ICES Journal of Marine Science

ICES Journal of Marine Science (2021), doi:10.1093/icesjms/fsab082

Contribution to the Themed Section: '*Patterns of Biodiversity of Marine Zooplankton Based on Molecular Analysis*'

The role of taxonomic expertise in interpretation of metabarcoding studies

Paula Pappalardo [1]*, Allen G. Collins [1,2], Katrina M. Pagenkopp Lohan [3], Kate M. Hanson [4], Sarit B. Truskey [1,5], William Jaeckle [6], Cheryl Lewis Ames [1,7], Jessica A. Goodheart [1], Stephanie L. Bush [1,8,9], Leann M. Biancani [1], Ellen E. Strong [1], Michael Vecchione [1,2], M. G. Harasewych [1], Karen Reed [1], Chan Lin [1], Elise C. Hartil [10], Jessica Whelpley[11], Jamie Blumberg [6], Kenan Matterson [12], Niamh E. Redmond [13], Allison Becker [13], Michael J. Boyle [14,†], and Karen J. Osborn [1,†]

[1]*Department of Invertebrate Zoology, Smithsonian National Museum of Natural History, Washington, DC, USA*
[2]*National Systematics Laboratory, NOAA's National Marine Fisheries Service, Smithsonian National Museum of Natural History, Washington, DC, USA*
[3]*Marine Disease Ecology Laboratory, Smithsonian Environmental Research Center, Edgewater, MD, USA*
[4]*Biology Department, Duke University, Durham, NC, USA*
[5]*Marine Science Center, Northeastern University, Nahant, MA, USA*
[6]*Biology Department, Illinois Wesleyan University, Bloomington, IL, USA*
[7]*International Integrative Research and Instruction, Graduate School of Agricultural Science, Tohoku University, Sendai, Japan*
[8]*Monterey Bay Aquarium Research Institute, Moss Landing, Midwater Ecology Lab, CA, USA*
[9]*Monterey Bay Aquarium, Animal Husbandry Department, Monterey, CA, USA*
[10]*School of Marine Sciences, University of Maine, Orono, ME, USA*
[11]*Whitney Laboratory for Marine Bioscience, University of Florida, St. Augustine, FL, USA*
[12]*Department of Biological, Geological, and Environmental Sciences, University of Bologna, Ravenna, Italy*
[13]*Smithsonian Institution Barcode Network, Smithsonian National Museum of Natural History, Washington, DC, USA*
[14]*Smithsonian Marine Station, Smithsonian National Museum of Natural History, Fort Pierce, FL, USA*

*Corresponding author: e-mail: paulapappalardo@gmail.com.
†The last two authors are co-senior authors.

Pappalardo, P., Collins, A. G., Pagenkopp Lohan, K. M., Hanson, K. M., Truskey, S. B., Jaeckle, W., Ames, C. L., Goodheart, J. A., Bush, S. L., Biancani, L. M., Strong, E. E., Vecchione, M., Harasewych, M. G., Reed, K., Lin, C., Hartil, E. C., Whelpley, J., Blumberg, J., Matterson, K., Redmond, N. E., Becker, A., Boyle, M. J., and Osborn, K. J. The role of taxonomic expertise in interpretation of metabarcoding studies. – ICES Journal of Marine Science, doi:10.1093/icesjms/fsab082.

Received 29 December 2020; revised 5 April 2021; accepted 5 April 2021.

The performance of DNA metabarcoding approaches for characterizing biodiversity can be influenced by multiple factors. Here, we used morphological assessment of taxa in zooplankton samples to develop a large barcode database and to assess the congruence of taxonomic identification with metabarcoding under different conditions. We analysed taxonomic assignment of metabarcoded samples using two genetic markers (COI, 18S V1–2), two types of clustering into molecular operational taxonomic units (OTUs, ZOTUs), and three methods for taxonomic assignment (RDP Classifier, BLASTn to GenBank, BLASTn to a local barcode database). The local database includes 1042 COI and 1108 18S (SSU) barcode sequences, and we added new high-quality sequences to GenBank for both markers, including 109 contributions at the

© International Council for the Exploration of the Sea 2021.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

species level. The number of phyla detected and the number of taxa identified to phylum varied between a genetic marker and among the three methods used for taxonomic assignments. Blasting the metabarcodes to the local database generated multiple unique contributions to identify OTUs and ZOTUs. We argue that a multi-marker approach combined with taxonomic expertise to develop a curated, vouchered, local barcode database increases taxon detection with metabarcoding, and its potential as a tool for zooplankton biodiversity surveys.

Keywords: barcode, biodiversity, Gulf Stream, holoplankton, meroplankton, metabarcode, zooplankton

Introduction

Recent studies suggest that metabarcoding could eventually replace traditional morphological methods for biodiversity surveys and ecosystem assessment (Lejzerowicz *et al.*, 2015; Aylagas *et al.*, 2016; Elbrecht *et al.*, 2017; Lobo *et al.*, 2017; Carew *et al.*, 2018). By using DNA-identification to analyse a mixture of unidentified species, metabarcoding offers the potential to dramatically reduce the time and cost of biodiversity surveys while increasing detection of taxa that are difficult to identify based on morphology alone (e.g. invasive species, Abad *et al.*, 2016; cryptic species, Aylagas *et al.*, 2016; parasites, Pagenkopp Lohan *et al.*, 2016). However, in many environments, it remains to be demonstrated that metabarcoding can perform better than traditional morphological assessments by taxonomic experts. The performance of metabarcoding approaches is highly dependent on multiple factors, including how the initial samples are processed (Ransome *et al.*, 2017; Pagenkopp Lohan *et al.*, 2019), the choice of primers (Lobo *et al.*, 2017), PCR profiles (Aylagas *et al.*, 2016; Clarke *et al.*, 2017), genetic markers (Pitz *et al.*, 2020), the completeness of the DNA sequence reference database (Lindeque *et al.*, 2013), and even the software used for taxonomic assignment (Bazinet and Cummings, 2012).

Even though traditional biodiversity surveys based on morphological assessments are time consuming and depend on the level of taxonomic expertise available, they may capture relevant details for population and community analyses (such as sex, life history stage, and relative abundance of taxa, Lindeque *et al.*, 2013). Most importantly, taxonomic experts can contribute DNA sequences to reference databases based on confidently identified and well-documented specimens deposited in a vouchering institution. When paired with metabarcoding, this traditional approach can dramatically reduce the number of unidentified (or misidentified) sequences (Ransome *et al.*, 2017), which in many studies, comprise a large part of a community sample (e.g. 93% of taxa for the COI marker in Cowart *et al.*, 2015). These unidentified sequences may actually represent species that are new to science; however, most often they represent known taxa that are not represented in current reference DNA-sequence databases. Therefore, experts agree that continued contribution of vouchered sequences to existing reference databases will be a major step to fulfil the promise of metabarcoding approaches (Cristescu, 2014; Cowart *et al.*, 2015; Bucklin *et al.*, 2016; Ransome *et al.*, 2017; Porter and Hajibabaei, 2018).

Several curated databases have been developed in parallel to the implementation of metabarcoding methods, emphasizing specific taxonomic groups or genetic markers (e.g. BOLD, Ratnasingham and Hebert, 2007; SILVA, Quast *et al.*, 2012; see Table 2 of Porter and Hajibabaei, 2018 for a summary of databases commonly used in metabarcoding studies). PR2 and Midori are additional curated databases that were not assessed in Porter and Hajibabaei (2018). The PR2 database includes rRNA sequences with a curated taxonomy structured to conform with Linnaean ranks (Guillou *et al.*, 2012). The Midori database includes sequences from GenBank with taxonomic information at the species level for 13 protein-coding genes (including COI) and two ribosomal RNA (lrRNA and srRNA; Machida *et al.*, 2017). The stringent filtering of GenBank data improves data reliability in Midori, but greatly reduces the representation of known species diversity (Machida *et al.*, 2017). While not often quantified, similarly uneven and incomplete representation of extant taxa should be expected for other curated databases.

The identification at any level from population to phylum of metabarcoding sequence profiles depends not only on the completeness of the reference database but also on the algorithm used. Following Bazinet and Cummings (2012), these algorithms can be classified as (i) similarity-based (e.g. programs using BLAST), (ii) composition-based (e.g. programs using Naive Bayes Classifier), and (iii) phylogeny-based (e.g. programs using maximum likelihood or Bayesian methods). Even within a category, the performance of different bioinformatics pipelines can vary widely when applied to the same dataset (Bazinet and Cummings, 2012; Pitz *et al.*, 2020). In addition, the resolution and error rate of taxonomic assignments can vary for different genetic markers (Richardson *et al.*, 2017; Pitz *et al.*, 2020). Opinions on which markers work best vary greatly even for the same taxonomic group (Bhadury *et al.*, 2006; Tang *et al.*, 2012). In studies targeting taxonomically diverse assemblages, there is agreement that a multi-marker approach is more effective (Cowart *et al.*, 2015; Bucklin *et al.*, 2016; Djurhuus *et al.*, 2018), but continued assessment of different methods and markers for taxonomic assignment remains critical to enable metabarcoding data for studies of biodiversity and ecological research.

A more fundamental methodological choice in any metabarcoding study is to determine how to generate the final set of sequences for taxonomic identification. Sequences are usually clustered into operational taxonomic units (OTUs), using a defined similarity threshold (typically 97%) that is likely to represent different species. But recent studies have suggested that individual unique amplicons (sequences amplified from an environmental DNA sample) make metabarcoding more reproducible and better represent the full biological diversity within environmental samples (Callahan *et al.*, 2017; Edgar, 2018). These unique sequences have been called amplicon sequence variants (ASVs) or zero-radius OTUs (ZOTUs) and capture both inter-specific and intra-specific variability. While the full impact of this choice on derived research questions is still being assessed, it affects the number of taxa detected (Schenk *et al.*, 2020).

Our study combined expert taxonomic knowledge with standard DNA barcoding and examines how metabarcoding results can be improved by using taxonomic expertise to develop a curated database of vouchered DNA barcodes for multiple genetic markers. We focused on a challenging yet commonly sampled assemblage, marine zooplankton communities. Given their broad taxonomic diversity, their roles as critical members of ocean food

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

webs, their many morphologically distinctive early life stages, and the extensive taxonomic expertise required for accurate morphological identifications, marine invertebrate zooplankton are a natural target for diversity surveys using a metabarcoding approach. Specifically, we compared the resulting taxonomic identification of metabarcoding samples analysed using: (i) different markers, i.e., either COI or 18S V1–2, (ii) different bioinformatic pipelines, and (iii) different operational taxonomic units (OTUs and ZOTUs). The different methods were compared among each other, since the true community composition is unknown. To accomplish these goals, the StreamCode Project (a collaborative effort led by researchers in the Smithsonian National Museum of Natural History), made multiple collections of plankton from a specific geographic area and developed one of the largest curated zooplankton databases that includes DNA barcodes. We analyse if this regional database can help to identify organisms to higher taxonomic levels such as phylum, and also if it can increase the taxonomic resolution necessary to classify zooplankton into ecological plankton type to identify permanent (holoplankton) or temporary (meroplankton) residents of the plankton.

Methods

StreamCode project plankton collection

Twelve plankton tow samples were collected from the Florida Current of the Gulf Stream off the Atlantic coast near Fort Pierce, Florida (27.45°N 79.95°W) in June and August 2017 between the surface and 77–145 m depth. To maximize the diversity of taxa recovered, we used both a circular mouth plankton net (mesh size 209 µm) and a modified mid-water trawl (1.0 m^2 square mouth net with 500 µm mesh). Upon collection and mixing, approximately one-quarter of each plankton sample was separated for metabarcoding analysis. Each tow-specific subsample designated for the metabarcoding analysis was concentrated by pouring it through a 50 µm mesh nylon filter, rinsed with ultracold 95% ethanol into a 50 ml polypropylene tube, stored on dry ice for transport to the Smithsonian Marine Station at Fort Pierce (SMSFP), and stored at −80°C until processing. The leftover portion of the mixed tow sample was diluted in at least 10 l of seawater, kept aerated and chilled during transport to SMSFP, and used for live sorting of individuals.

Taxonomic identification using morphology

Live plankton were hand-sorted based on taxonomic groups under stereomicroscopes at the SMSFP by a team of taxonomic experts and students. Animals were selectively picked with the goal of maximizing the diversity of samples collected from focal groups across 15 invertebrate phyla (between 1 and 200 individuals per focal group, Table 1). For some groups (pteropods, holoplanktonic polychaetes), selective sampling was able to capture most of the expected species diversity. For species-rich groups such as copepods, the sorting effort captured only a fraction of the expected species diversity, and experts targeted taxa not yet represented in GenBank. Live specimens were grouped by morphotype and classified to the lowest taxonomic level possible.

Individuals and populations of live specimens were photographed in the laboratory at SMSFP (see photography details and how to access images in the Supplementary material). The high-resolution images allowed us to verify the initial identifications made in the field. Tissue samples or whole animals were placed directly in 150 µl TD-M2 tissue buffer (Autogen, Inc., Hollister, MA) for extraction, PCR amplification and Sanger sequencing (see DNA Barcoding below). All morphological and image vouchers generated by the StreamCode project (Supplementary Table S1), as well as tissues and DNA aliquots, are available in the collections of the National Museum of Natural History (NMNH).

After each of the samples was sequenced for COI and 18S V1–2 regions, a two-pronged approach was taken to verify and refine the initial identifications: (i) BLAST and alignment based and (ii) phylogeny based. Both approaches were run on the Smithsonian Institution's High Performance Computing Cluster (https://doi.org/10.25572/SIHPC), see Supplementary material for additional details. For the BLAST approach (Altschul *et al.*, 1990), each sequence was compared against sequences in GenBank and personal reference libraries to verify taxon assignments. For the phylogeny-based approach, we combined target sequences, related sequences from GenBank, and sequences from personal libraries to build phylogenetic trees and to determine the placement of each target sequence in relation to known sequences. Based on the initial field IDs, results of the BLASTn searches, the phylogenetic trees, photographs, and voucher specimens, taxonomic experts assigned a final ID to each sample (Supplementary Table S1).

DNA barcoding

Detailed sequencing methods, primers used (Supplementary Table S2), and quality control protocols are provided in the Supplementary material. DNA extraction, amplification, and sequencing were performed in the Smithsonian Laboratories of Analytical Biology (LAB), NMNH. All 18S V1–2 and COI sequences produced in the current study, the StreamCode DNA Barcode Database, were uploaded to GenBank (https://www.ncbi.nlm.nih.gov/genbank/, NCBI BioProject PRJNA421480, see Supplementary Table S1 for accession numbers).

StreamCode DNA barcode database

To identify novel contributions of the StreamCode DNA barcode database to GenBank, we used the gap analysis tool developed by the Global Genome Initiative (GGI) for each refined ID (Supplementary Figure S1) at the rank of species and higher taxonomic levels (GGI Gap Analysis Tool, 2019; https://www.globalgeno.me/gaps/live), GGI tool last accessed on 29 October 2020. The live gap analysis tool searches GenBank for each name considering only sequences of barcode quality, defined as GenBank sequence records for COI, rbcL, matK, and/or ITS with a sequence length $>$ 500 bp that include information identifying a scientific name, a voucher specimen, and the amplification primer sequences (GGI Gap Analysis Tool, 2019). From the two markers we used for this study (COI and 18S), only COI is officially considered a barcode and included in the Gap Analysis Tool; however, we modified this tool with an in-house Python script to search for 18S sequences in GenBank separately for each Phylum.

DNA metabarcoding and bioinformatics protocol

DNA extraction, PCR amplification, and multiplex library preparation of pooled amplicons were performed at the SMSFP; final quantification and library-size determination steps for sequencing protocols were performed by staff at LAB. Raw metabarcoding data is available through NCBI Sequence Read Archive and can be found in NCBI BioProject PRJNA421480 (https://www.ncbi.nlm.

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

Table 1. Barcoding success.

Phylum	Target group	Number of morphospecies	Total number of individuals	COI barcodes	18S barcodes	Individuals barcoded for both	% success COI	% success 18S	% success both
Annelida	Class Polychaeta	57	160	107	115	93	66.88	71.88	58.13
Arthropoda	Subclass Copepoda	51	95	77	95	77	81.05	100	81.05
	Suborder Hyperiidea	60	202	177	132	122	87.62	65.35	60.4
	Miscelaneous arthropods	78	130	108	116	103	83.08	89.23	79.23
	Class Ostracoda	18	91	53	70	41	58.24	76.92	45.05
Brachiopoda	Phylum Brachiopoda	2	12	12	12	12	100	100	100
Bryozoa	Phylum Bryozoa	3	7	6	7	6	85.71	100	85.71
Chaetognatha	Phylum Chaetognatha	10	43	19	12	8	44.19	27.91	18.6
Chordata	Class Appendicularia	7	24	5	16	4	20.83	66.67	16.67
Chordata	Class Thaliacea	7	17	1	15	1	5.88	88.24	5.88
Cnidaria	Class Anthozoa	8	27	20	20	17	74.07	74.07	62.96
	Class Hydrozoa	38	103	58	86	57	56.31	83.5	55.34
	Class Scyphozoa	4	19	17	17	17	89.47	89.47	89.47
Ctenophora	Phylum Ctenophora	3	5	0	5	0	0	100	0
Echinodermata	Class Asteroidea	6	30	28	27	25	93.33	90	83.33
	Class Echinoidea	5	7	7	7	7	100	100	100
	Class Holothuroidea	5	13	12	13	12	92.31	100	92.31
	Class Ophiuroidea	11	26	17	23	17	65.38	88.46	65.38
Hemichordata	Phylum Hemichordata	6	16	8	6	6	50	37.5	37.5
Mollusca	Class Bivalvia	5	10	8	10	8	80	100	80
	Class Cephalopoda	10	20	17	15	15	85	75	75
	Class Gastropoda	51	96	76	73	66	79.17	76.04	68.75
	Order Pteropoda	33	209	177	181	163	84.69	86.6	77.99
Nemertea	Phylum Nemertea	1	1	0	0	0	0	0	0
Phoronida	Phylum Phoronida	1	1	1	1	1	100	100	100
Sipuncula	Phylum Sipuncula	19	35	31	34	31	88.57	97.14	88.57
Total		499	1399	1042	1108				

Total number of morphospecies, total number of StreamCode samples collected for each taxonomic group, details of number of samples for each genetic marker that yielded useful sequences, and percentage of samples that were successfully sequenced for COI and 18S V1–2. Cells in red highlight the groups with less than 75% success. Miscellaneous arthropods include non-hyperiid amphipods, and members of the orders Cumacea, Decapoda, Euphausiacea, Mysida, and Stomatopoda.

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

Table 2. Performance of methods to assign barcodes to phylum using the StreamCode DNA barcode database.

			RDP classifier		BLASTn to GenBank		BLASTn to StreamCode	
Marker	**StreamCode Phylum**	**N barcodes**	**% identified**	**% wrong**	**% identified**	**% wrong**	**% identified**	**% wrong**
COI	Annelida	107	6.5	14.3	22.4	0	83.2	0
	Arthropoda	415	78.8	0	74.2	0	90.1	0
	Brachiopoda	12	NA	NA	NA	NA	100	0
	Bryozoa	6	NA	NA	NA	NA	100	0
	Chaetognatha	19	89.5	5.9	84.2	0	84.2	0
	Chordata	6	NA	NA	NA	NA	66.7	0
	Cnidaria	95	74.7	0	72.6	0	86.3	0
	Echinodermata	64	89.1	0	64.1	0	84.4	0
	Hemichordata	8	12.5	100	NA	NA	87.5	0
	Mollusca	278	75.2	0	77	0	87.8	0
	Phoronida	1	NA	NA	NA	NA	NA	NA
	Sipuncula	31	35.5	0	74.2	0	77.4	0
18S V1–2	Annelida	115	88.7	8.8	99.1	1.8	100	0
	Arthropoda	413	87.2	0	96.9	0	99.5	0
	Brachiopoda	12	100	0	100	0	100	0
	Bryozoa	7	100	0	100	0	100	0
	Chaetognatha	12	91.7	0	100	0	100	0
	Chordata	31	100	0	100	0	100	0
	Cnidaria	123	94.3	0	100	0	100	2.4
	Ctenophora	5	100	0	100	0	100	0
	Echinodermata	70	87.1	0	100	0	100	0
	Hemichordata	6	100	0	100	0	100	0
	Mollusca	279	95	0	99.6	0	99.3	0
	Phoronida	1	100	0	100	0	100	100
	Sipuncula	34	88.2	0	100	0	97.1	0

Total number of barcodes tested, percentage of barcodes identified to phylum, and percentage of barcodes with a wrong assignment for each marker and StreamCode phylum. To accept an identification to phylum, the confidence threshold was equal to or larger than 0.90 with RDP Classifier, and percent similarity was equal to or higher than 85% for BLASTn. The RDP Classifier was used with the Midori database for COI and with the PR2 database for 18S. NA (not applicable) are instances where no assignment passed the confidence threshold.

nih.gov/bioproject/?term=PRJNA421480). Additional details on primers (Supplementary Table S3), sequencing methods and bioinformatics pipeline are provided in the Supplementary material.

Reads were merged and filtered with USEARCH 10.0 (Edgar, 2013). Post-merging, the allowable sequence range was 400–500 bp for 18S V1–2 and 300–400 bp for COI. Primer specific barcodes were demultiplexed in QIIME (Caporaso *et al.*, 2010). In USEARCH, primers were removed, and then unique sequences were identified, de-replicated, and sorted by abundance. Because the "species level" clustering threshold likely varies by taxonomic group, we took two clustering approaches: (i) clustering sequences into operational taxonomic units (OTUs) at 97% similarity after removing singletons and (ii) clustering sequences into zero-radius operational taxonomic units (ZOTUs), allowing at least four copies, which is recommended for smaller datasets (Edgar, 2016). The number of metabarcode sequences in each filtering step is detailed in Supplementary Table S4. The variability in the number of OTUs per phylum with an alternative OTU filtering (keeping only sequences with at least four copies) is presented in Supplementary Table S5.

Assigning taxonomy to the metabarcode sequences

We took three approaches to assign taxonomic identities to OTUs and ZOTUs: (i) taxonomic assignment using the RDP Classifier (Wang *et al.*, 2007) with Midori or PR2 as reference databases for COI and 18S, respectively; (ii) taxonomic assignment using BLASTn against the NCBI nt database (we refer to this approach as BLASTn-GenBank); (iii) taxonomic assignment using BLASTn against the StreamCode DNA barcode database (we refer to this approach as BLASTn-StreamCode).

The RDP classifier (Wang *et al.*, 2007) was implemented in QIIME (Caporaso *et al.*, 2010) for the taxonomic assignment of the 18S sequences, selecting PR2 11.1 as the reference database (Guillou *et al.*, 2012, PR2 database includes rRNA sequences with curated taxonomy). For the COI sequences, the RDP Classifier was implemented in the Midori server queried on 3 October 2020 (http://reference-midori.info/server.php; Leray *et al.*, 2018), using the Midori 2-LONGEST database as the reference (Machida *et al.*, 2017; filtered database of mitochondrial-encoded genes for metazoans). We accepted a phylum classification as correct when the confidence threshold was equal to or larger than 0.90. Based on the information available in Midori, we observed that using 0.90 as the confidence threshold for COI yields an error rate less than 1% for most phyla commonly found in zooplankton assemblages (http://reference-midori.info/download.php#; follow the path Archive/Leave-one-sequence-out_test_1.1/COI). We are not aware of similar specific guidelines for 18S using the PR2 database, so we applied the same confidence threshold to 18S.

For the BLASTn-GenBank approach, we assigned taxonomy to the OTUs and ZOTUs by running a BLASTn search with default options on 6th October 2020. We filtered out matches lacking specific taxonomic information using custom scripts in R (R Core Team, 2020, see Data Availability for code). For the BLASTn-StreamCode approach, we created a reference database with the StreamCode voucher-based barcodes for each marker and assigned taxonomy using BLASTn. A match to phylum was

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

accepted as correct if the percent similarity and sequence coverage were equal to or larger than 85% (Ransome *et al.*, 2017). At least for COI, Ransome *et al.* (2017) showed that an 85% similarity threshold provides a relatively small error rate (0.7%) when assigning sequences to phylum.

At the beginning of this project, we intended to compare agreements at various taxonomic levels, but we did not find standardized recommendations. Initial explorations with our dataset showed that we would need to generate specific thresholds of taxonomic assignments for each marker, method, and taxonomic group. For this reason, we focused our main analysis on the assignments to phylum.

Taxonomic matching between databases

To standardize the taxonomic names across the morphological assignments and the different databases consulted in the metabarcoding analysis (PR2, MIDORI, and GenBank), we followed the classification in the World Register of Marine Species (WoRMS, Horton *et al.*, 2020). All names corresponding to marine organisms were matched using the WoRMS online taxon match tool (http://www.marinespecies.org/aphia.php?p=match), last accessed in October 2020. For matches to non-marine organisms, we followed the NCBI taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy), last accessed in October 2020. When a phylum assignment of a sequence did not pass the corresponding thresholds for each method, the phylum was labelled "Unidentified."

Standardizing names across databases is important because the taxonomic differences between databases are not minor. To give one example, the PR2 database considers Urochordata to be a phylum, but the WoRMS classification considers Urochordata to be subphylum Tunicata (within the Phylum Chordata). There is some debate about the placement of Sipuncula as a subgroup of the phylum Annelida (Struck *et al.*, 2011; Parry *et al.*, 2016); to be consistent in our analysis, we followed the WoRMS classification, as of November 2020, that considers Sipuncula a phylum. In addition, there were a few cases in which the minimum taxonomic level reported in the output (by MIDORI or PR2) did not match the phylum assigned by WoRMS. When that happened, we ran independent BLASTn searches of each OTU/ZOTU and verified that the WoRMS assignment was correct. We enumerate those cases in the R code provided in the Dryad data package.

Validating methods for a taxonomic assignment using the StreamCode DNA barcode database

To validate the performance of the taxonomic assignment to phylum for the metabarcoding data, we ran the StreamCode COI and 18S barcodes using the same three methods for a taxonomic assignment detailed for the metabarcoding data: (i) RDP Classifier, (ii) BLASTn-GenBank, and (iii) BLASTn-StreamCode, without including the query sequence. As not all the barcodes were successfully identified at the quality thresholds specified, we calculated the percentage of barcodes that were identified, and from the barcodes identified, we calculated the percentage of wrong assignments to phylum.

Variability in the metabarcoding results

To examine the variability of taxonomic assignments associated with methodological choices when assessing a zooplankton assemblage, we compared sequence assignments to the phylum level resulting from the use of different genetic markers (COI, 18S V1–2), different types of clustering (OTUs, ZOTUs), and different approaches to assigning taxonomy (RDP Classifier with MIDORI/PR2, BLASTn-GenBank, and BLASTn-StreamCode). We calculated the number (and proportion) of OTUs/ZOTUs for each phylum, genetic marker, and approach for a taxonomic assignment.

Agreements and unique contributions among methods for assignment to phylum

For all the OTUs/ZOTUs, we quantified the agreement between methods for taxonomic assignment to phylum. We recorded whether all methods agreed, only two agreed, or zero agreed. Additionally, we kept records of how many agreements were to a specific phylum (e.g. Annelida) versus agreements in failure to identify a phylum (e.g. Unidentified).

For the taxonomic groups best studied by the StreamCode taxonomic experts (Annelida, Arthropoda, Cnidaria, Mollusca, and Sipuncula), we combined the barcodes and OTUs into a phylogenetic tree (details in Supplementary material). For the metabarcodes, we highlighted the cases in which two or three methods agreed in the assignment and unique cases in which only one method was able to assign the sequence to a phylum (we refer to these cases as unique contributions).

Plankton types

To investigate a categorical approach where it is necessary to match taxa with an ecological planktonic type, we classified OTUs/ZOTUs into meroplankton and holoplankton. Meroplankton consists of the life-history stages of benthic organisms that live temporarily in the water column (e.g. eggs, larvae, and adults in the case of medusozoan cnidarians), whereas holoplankton refers to organisms that live permanently within the water column. Although in some cases all members of a phylum belong to the same plankton type (e.g. all Sipuncula were meroplankton), for Annelida, Arthropoda, Chordata, Cnidaria, and Mollusca, we used customized taxonomic filters that involved taxonomic assignments below the level of phylum (i.e. Class, Order, Family). In this analysis, we excluded OTUs/ZOTUs assigned as "Unidentified" or as non-target phyla.

Results

StreamCode project

The collection, handling, morphological identification, and photographing of zooplankton samples involved approximately 1800 h of field work by 31 persons, or 58 person-hours. Of the 2260 specimens collected, representative individuals were catalogued ($n = 1529$; Supplementary Table S1), tissue sampled ($n = 1399$), photographed ($n = 1149$, Figure 1), and, when possible, a morphological voucher was deposited into the National Museum of Natural History collections ($n = 403$). Universal primers were effective for sequencing individuals of most taxonomic groups, although there was considerable variation in success between the two genetic markers (COI, 18S V1–2) across the different target groups (Table 1). The number of morphospecies identified by the taxonomic experts for each target taxonomic group is reported in Table 1. The refined ID for each sample, current classifications following the WoRMS taxonomy, voucher identification number (USNM), and GenBank accession numbers

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

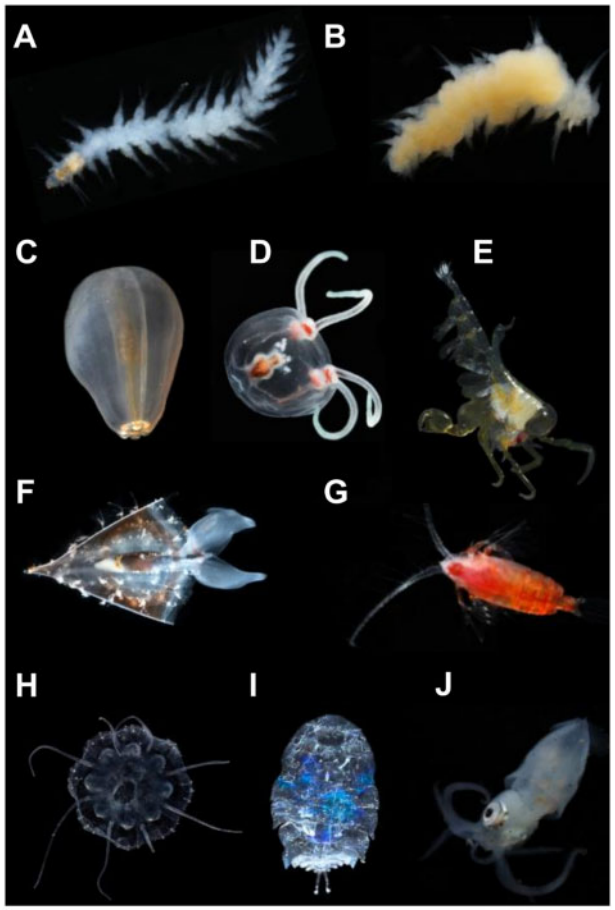

Figure 1. Examples of StreamCode specimens. (a) *Pelagobia* sp. (Polychaeta), USNM 1450035; (b) *Maupasia* sp. (Polychaeta), USNM 1450037; (c) *Peachia* sp. (Cnidaria), USNM 1448832; (d) *Cytaeis* sp. (Cnidaria), USNM 1447971; (e) *Phronima* sp. (Arthropoda), USNM 1450286; (f) *Clio recurva* (Mollusca), USNM 1448342; (g) *Euchirella curticauda* (Arthropoda), USNM 1448593; (h) *Otoporpa sp.* (Cnidaria), USNM 1448490; (i) *Copilia* sp. (Arthropoda), USNM 1448598; (j) *Abralia veranyi* (Mollusca), USNM 1447996. (a–d) Highlight of four StreamCode samples that represent new COI contributions of barcode quality at the genus level not previously represented in GenBank.

for the samples sequenced successfully are also included in Supplementary Table S1.

Contributions to GenBank

The gap analysis allowed us to identify taxa for which StreamCode contributed new COI or 18S sequences to GenBank as of 29 October 2020 (taxon names are detailed in Supplementary Figure S1). The COI results were specific to contributions of barcode quality (e.g. *Flaccisagitta enflata* will count as a new contribution because the available sequence in GenBank is shorter than 500 bp), but the 18S results represent any contribution for that genetic marker without a predefined quality filter. For COI, we contributed sequences for three classes that were previously not included in GenBank (Appendicularia and Thaliacea within phylum Chordata, and Phascolosomatidea within phylum Sipuncula). No new class-level contributions were made for 18S. Order level contributions were made for both COI and 18S within the phyla Echinodermata, Mollusca, and Sipuncula (these could be related to taxonomic backbone differences, see Supplementary Figure S1); and for COI within the phyla Annelida and Chordata (Supplementary Figure S1). We also contributed COI sequences for 33 families (from 6 phyla) previously not represented, and 11 for 18S; sequences were contributed for 66 genera (from 8 phyla) previously not represented for COI, and 33 for 18S; and sequences were contributed for 63 species previously not represented for COI, and 46 for 18S (Supplementary Figure S1). Overall, the largest number of sequence contributions were for the phyla Arthropoda, Cnidaria, and Mollusca.

Validating taxonomic assignment methods using the StreamCode DNA barcode database

The percentage of StreamCode barcodes identified to phylum ranged from 0 to 100%, depending on the genetic marker and the method used to identify taxa (Table 2). For both markers and most phyla, the BLASTn-GenBank method identified a larger percentage of barcodes than the RDP Classifier, and in most cases at a smaller or zero error rate (Table 2). Using the local database (BLASTn-StreamCode) improved the percentage of identified taxa for most phyla, with the largest improvement observed with COI for the phylum Annelida (Table 2). Not all the barcodes identified to a phylum were assigned to the correct phylum; the percentage of wrong assignments varied with marker and method but was in general larger for COI than for 18S. All of the StreamCode phyla were identified using the 18S barcodes, with only small errors in the assignment to phylum for Annelida and Cnidaria (Table 2), and a failure to correctly identify Phoronida using the BLASTn-StreamCode approach. Since there was only one StreamCode DNA barcode available for the phylum Phoronida, when that sequence was queried there were no other sequences from the phylum Phoronida for comparison.

Variability in metabarcoding results

The total number of OTUs/ZOTUs identified to phylum differed by marker, and methods used for a taxonomic assignment (Table 3). Regardless of the type of clustering or method for a taxonomic assignment, the proportion of unidentified taxa was smaller for 18S V1–2 than COI (18S: 15.3–40.8%; COI: 34.7–75.0%). In general, the BLASTn-GenBank method returned a smaller number of unidentified taxa and identified a larger number of non-target phyla when compared with the other methods (Table 3).

Phylum detection at the defined taxonomic thresholds varied by marker, type of clustering, and method for a taxonomic assignment. The phyla detected by all metabarcoding method combinations (and by the morphological analysis) were Annelida, Arthropoda, Bryozoa, Chaetognatha, Chordata, Cnidaria, Echinodermata, Mollusca, and Sipuncula. Other phyla were detected only by some combinations: Acanthocephala using 18S V1–2; Porifera using COI with BLASTn-GenBank and by morphological analysis; Nematoda using OTUs by BLAST-GenBank and RDP Classifier. By definition, the phyla not represented in the StreamCode DNA barcode database (Acanthocephala, Nematoda, Nemertea, and Platyhelminthes) were unidentified using the BLASTn-StreamCode approach.

The cases in which the only one of the methods used for a taxonomic assignment was able to identify a phylum (unique

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

Table 3. Summary of metabarcoding results.

Marker	Clustering	Taxa	RDP classifier	BLASTn-GenBank	BLASTn-StreamCode
COI	OTUs	Zooplankton	1661	1986	1234
		Unidentified	3283	2817	3712
		Non-target	2	143	0
	ZOTUs	Zooplankton	5168	5910	4646
		Unidentified	4080	3206	4603
		Non-target	1	133	0
18S V1–2	OTUs	Zooplankton	1892	1914	2935
		Unidentified	1468	1453	663
		Non-target	238	231	0
	ZOTUs	Zooplankton	2252	2570	2744
		Unidentified	781	526	707
		Non-target	418	355	0

Number of OTUs/ZOTUs identified to target zooplankton phylum, unidentified to phylum, or identified to non-target phylum for each genetic marker, type of clustering, and method used for taxonomic assignment.

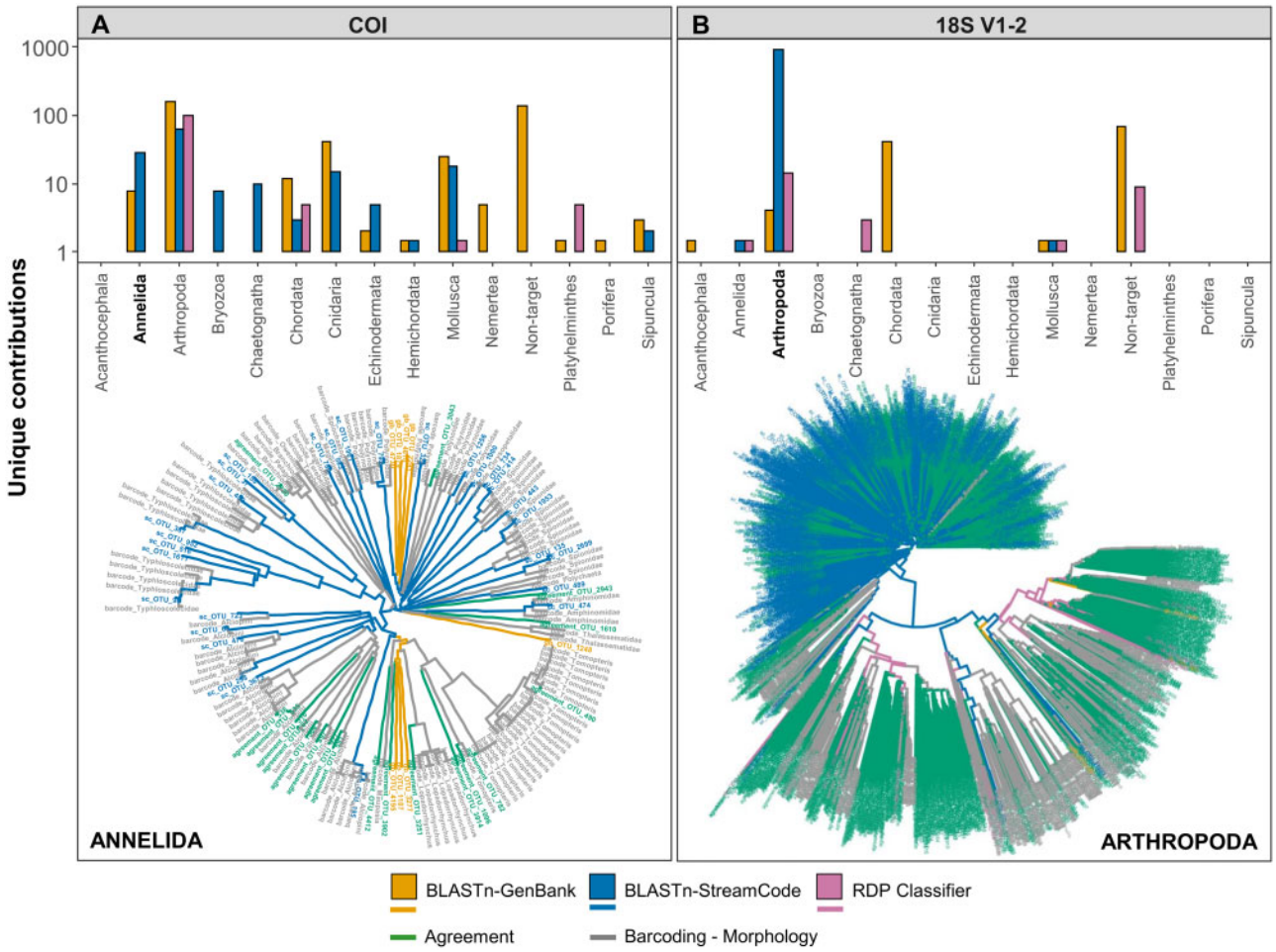

Figure 2. Unique OTUs contributions from each method for a taxonomic assignment (BLASTn-GenBank, BLASTn-StreamCode, RDP classifier), for (a) COI and (b) 18S V1–2. The RDP Classifier was used with the PR2 database for 18S and MIDORI 2 database for COI. Upper panels: counts for all the unique contributions to identify OTUs in each phylum. We added a constant of 0.5 when the number of unique contributions was 1, to be able to represent those values in the logarithmic scale. "Non-target" refers to taxa identified to phyla that do not belong to the target zooplankton groups. Lower panels: distance tree of all the barcodes and OTUs for Annelida and Arthropoda, coloured according to the identification method for each genetic marker sequence. When two or three metabarcoding methods agreed in the assignment to the phylum, we coded it "Agreement."

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

contributions) varied by taxonomic group and genetic marker (Figure 2). For COI OTUs, the BLASTn-StreamCode method generated more unique contributions for Annelida, Bryozoa, Chaetognatha, and Echinodermata (Figure 2a); whereas for 18S V1–2 OTUs, the BLASTn-StreamCode method generated the largest number of unique contributions for taxa within Arthropoda (Figure 2b). BLASTn-GenBank and RDP Classifier had more unique contributions within Chordata, Arthropoda (for COI) and non-target phyla; BLASTn-GenBank and BLASTn-StreamCode had the largest number of unique contributions for the phylum Cnidaria; and RDP Classifier (for COI) performed best to identify taxa from the phylum Platyhelminthes (Figure 2). A distance tree including StreamCode barcodes and OTUs for taxa assigned to Arthropoda and Annelida is presented in Figure 2 (lower panel), similar trees for both markers and other phyla are presented in Supplementary Figure S3. The unique contributions to identify ZOTUs were fairly similar and are presented in Supplementary Figure S2.

Arthropods (predominantly crustaceans) were the most diverse group in our plankton samples for all markers and clustering schemes (Figure 3a). Depending on the marker and clustering type, the second most diverse groups were chaetognaths, chordates, cnidarians, mollusks, or non-target phyla. (Supplementary Table S6, Figure 3b).

Combining all OTUs and ZOTUs to analyse agreement among methods, we observed that the three methods agreed on the phylum assignment in 55% of sequences for 18S and 68% of sequences for COI. Agreements between the two methods were observed in 42% of sequences for 18S and 32% of sequences for COI (Figure 4). The case of no agreement between methods was rare (0.1–3%). However, in many cases, the methods agreed only in their inability to identify a specimen to phylum, e.g., in 53% of the cases all three methods were unable to identify a phylum for COI at the specified thresholds (Figure 4). When all three methods assigned a specific phylum, 98.9% of 18S V1–2 sequences agreed on phylum and 99.8% of COI sequences had matching phylum assignments.

Plankton types

After removing unidentified matches or matches to non-target phyla, we found that most metabarcoding samples consisted of holoplankton, in agreement with morphological analyses

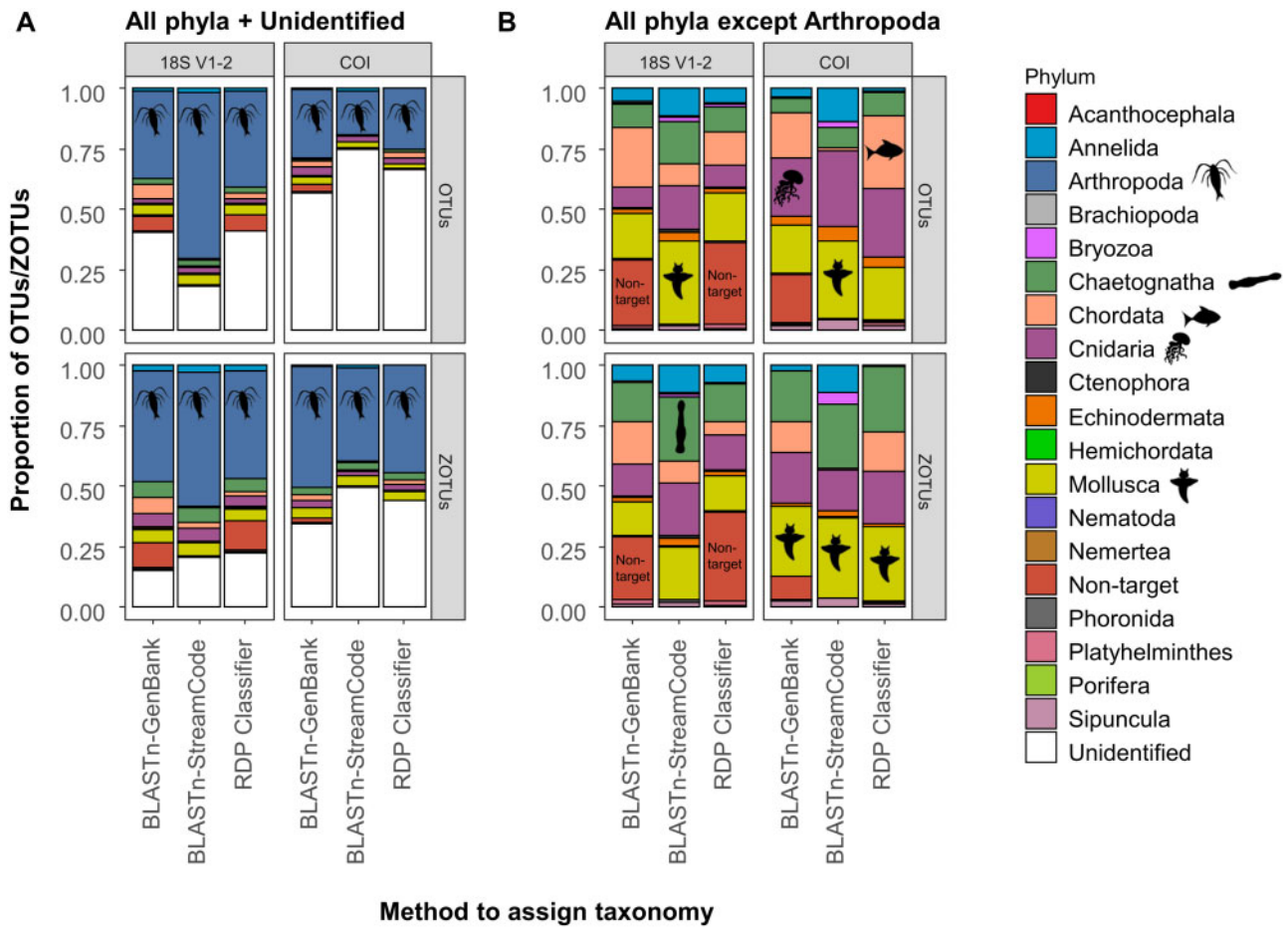

Figure 3. Variability in the proportion of OTUs/ZOTUs belonging to each phylum identified using different genetic markers (COI and 18S V1–2) and different approaches for taxonomic assignment (BLASTn-GenBank, BLASTn-StreamCode, RDP Classifier). (a) All samples (unidentified samples and samples identified to phylum). (b) Only samples identified to phylum without including the phylum Arthropoda. The RDP Classifier was used with the PR2 database for 18S and MIDORI database for COI. "Non-target" refers to taxa identified to phyla that do not belong to the target zooplankton groups, "Unidentified" are the OTUs/ZOTUs without a phylum assignment at the defined taxonomic thresholds. The more abundant taxa are highlighted with black silhouettes.

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

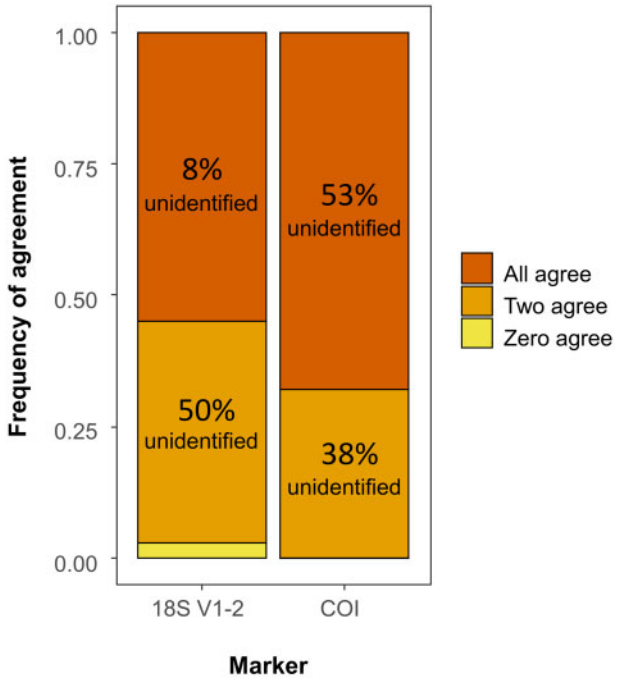

Figure 4. Frequency of agreement in the identification to phylum between the different methods for taxonomic assignment for each of the two genetic markers (18S V1–2 and COI). Colours separate the cases in which all methods agree, two methods agree, and zero methods agree. For the cases in which two or three methods agree, the percentage of agreement in the failure to identify a phylum is presented within the bar.

(Figure 5). Although found at a comparatively smaller proportion, a meroplankton component was detected across all plankton samples, with a larger proportion detected using morphology and the BLASTn-StreamCode method in the metabarcoding samples. For this analysis, we used the classification to lower taxonomic levels when available; however, many of the metabarcoding samples lacked enough resolution to classify taxa into holoplankton or meroplankton (coded N/A in Figure 5). For each combination of clustering and genetic marker, the BLASTn-StreamCode method produced the smallest number of taxa unassigned to plankton type (Figure 5). In Supplementary Figure S4, we detail the phyla composition for holoplankton and meroplankton identified using all methods.

Discussion

Biodiversity assessments performed with metabarcoding methods indicate astonishing levels of diversity in biological communities. Despite the promising potential of metabarcoding methods for standardized biodiversity surveys (Baird and Hajibabaei, 2012; Deiner *et al.*, 2017), our study highlights that even identification to phylum is not straightforward for understudied ecosystems with poor representation in DNA reference databases. Identifying metabarcoding samples at taxonomic levels below phylum is critical for ecological studies (e.g. biogeography, connectivity, functional groups, biological indicators, parasitism, invasive species), yet, our case study demonstrates how many metabarcoding samples cannot even be identified to phylum or plankton type. Because bias in taxonomic identification with metabarcoding propagates to downstream analyses, it is important to minimize errors in taxonomic assignment (Santoferrara, 2019). To minimize errors and improve taxonomic resolution, the implementation of metabarcoding methods and compilation of reference databases should be associated with taxonomic experts (Clarke *et al.*, 2017; Porter and Hajibabaei, 2018). By integrating traditional taxonomic assessment, barcoding, and metabarcoding, we have highlighted how a targeted effort to develop a regional DNA barcode database can help improve metabarcoding taxonomic assignments. Moreover, we showed that the chances of correctly identifying taxa belonging to different marine zooplankton phyla varied by genetic marker, the type of clustering, and the approach used for taxonomic assignment.

During the short timeframe of the StreamCode project (one field season, two collection weeks), a small collaboration of taxonomic experts made a large contribution of barcode quality sequences to GenBank (1108 for 18S and 1042 for COI belonging to 15 phyla) as well as images and voucher information (Supplementary Table S1). Hopefully, this encourages similar work that targets taxa underrepresented in reference databases. Local bioblitz efforts (Plumb, 2014, where there is intense biodiversity sampling in a small region carried out by experts and volunteers) coupled with barcoding efforts are becoming popular because they are essential for enhancing reference databases. Large international efforts such as the Marine Barcode of Life project have contributed voucher information and barcode sequences to the Barcode Of Life (BOLD) system (Puillandre *et al.*, 2012). Barcoding efforts led by non-taxonomists would benefit from including taxonomic experts at the beginning of the project, incorporating their expertise for sampling design and collection of samples (DeWalt, 2011), and also for subsequent quality control checks and determination of unidentified taxa. Both targeted efforts and large collaborations are critical to improving the quality and taxonomic coverage of reference DNA databases, which in turn will make metabarcoding a more efficient tool for characterizing communities captured in environmental samples.

Developing local DNA sequence libraries helps to overcome problems associated with incomplete publically available DNA reference databases (and with incorrect entries), one of the current limitations of resolving taxonomic assignment with metabarcoding (Porter and Hajibabaei, 2018). Every new sequence counts because the addition of even a few species to a local reference database can improve metabarcoding taxonomic assignments substantially (Abad *et al.*, 2016); so when a large local barcode database is used, the improvement in taxonomic identification can be immense (e.g. up to 7-fold increase in Ransome *et al.*, 2017). In our study, by using the StreamCode DNA barcode database (BLASTn-StreamCode method), we identified more unique taxa to phyla that are likely less represented in reference databases (e.g. Annelida, Bryozoa, Chaetognatha; the first two also identified by Ransome *et al.*, 2017 as under-represented taxa). Working toward a comprehensive genetic inventory of marine life will not only reduce the number of unidentified sequences in metabarcode analyses, but it will also help focus taxonomic efforts by revealing undescribed species in poorly studied groups.

To avoid the problem of inadequately identified records (e.g. "environmental sample" or "uncultured metazoan"), researchers that apply metabarcoding methods can opt for curated databases that include only sequences that have passed a quality check and have complete taxonomic information. In some cases, curated databases also require sequences to be associated with voucher specimens (e.g. BOLD, http://barcodinglife.org/), which allows

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

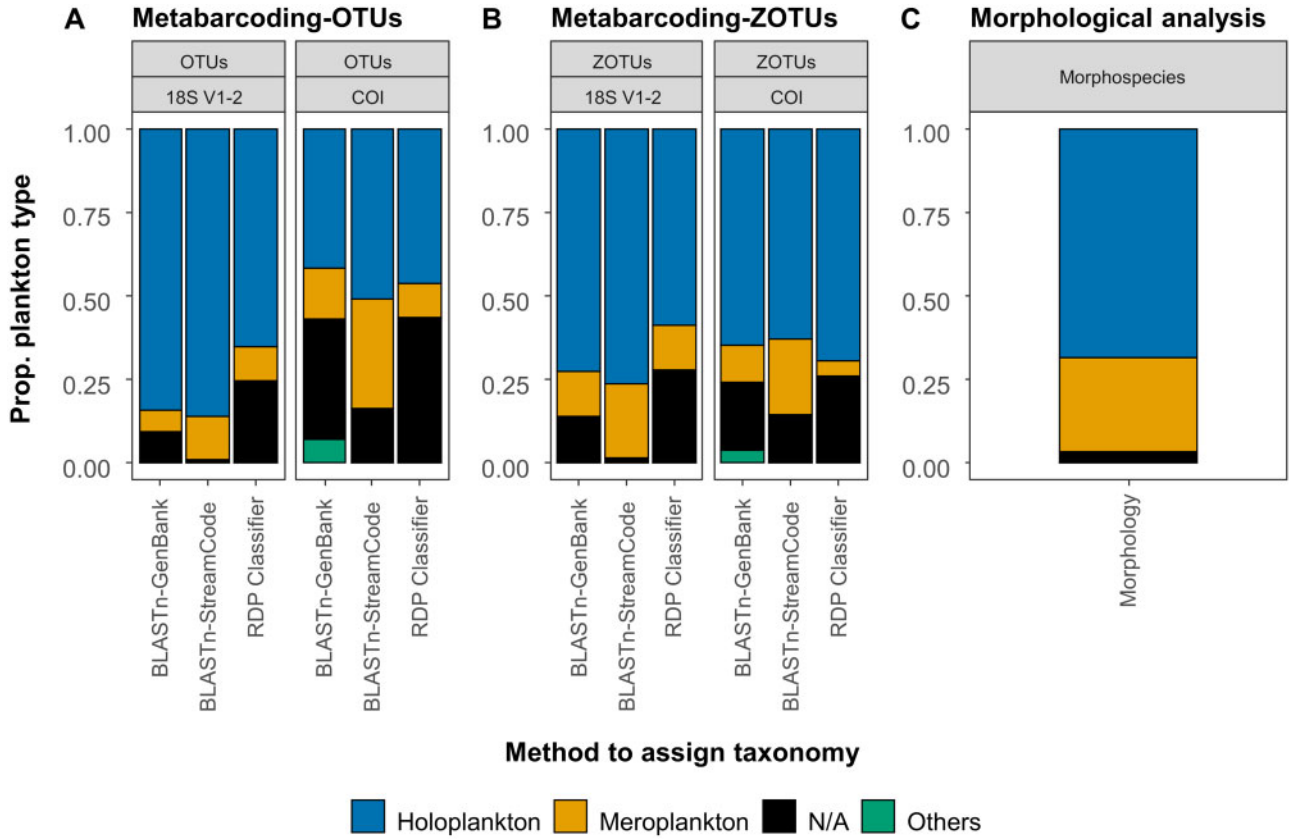

Figure 5. Proportion of taxa identified as members of the holoplankton, meroplankton, or others (mixed life histories within a group, or a benthic group) for the metabarcoding results with two genetic markers (COI, 18S V1–2) and three methods for taxonomic assignment, combined with the results from the morphological analysis. (a) Metabarcoding-OTUs, (b) metabarcoding-ZOTUs, and (c) morphological analysis. N/A indicates samples that lacked enough taxonomic resolution to be classified into holoplankton or meroplankton.

updates to identifications as taxonomic information continues to grow. Even though curated databases are not comprehensive, they are continuously improving and there are many efforts underway to develop curated databases for different taxonomic groups and genetic markers (EukRef, http://eukref.org/; MetaZooGene, https://metazoogene.org/; PR2, https://pr2-database.org/). Despite the higher reliability of data from curated databases, for a coarse identification of marine zooplankton taxa to phylum, we showed that the BLASTn-GenBank approach outperformed PR2 and Midori databases in assigning a phylum to most OTUs. This probably happens because PR2 and Midori: (i) lag behind GenBank in incorporating new sequences; (ii) include only sequences identified to genus or species level. Finally, the BLASTn-Streamcode method using our local database outperformed GenBank in some groups, likely because of our more complete coverage of relevant taxa (e.g. Annelida: Polychaeta). Additionally, we have observed that the StreamCode DNA barcode database improves resolution at lower taxonomic levels as well (Pappalardo *et al.*, 2020; ESA Abstract).

Importantly, given that we detected some phyla with only one genetic marker, we support the idea that a multi-marker approach can be more effective in studies with a broader scope (Cowart *et al.*, 2015; Bucklin *et al.*, 2016; Leray and Knowlton, 2016; Djurhuus *et al.*, 2018). Both when we tested the StreamCode barcodes (Table 2) and when we analysed the metabarcoding data (Table 3), we found a smaller proportion of unidentified sequences for 18S V1–2 than for COI, regardless of the method used for a taxonomic assignment. Other studies using COI have also found a large number of unidentified samples (Leray and Knowlton, 2015; Ransome *et al.*, 2017). To better reflect marine biodiversity, some authors recommend a first step that involves a conserved marker useful for coarse taxonomic assignments (e.g. 18S V1–2 or V9), and a second step that targets a highly variable marker (e.g. COI) to provide finer taxonomic resolution (Leray and Knowlton, 2016). In our study, there were also differences between markers in the relative taxonomic composition detected (as also observed by Cowart *et al.*, 2015; Djurhuus *et al.*, 2018), and in the proportion of taxa that we were able to classify into plankton type. Both of these sources of variation could have large impacts on the interpretation of community comparisons in ecological studies. Regardless of whether different genetic markers can vary in their specificity to detect some taxa, or the ability to identify samples with different taxonomic resolution, multiple markers may still reflect similar trends in spatial diversity (Pitz *et al.*, 2020).

An important consideration when using a multi-marker approach is to apply a common classification scheme across markers and DNA reference databases. In our marine zooplankton-focused analyses, we standardized the classification scheme used by Midori (based on NCBI) and the PR2 database (custom classification defined by experts) to that in WoRMS. Even though some aspects of the current WoRMS classification are still under debate

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

(e.g. placement of Sipuncula as a phylum), their large editorial group reviews newer publications and updates their classification when there is enough support. WoRMS also provides a useful tool for matching taxon names (https://www.marinespecies.org/aphia.php?p=match) and overall, is the most current and authoritative resource for the classification of marine taxa. It would be ideal if genetic data resources (e.g. NCBI, BoLD) could adopt a global taxonomic system (such as the Catalogue of Life, https://www.catalogueoflife.org/), provided that the classification hierarchy is maintained for all groups with the best data sources. For example, even though the Catalogue of Life follows WoRMS for marine organisms, their online website was not updated with the current WoRMS edition at the time of submitting this article (December, 2020). In the meantime, it falls on researchers to address taxonomic standards in multi-marker studies. Individual researchers and research groups could still adopt their own taxonomies, even rankless ones (given the arbitrary nature of Linnean ranks) such as PhyloCode (i.e. PhyloCode, http://phylonames.org/code/).

Our data also indicate that exploring different clustering methods can improve taxon detection. When a phylum was detected only in some of the combinations of clustering, genetic marker, and method for taxonomic assignment (e.g. Acanthocephala, Porifera, Hemichordata, see Supplementary Table S6), it was usually represented by a few (sometimes only one) OTUs or ZOTUs. We found two possible explanations for these differences in phylum detection. Differences could appear because of the different filtering algorithms used to generate OTUs and ZOTUs. For example, Schenk *et al.* (2020) found differences between OTUs and ASVs (=ZOTUs) in metabarcoding of freshwater nematode communities and suggested as an explanation the more stringent filtering in the pipeline prior to generating ASVs. In addition, many of the taxa found using only OTUs by Schenk *et al.* (2020) were detected with a small percentage of sequence reads. For our data, another explanation relates to the pre-defined confidence thresholds to assign a phylum. We noticed cases in which for the raw data a phylum was assigned for both OTUs and ZOTUs, but with different confidence thresholds, and only one of the clustering schemes passed the pre-defined taxonomic thresholds. Singletons are commonly filtered out in metabarcoding studies given the likelihood that they represent artefacts, but some authors argue that they could be important for the detection of rare species (Brown *et al.*, 2015). Future research using mock communities could contribute to our understanding of these topics; in the meantime, we suggest implementing different clustering approaches if taxon identification is important.

Recommended confidence thresholds for accurate taxonomic assignment depend on the completeness of the reference database for the focal taxon (Porter and Hajibabaei, 2018). We followed recommended guidelines to assign taxa to phylum (e.g. Ransome *et al.*, 2017; Leray *et al.*, 2018). However, there are no agreed-upon criteria at lower taxonomic levels. For example, for metabarcoding of 18S using the CREST LCAClassifier algorithm, Lanzén *et al.* (2012) used minimum similarities between related taxa and cross-validation from reference datasets to propose specific thresholds for each taxonomic level (phylum: 80%, class: 85%, order: 90%, family: 95%, genus: 97%, and species: 99%); whereas Leasi *et al.* (2018) used BLAST when analysing the 18S V9 region and chose different thresholds based on the literature and analysis of mock communities (phylum: 90%, family: 93%, species: >97%). For metabarcoding of COI, Elbrecht *et al.* (2017) mentioned specific thresholds for different taxonomic levels as a "rough proxy" (order: 85%, family: 90%, genus: 95%, and species: 98%), but did not specify how the thresholds were determined. In metabarcoding samples containing taxa from multiple phyla, uniform thresholds across different phyla may not be advisable, due to different evolutionary rates in different taxonomic groups. One example of this variation in COI can be seen in the range of confidence thresholds needed by RDP Classifier to obtain a correct assignment for each taxonomic level in different phyla of marine organisms (results from Midori "leave-one-out-test", available in http://reference-midori.info/download.php#). More research is needed on how to define confidence thresholds for assignment to lower taxonomic levels when using metabarcoding methods, and how those decisions affect results.

Increasingly, the computational methods being developed for analysing metabarcode data focus on "micro-scale" variation, such as strain-level variation (e.g. UNOISE, DADA2). These algorithms were developed with the primary goal of examining bacterial diversity and community structure, as bacteria are not traditionally categorized as species. In fact, many microbiologists explore strain-level variation in bacteria, as it can be used to elucidate differences in virulence, growth patterns, etc. In contrast, researchers examining metazoan communities to assess community diversity, identify potential invasive species, and examine trophic-level interactions, are generally more interested in examining inter-specific diversity. Thus, extending these "micro-scale" analytical tools to examination of metazoan sequences for inter-specific diversity, particularly without somewhat arbitrarily assigning a species-level sequence threshold for all taxa, can require substantial additional work. On the simpler end of the spectrum, this includes building phylogenetic trees to confirm "species-level" clustering, and on the other end involves incorporating sophisticated species delimitation methods (e.g. multi-rate Poisson tree processes, Kapli *et al.*, 2017). Development of analytical tools that can be used to explore metazoan diversity, particularly taking into account sequence variation across taxonomic groups, is needed.

Conclusion

If identification of different taxa is critical for a metabarcoding study, we recommend: (i) using multiple genetic markers, (ii) implementing multiple methods for a taxonomic assignment, (iii) clustering the data into different types of molecular operational units, and more importantly, (iv) collaborating with taxonomists to develop a regional database of the groups of interest, especially if they are underrepresented in reference databases. This multi-analysis approach can enhance taxon detection and increase confidence in the results. There are already many such collaborations underway that are collectively populating reference databases and will improve the performance of taxonomic assignment in biodiversity surveys using metabarcoding. In addition, new tools and pipelines are continuously being developed, many of them in open-access platforms. To improve taxonomic assignments at lower taxonomic levels (species, genus, family), we think that future research should aim to develop taxon-specific thresholds for different genetic markers, to account for the different evolutionary rates in different taxonomic groups.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

Funding

This work was funded through a Smithsonian National Museum of Natural History (NMNH) Associate Director for Science Core Grant (2017), with additional support from a Smithsonian Global Genome Initiative grant (GGI-Rolling-2017-109a). We also wish to acknowledge funding and technical support from the Smithsonian Institution Barcode Network (SIBN, FY2017 Award cycle). Katrina M. Pagenkopp Lohan was funded as a Robert and Arlene Kogod Secretarial Scholar.

Acknowledgements

Genetic benchwork and sequencing was completed at the Smithsonian NMNH Laboratories of Analytical Biology (LAB). We thank NMNH specialists Rebecca Dikow, Mike Trizna, and Matthew Kweskin for their assistance with the computations using the Smithsonian High Performance Cluster (SI/HPC, Smithsonian Institution, https://doi.org/10.25572/SIHPC), and Amanda Devine for assistance using the GGI Gap Analysis Tool. We also thank Museum Specialist Abigail Reft for her assistance with the cephalopods data, and Julia Steier for her helpful editorial review on the final version. We thank the crew of the M/V *Thunderforce* for technical support during zooplankton collections. Additionally, we are grateful to Smithsonian Marine Station laboratory and field support staff, Sherry Reed, Woody Lee, David Branson, and Scott Jones for their assistance, to NMNH administrative support Carol Youmans, Marisol Arciniega-Melendez, and Joan Kaminski, and to Lisa Comer and Mark Lehtonen for collections management support. This publication is Smithsonian Marine Station contribution no. 1156.

Author's contributions

MJB and KJO conceived the StreamCode project idea, secured the initial funding and led data collection in the field. KJO developed the workflows and record keeping for the fieldwork and secured additional funding. All authors excluding AB, KM, NER, PP, and KMLP, participated in the fieldwork and morphological classification. AGC, KMH, KJO, MJB, MV, PP, and WJ, performed quality controls on refined IDs. SBT generated the 18S V1–2 barcodes, KM and NER generated the COI barcodes. MJB processed all samples for metabarcoding. KMLP developed and implemented the bioinformatics pipeline, KMLP and PP worked on taxonomic assignments for sequence data. AB contributed with the gap analysis and metadata of StreamCode barcodes. PP designed and conducted data analyses and wrote the first draft. KJO, KMH, KMLP, and AGC, contributed critically to the data analysis and interpretation of results. AGC, EES, JAG, KJO, KMH, KMLP, MJH, MJB, MV, NER, SLB, and WJ provided feedback on the manuscript. All authors approved the final manuscript.

Data availability

The StreamCode DNA barcode sequences underlying this article are available in the GenBank Nucleotide Database at https://www.ncbi.nlm.nih.gov/genbank/, and can be accessed with the accession numbers provided in Supplementary Table S1; sequences are also associated to NCBI BioProject PRJNA421480. The StreamCode metabarcoding raw data is available through the NCBI Sequence Read Archive and included in the NCBI BioProject PRJNA421480. The raw data includes the V9 region of 18S that was not part of this analysis. The R code used for the analyses, and expanded StreaCode data including the sampling locations is available in the Dryad Digital Repository at DOI https://doi.org/10.5061/dryad.tdz08kpzx.

References

Abad, D., Albaina, A., Aguirre, M., Laza-Martínez, A., Uriarte, I., Iriarte, A., Villate, F., *et al.* 2016. Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. Marine Biology, 163: 149.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. 1990. Basic local alignment search tool. Journal of Molecular Biology, 215: 403–410.

Aylagas, E., Borja, Á., Irigoien, X., and Rodríguez-Ezpeleta, N. 2016. Benchmarking DNA metabarcoding for biodiversity-based monitoring and assessment. Frontiers in Marine Science, 3: http://journal.frontiersin.org/Article/10.3389/fmars.2016.00096/abstract (last accessed 11 January 2020).

Baird, D. J., and Hajibabaei, M. 2012. Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. Molecular Ecology, 21: 2039–2044.

Bazinet, A. L., and Cummings, M. P. 2012. A comparative evaluation of sequence classification programs. BMC Bioinformatics, 13: 92.

Bhadury, P., Austen, M., Bilton, D., Lambshead, P., Rogers, A., and Smerdon, G. 2006. Development and evaluation of a DNA-barcoding approach for the rapid identification of nematodes. Marine Ecology Progress Series, 320: 1–9.

Brown, E. A., Chain, F. J. J., Crease, T. J., MacIsaac, H. J., and Cristescu, M. E. 2015. Divergence thresholds and divergent biodiversity estimates: can metabarcoding reliably describe zooplankton communities? Ecology and Evolution, 5: 2234–2251.

Bucklin, A., Lindeque, P. K., Rodriguez-Ezpeleta, N., Albaina, A., and Lehtiniemi, M. 2016. Metabarcoding of marine zooplankton: prospects, progress and pitfalls. Journal of Plankton Research, 38: 393–400.

Callahan, B. J., McMurdie, P. J., and Holmes, S. P. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. The ISME Journal, 11: 2639–2643.

Caporaso, J. G., Kuczynski, J., and Knight, R. 2010. QIIME allows analysis of high-throughput community sequencing data. Nature Methods, 7: 335–336.

Carew, M. E., Kellar, C. R., Pettigrove, V. J., and Hoffmann, A. A. 2018. Can high-throughput sequencing detect macroinvertebrate diversity for routine monitoring of an urban river? Ecological Indicators, 85: 440–450.

Clarke, L. J., Beard, J. M., Swadling, K. M., and Deagle, B. E. 2017. Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. Ecology and Evolution, 7: 873–883.

Cowart, D. A., Pinheiro, M., Mouchel, O., Maguer, M., Grall, J., Miné, J., and Arnaud-Haond, S. 2015. Metabarcoding is powerful yet still blind: a comparative analysis of morphological and molecular surveys of seagrass communities. PLoS One, 10: e0117562.

Cristescu, M. E. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends in Ecology & Evolution, 29: 566–571.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., *et al.* 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Molecular Ecology, 26: 5872–5895.

DeWalt, R. E. 2011. DNA barcoding: a taxonomic point of view. Journal of the North American Benthological Society, 30: 174–181.

Djurhuus, A., Pitz, K., Sawaya, N. A., Rojas-Márquez, J., Michaud, B., Montes, E., Muller-Karger, F., *et al.* 2018. Evaluation of marine zooplankton community structure through environmental DNA

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021

metabarcoding: metabarcoding zooplankton from eDNA. Limnology and Oceanography: Methods, 16: 209–221.

Edgar, R. C. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Nature Methods, 10: 996–998.

Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon sequencing. bioRxiv, 081257. https://www.semanticscholar.org/paper/UNOISE2%3A-improved-error-correction-for-Illumina-16S-Edgar/a9867e89687a6fb581e664c8880e7438af8c8b5a

Edgar, R. C. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. Bioinformatics, 34: 2371–2375.

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., and Leese, F. 2017. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods in Ecology and Evolution, 8: 1265–1275.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., *et al.* 2012. The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research, 41: D597–D604.

Horton, T., Kroh, A., Ahyong, S., Bailly, N., Boyko, C. B., Brandão, S. N., Gofas, S., *et al.* 2020. World Register of Marine Species (WoRMS). WoRMS Editorial Board. https://www.marinespecies.org.

Kapli, P., Lutteropp, S., Zhang, J., Kobert, K., Pavlidis, P., Stamatakis, A., and Flouri, T. 2017. Multi-rate Poisson tree processes for single--locus species delimitation under maximum likelihood and Markov chain Monte Carlo. Journal of Plankton Research, 41: 571–582.

Lanzén, A., Jørgensen, S. L., Huson, D. H., Gorfer, M., Grindhaug, S. H., Jonassen, I., Øvreås, L., *et al.* 2012. CREST – classification resources for environmental sequence tags. PLoS One, 7: e49334.

Leasi, F., Sevigny, J. L., Laflamme, E. M., Artois, T., Curini-Galletti, M., de Jesus Navarrete, A., Di Domenico, M., *et al.* 2018. Biodiversity estimates and ecological interpretations of meiofaunal communities are biased by the taxonomic approach. Communications Biology, 1: 112.

Lejzerowicz, F., Esling, P., Pillet, L., Wilding, T. A., Black, K. D., and Pawlowski, J. 2015. High-throughput sequencing and morphology perform equally well for benthic monitoring of marine ecosystems. Scientific Reports, 5: 13932.

Leray, M., and Knowlton, N. 2015. DNA barcoding and metabarcoding of standardized samples reveal patterns of marine benthic diversity. Proceedings of the National Academy of Sciences, 112: 2076–2081.

Leray, M., and Knowlton, N. 2016. Censusing marine eukaryotic diversity in the twenty-first century. Philosophical Transactions of the Royal Society B: Biological Sciences, 371: 20150331.

Leray, M., Ho, S.-L., Lin, I.-J., and Machida, R. J. 2018. MIDORI server: a webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. Bioinformatics, 34: 3753–3754.

Lindeque, P. K., Parry, H. E., Harmer, R. A., Somerfield, P. J., and Atkinson, A. 2013. Next generation sequencing reveals the hidden diversity of zooplankton assemblages. PLoS One, 8: e81327.

Lobo, J., Shokralla, S., Costa, M. H., Hajibabaei, M., and Costa, F. O. 2017. DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. Scientific Reports, 7: 15618.

Machida, R. J., Leray, M., Ho, S.-L., and Knowlton, N. 2017. Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. Scientific Data, 4: 170027.

Pagenkopp Lohan, K., Campbell, T. L., Guo, J., Wheelock, M., DiMaria, R. A., and Geller, J. B. 2019. Intact vs. homogenized subsampling: testing impacts of pre-extraction processing of multi-species samples on invasive species detection. Management of Biological Invasions, 10: 324–341.

Pagenkopp Lohan, K. M., Fleischer, R. C., Carney, K. J., Holzer, K. K., and Ruiz, G. M. 2016. Amplicon-based pyrosequencing reveals high diversity of protistan parasites in ships' ballast water: implications for biogeography and infectious diseases. Microbial Ecology, 71: 530–542.

Pappalardo, P., Pagenkopp Lohan, K. M., Boyle, M. J., Collins, A. G., Hanson, K. M., Truskey, S. B., Jaekle, W., *et al.* 2020. Improving taxonomic assignment of DNA metabarcoding with taxonomic expertise. Ecological Society of America Annual Meeting, Aug 3-6. https://eco.confex.com/eco/2020/meetingapp.cgi/Paper/87812.

Parry, L. A., Edgecombe, G. D., Eibye-Jacobsen, D., and Vinther, J. 2016. The impact of fossil data on annelid phylogeny inferred from discrete morphological characters. Proceedings of the Royal Society B. Biological Sciences, 283: 20161378.

Pitz, K. J., Guo, J., Johnson, S. B., Campbell, T. L., Zhang, H., Vrijenhoek, R. C., Chavez, F. P., *et al.* 2020. Zooplankton biogeographic boundaries in the California Current System as determined from metabarcoding. PLoS One, 15: e0235159.

Porter, T. M., and Hajibabaei, M. 2018. Scaling up: a guide to high-throughput genomic approaches for biodiversity analysis. Molecular Ecology, 27: 313–338.

Puillandre, N., Bouchet, P., Boisselier-Dubayle, M.-C., Brisset, J., Buge, B., Castelin, M., Chagnoux, S., *et al.* 2012. New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Molecular Ecology Resources, 12: 396–402.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., *et al.* 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research, 41: D590–D596.

R Core Team. 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ransome, E., Geller, J. B., Timmers, M., Leray, M., Mahardini, A., Sembiring, A., Collins, A. G., *et al.* 2017. The importance of standardization for biodiversity comparisons: a case study using autonomous reef monitoring structures (ARMS) and metabarcoding to measure cryptic diversity on Mo'orea coral reefs, French Polynesia. PLoS One, 12: e0175066.

Ratnasingham, S., and Hebert, P. D. N. 2007. BOLD: the barcode of life data system (http://www.barcodinglife.org). Molecular Ecology Notes, 7: 355–364.

Richardson, R. T., Bengtsson-Palme, J., and Johnson, R. M. 2017. Evaluating and optimizing the performance of software commonly used for the taxonomic classification of DNA metabarcoding sequence data. Molecular Ecology Resources, 17: 760–769.

Santoferrara, L. F. 2019. Current practice in plankton metabarcoding: optimization and error management. Journal of Plankton Research, 41: 571–582.

Schenk, J., Kleinbölting, N., and Traunspurger, W. 2020. Comparison of morphological, DNA barcoding, and metabarcoding characterizations of freshwater nematode communities. Ecology and Evolution, 10: 2885–2899.

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., *et al.* 2011. Phylogenomic analyses unravel annelid evolution. Nature, 471: 95–98.

Tang, C. Q., Leasi, F., Obertegger, U., Kieneke, A., Barraclough, T. G., and Fontaneto, D. 2012. The widely used small subunit 18S rDNA molecule greatly underestimates true diversity in biodiversity surveys of the meiofauna. Proceedings of the National Academy of Sciences, 109: 16208–16212.

Wang, Q., Garrity, G. M., Tiedje, J. M., and Cole, J. R. 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology, 73: 5261–5267.

Handling editor: William Grant

Downloaded from https://academic.oup.com/icesjms/advance-article/doi/10.1093/icesjms/fsab082/6277123 by University of Connecticut user on 08 July 2021