UCLUST (pairwise alignment method). Highly variable regions were removed using *filter_alignment.py*. Next, aligned representative sequences were used to construct a phylogeny using Fasttree (Price et al., 2009) and rooted using the midpoint method (*make_phylogeny.py*). VSearch (Rognes et al., 2016) was implemented using QIIME2 v2019.4 (Bolyen et al., 2018), whereby quality-filtered reads were clustered at 99% using *cluster-features-open-reference*. Singletons were filtered out from the OTU table and chimeric sequences identified using the *de novo* option; both chimeric and borderline chimeric sequences were removed from the OTU table. The resulting representative sequences were aligned using MAFFT (Katoh et al., 2002), ambiguous regions were masked, and the resulting aligned sequences were used to infer a phylogenetic tree using Fasttree (*align-to-tree-mafft-fasttree*, with rooting using the midpoint method). Discrepancies between UCLUST and VSearch workflows were a result of algorithm changes between the QIIME1 and QIIME2 pipelines (e.g., MAFFT replacing the depreciated Pynast aligner in QIIME1, and UCLUST being replaced with VSearch in QIIME2). The UCLUST and VSearch workflows thus aimed to carry out the same approximate bioinformatics steps in QIIME1 vs. QIIME2, respectively. It was, therefore, crucial to assess whether these pipeline changes in QIIME2 would result in noticeable discrepancies in downstream results and biological and ecological interpretations.

For ASV workflows, DADA2 (Callahan et al., 2016) and Deblur (Amir et al., 2017) were both executed within QIIME2 v2019.4. DADA2 was implemented directly on the demultiplexed FASTQ files since this algorithm incorporates quality-filtering of raw Illumina reads. The DADA2 error-correcting algorithm was run using default parameters, except for the truncating and trimming parameters. Forward and reverse reads were truncated at 220 bp and 236 bp for the HC dataset and at 232 bp and 253 bp for the LC dataset (median PHRED score ≥30). Unlike DADA2, Deblur does not differentiate between forward and reverse reads. Therefore, reads were initially merged using the *Vsearch join-pairs* command and quality-filtered according to the *q-score-joined* protocol. Deblur was implemented on the merged quality-controlled reads using default parameters. Reads were trimmed at 360 bp and 350 bp for the HC and LC dataset, respectively, thus allowing us to maximize the initial number of reads for the Deblur pipeline. Deblur internally removes chimeric and error-prone reads on a sample-by-sample basis. A tree was generated for the DADA2 and Deblur datasets using the *align-to-tree-mafft-fasttree* workflow in QIIME2 as described above (see VSearch method).

2.3 | Bioinformatics analysis

All downstream analyses were conducted in RStudio (Team, 2017) using the phyloseq (McMurdie & Holmes, 2013), vegan (Oksanen, 2011), ggplot2 (Wickham, 2009), and ggpubr (Kassambara, 2018) packages. First, we recorded the overall number of reads and OTUs/ASVs retained by each pipeline (Tables S1and S2), including comparisons between individual samples and datasets (HC and LC). To assess whether the number of MOTUs was correlated with the number of reads retained by each pipeline, we estimated correlations (R^2 and *p*-value) using Pearson's correlation coefficient (Figure 2).

Second, we aimed to evaluate whether each pipeline resulted in similar alpha- and beta-diversity metrics across Arctic subregions (HC dataset), these representing well-defined geographic areas/habitats and known to be under the influence of specific oceanographic conditions (e.g., Mackenzie River plume). For nematode metabarcoding samples (LC dataset), we grouped specimens representing the four major (most common) nematode families *viz*. Chromadoridae, Comesomatidae, Desmoscolecidae, and Oxystominidae, thus providing enough "replicates" for further statistical analyses focusing on ecological patterns. Importantly, nematode species belonging to the same family often belong to the same trophic group (Table 1).

Alpha diversity was calculated in phyloseq using three different indices (observed MOTUs, Simpson, and Shannon) and visualized using ggplot2 and ggpubr (Figures S2,S3, and S4). Kruskal–Wallis (KW) analysis was used to test for significant differences in alpha diversity among Arctic subregions and nematode groups. Pairwise comparisons were performed using the Wilcox test when significant differences among groups were detected. Principal coordinate analysis (PCoA) was carried out to assess beta diversity using three different metrics: Bray–Curtis similarity, weighted Unifrac, and unweighted Unifrac (Lozupone & Knight, 2005). The HC dataset was rarefied at 1000 sequences per sample, resulting in five samples that were excluded from the UCLUST, DADA2, and Deblur pipelines (Table S1). The LC dataset was also rarefied at 1000 sequences per sample, resulting in 22 samples that were excluded from the Deblur pipeline (Table S2). Finally, a Procrustes analysis was used to determine concordance among the MOTU pipelines and beta-diversity metrics. Procrustes were visualized using a combination of vegan and ggplot2 packages. The Procrustes results were further assessed using PROTEST with 9999 permutations (Jackson, 1995). Procrustes and PROTEST were implemented on rarefied datasets (Table 2). For each pairwise comparison using Procrustes/PROTEST, samples not present in both pipelines were excluded from the analysis (i.e., samples that did not meet the minimum rarefaction threshold). The impact of bioinformatics parameters and distance indices was assessed using the returned R and M^{12} values from each pairwise comparison. Outputs are highly concordant if two algorithms displayed a high R value alongside a low M^{12} value, whereas the opposite (i.e., low R value alongside a high M^{12} value) characterizes them as being more discordant.

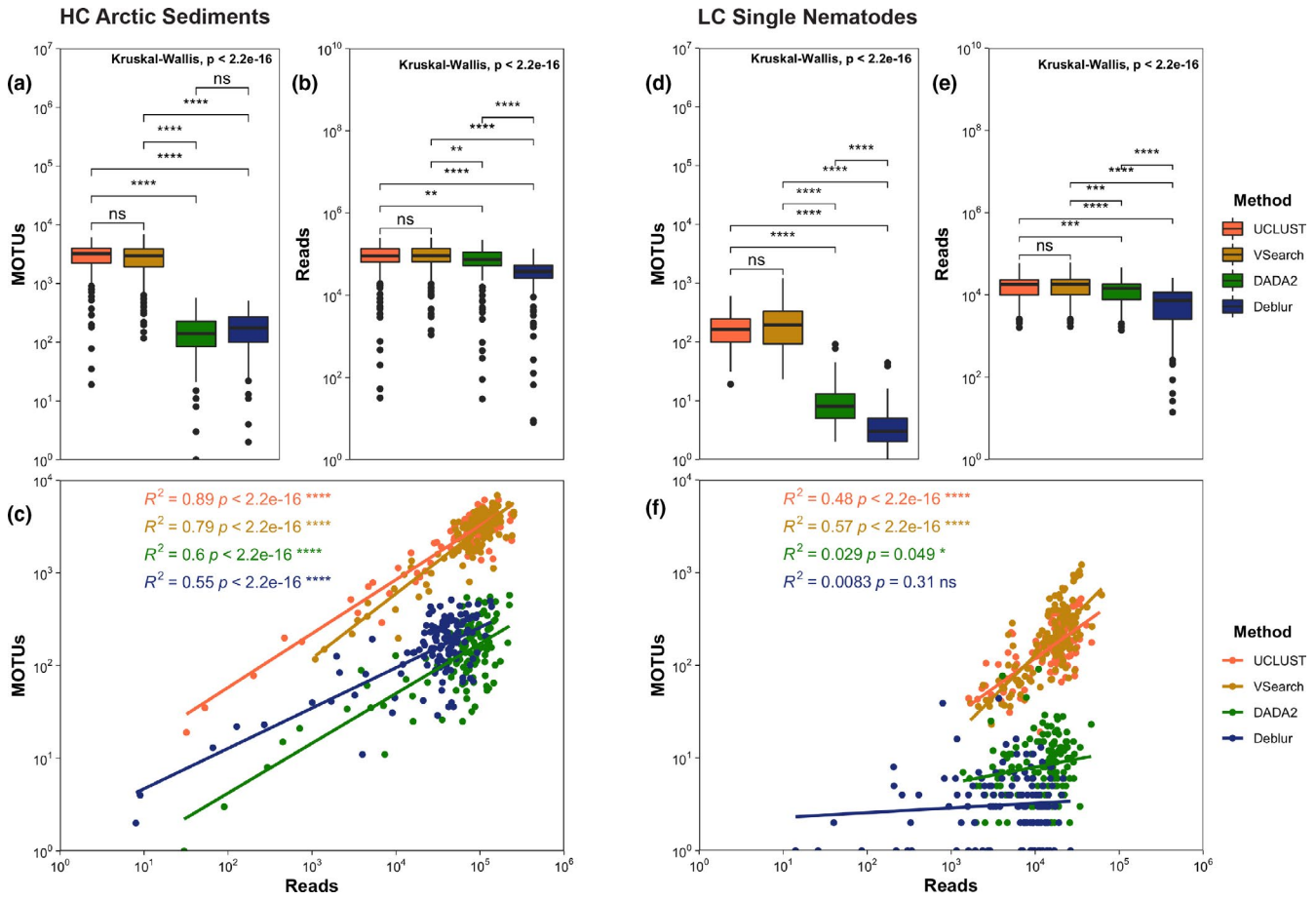


FIGURE 2 Number of MOTUs and reads retained by UCLUST, VSearch, DADA2, and Deblur. Median number of MOTUs (a, d) and reads (b, e) for HC and LC datasets, respectively. KW analysis was used to test for significant differences among bioinformatics pipelines for HC and LC datasets. Relationship between the number of MOTUs and number of reads (c, f) across the four bioinformatics pipelines for HC and LC datasets, respectively

TABLE 2 Results from the PROTEST/Procrustes analysis

	HC dataset						LC dataset					
	Unweighted-unifrac		Weighted-unifrac		Bray–Curtis		Unweighted-unifrac		Weighted-unifrac		Bray–Curtis	
Methods	R	M^{12}	R	M^{12}	R	M^{12}	R	M^{12}	R	M^{12}	R	M^{12}
UCLUST—VSearch	0.94	0.12	0.90	0.19	**0.96**	**0.08**	0.88	0.22	0.70	0.51	**0.97**	**0.05**
UCLUST—DADA2	0.90	0.19	0.92	0.15	**0.97**	**0.06**	0.77	0.41	0.74	0.45	**0.98**	**0.04**
UCLUST—Deblur	0.88	0.22	0.93	0.13	**0.96**	**0.08**	0.66	0.57	0.73	0.48	**0.98**	**0.04**
VSearch—DADA2	0.90	0.19	0.92	0.16	**0.95**	**0.11**	0.79	0.38	0.92	0.15	**0.95**	**0.08**
VSearch—Deblur	0.88	0.22	**0.93**	**0.13**	**0.93**	**0.13**	0.69	0.52	0.85	0.29	**0.95**	**0.09**
DADA2—Deblur	0.89	0.20	0.95	0.10	**0.98**	**0.03**	0.81	0.35	0.93	0.13	**0.99**	**0.01**

For each dataset (i.e., distance index), the highest values of concordance (R: correlation coefficient and M^{12}: goodness-of-fit statistic) are highlighted in bold. Beta diversity was estimated by rarefying data matrices at 1000 reads per sample. For all comparisons, *p*-value was always significant ($p < 0.01$).

Taxonomy was assigned for each MOTU table using QIIME2 v2019.4. For UCLUST, FASTA files containing OTU representative sequences were reformatted for compatibility with QIIME2 v2019.4. For all four pipelines, taxonomy was assigned using BLAST + with a minimum of 90% sequence identity, which is generally the cutoff used for "Phylum" level assignments (Bik,

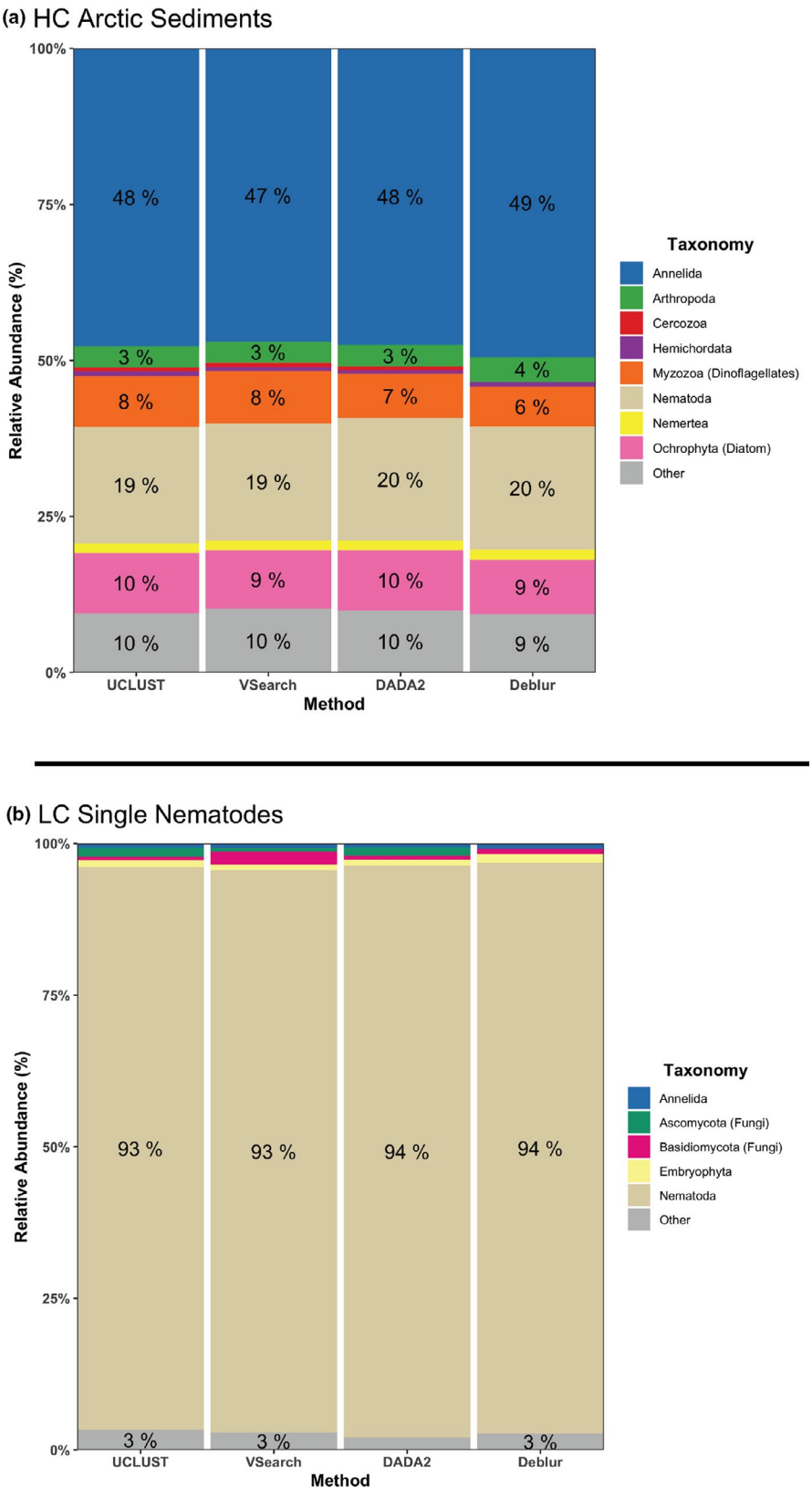


FIGURE 3 Barplots showing the most abundant taxa (i.e., relative abundance >0.5) across the different bioinformatics pipelines for (a) HC and (b) LC datasets. Lower abundant taxa were grouped in the category "Other." Taxonomic rank (phylum level) is the same for both HC and LC datasets. Note that the most abundant taxa do not necessarily match across pipelines

Sung et al., 2012; Creer et al., 2010). Our reference database of full-length 18S rRNA sequences was the QIIME-formatted SILVA 132 database with additional curated 18S rRNA sequences, as described in Pereira et al. (2020). The impact of the MOTU pipelines on taxonomic assignment was further explored by generating barplots with ggplot2 of taxonomic groups with abundances >0.5% (collapsed at phylum level for both HC and LC datasets; Figure 3).

3 | RESULTS

3.1 | Comparison of MOTUs recovered across bioinformatics pipelines

The total number of processed reads and MOTUs, including summary statistics (i.e., mean, median, standard deviation, and minimum/maximum values) obtained across the four pipelines for the HC and LC datasets are presented in Tables S1 and S2, respectively. For processed Illumina reads, significant differences were detected among pipelines (Global KW, <0.01, LC $X^2 = 111.12$; HC $X^2 = 110.29$; Figure 2), except when comparing UCLUST vs. VSEARCH (not significant in either the HC or LC dataset; Figure 2b,e). Furthermore, while three pipelines resulted in a similar number of final processed reads (UCLUST, VSearch, and DADA2), the Deblur algorithm returned far fewer reads and recovered only about 43–57% of the processed reads seen in the other pipelines (Tables S1 and S2).

The median number of MOTUs produced by OTU pipelines (UCLUST and VSearch) was at least 15-fold higher than the number of ASVs resulting from either DADA2 or Deblur (KW <0.01; Figure 2a,d). For both datasets, there were no significant differences in the median number of MOTUs produced by UCLUST and VSearch. However, we observed subtle differences in OTU membership, which includes the overall number of OTUs and OTU frequency/abundance. In this sense, VSearch consistently returned a higher proportion of "rare" OTUs (i.e., containing ≤10 reads per OTU; Table 3). This higher level of rare OTUs was much more pronounced for the LC dataset, with VSearch returning 39,967 more rare OTUs than UCLUST (a 266% increase; Table 3). For the HC dataset, VSearch exhibited 27,538 more rare OTUs than UCLUST (a 137% increase; Table 3). The MOTU patterns for ASV algorithms showed stark differences across the HC and LC datasets. For the Deblur pipeline, the median number of ASVs observed in the LC dataset was significantly lower than DADA2 (*p*-value ≤0.001, Figure 2d). In contrast, Deblur returned a higher median number of ASVs in the HC dataset, but not significantly different from DADA2 (Figure 2a). Although the number of Deblur ASVs appeared to be impacted by data complexity, the Deblur pipeline always returned lower total numbers of processed reads and MOTUs in each overall dataset compared with DADA2 (Table 3).

When looking at subsets (i.e., Arctic subregions and nematode families for the HC and LC datasets, respectively; Figure S2 and S3), these patterns of recovered reads and MOTUs were highly consistent across bioinformatics pipelines, particularly for the HC dataset. For example, no significant differences ($p > 0.05$) were detected among Arctic subregions for both MOTUs and retained reads (Figure S2). For the LC dataset, we also observed an overall agreement across algorithms for the number of reads, except for Deblur which differed from the others (e.g., family Oxystominidae; Figure S3). However, for MOTUs, we observed changes (i.e., from non-significant to significant) within and between algorithm classes in the LC dataset (Figure S3).

We also detected a significant positive correlation between the number of MOTUs and retained reads in both the OTU and ASV algorithms, except for Deblur in the LC dataset ($R^2 = 0.01$, *p*-value = 0.31). Although all four pipelines showed this relationship (Figure 2c,f), correlations were more moderate for methods estimating ASVs. This correlation between MOTUs and retained reads may be impacted by community complexity, sequencing depth, or both.

3.2 | MOTU rank abundance and taxonomy profiles

Given the significant differences in total number of reads and MOTUs recovered across algorithm classes (Tables S1 and S2), we next sought to investigate the source of this discrepancy and assess the potential implications for biological interpretations of environmental metabarcoding datasets. Rank-abundance curves were generated for MOTUs recovered across all four pipelines in the HC and LC datasets (Figure 4a,b). In both datasets, the OTU algorithms (UCLUST and Vsearch) exhibited a typical L-shaped rank-abundance curve with a steep gradient, indicating that top ranked OTUs had much higher abundances than the "long tail" of rare OTUs. OTU datasets were almost entirely dominated by rare MOTUs, representing anywhere from 75.8% to 96.5% of the entire dataset (Table 3). In contrast, the ASV algorithms showed S-shaped (DADA2) and C-shaped (Deblur) rank-abundance curves with much more gentle slopes. Overall, these patterns seem not to be impacted by dataset complexity.

Both DADA2 and Deblur strongly reduced the collection of rare MOTUs (e.g., the "long tail") commonly found in metabarcoding datasets (Figure 4, Table 3). The Deblur pipeline returned the lowest proportion of rare MOTUs across both datasets (i.e., only 4.4% and 6.8% of the HC and LC datasets, respectively), while DADA2

TABLE 3 Total number of processed reads, MOTUs, and rare MOTUs (≤10 reads) for each dataset and pipeline

Dataset	Metric	UCLUST	VSearch	DADA2	Deblur
HC arctic sediments	Total read number	11,902,088	12,196,959	9,986,980	5,229,272
	Total MOTUs	97,600	114,957	5,409	2,977
	% of MOTUs≤ 10 reads(number)	**75.8%(73,979)**	**88.3%(101,517)**	**24.0%(1,296)**	**4.4%(131)**
LC single nematodes	Total read number	9,926,235	10,108,666	8,060,929	4,590,872
	Total MOTUs	27,920	66,381	1,972	555
	% MOTUs≤ 10 reads(number)	**86.4%(24,113)**	**96.5%(64,080)**	**12.4%(244)**	**6.8%(38)**

Bold text indicates the percentage and number of rare MOTUs in the HC and LC datasets.

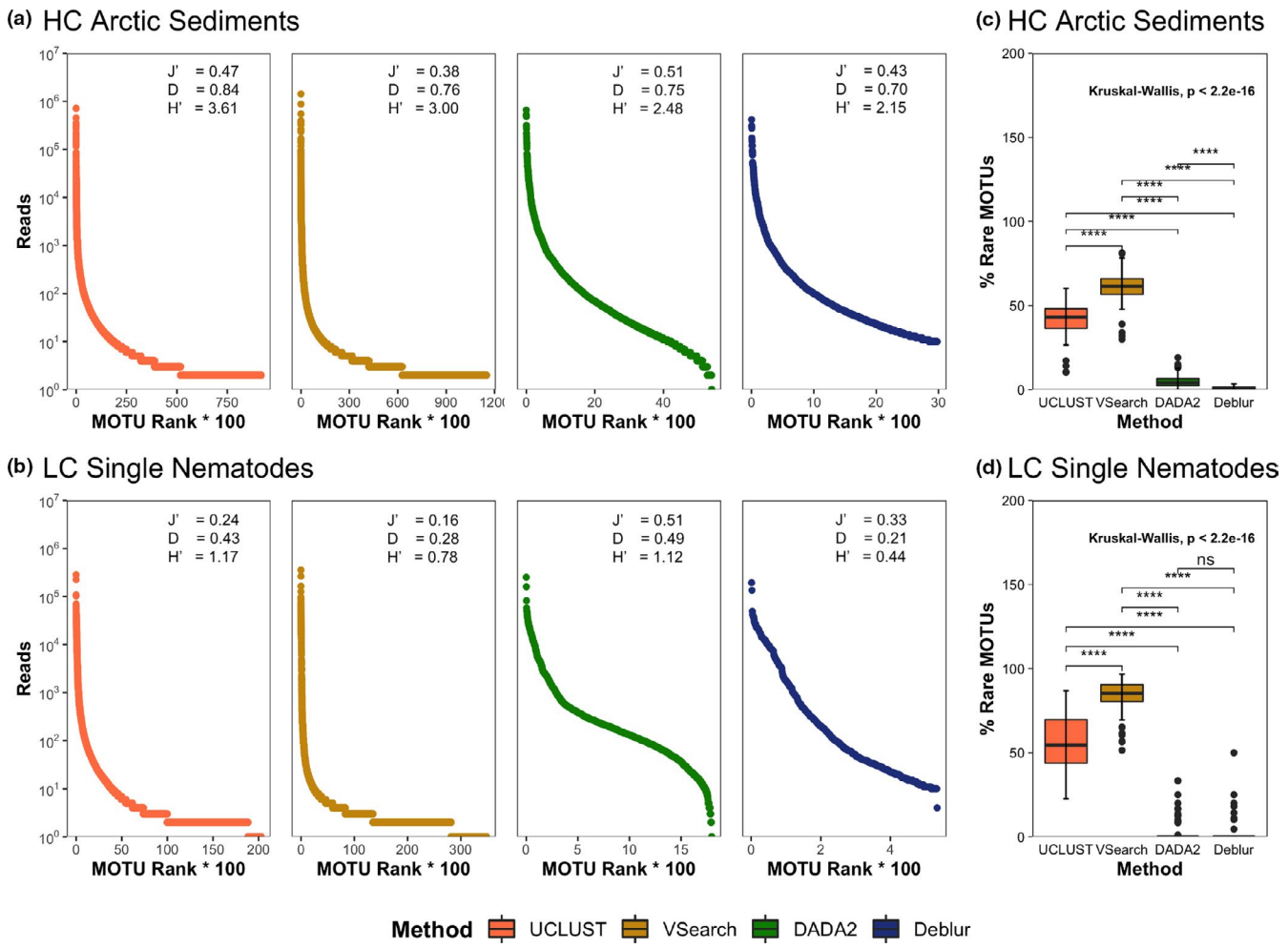


FIGURE 4 Ranked MOTU distribution showing "Head–Tail" patterns across pipelines. Rank-abundance curves for the HC (a) and LC (b) datasets. In the x-axis, the number of ranked MOTUs is multiplied by a factor of 100. Alpha-diversity indices including Shannon (H'), Simpson (D), and Pielou's evenness (J') are also given for each method/dataset. For rare MOTUs (i.e., containing ≤10 reads), KW analysis was used to test for significant differences among bioinformatics pipelines for HC (c) and LC (d) datasets

exhibited slightly higher proportions of rare MOTUs (24% and 12.4% of OTUs in the HC and LC datasets, respectively). This difference between ASV algorithms is also apparent in the MOTU rank-abundance curves (Deblur having a shorter tail of rare MOTUs; Figure 4). We also noted that dataset complexity appears to impact the percentage of rare MOTUs recovered in ASV pipelines: the HC dataset contained a higher proportion of rare MOTUs in DADA2 (as expected given the presumed higher sample biodiversity), while Deblur returned a higher proportion of rare MOTUs in the LC dataset, suggesting that dataset complexity may be a confounding factor for the Deblur algorithm performance. Overall, there appeared to be a higher evenness among MOTUs produced from ASV algorithms (DADA2 and Deblur), which is reflected by the slope of the curves (i.e., gentle slopes indicating less dominance of fewer top ranked ASVs) and the fewer low-ranked ASVs (Figure 4). Moreover, deeper exploration of the LC dataset (i.e., four major nematode families) confirmed that the behavior between methods within the same MOTU class is more similar than when comparing across classes and that OTU clustering algorithms tend to produce longer low-abundance tails (Figure S5).

Although the shape of rank-abundance curves varied across methods in the HC and LC datasets, the abundance of the dominant MOTU (i.e., the first rank MOTU) was consistent across all pipelines, and about an order of magnitude higher than the second most abundant MOTU (Figure 4). In the LC dataset where we could evaluate the MOTU profile of individual nematodes, the nucleotide sequence of the most abundant (Head) MOTU remained identical across all four algorithms (Table 4). Moreover, we observed that Head MOTUs in ASV pipelines showed a 10–20% reduction in relative abundance, presumably because ASVs are a less inclusive "cloud" of sequence reads. When examining Head MOTUs in the HC dataset, however, congruence was only detected between algorithms of the same MOTU class, which also led to some different taxonomic assignments (Table 4). The primary difference across pipelines remained the curve shape and length of the "long tail" of rare MOTUs, with algorithm class having the strongest influence (OTU vs. ASV pipelines). Similar patterns were also apparent in subgroups of the LC dataset, where OTU algorithms appeared to produce a longer and more prominent tail of rare MOTUs (e.g., Chromadoridae, Figure

TABLE 4 Ten most abundant MOTUs for each bioinformatics pipeline in the HC and LC datasets. UCLUST taxonomy and relative abundance (RA %) are used as a reference for comparison with the other three methods

Method	MOTU	RA (%)	HC Dataset – Taxonomy	Vsearch (RA %)	Dada2 (RA %)	Deblur (RA %)
UCLUST	MOTU1	6.0	Annelida; Polychaeta; Scolecida; Spionida	OTU1 (11.69)	ASV2 (5.37)	ASV2 (5.91)
	MOTU2	6.0	Annelida; Polychaeta; Palpata; Phyllodocida	OTU2 (7.20)	ASV1 (6.64)	ASV1 (7.96)
	MOTU3	3.8	Ochrophyta; Diatomea; Bacillariophytina; Mediophyceae; Chaetoceros	OTU3 (4.5)	ASV4 (4.83)	ASV3 (5.77)
	MOTU4	3.0	Annelida; Polychaeta; Scolecida; Spionida	—	**ASV3 (5.0)**	**ASV4 (5.26)**
	MOTU5	2.9	Nematoda; Chromadorea; Araeolaimida; Comesomatidae; *Sabatieria*; *Sabatieria* sp.	OTU4 (3.25)	ASV6 (2.95)	ASV6 (3.15)
	MOTU6	2.7	Dinoflagellata; Dinophyceae; Peridiniphycidae; Gonyaulacales; *Alexandrium*	OTU6 (2.78)	ASV5 (3.17)	ASV5 (3.31)
	MOTU7	2.5	Annelida; Polychaeta; Scolecida; Spionida	—	—	—
	MOTU8	2.3	Annelida; Polychaeta; Scolecida; Spionida	OTU7 (2.67)	ASV7 (2.66)	ASV7 (2.90)
	MOTU9	1.9	Annelida; Polychaeta; Scolecida; Spionida	OTU9 (2.05)	ASV8 (2.16)	ASV8 (2.79)
	MOTU10	1.8	Annelida; Polychaeta; Scolecida; Spionida	OTU5 (3.04)	ASV9 (2.15)	ASV9 (2.33)
Method	**MOTU**	**RA (%)**	**Taxonomy – LC Dataset**	**Vsearch (RA %)**	**Dada2 (RA %)**	**Deblur (RA %)**
UCLUST	MOTU1	7.13	Nematoda; Chromadorea; Chromadorida; Chromadoridae	OTU1 (8.83)	ASV1 (8.01)	ASV1 (12.12)
	MOTU2	5.69	Nematoda; Chromadorea; Araeolaimida; Comesomatidae; *Sabatieria*; *Sabatieria* sp.	OTU2 (6.55)	ASV2 (5.09)	ASV2 (8.46)
	MOTU3	2.71	Nematoda; Chromadorea; Desmodorida	OTU4 (3.15)	ASV3 (2.6)	ASV3 (3.09)
	MOTU4	2.59	Nematoda; Chromadorea; Araeolaimida; Comesomatidae; *Sabatieria*; *Sabatieria* sp.	OTU3 (4.08)	ASV4 (1.84)	ASV4 (2.6)
	MOTU5	1.76	Nematoda; Chromadorea; Araeolaimida; Comesomatidae	OTU9 (1.78)	ASV5 (1.61)	ASV5 (2.45)
	MOTU6	1.67	Nematoda; Enoplea; Enoplida; Oxystominidae	OTU6 (2.16)	ASV9 (1.52)	—
	MOTU7	1.62	Nematoda; Chromadorea; Araeolaimida; Comesomatidae	OTU8 (1.84)	ASV6 (1.56)	ASV6 (2.41)
	MOTU8	1.6	Nematoda; Chromadorea; Desmoscolecida; Desmoscolecidae; *Desmoscolex*	—	ASV10 (1.34)	ASV8 (2.12)
	MOTU9	1.59	Nematoda; Enoplea; Enoplida; Thoracostomopsidae	OTU5 (2.41)	—	ASV10 (1.77)
	MOTU10	1.48	Nematoda; Chromadorea; Araeolaimida; Comesomatidae; *Sabatieria*; *Sabatieria* sp.	OTU7 (1.86)	—	—

MOTUs that agreed in the taxonomic assignment across pipelines but differed in one nucleotide are highlighted in bold.
(—) It indicates that no UCLUST correspondent MOTU was detected for that specific pipeline.

S5). Interesting, for the highest-ranked MOTUs (i.e., the top 10 most abundant), the different algorithms behaved fairly similar, except for Deblur (family Oxystominidae; Figure S5). Despite these observed changes in MOTU rank-abundance curves, the relative abundance of major taxonomic groups (phylum level) remained fairly stable across all four bioinformatics pipelines in both the LC and HC datasets (i.e., percentage values in Figure 3). Small differences were only observed for the Deblur method where some of the low abundant taxa were not recovered (e.g., Cercozoa and Ascomycota in the HC and LC datasets, respectively). Furthermore, we observed that for the LC dataset, which is largely dominated by Nematoda, visualizing the data at higher taxonomic resolution (e.g., family level) led to greater differences across methods (e.g., a 4% increase in Comesomatidae and a 7% reduction in Oxystominidae in the Deblur method) and the absence of some low abundant taxa, which varied according to pipeline (e.g., Linhomoeidae in DADA2 outputs; Microlaimidae in Deblur outputs; data not shown).

3.3 | Alpha-diversity metrics across pipelines

Alpha-diversity metrics were calculated for both HC and LC datasets, using MOTU tables resulting from all four bioinformatics pipelines. To assess fine-scale differences, each metabarcoding study

was analyzed according to biologically relevant sample groupings including six Arctic geographic subregions (HC dataset), and four well-sampled nematode families (LC dataset). As observed in the full datasets (Figure 2), the number of OTUs recovered for each HC and LC data subset remained about 15-fold higher than the number of ASVs (Figures S2 and S3) across all subgroups, indicating that OTU pipelines consistently inflate MOTU diversity at all levels of a metabarcoding dataset. With respect to Shannon and Simpson diversity (Figure S4), there were typically no significant differences among subgroups (i.e., Arctic subregion or nematode family), except for the DADA2 pipeline in the LC dataset. Thus, the same subgroups were consistently recovered as having the lowest and highest median values of alpha diversity regardless of the MOTU pipeline (e.g., for the HC dataset Shannon Diversity metric, Beaufort Shelf and Camden Bay exhibited the lowest median value, while Bank Islands exhibited the highest median value Figure S4).

Conversely, bioinformatics pipelines did influence the absolute alpha diversity of each specific subgroup (Figures 5 and 6). For example, ASV pipelines resulted in significant reductions in Shannon diversity for the HC Arctic sediments (Figure 5), as opposed to Simpson diversity which was overall more stable across pipelines (although we did observe significant differences across methods for Arctic regions with >30 samples, Amundsen Gulf, Chukchi Sea, and Mackenzie River Plume; Figure 6). For the LC dataset, Shannon and Simpson indices were more consistent, although bioinformatics pipelines appeared to have disproportionately higher effects on the calculated diversity of each nematode family. Almost all pipeline comparisons resulted in significant differences in alpha diversity for each LC dataset subgroup (Figures 5 and 6). Surprisingly, the difference between UCLUST and DADA2 was rarely significant, with these two algorithms returning similar levels of Shannon and Simpson diversity despite representing different MOTU algorithm classes. Both dataset complexity and the number of samples representing a specific subgroup appeared to be responsible for the significant differences observed in these diversity indices.

3.4 | Beta-diversity metrics across pipelines

Beta-diversity metrics were also calculated for both HC and LC datasets using results from all four pipelines and based on the Bray–Curtis, unweighted Unifrac, and weighted Unifrac indices (rarefied at 1000 sequences per sample). The Procrustes analysis always revealed significant concordance ($p < 0.01$) between pipelines regardless of the dataset or distance index (Table 2, Figure 7). Our results suggest that beta-diversity metrics, as well as dataset complexity, can both impact the level of concordance observed in beta-diversity analyses across bioinformatics pipelines. For both the HC and LC datasets, Bray–Curtis similarity exhibited high concordance across all pairwise comparisons of bioinformatics pipelines, especially between DADA2 and Deblur. Apparently, neither algorithm class (OTU vs. ASV generation) nor dataset complexity (HC vs. LC dataset) impacted the beta-diversity patterns recovered with the Bray–Curtis metric as observed by the relatively similar statistical values and the length of vectors connecting data points (Table 2, Figure 7). In contrast, dataset complexity had a clear influence on the level of concordance observed for both weighted and unweighted Unifrac metrics. In the LC dataset, most pairwise Unifrac comparisons exhibited low concordance (e.g., $R < 0.8$ and $M^{12} > 0.2$; Table 2), except for UCLUST-Vsearch (unweighted Unifrac), VSearch-DADA2, and DADA2-Deblur (both weighted Unifrac), which exhibited high concordance. In contrast, all pairwise comparisons across both weighted/unweighted Unifrac metrics were recovered as highly concordant in the HC dataset.

4 | DISCUSSION

This study illustrates the differential influence of four bioinformatics pipelines on eukaryotic 18S rRNA metabarcoding studies. We focused on UCLUST, VSearch, DADA2, and Deblur since these workflows represent the most commonly applied approaches in the 18S rRNA metabarcoding literature. We did not attempt an exhaustive assessment of all software tools that exist for the delimitation of MOTUs (Boyer et al., 2016; Edgar, 2016; Eren et al., 2015; Mahé et al., 2014; Schloss et al., 2009). However, most MOTU algorithms process metabarcoding reads in ways that are conceptually similar to our four chosen workflows, and thus, our results should be broadly generalizable across studies.

Algorithm class (OTU vs. ASV delimitation) appeared to be the strongest driver of the observed differences in both the HC and LC datasets, having a clear impact on both the number of quality-processed reads and MOTUs (Figure 2), the shape of MOTU rank-abundance curves (Figure 4a,b), and the proportion of "rare biosphere" taxa represented by low-abundance MOTUs (Figure 4c,d, Table 3). Such differences are explained by the behavior of each algorithm with ASV algorithm class tending to be more stringent (e.g., due to modelling of sequence errors) and less prone to produce false positives when compared to OTU algorithm class (Amir et al., 2017; Caruso et al., 2019). According to Antich et al. (2021), however, clustering (OTUs) and denoising (ASVs) accomplish different functions, the former seeking to recover meaningful "species-level entities" and the latter seeking to recover "correct sequences," and therefore such differences should be expected.

Dataset complexity also impacted bioinformatics outputs in subtle and unexpected ways. For example, LC datasets appeared to introduce an element of randomness and stochasticity into the downstream ecological analysis (e.g., beta-diversity metrics; Table 2,

FIGURE 5 Shannon diversity index for (a) HC and (b) LC datasets. KW analysis was used to test for significant differences among bioinformatics pipelines for each subregion (HC dataset) or nematode family (LC dataset), separately. Number of samples (specimens for the LC dataset) representing each subgroup is given in Table 1

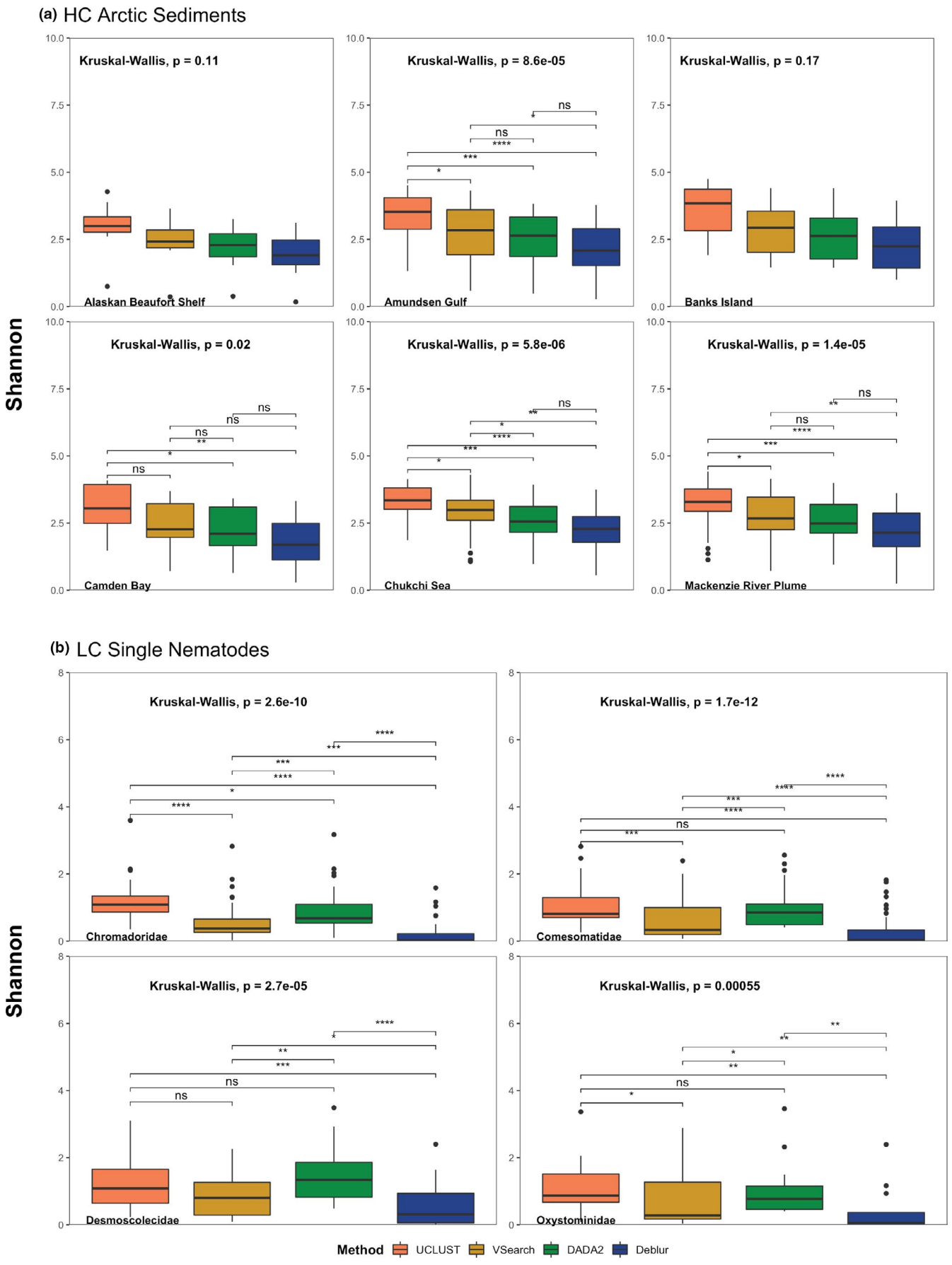

(a) HC Arctic Sediments
Shannon
Kruskal-Wallis, p = 0.11
Alaskan Beaufort Shelf
Kruskal-Wallis, p = 8.6e-05
Amundsen Gulf
Kruskal-Wallis, p = 0.17
Banks Island
Kruskal-Wallis, p = 0.02
Camden Bay
Kruskal-Wallis, p = 5.8e-06
Chukchi Sea
Kruskal-Wallis, p = 1.4e-05
Mackenzie River Plume
(b) LC Single Nematodes
Shannon
Kruskal-Wallis, p = 2.6e-10
Chromadoridae
Kruskal-Wallis, p = 1.7e-12
Comesomatidae
Kruskal-Wallis, p = 2.7e-05
Desmoscolecidae
Kruskal-Wallis, p = 0.00055
Oxystominidae
Method UCLUST VSearch DADA2 Deblur

Figure 7), and the specific influence of algorithm parameters became more extreme for metabarcoding samples with simpler community structure (e.g., patterns of recovered reads and MOTUs for the LC dataset; Table 3). In fact, Siegwald et al. (2017) pointed out that the analysis of low complexity datasets is often a difficult task, perhaps because it contains a higher number of low-abundance taxa (i.e., rare MOTUs).

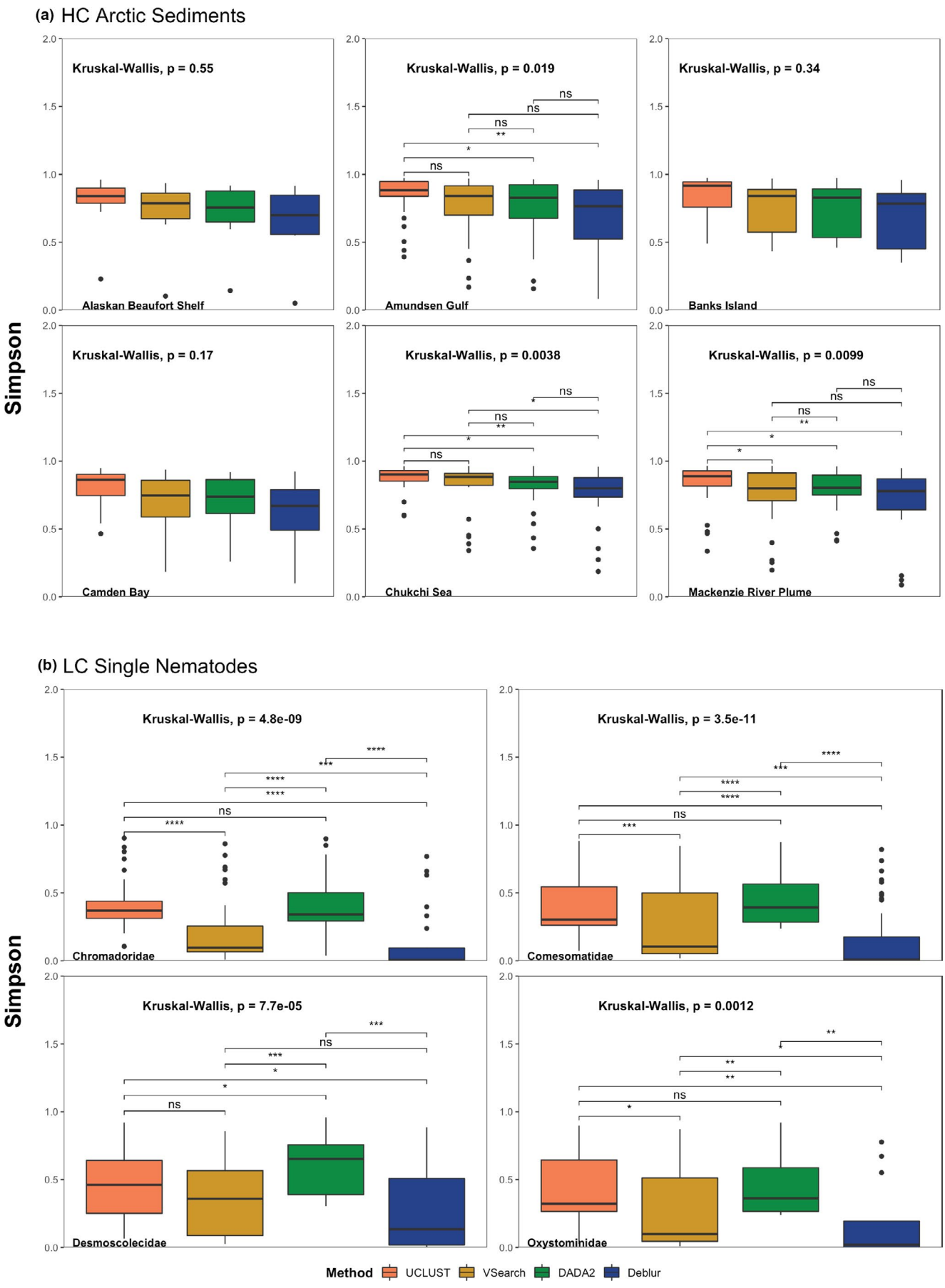
In our LC dataset (where per-sample biodiversity consisted of one nematode and a small number of host-associated taxa and gut contents; [Pereira et al., 2020]), the VSearch algorithm returned 2.4x as many OTUs compared with UCLUST despite these two pipelines being in the same MOTU class (Table 3). Furthermore, we observed that the "long tail" of rare MOTUs was even more pronounced in the VSearch pipeline, especially for the LC dataset (i.e., 96.5% of all OTUs). These trends appear to be related to the underlying functions of the two OTU algorithms. While UCLUST utilizes a heuristic method to find read alignments with the best score (via implementation of USEARCH; [Edgar, 2010]), the VSearch algorithm instead generates the full alignment vectors during MOTU generation (via global pairwise alignments using the Needleman–Wunsch algorithm; [Rognes et al., 2016]). Therefore, it is possible that VSearch produces tighter clusters (i.e., with smaller sequence divergence) leading to a higher number of OTUs and including more rare MOTUs. Previous benchmark studies have demonstrated that UCLUST is likely to produce looser clusters, especially at lower similarity thresholds (Ghodsi et al., 2011). Despite the higher proportion of rare MOTUs in VSearch outputs, previous studies have indicated that VSearch exhibits equal or higher accuracy and returns more stable OTU clusters (that are unlikely to disappear when the data is subsetted or increased) compared with the UCLUST algorithm when performing *de novo* clustering (Jackson et al., 2016; Westcott & Schloss, 2015). The "long tail" of rare MOTUs appears to be an inherent feature of out-picking pipelines where sequences are clustered via pairwise sequence identity. However, our results suggest that the relative proportion of rare MOTUs (and the dominance of this "long tail" in any given metabarcoding dataset) will vary according to algorithm choice and the underlying dataset complexity, thus supporting previous findings (Siegwald et al., 2017, 2019).

In ASV approaches, the Deblur algorithm also appeared to be especially sensitive to dataset complexity. In our LC dataset, the contribution of rare MOTUs was higher than expected in the Deblur pipeline (6.8% of ASVs, which was ~1.5× more than in the HC dataset; Table 3). Similarly, read error correction was markedly more conservative in our LC dataset, where Deblur showed a significant reduction in the number of ASVs recovered compared with DADA2 (~15% reduction; Figure 2d). In contrast, the difference between the two ASV pipelines was not significant in the HC dataset, with Deblur returning a higher number of MOTUs than DADA2; Figure 2a). However, Deblur always returned a lower *total* number of reads and MOTUs for the overall HC and LC datasets compared with DADA2 (Tables S1 and S2). For mock communities, Prodan et al. (2020) reported that Deblur conserved only about 50% of the initial read number input, whereas other pipelines were above 70%. Accordingly, the authors related Deblur's low conversion rate to the count-subtraction nature of the algorithm before the ASV estimation. This discrepancy between the total vs. mean number of recovered reads/MOTUs appears to stem from the fundamentally distinct ways in which DADA2 and Deblur perform quality checks and error correction on raw Illumina data (Amir et al., 2017; Callahan et al., 2016).

DADA2 merges raw Illumina reads after its ASV-generation algorithm (i.e., after trimming, truncating, and denoising), while Deblur merges reads as a first step before subsequent quality checks are performed. Most importantly, Deblur performs all error correction and ASV-generation steps on a sample-by-sample basis. The per-sample focus was a conscious choice by the Deblur developers, who designed this ASV algorithm to maximize computational speed and maintain the ability to be highly parallelizable when required for large datasets. However, this software design can erroneously remove sequences considered as "rare MOTUs" by only viewing slices of a large HTS dataset. DADA2 instead generates an error-correction model by considering all samples and sequences at once (e.g., with the underlying assumption that each Illumina run has a unique profile of sequencing artifacts that can be eliminated by comparing abundant vs. rare reads). As a result, DADA2 is much more computationally intensive and requires more computational time. Our results suggest that the inherent features of the Deblur algorithm produce less consistent outputs that can be heavily influenced by the biological community complexity built into a metabarcoding study, and potentially impacting the ability of statistical tests to detect significant differences between groups (e.g., subregions and nematode families in our study). Existing ASV algorithms show wide divergence in algorithm design and function and therefore must be chosen with care.

We observed a surprising level of stability for some biological patterns recovered in downstream ecological analyses. For example, the relative abundance of major taxonomic groups is generally preserved across bioinformatics pipelines regardless of dataset complexity (Figure 3). Although MOTU clustering algorithm did lead to some significant differences in alpha-diversity comparisons (Figures 5 and 6), these did not appear to impact biological interpretations within a specific dataset and method; that is, the most diverse subgroup (Arctic subregion or nematode family) was rarely affected in Simpson and Shannon diversity metrics (Figure S4). A previous study by Jackson et al. (2016) confirmed that the absolute values of alpha-diversity indices cannot be compared across bioinformatics methods, and such differences in diversity estimates can be effectively eliminated by collapsing MOTUs according to

FIGURE 6 Simpson diversity index for (a) HC and (b) LC datasets. KW analysis was used to test for significant differences among bioinformatics pipelines for each subregion (HC dataset) or nematode family (LC dataset), separately. Number of samples (specimens for the LC dataset) representing each subgroup is given in Table 1


(a) HC Arctic Sediments
Simpson
Kruskal-Wallis, p = 0.55
Alaskan Beaufort Shelf
Kruskal-Wallis, p = 0.019
Amundsen Gulf
Kruskal-Wallis, p = 0.34
Banks Island
Kruskal-Wallis, p = 0.17
Camden Bay
Kruskal-Wallis, p = 0.0038
Chukchi Sea
Kruskal-Wallis, p = 0.0099
Mackenzie River Plume
(b) LC Single Nematodes
Simpson
Kruskal-Wallis, p = 4.8e-09
Chromadoridae
Kruskal-Wallis, p = 3.5e-11
Comesomatidae
Kruskal-Wallis, p = 7.7e-05
Desmoscolecidae
Kruskal-Wallis, p = 0.0012
Oxystominidae
Method UCLUST VSearch DADA2 Deblur

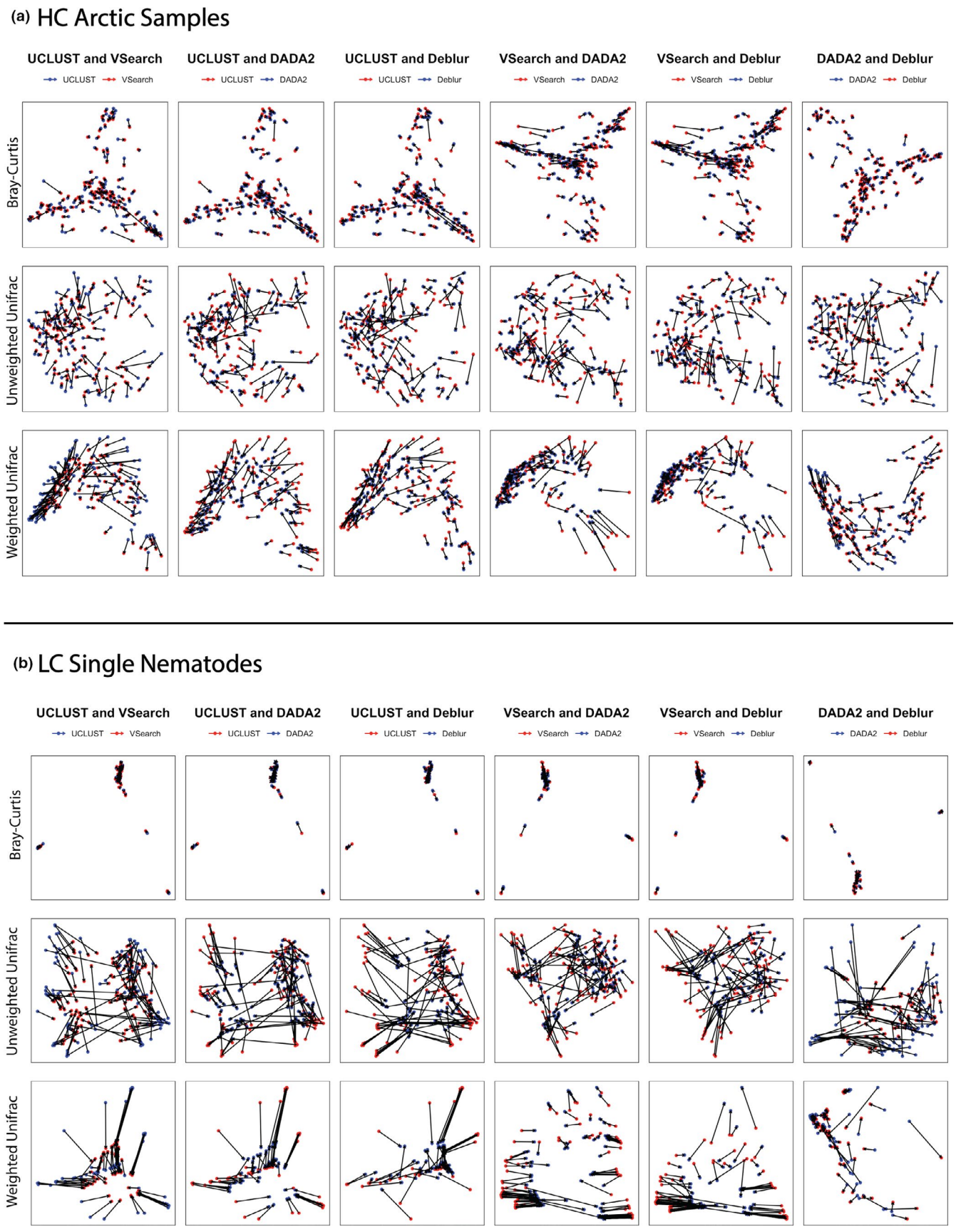


FIGURE 7 Procrustes analysis for beta diversity. A Procrustes ordination plot based on PCoA coordinates was built for each beta-diversity index (e.g., Bray–Curtis, unweighted Unifrac, and weighted Unifrac). The color underneath each method refers to their color on the plot. Longer lines connecting samples indicate higher discordance in beta-diversity patterns, whereas short lines (or no visible lines) denote the highest concordance

their taxonomy assignments. However, we note that MOTU taxonomy assignments will also be impacted by different bioinformatics methods (e.g., using BLAST searches vs. Bayesian classifier tools) and the completeness of reference databases (Holovachov et al., 2017; Macheriotou et al., 2019), and both of these factors typically vary depending on the chosen metabarcoding workflow and target taxon. When assessing the beta diversity among subgroups based on the Bray–Curtis coefficient, we also found high concordance across all methods in the HC and LC datasets (Figure 7), presumably because the Bray–Curtis coefficient as a resemblance measure has valid underlying conditions (e.g., independence of joint absence) and tends to capture the important assemblage relationships (Clarke et al., 2006). Taken together, our results indicate that many overarching biological patterns within 18S rRNA metabarcoding datasets remain consistent across distinct bioinformatics pipelines, regardless of dataset complexity. Therefore, beta-diversity patterns appear to be largely unaffected by the overall number of MOTUs (Table 3) or the number of rare MOTUs as depicted in the head–tail curves (Figure 4).

Downstream analyses that took rare MOTUs into account were notably less consistent across the four pipelines, especially for the LC dataset. Procrustes analysis of beta-diversity patterns revealed low concordance across bioinformatics pipelines for the LC dataset when the Unifrac distance was applied (Table 2). Since Unifrac relies on the topology of a phylogenetic tree to calculate the evolutionary distance of MOTUs between samples (Lozupone & Knight, 2005), a low number of MOTUs with skewed relative abundance profiles may heavily impact beta-diversity analyses across bioinformatics pipelines (e.g., a few dominant MOTUs containing the vast majority of reads). We believe that LC datasets may be especially sensitive to sample rarefaction employed before Unifrac calculations, which potentially increases the stochasticity of the MOTUs being subsampled and makes beta-diversity comparisons more prone to random effects. From a biological perspective, ASV pipelines offer significant advantages for downstream ecological analyses where rarefaction is employed as they tend to be more stable regardless of the rarefaction depth (Prodan et al., 2020). Furthermore, they strongly reduce the likelihood that sequencing artifacts and erroneous reads will be incorporated into diversity metrics that emphasize changes in the "rare biosphere" of low-read MOTUs. For alpha diversity specifically, other methods may also be considered when comparing sample groups (e.g., distance which uses mean-pairwise distance [MPD] using a Bayesian approach; Hackmann, 2020).

For metabarcoding studies, sequencing technologies and bioinformatics pipelines will inevitably continue to evolve. Our prime consideration was to assess the stability of biological inferences across historical (and future) shifts in computational workflows for MOTU generation. For published studies relying on cluster-based OTU methods, the recent shift from QIIME1 (where UCLUST is the default algorithm; (Caporaso et al., 2010) to QIIME2 (where VSearch is the default algorithm; (Bolyen et al., 2019) is not likely to have a significant impact on the downstream biological conclusions within this class of algorithm. OTU pipelines report similar (albeit highly inflated) levels of biodiversity, and results from UCLUST and VSearch were generally consistent regardless of study design and dataset complexity (HC vs LC datasets; Figures 2, 4–6). Major patterns of taxon abundance (Figure 3) and sample grouping (ordinations based on Bray–Curtis coefficient; Figure 7) were also unaffected by algorithm class.

With the evolution of new algorithms in recent years, the metabarcoding and microbial ecology communities have rapidly shifted toward ASVs as a more stable and objective type of MOTU where reference sequences can be directly compared across studies (Callahan et al., 2017). OTUs are somewhat arbitrary "clouds" of sequence reads, and OTU membership can be heavily influenced by the specific parameters of the underlying algorithm, a phenomenon also shared by 16S rRNA datasets (Abellan-Schneyder et al., 2021; He et al., 2015; Jackson et al., 2016). Surely, databases for archaea/bacteria are more complete when compared to eukaryotes, and this can potentially alleviate issues during OTU clustering and taxonomic assignments (Brandt et al., 2021). Furthermore, OTUs are dataset specific and not directly comparable across studies (Callahan et al., 2017). Our results confirm the significant biological advantages of ASV-generation algorithms: the improved error correction eliminates the artefactual "long tail" of rare sequences while maintaining species-specific barcodes ("Head" MOTUs with high relative abundance; (Porazinska et al., 2010). More importantly, ASVs do not completely eliminate rare MOTUs, and the remaining low-abundance reads can provide important biological insights regarding ecological interactions and population-level variation (e.g., patterns of intragenomic rRNA variation [Pereira et al., 2020; Qing et al., 2020], gut contents, and host-associated microbiome taxa [Schuelke et al., 2018]).

We specifically recommend the DADA2 pipeline for eukaryotic metabarcoding studies, as the resulting ASV dataset appears to represent the best approximation of real biological patterns (and the number of distinct species present), especially for LC communities where the sequencing effort has likely reached saturation (Macheriotou et al., 2019; Pereira et al., 2020). The Deblur algorithm should be used with caution, however, since Deblur's optimization for fast computational speed has resulted in a tradeoff whereby biologically valid MOTUs (true positives) are overzealously eliminated from metabarcoding datasets. Further analysis of already published 18S rRNA metabarcoding studies using the DADA2 pipeline may reveal compelling ecological and evolutionary insights for diverse eukaryotic taxa on a global scale given the increased biological accuracy of this method (e.g., shorter tail of rare MOTUs, less likely to be impacted by data complexity).

The present study focused on a metabarcoding locus optimized for microbial metazoa, targeting the V1–V2 regions of the 18S rRNA gene (Creer et al., 2010). It remains to be seen whether the observed bioinformatics patterns will extend to other rRNA loci and protein-coding genes. For soil nematodes specifically, Kenmotsu et al. (2020) suggested that the amplification of regions V7-V9 (located at 3' end of the 18S rRNA gene) followed by the analysis using DADA2 may produce the most realistic results. Future research efforts should

evaluate computational pipelines using other common metabarcoding genes (COI, 12S, rbcl, etc.) and 18S rRNA hypervariable regions (V4, V9, etc.) thus incorporating broader assessments of both nuclear and mitochondrial loci (e.g., PEMA pipeline; [Zafeiropoulos et al., 2020]). How dataset complexity influences bioinformatics outputs is likely to be distinct for metabarcoding loci such as COI, where Illumina datasets will inherently contain a high level of both population-level (e.g., gene haplotypes) and species-level ("Head" MOTUs) genetic variation. Although DADA2 performed well with both HC and LC datasets and therefore should be the chosen method for 18S rRNA metabarcoding studies, our study also supports the long-held assumption that major biological and ecological patterns will tend to emerge from a well-designed metabarcoding study (Xiong & Zhan, 2018; Zinger et al., 2019), regardless of the underlying bioinformatics pipeline used for data analysis. Summarizing MOTU datasets by taxonomy assignments (e.g., collapsing MOTUs at the genus or species level) appears to be a robust approach for avoiding any discrepancies that may arise across different computational pipelines (Jackson et al., 2016), and this approach may help to alleviate some of the observed difficulties in analyzing rare MOTUs and LC datasets. However, such a taxonomy-dependent approach requires a comprehensive database of reference DNA barcodes (Ruppert et al., 2019). Unfortunately, current reference databases for common eukaryotic metabarcoding loci are comparatively sparse and patchy (versus 16S rRNA gene databases for bacteria/archaea where most known genera are well represented; Bik, Porazinska et al., 2012; Macheriotou et al., 2019; Pereira et al., 2020). Improvement of eukaryotic reference databases, combined with a movement toward phylogeny-based biodiversity analyses, will help to further improve the ecological and evolutionary metrics that can be applied to metabarcoding studies.

ACKNOWLEDGMENTS

The authors would **like to thank Taruna Schuelke and Alexis Walker for their help and assistance in generating the original metabarcoding datasets used as the basis for this study. Funding for this study was provided by the** North Pacific Research Board (NPRB project 1303) and institutional startup funding from the University of Georgia. This research was made possible by a grant from the Gulf of Mexico Research Initiative. Data are publicly available through the Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC) at https://data.gulfresearchinitiative.org (doi: R5.x272.000:0004).

CONFLICT OF INTEREST

None of the authors have competing interests to declare.

AUTHOR CONTRIBUTIONS

ADS, TJP, and HMB conceived the ideas and designed the study methodology. SMH collected marine sediment samples and provided intellectual input on hypothesis testing and data analysis workflows. ADS collated and analyzed the data. ADS, TJP, and HMB led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DATA AVAILABILITY STATEMENT

Raw Illumina reads from Arctic marine sediments (HC dataset) are deposited in the NCBI Sequence Read Archive (BioProject PRJNA723923 and SRA accession SUB9498354). For the Low Complexity (LC) single-nematode dataset, raw Illumina reads are also available in the NCBI Sequence Read Archive (BioProject PRJNA422296 and SRA accession SRP128131). Additionally, full-length 18S rRNA gene sequences were generated via Sanger sequencing for most nematode specimens used in the LC dataset and deposited on GenBank (Accession Numbers: MN250033–MN250142). Primer constructs for both HC and LC dataset have been deposited in FigShare (https://doi.org/10.6084/m9.figshare.5701090). QIIME mapping files, final MOTU tables, and all scripts used for processing and analyzing the data are available on GitHub (https://github.com/BikLab/OTU-ASV-euk-benchmarking).

ORCID

Alejandro De Santiago https://orcid.org/0000-0001-9086-3050
Tiago José Pereira https://orcid.org/0000-0002-6424-2848
Sarah L. Mincks https://orcid.org/0000-0002-3328-712X
Holly M. Bik https://orcid.org/0000-0002-4356-3837

REFERENCES

Abellan-Schneyder, I., Matchado, M. S., Reitmeier, S., Sommer, A., Sewald, Z., Baumbach, J., List, M., & Neuhaus, K. (2021). Primer, pipelines, parameters: Issues in 16s rRNA gene sequencing. *Msphere*, *6*(1), e01202–20. https://doi.org/10.1128/mSphere.01202-20

Akita, S., Takano, Y., Nagai, S., Kuwahara, H., Kajihara, R., Tanabe, A. S., & Fujita, D. (2019). Rapid detection of macroalgal seed bank on cobbles: Application of DNA metabarcoding using next-generation sequencing. *Journal of Applied Phycology*, *31*(4), 2743–2753. https://doi.org/10.1007/s10811-018-1730-9

Alberdi, A., Aizpurua, O., Gilbert, M. T. P., & Bohmann, K. (2018). Scrutinizing key steps for reliable metabarcoding of environmental samples. *Methods in Ecology and Evolution/British Ecological Society*, *9*(1), 134–147. https://doi.org/10.1111/2041-210x.12849

Amir, A., McDonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Zech Xu, Z., Kightley, E. P., Thompson, L. R., Hyde, E. R., Gonzalez, A., & Knight, R. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *Msystems*, *2*(2), e00191–16. https://doi.org/10.1128/mSystems.00191-16

Anslan, S., Nilsson, R. H., Wurzbacher, C., Baldrian, P., Tedersoo, L., & Bahram, M. (2018). Great differences in performance and outcome of high-throughput sequencing data analysis platforms for fungal metabarcoding. *MycoKeys*, *39*, 29–40. https://doi.org/10.3897/mycokeys.39.28109

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, *22*(1), 177. https://doi.org/10.1186/s12859-021-04115-6

Arribas, P., Andújar, C., Hopkins, K., Shepherd, M., & Vogler, A. P. (2016). Metabarcoding and mitochondrial metagenomics of endogean arthropods to unveil the mesofauna of the soil. *Methods in Ecology and Evolution / British Ecological Society*, *7*(9), 1071–1081. https://doi.org/10.1111/2041-210x.12557

Bálint, M., Schmidt, P.-A., Sharma, R., Thines, M., & Schmitt, I. (2014). An Illumina metabarcoding pipeline for fungi. *Ecology and Evolution*, *4*(13), 2642–2653. https://doi.org/10.1002/ece3.1107

Beentjes, K. K., Arjen, G. C., Schilthuizen, M., Hoogeveen, M., & van der Hoorn, B. B. (2019). The effects of spatial and temporal replicate sampling on eDNA metabarcoding. *PeerJ*, *7*, e7335. https://doi.org/10.7717/peerj.7335

Bell, K. L., Loeffler, V. M., & Brosi, B. J. (2017). An rbcL reference library to aid in the identification of plant species mixtures by DNA metabarcoding. *Applications in Plant Sciences*, *5*(3), 1600110. https://doi.org/10.3732/apps.1600110

Bik, H. M., Fournier, D., Sung, W., Bergeron, R. D., & Thomas, W. K. (2013). Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS One*, *8*(10), e78230. https://doi.org/10.1371/journal.pone.0078230

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., & Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution*, *27*(4), 233–243. https://doi.org/10.1016/j.tree.2011.11.010

Bik, H. M., Sung, W., De Ley, P., Baldwin, J. G., Sharma, J., Rocha-Olivares, A., & Thomas, W. K. (2012). Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. *Molecular Ecology*, *21*(5), 1048–1059. https://doi.org/10.1111/j.1365-294X.2011.05297.x

Blaxter, M. (2016). Imagining Sisyphus happy: DNA barcoding and the unnamed majority. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150329. https://doi.org/10.1098/rstb.2015.0329

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., Vanfleteren, J. R., Mackey, L. Y., Dorris, M., Frisse, L. M., Vida, J. T., & Thomas, W. K. (1998). A molecular evolutionary framework for the phylum Nematoda. *Nature*, *392*, 71–75. https://doi.org/10.1038/32160

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R., & Abebe, E. (2005). Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1462), 1935–1943. https://doi.org/10.1098/rstb.2005.1725

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C. C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Caporaso, J. G. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nature Biotechnology*, *37*(8), 852–857. https://doi.org/10.1038/s41587-019-0209-9

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodríguez, A. M., Chase, J., … Gregory Caporaso, J. (2018). *QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science* (No. e27295v2). PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.27295v2

Boyer, F., Mercier, C., Bonin, A., Le Bras, Y., Taberlet, P., & Coissac, E. (2016). obitools: A unix-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, *16*(1), 176–182. https://doi.org/10.1111/1755-0998.12428

Bradley, I. M., Pinto, A. J., & Guest, J. S. (2016). Design and evaluation of Illumina MiSeq-compatible, 18S rRNA gene-specific primers for improved characterization of mixed phototrophic communities. *Applied and Environmental Microbiology*, *82*(19), 5878–5891. https://doi.org/10.1128/AEM.01630-16

Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, *21*(6), 1904–1921. https://doi.org/10.1111/1755-0998.13398

Brannock, P. M., & Halanych, K. M. (2015). Meiofaunal community analysis by high-throughput sequencing: Comparison of extraction, quality filtering, and clustering methods. *Marine Genomics*, *23*, 67–75. https://doi.org/10.1016/j.margen.2015.05.007

Cahill, A. E., Pearman, J. K., Borja, A., Carugati, L., Carvalho, S., Danovaro, R., Dashfield, S., David, R., Féral, J.-P., Olenin, S., Šiaulys, A., Somerfield, P. J., Trayanova, A., Uyarra, M. C., & Chenuil, A. (2018). A comparative analysis of metabarcoding and morphology-based identification of benthic communities across different regional seas. *Ecology and Evolution*, *8*(17), 8908–8920. https://doi.org/10.1002/ece3.4283

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643. https://doi.org/10.1038/ismej.2017.119

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583. https://doi.org/10.1038/nmeth.3869

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, *7*(5), 335–336. https://doi.org/10.1038/nmeth.f.303.

Caruso, V., Song, X., Asquith, M., & Karstens, L. (2019). Performance of microbiome sequence inference methods in environments with varying biomass. *Msystems*, *4*, e00163-18. https://doi.org/10.1128/msystems.00163-18

Clarke, K. R., Somerfield, P. J., & Chapman, M. G. (2006). On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. *Journal of Experimental Marine Biology and Ecology*, *330*(1), 55–80. https://doi.org/10.1016/j.jembe.2005.12.017

Creer, S., Fonseca, V. G., Porazinska, D. L., Giblin-Davis, R. M., Sung, W., Power, D. M., Packer, M., Carvalho, G. R., Blaxter, M. L., Lambshead, P. J. D., & Thomas, W. K. (2010). Ultrasequencing of the meiofaunal biosphere: Practice, pitfalls and promises. *Molecular Ecology*, *19*(Suppl 1), 4–20. https://doi.org/10.1111/j.1365-294X.2009.04473.x

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, *6*(1), 226. https://doi.org/10.1186/s40168-018-0605-2.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Dell'Anno, A., Carugati, L., Corinaldesi, C., Riccioni, G., & Danovaro, R. (2015). Unveiling the biodiversity of deep-sea nematodes through metabarcoding: Are we ready to bypass the classical taxonomy? *PLoS One*, *10*(12), e0144928. https://doi.org/10.1371/journal.pone.0144928

Djemiel, C., Dequiedt, S., Karimi, B., Cottin, A., Girier, T., El Djoudi, Y., Wincker, P., Lelièvre, M., Mondy, S., Chemidlin Prévost-Bouré, N., Maron, P.-A., Ranjard, L., & Terrat, S. (2020). BIOCOM-PIPE: a new user-friendly metabarcoding pipeline for the characterization of microbial diversity from 16S, 18S and 23S rRNA gene amplicons. *BMC Bioinformatics*, *21*(1), 492. https://doi.org/10.1186/s12859-020-03829-3

Djurhuus, A., Pitz, K., Sawaya, N. A., Rojas-Márquez, J., Michaud, B., Montes, E., Muller-Karger, F., & Breitbart, M. (2018). Evaluation of

marine zooplankton community structure through environmental DNA metabarcoding. *Limnology and Oceanography, Methods/ASLO*, *16*(4), 209–221. https://doi.org/10.1002/lom3.10237

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, *26*(19), 2460–2461. https://doi.org/10.1093/bioinformatics/btq461

Edgar, R. C. (2016). UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing (p. 081257). https://doi.org/10.1101/081257

Edgar, R. C. (2017). Accuracy of microbial community diversity estimated by closed- and open-reference OTUs. *PeerJ*, *5*(e3889), e3889. https://doi.org/10.7717/peerj.3889

Edgar, R. C. (2018). Updating the 97% identity threshold for 16S ribosomal RNA OTUs. *Bioinformatics*, *34*(14), 2371–2375. https://doi.org/10.1093/bioinformatics/bty113

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200. https://doi.org/10.1093/bioinformatics/btr381

Elbrecht, V., & Leese, F. (2017). Validation and development of COI metabarcoding primers for freshwater macroinvertebrate bioassessment. *Frontiers of Environmental Science & Engineering in China*, *5*, 11. https://doi.org/10.3389/fenvs.2017.00011

Elbrecht, V., Vamos, E. E., Steinke, D., & Leese, F. (2018). Estimating intraspecific genetic diversity from community DNA metabarcoding data. *PeerJ*, *6*, e4644. https://doi.org/10.7717/peerj.4644

Eren, A. M., Morrison, H. G., Lescault, P. J., Reveillaud, J., Vineis, J. H., & Sogin, M. L. (2015). Minimum entropy decomposition: unsupervised oligotyping for sensitive partitioning of high-throughput marker gene sequences. *The ISME Journal*, *9*(4), 968–979. https://doi.org/10.1038/ismej.2014.195

Escudié, F., Auer, L., Bernard, M., Mariadassou, M., Cauquil, L., Vidal, K., Maman, S., Hernandez-Raquet, G., Combes, S., & Pascal, G. (2018). FROGS: Find, rapidly, OTUs with galaxy solution. *Bioinformatics*, *34*(8), 1287–1294. https://doi.org/10.1093/bioinformatics/btx791

Fidler, G., Tolnai, E., Stagel, A., Remenyik, J., Stundl, L., Gal, F., Biro, S., & Paholcsek, M. (2020). Tendentious effects of automated and manual metagenomic DNA purification protocols on broiler gut microbiome taxonomic profiling. *Scientific Reports*, *10*, https://doi.org/10.1038/s41598-020-60304-y

Fonseca, V. G., Sinniger, F., Gaspar, J. M., Quince, C., Creer, S., Power, D. M., Peck, L. S., & Clark, M. S. (2017). Revealing higher than expected meiofaunal diversity in Antarctic sediments: A metabarcoding approach. *Scientific Reports*, *7*(1), 6094. https://doi.org/10.1038/s41598-017-06687-x

Geisen, S., Snoek, L. B., ten Hooven, F. C., Duyts, H., Kostenko, O., Bloem, J., Martens, H., Quist, C. W., Helder, J. A., & der Putten, W. H. (2018). Integrating quantitative morphological and qualitative molecular methods to analyse soil nematode community responses to plant range expansion. *Methods in Ecology and Evolution*, *9*, 1366–1378. https://doi.org/10.1111/2041-210x.12999

Ghodsi, M., Liu, B., & Pop, M. (2011). DNACLUST: accurate and efficient clustering of phylogenetic marker genes. *BMC Bioinformatics*, *12*, 271. https://doi.org/10.1186/1471-2105-12-271

Giebner, H., Langen, K., Bourlat, S. J., Kukowka, S., Mayer, C., Astrin, J. J., Misof, B., & Fonseca, V. G. (2020). Comparing diversity levels in environmental samples: DNA sequence capture and metabarcoding approaches using 18S and COI genes. *Molecular Ecology Resources*, *20*(5), 1333–1345. https://doi.org/10.1111/1755-0998.13201

Group, C. P. W., Hollingsworth, P. M., Forrest, L. L., Spouge, J. L., Hajibabaei, M., & Ratnasingham, S., van der Bank, M., Chase, M. W., Cowan, R. S., Erickson, D. L., & Fazekas, A. J. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, *106*(31), 12794–12797. http://www.pnas.org/content/106/31/12794.short

Hackmann, T. J. (2020). Accurate estimation of microbial sequence diversity with distanced. *Bioinformatics*, *36*(3), 728–734. https://doi.org/10.1093/bioinformatics/btz668

He, Y., Caporaso, J. G., Jiang, X.-T., Sheng, H.-F., Huse, S. M., Rideout, J. R., Edgar, R. C., Kopylova, E., Walters, W. A., Knight, R., & Zhou, H.-W. (2015). Stability of operational taxonomic units: An important but neglected property for analyzing microbial diversity. *Microbiome*, *3*, 20. https://doi.org/10.1186/s40168-015-0081-x

Hebert, P. D. N., Cywinska, A., Ball, S. L., & deWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings. Biological Sciences/the Royal Society*, *270*(1512), 313–321. https://doi.org/10.1098/rspb.2002.2218

Holovachov, O., Haenel, Q., Bourlat, S. J., & Jondelius, U. (2017). Taxonomy assignment approach determines the efficiency of identification of OTUs in marine nematodes. *Royal Society Open Science*, *4*(8), 170315. https://doi.org/10.1098/rsos.170315

Hornung, B. V. H., Zwittink, R. D., & Kuijper, E. J. (2019). Issues and current standards of controls in microbiome research. *FEMS Microbiology Ecology*, *95*(5), https://doi.org/10.1093/femsec/fiz045

Jackson, D. A. (1995). PROTEST: A PROcrustean Randomization TEST of community environment concordance. *Écoscience*, *2*(3), 297–303. https://doi.org/10.1080/11956860.1995.11682297

Jackson, M. A., Bell, J. T., Spector, T. D., & Steves, C. J. (2016). A heritability-based comparison of methods used to cluster 16S rRNA gene sequences into operational taxonomic units. *PeerJ*, *4*, e2341. https://doi.org/10.7717/peerj.2341

Kassambara, A. (2018). ggpubr:"ggplot2" based publication ready plots (Version 0.1. 7). *Obtido Desde*. https://CRAN.R-Project.Org/Package=Ggpubr

Katoh, K., Misawa, K., Kuma, K.-I., & Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, *30*(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Kenmotsu, H., Uchida, K., Hirose, Y., & Eki, T. (2020). Taxonomic profiling of individual nematodes isolated from copse soils using deep amplicon sequencing of four distinct regions of the 18S ribosomal RNA gene. *PLoS One*, *15*(10), e0240336. https://doi.org/10.1371/journal.pone.0240336

Koziol, A., Stat, M., Simpson, T., Jarman, S., DiBattista, J. D., Harvey, E. S., Marnane, M., McDonald, J., & Bunce, M. (2019). Environmental DNA metabarcoding studies are critically affected by substrate selection. *Molecular Ecology Resources*, *19*(2), 366–376. https://doi.org/10.1111/1755-0998.12971

Kumar, V., Dickey, A. M., Seal, D. R., Shatters, R. G., Osborne, L. S., & McKenzie, C. L. (2017). Unexpected high intragenomic variation in two of three major pest thrips species does not affect ribosomal internal transcribed spacer 2 (ITS2) utility for thrips identification. *International Journal of Molecular Sciences*, *18*(10), 2100. https://doi.org/10.3390/ijms18102100

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., & Storey, J. D. (2012). The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*, *28*(6), 882–883. https://doi.org/10.1093/bioinformatics/bts034

Leray, M., & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150331. https://doi.org/10.1098/rstb.2015.0331

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*, 34. https://doi.org/10.1186/1742-9994-10-34

Lindner, D. L., Carlsen, T., Henrik Nilsson, R., Davey, M., Schumacher, T., & Kauserud, H. (2013). Employing 454 amplicon pyrosequencing to

reveal intragenomic divergence in the internal transcribed spacer rDNA region in fungi. *Ecology and Evolution*, *3*(6), 1751–1764. https://doi.org/10.1002/ece3.586

Lozupone, C., & Knight, R. (2005). UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, *71*(12), 8228–8235. https://doi.org/10.1128/AEM.71.12.8228-8235.2005

Macheriotou, L., Guilini, K., Bezerra, T. N., Tytgat, B., Nguyen, D. T., Phuong Nguyen, T. X., Noppe, F., Armenteros, M., Boufahja, F., Rigaux, A., Vanreusel, A., & Derycke, S. (2019). Metabarcoding free-living marine nematodes using curated 18S and CO1 reference sequence databases for species-level taxonomic assignments. *Ecology and Evolution*, *9*(3), 1211–1226. https://doi.org/10.1002/ece3.4814

Machida, R. J., Kweskin, M., & Knowlton, N. (2012). PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. *PLoS One*, *7*(4), e35887. https://doi.org/10.1371/journal.pone.0035887

Magoč, T., & Salzberg, S. L. (2011). FLASH: Fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, *27*(21), 2957–2963. https://doi.org/10.1093/bioinformatics/btr507

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, *2*, e593. https://doi.org/10.7717/peerj.593

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2019). microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. *Environmental DNA*, *1*, 14–25. https://doi.org/10.1002/edn3.11

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, *8*(4), e61217. https://doi.org/10.1371/journal.pone.0061217

Mincks, S. L., Pereira, T. J., Sharma, J., Blanchard, A. L., & Bik, H. M. (2021). Composition of marine nematode communities across broad longitudinal and bathymetric gradients in the Northeast Chukchi and Beaufort Seas. *Polar Biology*, *44*(1), 85–103. https://doi.org/10.1007/s00300-020-02777-1

Murray, D. C., Coghlan, M. L., & Bunce, M. (2015). From benchtop to desktop: Important considerations when designing amplicon sequencing workflows. *PLoS One*, *10*(4), e0124671. https://doi.org/10.1371/journal.pone.0124671

Mysara, M., Njima, M., Leys, N., Raes, J., & Monsieurs, P. (2017). From reads to operational taxonomic units: An ensemble processing pipeline for MiSeq amplicon sequencing data. *GigaScience*, *6*, giw017. https://doi.org/10.1093/gigascience/giw017

Nearing, J. T., Douglas, G. M., Comeau, A. M., & Langille, M. G. I. (2018). Denoising the Denoisers: An independent evaluation of microbiome sequence error-correction approaches. *PeerJ*, *6*, e5364. https://doi.org/10.7717/peerj.5364

Oksanen, J. (2011). Vegan: Community ecology package. R package ver. 2.0-2. Http://CRAN. R-Project.Org/Package=Vegan. https://ci.nii.ac.jp/naid/20001510490/

Pauvert, C., Buée, M., Laval, V., Edel-Hermann, V., Fauchery, L., Gautier, A., Lesur, I., Vallance, J., & Vacher, C. (2019). Bioinformatics matters: The accuracy of plant and soil fungal community data is highly dependent on the metabarcoding pipeline. *Fungal Ecology*, *41*, 23–33. https://doi.org/10.1016/j.funeco.2019.03.005

Pawłowska, J., Wollenburg, J. E., Zajączkowski, M., & Pawlowski, J. (2020). Planktonic foraminifera genomic variations reflect paleoceanographic changes in the Arctic: Evidence from sedimentary ancient DNA. *Scientific Reports*, *10*(1), 15102. https://doi.org/10.1038/s41598-020-72146-9

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C., Bowser, S. S., Cepicka, I., Decelle, J., Dunthorn, M., Fiore-Donno, A. M., Gile, G. H., Holzmann, M., Jahn, R., Jirků, M., Keeling, P. J., Kostka, M., Kudryavtsev, A., Lara, E., … de Vargas, C. (2012). CBOL protist working group: Barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, *10*(11), e1001419. https://doi.org/10.1371/journal.pbio.1001419

Pereira, T. J., De Santiago, A., Schuelke, T., Hardy, S. M., & Bik, H. M. (2020). The impact of intragenomic rRNA variation on metabarcoding-derived diversity estimates: A case study from marine nematodes. *Environmental DNA*, *2*(4), 519–534. https://doi.org/10.1002/edn3.77

Porazinska, D. L., Giblin-Davis, R. M., Esquivel, A., Powers, T. O., Sung, W., & Thomas, W. K. (2010). Ecometagenetics confirm high tropical rainforest nematode diversity. *Molecular Ecology*, *19*(24), 5521–5530. https://doi.org/10.1111/j.1365-294X.2010.04891.x

Porazinska, D. L., Giblin-Davis, R. M., Sung, W., & Thomas, W. K. (2010). Linking operational clustered taxonomic units (OCTUs) from parallel ultra sequencing (PUS) to nematode species. *Zootaxa*, *2427*(1), 55. https://doi.org/10.11646/zootaxa.2427.1.6

Price, M. N., Dehal, P. S., & Arkin, A. P. (2009). FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, *26*(7), 1641–1650. https://doi.org/10.1093/molbev/msp077

Prodan, A., Tremaroli, V., Brolin, H., Zwinderman, A. H., Nieuwdorp, M., & Levin, E. (2020). Comparing bioinformatic pipelines for microbial 16S rRNA amplicon sequencing. *PLoS One*, *15*(1), e0227434. https://doi.org/10.1371/journal.pone.0227434

Qing, X., Bik, H., Yergaliyev, T. M., Gu, J., Fonderie, P., Brown-Miyara, S., Szitenberg, A., & Bert, W. (2020). Widespread prevalence but contrasting patterns of intragenomic rRNA polymorphisms in nematodes: Implications for phylogeny, species delimitation, and life history inference. *Molecular Ecology Resources*, *20*, 318–332. https://doi.org/10.1111/1755-0998.13118

R Core Team (2017). R: A language and environment for statistical computing. *R Found. Stat. Comput. Vienna, Austria. URL* http://Www.R-Project.Org/. Page R Foundation for Statistical Computing.

Rideout, J. R., He, Y., Navas-Molina, J. A., Walters, W. A., Ursell, L. K., Gibbons, S. M., Chase, J., McDonald, D., Gonzalez, A., Robbins-Pianka, A., Clemente, J. C., Gilbert, J. A., Huse, S. M., Zhou, H.-W., Knight, R., & Caporaso, J. G. (2014). Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ*, *2*, e545. https://doi.org/10.7717/peerj.545

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

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Salter, S. J., Cox, M. J., Turek, E. M., Calus, S. T., Cookson, W. O., Moffatt, M. F., Turner, P., Parkhill, J., Loman, N. J., & Walker, A. W. (2014). Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology*, *12*, 87. https://doi.org/10.1186/s12915-014-0087-z

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Schnell, I. B., Bohmann, K., & Gilbert, M. T. P. (2015). Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. *Molecular Ecology Resources*, *15*(6), 1289–1303. https://doi.org/10.1111/1755-0998.12402

Schuelke, T., Pereira, T. J., Hardy, S. M., & Bik, H. M. (2018). Nematode-associated microbial taxa do not correlate with host phylogeny,

geographic region or feeding morphology in marine sediment habitats. *Molecular Ecology*, *27*(8), 1930–1951. https://doi.org/10.1111/mec.14539

Siegwald, L., Caboche, S., Even, G., Viscogliosi, E., Audebert, C., & Chabé, M. (2019). The impact of bioinformatics pipelines on microbiota studies: Does the analytical "microscope" affect the biological interpretation? *Microorganisms*, *7*(10), 393. https://doi.org/10.3390/microorganisms7100393

Siegwald, L., Touzet, H., Lemoine, Y., Hot, D., Audebert, C., & Caboche, S. (2017). Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. *PLoS One*, *12*(1), e0169563. https://doi.org/10.1371/journal.pone.0169563

Terrat, S., Djemiel, C., Journay, C., Karimi, B., Dequiedt, S., Horrigue, W., Maron, P.-A., Chemidlin Prévost-Bouré, N., & Ranjard, L. (2020). ReClustOR: A re-clustering tool using an open-reference method that improves operational taxonomic unit definition. *Methods in Ecology and Evolution*, *11*(1), 168–180. https://doi.org/10.1111/2041-210x.13316

Toju, H., & Baba, Y. G. (2018). DNA metabarcoding of spiders, insects, and springtails for exploring potential linkage between above- and below-ground food webs. *Zoological Letters*, *4*, 4. https://doi.org/10.1186/s40851-018-0088-9

Turner, C. R., Uy, K. L., & Everhart, R. C. (2015). Fish environmental DNA is more concentrated in aquatic sediments than surface water. *Biological Conservation*, *183*, 93–102. https://doi.org/10.1016/j.biocon.2014.11.017

Westcott, S. L., & Schloss, P. D. (2015). De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ*, *3*, e1487. https://doi.org/10.7717/peerj.1487

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*, 2nd ed. Springer Publishing Company, Incorporated. https://dl.acm.org/citation.cfm?id=1795559

Wolf, D. I., & Vis, M. L. (2020). Stream algal biofilm community diversity along an acid mine drainage recovery gradient using multimarker metabarcoding. *Journal of Phycology*, *56*(1), 11–22. https://doi.org/10.1111/jpy.12935

Xiong, W., & Zhan, A. (2018). Testing clustering strategies for metabarcoding-based investigation of community-environment interactions. *Molecular Ecology Resources*, *18*(6), 1326–1338. https://doi.org/10.1111/1755-0998.12922

Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavloudi, C., & Pafilis, E. (2020). PEMA: A flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience*, *9*(3), giaa022. https://doi.org/10.1093/gigascience/giaa022

Zhao, F., Filker, S., Xu, K., Li, J., Zhou, T., & Huang, P. (2019). Effects of intragenomic polymorphism in the SSU rRNA gene on estimating marine microeukaryotic diversity: A test for ciliates using single-cell high-throughput DNA sequencing. *Limnology and Oceanography, Methods / ASLO*, *17*(10), 533–543. https://aslopubs.onlinelibrary.wiley.com/doi/abs/10.1002/lom3.10330?casa_token=UB6DoSCGRIwAAAAA:kfyKTrTzNRtNcAFnEa0e_bH8KDjU7v61kROIxW6r6AdUi_Tgs1b4xqR7slr5VZ1QG8XzCl203jmo0A

Zinger, L., Bonin, A., Alsos, I. G., Bálint, M., Bik, H., Boyer, F., Chariton, A. A., Creer, S., Coissac, E., Deagle, B. E., De Barba, M., Dickie, I. A., Dumbrell, A. J., Ficetola, G. F., Fierer, N., Fumagalli, L., Gilbert, M. T. P., Jarman, S., Jumpponen, A., … Taberlet, P. (2019). DNA metabarcoding-Need for robust experimental designs to draw sound ecological conclusions. *Molecular Ecology*, *28*(8), 1857–1862. https://doi.org/10.1111/mec.15060

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: De Santiago, A., Pereira, T. J., Mincks, S. L., & Bik, H. M. (2022). Dataset complexity impacts both MOTU delimitation and biodiversity estimates in eukaryotic 18S rRNA metabarcoding studies. *Environmental DNA*, 4, 363–384. https://doi.org/10.1002/edn3.255