Supporting Information for: “The role of taxonomic expertise in interpretation of metabarcoding studies”

Pappalardo et al., ICES Journal of Marine Science

Overview

We provide additional details on methods and expanded results. In the companion Dryad Data Package https://doi.org/10.5061/dryad.tdz08kpzx (https://doi.org/10.5061/dryad.tdz08kpzx) you can find this Supplementary Material as an HTML file *which allows for the feature of interactive tables*. In the HTML version, Table S1 and Table S6 can be scrolled up and down and left to right, displaying all the data. In this pdf, Table S1 and Table S6 are only partially displayed; both tables are also available in the “Results” folder included in the data package. In addition, the data package contains data and code used for the analysis.

Sections

1. Additional Information on methods

```
1.1 Information on StreamCode samples

    * Table S1: Information on StreamCode samples

1.2 DNA Barcoding methods

    * Table S2: Primers used for DNA barcoding

1.3 DNA Metabarcoding methods

    * Table S3: Primers used for DNA metabarcoding

1.4 Bioinformatics pipeline

    * Table S4: Number of metabarcode sequences in each filtering step

    * Table S5: Number of metabarcode sequences by phylum with two types of OTU filtering
```

2. Additional Results

3. Additional References

1. Additional information on Methods

1.1 Information on StreamCode project samples

Field sampling. The contents within the codend of each plankton tow were placed into a rectangular plastic bin of seawater on board the research vessel. The contents were then stirred, to make sure the tow sample was well mixed, and approximately one quarter of those contents were collected from the bin with a plastic scoop, representing a subsample of the plankton tow for metabarcoding analysis. The scooped portion was sieved to concentrate it, and rinsed with prechilled 95% ethanol into a 50 ml polypropylene falcon tube. The tube was stored on dry ice for transport to the Smithsonian Marine Station at Fort Pierce and stored at -80˚C until processing. The other portion of each tow was diluted in 10L buckets with seawater, kept aerated and chilled for their transport to the marine station and live sorting of the individuals. This later portion was used for the morphological analysis and to separate individuals for barcoding.

Laboratory analysis. The sample used for barcoding and morphological analysis was subsampled in various ways to maximize diversity of plankton examined (skimmed from the top, picked rare specimens, subsampled the settled, middle and well-mixed portions). The objective was to sample as much of the diversity as possible. When it was not feasible to process all specimens from a particular group, e.g. copepods, ostracods, each taxonomic group expert focused on taxa that were less likely to be represented in GenBank. For example, copepods are very abundant and tend to be the most represented group in plankton samples; our taxonomic expert spend a bit more time focusing on a few families of small-bodied copepods found in tropical areas (Corycaeidae and Oncaeidae). For other groups like pteropods and polychaetes, it was possible to capture most of their diversity by selective sampling.

Live specimens were grouped by morphotype and classified to the lowest taxonomic level possible at the SMSFP-NMNH. Live specimens were photographed in the laboratory using a vertically mounted Canon EOS 5DS R with either an MP-E 65 mm f2.8 1-5x or EF 100 mm f2.8 macro lens and two strobes or a Jenoptik PROKYON camera and Zeiss Stemi 508 stereo microscope imaging system. The initial taxonomic classification was later double-checked by examining specimens’ high-resolution images. In this process, samples that were identified as

belonging to the same population were linked by their voucher numbers. The images can be accessed via the SI-NMNH Invertebrate Zoology Collections Database at https://collections.nmnh.si.edu/search/iz/ (https://collections.nmnh.si.edu/search/iz/). Each sample has a unique identifier in the SI-NMNH system (USNM column in Table S1).

Final taxonomic assignment. To verify and refine the taxonomic identifications of samples with COI and 18S V1-2 sequences available, we complemented the morphological assessment with a BLAST search and we constructed phylogenetic trees to analyze the placement of each taxa:

- BLAST search: BLASTn searches were run using the NCBI nucleotide (nt) database with the commands **blastn -task blastn** and keeping default values for other settings. The first ten entries were manually checked for taxonomic consistency; if there was no clear taxonomic match (i.e., “environmental sample” or “uncultured eukaryote”), the search was run again from the NCBI portal (https://blast.ncbi.nlm.nih.gov/Blast.cgi (https://blast.ncbi.nlm.nih.gov/Blast.cgi)) excluding environmental samples from the search.
- Phylogenetic trees: We generated trees using default options in the PhyML or RaxML plug-ins in Geneious (Geneious Prime® 2019.0.4) or in the Smithsonian Institution’s High Performance Computing Cluster. When an ID could be improved based on genetic similarity, the experts looked at the origin of the matching sequences and all the information available for that group to make the final judgment on lowest taxonomic placement for each sample. In cases where identification to a full binomial name was not possible and the uncorrected COI difference exceeded 3%, we assigned unique clade names (letters).

Because the experts focused on many groups that are understudied, not all the samples were resolved to species level. As our gap analysis showed, even for samples in which the refined IDs were at higher taxonomic levels, we found novel contributions to GenBank. The StreamCode samples that were not identified to the species level may continue to be studied more in the future, and if more specific identifications are made, we will update the taxonomic information in GenBank.

Table S1: Information on StreamCode samples. Zooplankton samples were collected from the Gulf stream off Fort Pierce. Two locations were visited in June, 2017 and two in August, 2017. USNM numbers identify the morphological vouchers for the specimens deposited at the National Museum of Natural History. For each sample, we marked with an “X” when an extract voucher, a morphology voucher, or a photo were available. The refined ID is the current taxonomic assignment for each sample based on expert assessment. The Taxonomy column includes Phylum;Class;Order;Family;Genus (when applicable) for each refined ID following the WoRMS taxonomic hierarchy (http://www.marinespecies.org/index.php (http://www.marinespecies.org/index.php)). If any of these samples are identified to lower taxonomic levels in the future, we will update their classification in GenBank. The original dataset (and corresponding metadata) is available as the *StreamCode_data.csv* file in the Dryad data repository associated to this publication, and has additional information on location, collection method, and depth. This full table is also avi

USNM	target.group	refinedID	finalID.WoRMS	Taxonomy	extr
1449980	mollusks_cephalopods	Abralia veranyi	Abralia (Asteroteuthis) veranyi	Mollusca ; Cephalopoda ; Oegopsida ; Enoploteuthidae ; Abralia	X
1447996	mollusks_cephalopods	Abralia veranyi	Abralia	Mollusca ;	X

USNM	target.group	refinedID	finalID.WoRMS	Taxonomy	extr
			verany)	Oegopsida ; Enoploteuthidae ; Abralia	
1449981	mollusks_cephalopods	Abraliopsis atlantica	Abraliopsis (Pfefferiteuthis) atlantica	Mollusca ; Cephalopoda ; Oegopsida ; Enoploteuthidae ; Abraliopsis	X
1450505	cnidarians_hydrozoans	Abylopsis eschscholtzii	Abylopsis eschscholtzii	Cnidaria ; Hydrozoa ; Siphonophorae ; Abylidae ; Abylopsis	X
1450513	cnidarians_hydrozoans	Abylopsis eschscholtzii	Abylopsis eschscholtzii	Cnidaria ; Hydrozoa ; Siphonophorae ; Abylidae ; Abylopsis	X
1448253	cnidarians_hydrozoans	Abylopsis tetragona	Abylopsis tetragona	Cnidaria ; Hydrozoa ; Siphonophorae ; Abylidae ; Abylopsis	X
1449939	arthropods_misc	Acanthephyridae	Acanthephyridae	Arthropoda ; Malacostraca ; Decapoda ; Acanthephyridae ;	X

The number of samples collected was 1529, the total number of specimens was 2260, the number of tissue/extracts samples available is 1399, the number of photos available is 1149, and the number of morphological vouchers is 403.

1.2 DNA Barcoding methods

DNA extraction, amplification, and sequencing were performed in the Laboratories of Analytical Biology (LAB), NMNH. Total genomic DNA was extracted from tissue samples (or whole animals when the specimens were minute) using the AutoGenPrep 965 high-throughput DNA extractor (AutoGen) following the manufacturer’s instructions for animal tissue extraction. This included two ethanol wash steps and a final elution in 100 µl of AutoGen R9 reagent solution. A ~655 bp region at the 5’ end of cytochrome oxidase-c subunit I (COI) was amplified using multiple primer sets (Table 1). A ~450 bp region that included the first two variable portions near the 5’ end of the small subunit of the ribosomal gene complex (18S, regions V1-2) was amplified using primers SSU_F04 and SSU_R22 (Table S2).

Polymerase chain reaction (PCR) was carried out in 10 µl reactions for each sample. For COI, initial amplification was attempted using the primer set jgLCO1490/jgHCO2198 and GoTaq® Mastermix (2X; Promega, Madison, WI). Each 10 µl sample reaction consisted of the following: 5 µl of GoTaq® Mastermix (Promega Inc.), 3.2 µl of sterile water, 0.3 µl of each forward and reverse primers (10mM concentration), 0.1 µl of Magnesium Chloride (50mM concentration, part of the BIOLASE DNA Polymerase (BIO-21066) made by Bioline in 2018) , 0.1 µl BSA (New

England Biolabs, 20 mg/ml) and 1 µl of template DNA. The PCR thermocycling protocol consisted of the following steps: initial denaturation at 95°C for 5 min; 4 cycles of 94ºC for 30 s, 50°C for 45 s and 72°C for 60 s; then 34 cycles of 94ºC for 30 s, 45°C for 45 s and 72°C for 60 s; and the final extension at 72°C for 8 min. If amplification was not successful, sterile water was reduced to 2.2 µl and template DNA was increased to 2 µl. If amplification was still not successful, primer set dgLCO1490/dgHCO2198 was used and no MgCl was added. The PCR thermocycling protocol consisted of the following steps: initial denaturation at 95°C for 5 min; 35 cycles of 94ºC for 45 s, 50°C for 45 s and 72°C for 60 s; and the final extension at 72°C for 5 min. For PCR of18s V1-2 regions, each 10 µl sample reaction consisted of the following: 0.1 µl Taq polymerase (5u/µl, GoTaq, Promega, Inc.), 5.95 µl of sterile water, 0.3 µl of both SSU_F04 and SSU_R22 primers (10mM concentrations), 0.50 µl of dNTPs, 1 µl of 5X Buffer, 0.60 µl of 50 mM Magnesium Chloride, 0.25 µl BSA (New England Biolabs, 20 mg/ml), and 1 µl of template DNA. The PCR thermocycling protocol consisted of the following: initial denaturation at 95° for 2 min; 35 cycles of 95° for 60 s, 57° for 45 s, and 72° for 180 s; and the final extension at 72° for 10 min.

PCR products were visualized using 1.5% agarose gel electrophoresis and purified with USB ExoSAP-IT following the manufacturer’s protocol (Affymetrix, Santa Clara, CA). Purified PCR products were then used in cycle sequencing with BigDye® Terminator (Life Technologies, Carlsbad, CA) chemistry using the following thermal cycling profile: 4 min initial denaturation at 96 °C, 30 cycles of 10 s at 95 °C, 30 s at 50 °C, and 4 min at 60°C. Purification of cycle sequencing products was performed with Sephadex® G-50 gel column filtration (GE Healthcare Life Sciences, Pittsburgh, PA). Purified PCR products were sequenced using a 3730xl DNA analyzer (Applied Biosystems, Inc., Waltham, MA). Forward and reverse raw traces were assembled into contigs using the ‘De Novo Assemble’ function in Geneious v9 and v11 (Biomatters Ltd., Auckland, New Zealand) , where primers and poor quality 3’ and 5’ ends were automatically trimmed. For COI, further quality control was performed by creating multiple sequence alignments of consensus sequences at various taxonomic levels (MUSCLE alignment option and default settings in Geneious). These were used to check for length variation, presence of indel regions and/or stop codons, and variability between closely related taxa. Questionable sequences were run through the BLASTn search algorithm (Altschul et al., 1990) to check for contamination.

All 18S and COI sequences produced in the current study were uploaded to GenBank (https://www.ncbi.nlm.nih.gov/genbank/ (https://www.ncbi.nlm.nih.gov/genbank/), see Table S1 for accession numbers). Sequences can also be found in NCBI BioProject PRJNA421480 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA421480 (https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA421480)). We refer to these sequences as the StreamCode DNA Barcode Database.

Table S2: Primers used for DNA barcoding.

marker	name	direction	sequence	reference
COI	jgLCO1490	Forward	TNTCNACNAAYCAYAARGAYATTGG	Geller et al., 2013
	jgHCO2198	Reverse	TANACYTCNGGRTGNCCRAARAAYCA	Geller et al., 2013
	dgLCO1490	Forward	GGTCAACAAATCATAAAGAYATYGG	Meyer, 2013
	dgHCO2198	Reverse	TAAACTTCAGGGTGACCAAARAAYCA	Meyer, 2013
18S V1-2	18s_SSU_F04	Forward	GCTTGTCTCAAAGATTAAGCC	Blaxter et al., 1998
	18s_SSU_R22	Reverse	GCCTGCTGCCTTCCTTGGA	Blaxter et al., 1998

1.3 DNA Metabarcoding methods

Approximately 3-4 g per subsample of concentrated plankton were ground into a paste by mortar and pestle, and processed for DNA extraction and purification with a DNeasy PowerMax Soil Kit (Qiagen) following Leray and Knowlton (2015). Proteinase K (Bioline, USA) was added (0.4 mg/mL) to the PowerMax powerbead solution C1 containing ground plankton, which was incubated at 56˚C for 1.0 hour while rocking. Additional purification and clean-up steps followed standard PowerMax Soil Kit protocols. DNA extractions were quantified with a Qubit dsDNA HS Assay kit and fluorometer 2.0 (Invitrogen).

PCR amplifications were performed for each of two gene fragments (~313-bp COI, ~450-bp 18S V1-2) using two sets of four individually tagged (tailed) PCR primer pairs (tailed-mlCOIint / tailed-jgHCO; tailed-SSU_F04 / tailed-SSU_R22; Table S3) respectively. Each gene-specific fragment was amplified with triplicate PCR reactions for plankton DNA subsamples 1-6, then repeated or simultaneously run for subsamples 7-12. Each 20-µL PCR reaction mixture included 2.0 µl of Clontech 10X Advantage 2 PCR buffer, 1.0 µl each of 10 µM tailed forward and reverse primers, 1.4 µl of 10 mM dNTPs, 0.4 µl of Clontech Advantage 2 Polymerase Mix, 10 ng of purified DNA, and nuclease-free H2O. PCR cycling parameters included initial denaturation for 2.0 min at 95˚C; followed by 5 cycles of denaturation for 30 sec at 95˚C, annealing for 45 sec at 54˚C (COI) or 52˚C (18S V1-2), and extension for 45 sec at 68˚C, followed by 35 similar cycles with lower annealing at 48˚C for each primer pair, and a final extension for 5.0 min at 68˚C. Triplicate gene-specific products were pooled for each of the 12 plankton samples, purified with SPRI magnetic beads (Illumina) and quantified (Qubit, Invitrogen). To standardize potential numbers of sequence reads per sample, equimolar amounts of each PCR product were pooled into two sets of subsamples 1-6 and 7-12 for each gene fragment (2 pools x 2 genes = 4 pooled samples). Each of the four samples were then processed through end-repair, A-tailing and index-adapter ligation protocols with TruSeq DNA PCR-free technology (Illumina). The combined strategy of tailed PCR primers and indexed adapters produced a ‘hierarchical tagging approach’ (Leray & Knowlton, 2015) for multiplex sequencing of metabarcode amplicon libraries containing multiple samples with or without different gene loci. Each of the six libraries were then purified, quantified and diluted to 10ng/µl (COI and 18S V1-2 libraries were pooled) prior to submission for sequencing. Metabarcoding library preparation and construction followed protocols as per Leray (Methods in Biodiversity, Friday Harbor Laboratories, 2016; available upon request). Final quantification and library-size determination steps were performed by staff at LAB where the metabarcoding library templates were amplified and sequenced on an Illumina MiSeq platform with MiSeq Reagent Kit v3 (COI, 18S V1-2).

Table S3: Primers used for DNA metabarcoding.

marker	name	direction	sequence	reference
COI	mlCOIInt	Forward	GGWACWGGWTGAACWGTWTAYCCYCC	Leray et al., 2013
	jgHCO	Reverse	TAIACYTCIGGRTGICCRAARAAYCA	Leray et al., 2013
18S V1-2	SSU_F04	Forward	GCTTGTCTCAAAGATTAAGCC	Blaxter et al., 1998
	SSU_R22	Reverse	GCCTGCTGCCTTCCTTGGA	Blaxter et al., 1998

1.4 Bioinformatics pipeline

We merged forward and reverse reads using USEARCH v10.0 (Edgar, 2013) with a maximum of 20 (18S V1-2) and 30 (COI) differences allowed in the overlapping region (-fastq_mergepairs), because larger numbers of differences are recommended for merge areas that are >100bp. Post-merging, the allowable sequence range was 400-500bp for 18S V1-2 and 300-400bp for COI. To avoid low quality sequences, all sequences with maximum expected error rates >1 (-fastq_filter) were removed. In QIIME (Caporaso et al., 2010), primer specific barcodes were demultiplexed, allowing for 0 errors in the barcode sequence. Only sequences with zero errors in the primer (for 18S) were also used and primers were removed (-fastx_truncate) in USEARCH. For each dataset, unique sequences were identified and de-replicated (-fastx_uniques), then sequences were sorted by abundance (-sortbysize). To explore how different clustering methods affect taxa detection, we used two clustering approaches:

1. **OTUs**: cluster sequences into operational taxonomic units (OTUs) at 97% similarity (no denoising, singletons removed, chimera detection and removal).
2. **ZOTUs**: cluster sequences into zero-radius operational taxonomic units (ZOTUs, -unoise), allowing for a minimum of four copies, which is recommended for smaller datasets (Edgar, 2016), chimera detection and removal, and denoising of sequences.

Here we illustrate how many reads were kept after each step in the metabarcoding analysis pipeline. The final column is the total number of reads that passed the filtering steps and thus is the number of high quality reads that were clustered into otus and zotus for each marker.

Table S4: Number of metabarcode sequences in each filtering step

Marker	RawReads	MergedReads	PostFilter	SecondDemultiplex
COI	7,153,298	5,983,498	5,727,512	5,132,686
18S V1-2	7,577,617	5,925,813	5,596,052	5,112,366

Comparing two types of OTU filtering: To explore how the type of OTU filtering affects the number of OTUs for each phylum, we created two different files. One was generated following the best-practice guidelines for OTUs and removing singletons (the approach we followed in the main analysis); the other was generated simulating the filter used for the construction of ZOTUs that requirea a minimum of four sequences. While inspecting the data we observed different taxa identified with the two approaches. There are differences in the classes found with the two different approaches. Removing *only singletons* allows detection of some taxa that is filtered out in the *minimum of 4 sequences* approach. Some of the differences were finding Class Holothuroidea, Class Ophiuroidea, Class Mammalia (not part of our target groups), and orders Ascaridida, Aspidochirotida, Echinostomida, Euryalida, and Ophiurida, when filtering *only singletons*. Below the number of OTUs with the two types of filtering:

Table S5: Number of metabarcode sequences by phylum with two types of OTU filtering

phylum.mid	n.otus.singletons	n.otus.4minimun
Annelida	12	12
Arthropoda	1,248	753

phylum.mid	n.otus.singletons	n.otus.4minimun
Brachiopoda	1	1
Bryozoa	3	2
Chaetognatha	39	28
Chordata	127	108
Cnidaria	118	88
Echinodermata	17	9
Mollusca	90	74
Nematoda	1	
Non-target	2	1
Platyhelminthes	5	5
Unidentified	3,283	1,945

2. Additional Results

2.1 Gap Analysis detailed results

To make sure we did not overestimate our contributions because of synonyms, for each taxa we searched for the refined ID name and also the accepted name in WoRMS by October, 2020. In Fig. S1 we presented the accepted names, and Table S1 has the correspondence between refined ID and accepted names.

Annelida

	Name	COI	18S
Order	Echiuroidea	X	
Family	Alciopini	X	
	Lopadorrhynchidae	X	
	Thalassematidae	X	
	Tomopteridae	X	
Genus	Branchiomma	X	
	Lopadorrhynchus	X	
	Maupasia*	X	X
	Pelagobia	X	
	Tomopteris	X	
	Vanadis	X	
Species	Vanadis formosa	X	

Chaetognatha

	Name	COI	18S
Genus	Flaccisagitta	X	
Species	Flaccisagitta enflata	X	

Chordata

	Name	COI	18S
Order	Copelata	X	
	Salpida	X	
	Oikopleuridae	X	
	Salpidae	X	
Genus	Cyclosalpa	X	

Echinodermata

	Name	COI	18S
Order	Amphilepidida*	X	X
	Camarodonta*	X	X
	Holothuriida*	X	X
	Ophiacanthida*	X	X
Family	Mithrodiidae		X
	Oreasteridae		X
Genus	Mithrodia		X
Species	Mithrodia clavigera		X
	Tripneustes ventricosus		X

Sipuncula

	Name	COI	18S
Order	Aspidosiphonida*	X	X
	Phascolosomatida*	X	X
Family	Aspidosiphonidae	X	
	Golfingiidae	X	
	Phascolosomatidae	X	
Genus	Aspidosiphon	X	
	Phascolosoma	X	
	Siphonosoma	X	
	Xenosiphon	X	
Species	Aspidosiphon laevis	X	
	Siphonosoma vastum	X	
	Sipunculus polymyotus	X	

Arthropoda

	Name	COI	18S
Family	Dairellidae*	X	X
	Ethusidae	X	X
	Euryplacidae	X	
	Iulopididae	X	
	Lestrigonidae	X	
	Luciferidae		X
	Lycaeidae	X	
	Parascelidae	X	
	Pseudorhombilidae	X	
Genus	Acanthocarpus	X	X
	Conchoecetta		X
	Cranocephalus	X	
	Cronius		X
	Cryptosoma	X	X
	Dairella*	X	X
	Ethusa	X	
	Euconchoecia		X
	Frevillea	X	

Cnidaria

	Name	COI	18S
Order	Coronatae	X	
	Spirularia*	X	X
Family	Abylidae	X	
	Agalmatidae	X	
	Cytaeididae	X	
	Diphyidae	X	
	Geryoniidae	X	
	Haloclavidae	X	
	Nausithoidae	X	
	Physophoridae	X	
	Prayidae	X	
	Sphaerocorynidae	X	
Genus	Abylopsis	X	
	Agalma*	X	
	Amphicaryon	X	
	Athorybia	X	
	Bassia	X	

Mollusca

	Name	COI	18S
Order	Venerida*	X	X
Family	Atlantidae		X
	Cliidae	X	X
	Cymatiidae	X	X
	Hipponicidae	X	
	Litiopidae	X	
	Pleurobranchaeidae	X	X
	Pterotracheidae		X
	Tonnidae	X	
	Triphoridae		X
Genus	Abralia	X	
	Atlanta		X
	Bursa	X	
	Cerithium	X	
	Coralliophila	X	
	Cymbulia		X
	Cyphoma	X	X
	Firoloida		X
	Gutturnium	X	X
	Litiopa	X	X

	Gardinerosergia	X	X
	Halocypris		X
	Haloptilus	X	
	Hyperioides	X	
	Iulopis	X	
	Leptocotis	X	
	Lucifer		X
	Macrocypridina	X	X
	Micropanope	X	
	Mikroconchoecia		X
	Orthoconchoecia		X
	Pachos	X	
	Paraconchoecia		X
	Parascelus	X	
	Phronimopsis	X	X
	Porroecia		X
	Scina	X	
	Simorhynchotus	X	
	Thyropus	X	
	Triconia	X	
Species	Acanthocarpus alexandri	X	X
	Brachyscelus globiceps	X	
	Callinectes ornatus		X
	Candacia bipinnata	X	
	Clibanarius tricolor	X	X
	Conchoecetta acuminata		X
	Euchirella messinensis		X
	Euconchoecia chierchiae		X
	Gardinerosergia splendens	X	X
	Halocypris inflata		X
	Iulopis loveni	X	
	Leptocotis tenuirostris	X	
	Nematoscelis atlantica		
	Orthoconchoecia atlantica		X
	Orthoconchoecia secernenda		X
	Panulirus guttatus	X	X
	Paraconchoecia oblonga		X
	Parascelus edwardsi*	X	X
	Pareucalanus sewelii	X	
	Phronima stebbingi*	X	
	Phronimopsis spinifera	X	X
	Primno evansi	X	
	Sapphirina darwinii	X	
	Sapphirina metallina	X	X
	Simorhynchotus antennarius	X	
	Thyropus sphaeroma	X	
	Triconia conifera	X	

	Bassia	X	
	Chelophyes	X	
	Cytaeis	X	
	Enneagonum	X	X
	Eudoxoides	X	
	Hybocodon	X	
	Laodicea	X	
	Liriope	X	
	Nausithoe	X	
	Otoporpa*	X	X
	Peachia	X	
	Physophora	X	
	Slabberia	X	
	Sphaerocoryne	X	
	Turritopsis	X	
Species	Abylopsis eschscholtzii	X	
	Abylopsis tetragona	X	
	Agalma okenii	X	
	Amphicaryon ernesti	X	X
	Athorybia rosacea	X	
	Bassia bassensis	X	X
	Chelophyes appendiculata	X	
	Enneagonum hyalinum	X	X
	Eudoxoides mitra	X	
	Eudoxoides spiralis	X	X
	Liriope tetraphylla	X	X
	Nausithoe maculata*	X	X
	Nausithoe marginata*	X	X
	Nausithoe punctata	X	
	Palythoa caribaeorum	X	
	Palythoa heliodiscus	X	

	Lobatus*	X	X
	Macrotritopus	X	X
	Marshallora		X
	Ornithoteuthis	X	
	Oxygyrus		X
	Phrontis		X
	Pinna	X	
	Polygona	X	
	Pterotrachea		X
	Scaeurgus	X	X
	Telodiacria*	X	X
	Tonna	X	
Species	Abralia veranyi	X	
	Abraliopsis atlantica	X	X
	Atlanta inclinata		X
	Atlanta selvagensis	X	X
	Bursa granularis	X	
	Bursa thomae	X	X
	Cavolinia inflexa	X	X
	Coralliophila caribaea	X	X
	Creseis acicula		X
	Cyphoma gibbosum	X	X
	Diacavolinia longirostris		X
	Firoloida desmarestia	X	X
	Gutturnium muricinum	X	X
	Illex illecebrosus	X	X
	Limacina lesueurii	X	
	Limacina trochiformis	X	X
	Litiopa melanostoma	X	X
	Macrotritopus defilippi	X	X
	Octopoteuthis megaptera	X	
	Ornithoteuthis antillarum	X	
	Peracle reticulata	X	
	Phrontis alba		X
	Pinna carnea	X	
	Pneumoderma violaceum		X
	Polygona infundibulum	X	
	Pterotrachea coronata		X
	Pterotrachea hippocampus		X
	Scaeurgus unicirrhus	X	X
	Strombus alatus	X	X
	Lobatus raninus	X	X
	Tonna galea	X	

Figure S1: Gap analysis results for COI and 18S. StreamCode’s contribution of sequences (of barcode quality for COI) to GenBank for taxa not previously represented as of October 29, 2020. Contributions are quantified by phylum and taxonomic level. Note that not all of the StreamCode samples were identified to species level. The asterisks highlight names that were not found in the NCBI taxonomic framework. For higher taxonomic levels, this means we can’t discard a sequence that is available in GenBank but registered under an alternative name.

2.2 Metabarcoding OTU/ZOTU table by phylum

Table S6: Number of OTUs/ZOTUs by phylum for each genetic marker and method used for taxonomic assignment

phylum	clustering	marker	RDP.Classifier	BLASTn.StreamCode	BLASTn.GenBank
Acanthocephala	OTUs	18S V1-2	NA	NA	1
Acanthocephala	ZOTUs	18S V1-2	1	NA	1

phylum	clustering	marker	RDP.Classifier	BLASTn.StreamCode	BLASTn.GenBank
Annelida	OTUs	COI	4	46	27
Annelida	OTUs	18S V1-2	42	51	46
Annelida	ZOTUs	COI	5	122	30
Annelida	ZOTUs	18S V1-2	77	95	86
Arthropoda	OTUs	COI	1247	894	1411
Arthropoda	OTUs	18S V1-2	1424	2489	1290
Arthropoda	ZOTUs	COI	4123	3567	4639
Arthropoda	ZOTUs	18S V1-2	1534	1920	1570
Brachiopoda	OTUs	18S V1-2	2	2	2
Brachiopoda	ZOTUs	18S V1-2	4	5	6
Bryozoa	OTUs	COI	3	8	3
Bryozoa	OTUs	18S V1-2	9	8	9

2.3 Unique contributions of each method for taxonomic assignment

Counts of unique identifications by each method

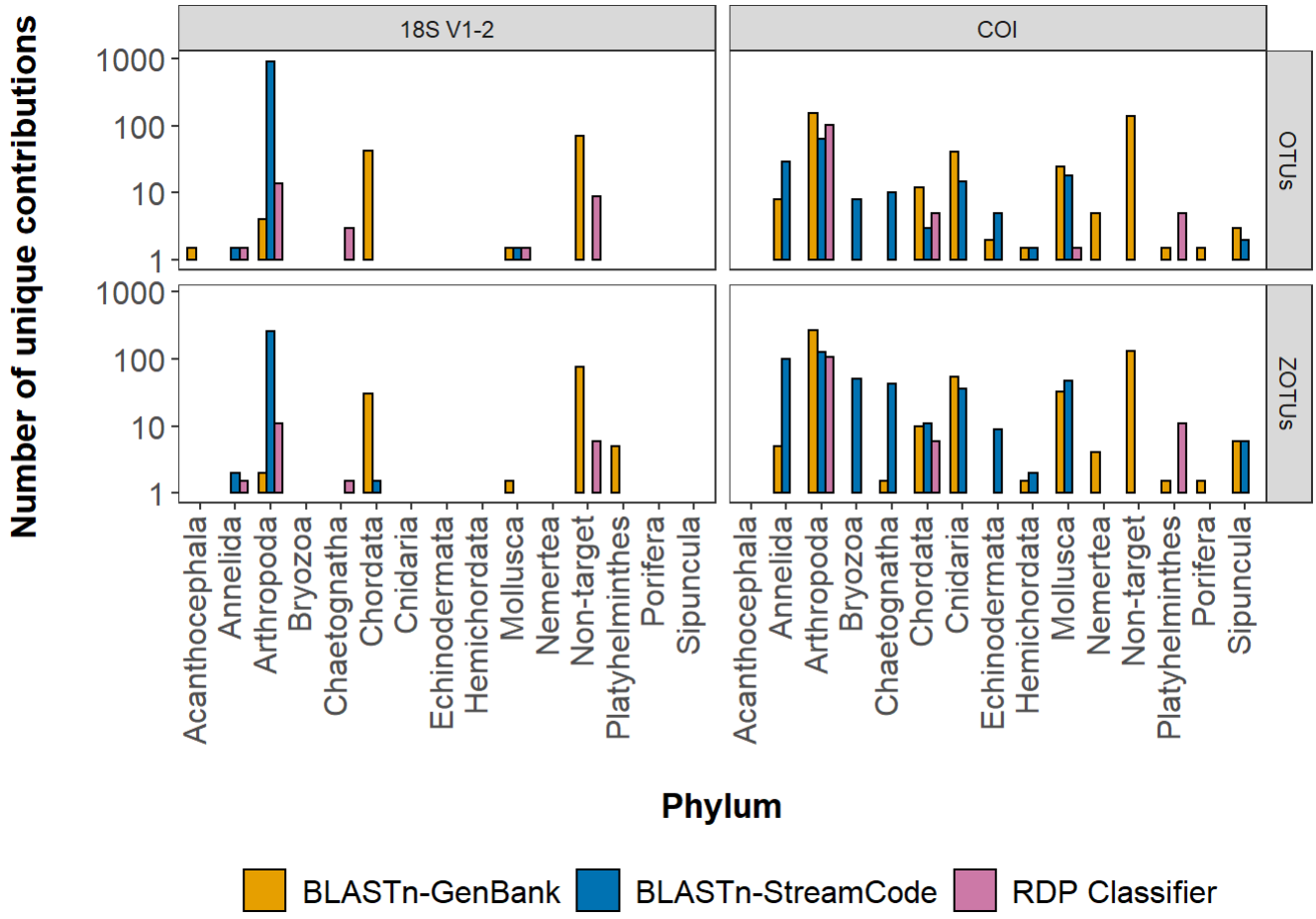
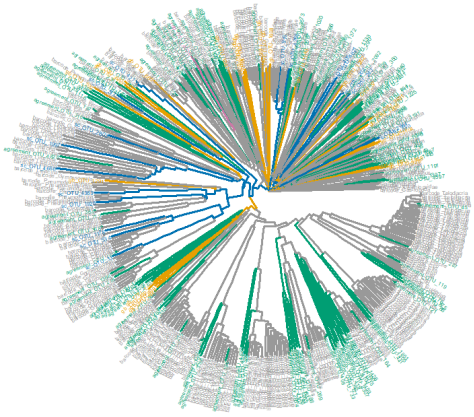


Figure S2: Unique contributions of each method for taxonomic assignment (BLASTn-GenBank, BLASTn-StreamCode, RDP Classifier), for the two genetic markers (COI and 18S V1-2) and two clustering approaches (OTUs and ZOTUs). The RDP Classifier was used with the PR2 database for 18S and MIDORI 2 database for COI. We added a constant of 0.5 when the number of OTUs/ZOTUs 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.

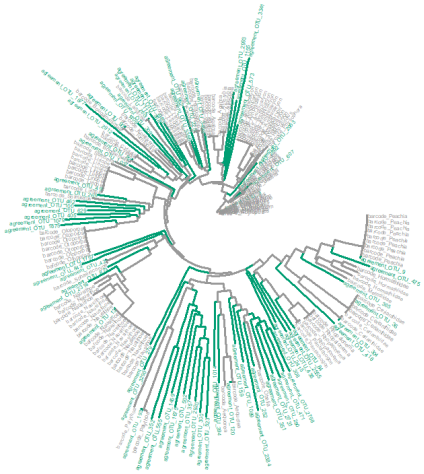
Trees combining barcodes and metabarcodes

18S V1-2 : Mollusca

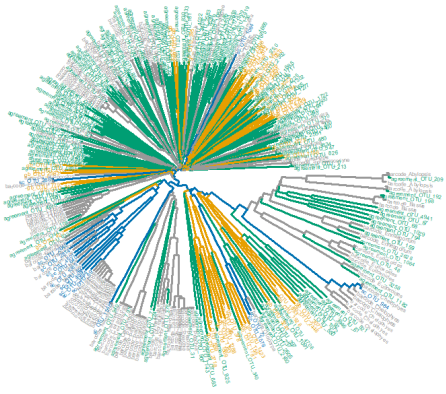
COI : Mollusca



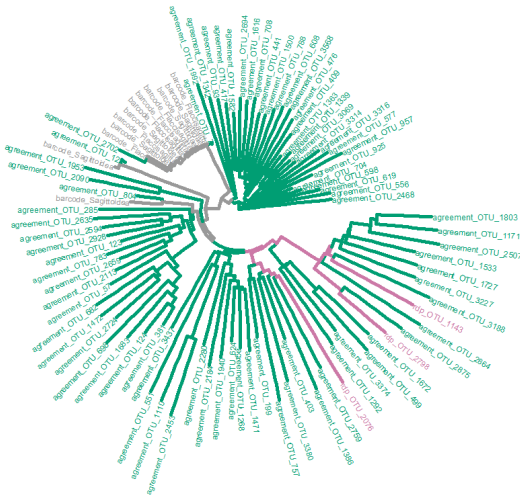
18S V1-2 : Cnidaria



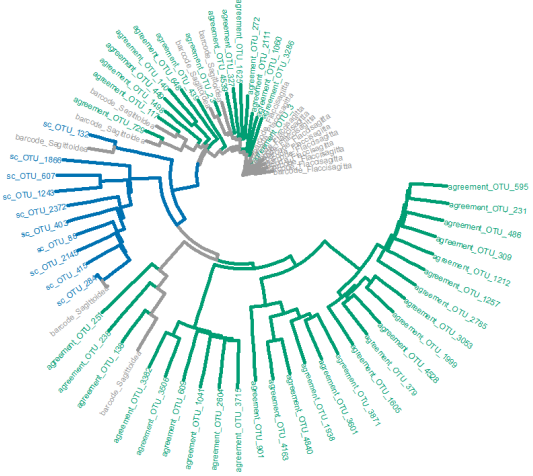
COI : Cnidaria



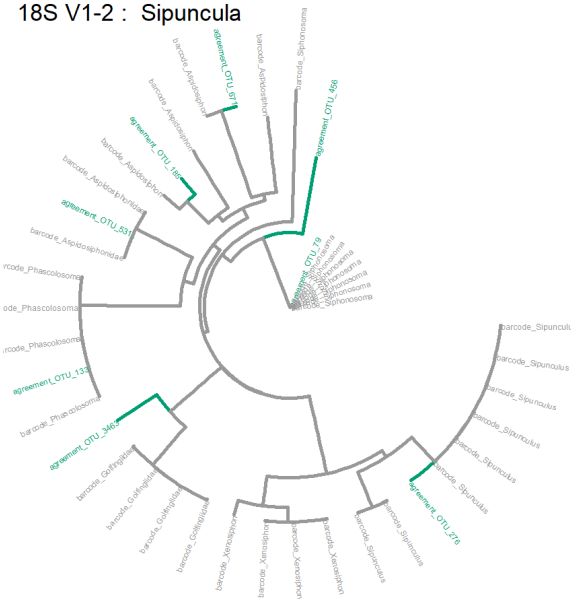
18S V1-2 : Chaetognatha



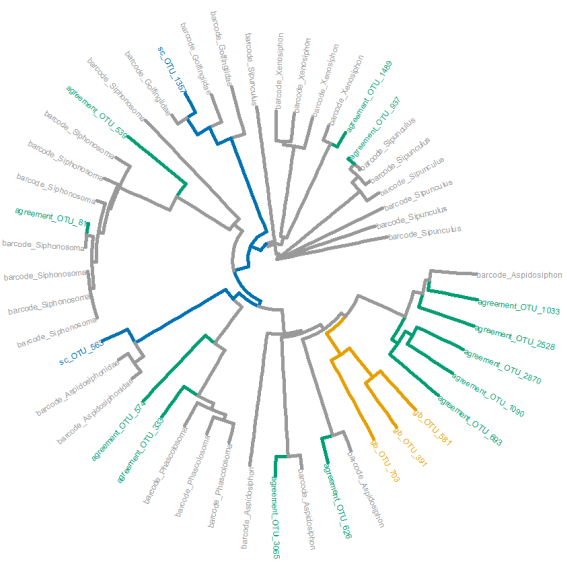
COI : Chaetognatha



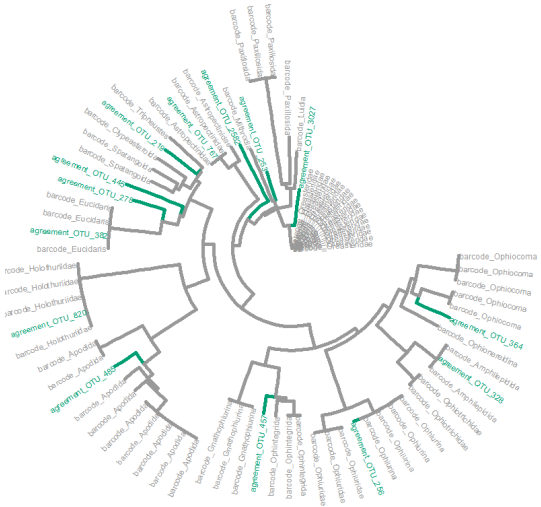
18S V1-2 : Sipuncula



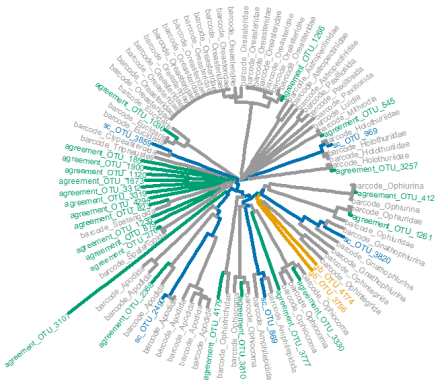
COI : Sipuncula



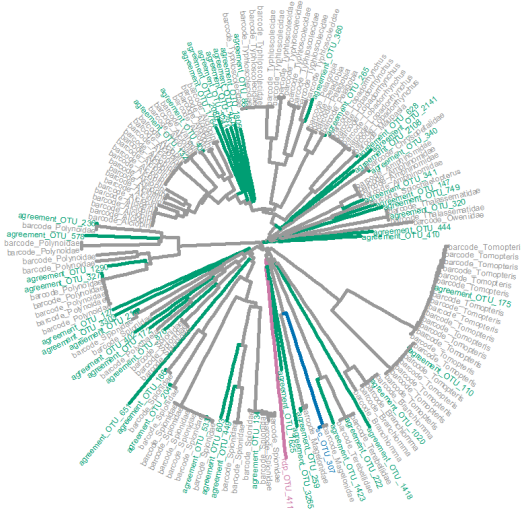
18S V1-2 : Echinodermata



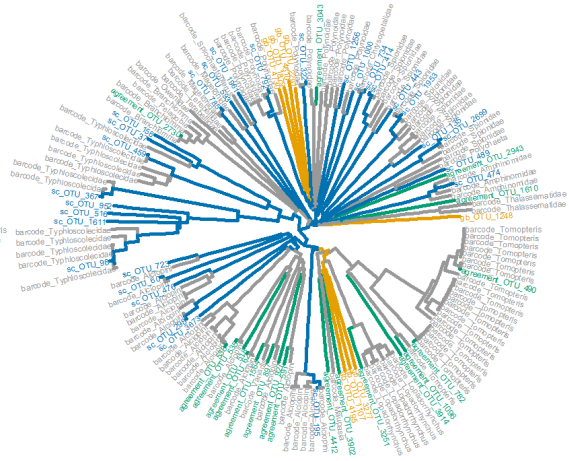
COI : Echinodermata



18S V1-2 : Annelida



COI : Annelida



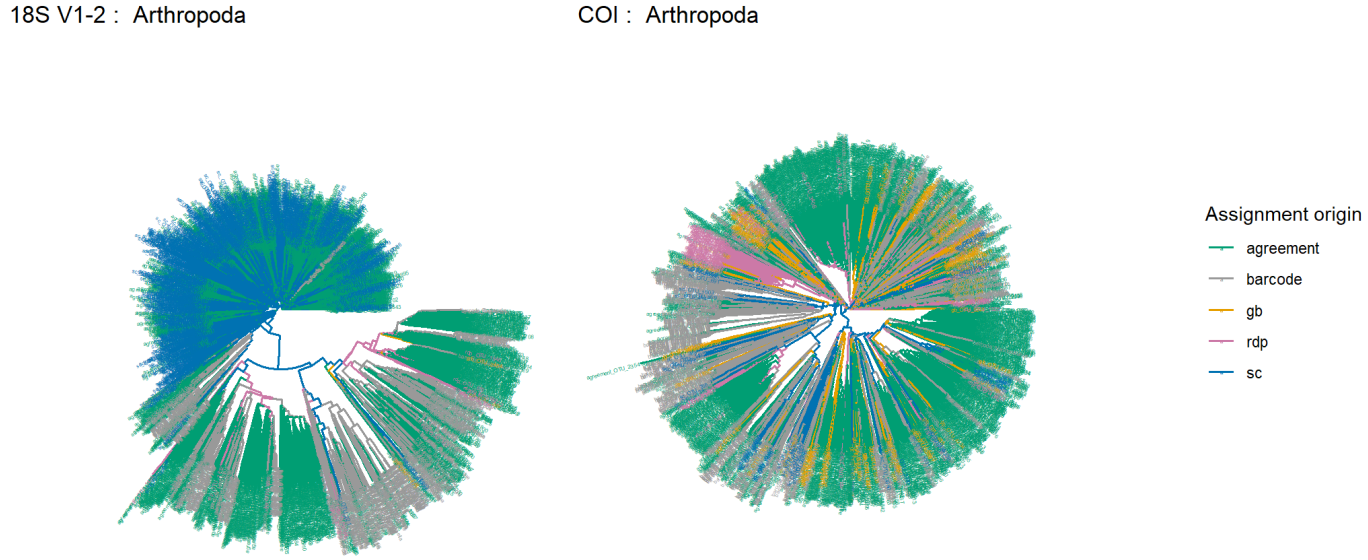


Figure S3: Distance trees highlighting the unique contributions of each method for taxonomic assignment (BLASTn-GenBank = gb, BLASTn-StreamCode = sc, RDP Classifier = rdp), for the identification of OTUs with two genetic markers (COI and 18S V1-2). The RDP Classifier was used with the PR2 database for 18S and MIDORI 2 database for COI. When the methods agree in the classification to phylum we color code it as “Agreement”. We also included the sequences of the StreamCode barcodes that were identified using morphology and expert judgment.

2.4 Holoplankton and meroplankton by phylum

Here we present the relative phyla composition for holoplankton and meroplankton (filtering out N/A and the few entries classified as “Others”):

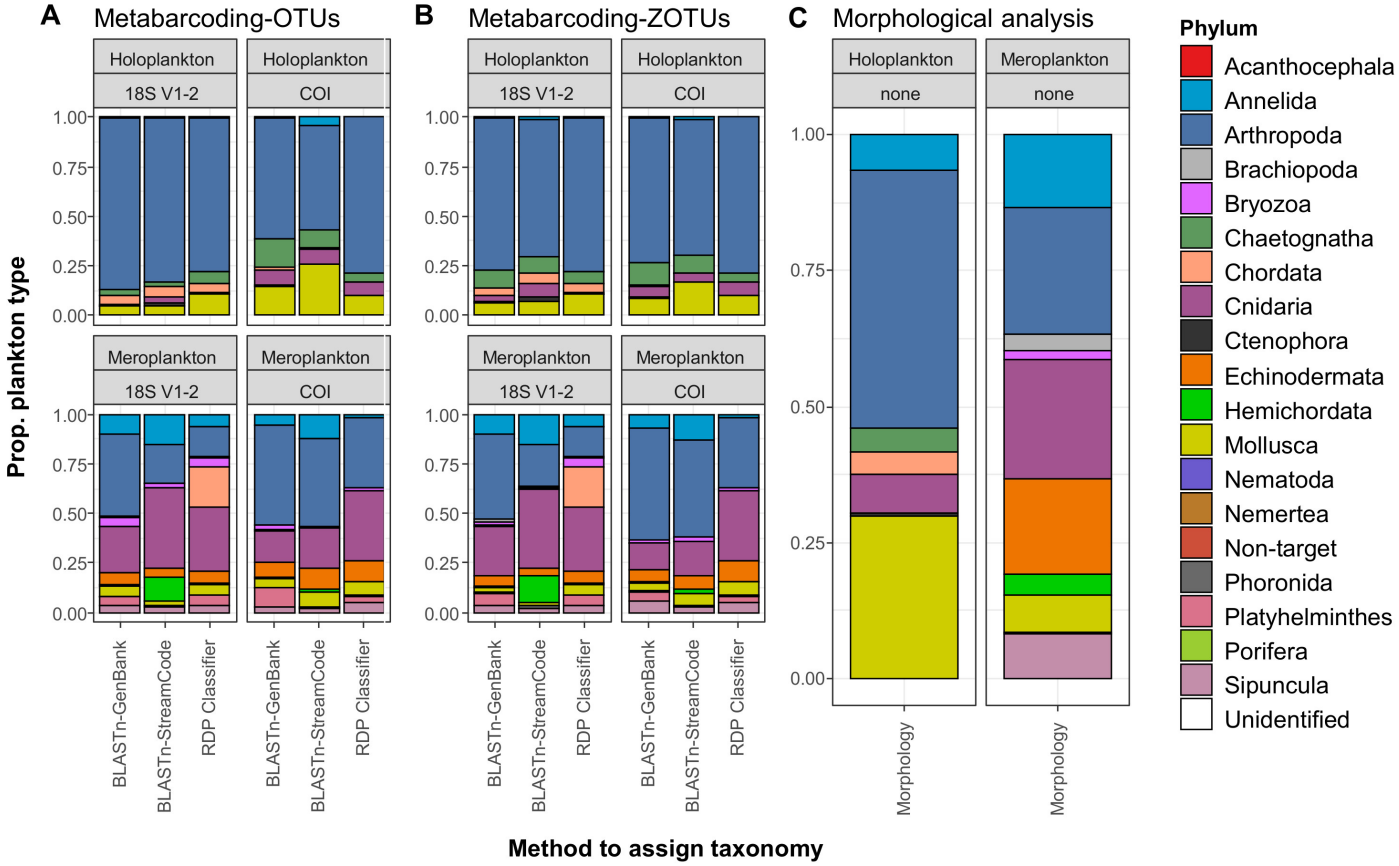


Figure S4: Phyla composition for holoplankton and meroplankton. Proportion of each phylum identified as holoplankton or meroplankton for the metabarcoding and morphology results. Taxa not identified to plankton type (“N/A”) or identified as “Others” are not included. A) Metabarcoding OTUs, B) Metabarcoding ZOTUs, C) Morphology.

3. References

(Geller *et al.*, 2013)(Blaxter *et al.*, 1998)(Leray *et al.*, 2013)(Caporaso *et al.*, 2010)(Edgar, 2013)(Meyer, 2003) (Deagle *et al.*, 2014)

Blaxter, M. L., De Ley, P., Garey, J. R., Liu, L. X., Scheldeman, P., Vierstraete, A., and Vanfleteren, J. R. *et al.* 1998. A molecular evolutionary framework for the phylum Nematoda. Nature, 392: 71–75. http://www.nature.com/articles/32160 (http://www.nature.com/articles/32160) (Accessed 3 June 2020).

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

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., and Taberlet, P. 2014. DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: Not a perfect match. Biology Letters, 10: 20140562. https://royalsocietypublishing.org/doi/10.1098/rsbl.2014.0562 (https://royalsocietypublishing.org/doi/10.1098/rsbl.2014.0562) (Accessed 23 October 2019).

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

Geller, J., Meyer, C., Parker, M., and Hawk, H. 2013. Redesign of PCR primers for mitochondrial cytochrome *c* oxidase subunit I for marine invertebrates and application in all-taxa biotic surveys. Molecular Ecology Resources, 13: 851–861. http://doi.wiley.com/10.1111/1755-0998.12138 (http://doi.wiley.com/10.1111/1755-0998.12138) (Accessed 8 January 2020).

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., and Boehm, J. T. *et al.* 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. http://frontiersinzoology.biomedcentral.com/articles/10.1186/1742-9994-10-34 (http://frontiersinzoology.biomedcentral.com/articles/10.1186/1742-9994-10-34) (Accessed 11 January 2020).

Meyer, C. P. 2003. Molecular systematics of cowries (Gastropoda: Cypraeidae) and diversification patterns in the tropics. Biological Journal of the Linnean Society, 79: 401–459. https://academic.oup.com/biolinnean/article-lookup/doi/10.1046/j.1095-8312.2003.00197.x (https://academic.oup.com/biolinnean/article-lookup/doi/10.1046/j.1095-8312.2003.00197.x) (Accessed 3 June 2020).