Received: 16 December 2020 | Revised: 28 January 2021 | Accepted: 29 January 2021

DOI: 10.1002/edn3.187

METHOD

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

WILEY

Mitohelper: A mitochondrial reference sequence analysis tool for fish eDNA studies

Shen Jean Lim[1,2] | **Luke R. Thompson**[2,3]

[1]Rosenstiel School of Marine and Atmospheric Science, Cooperative Institute for Marine and Atmospheric Studies, University of Miami, Miami, FL, USA

[2]Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, FL, USA

[3]Northern Gulf Institute, Mississippi State University, Mississippi State, MS, USA

Correspondence
Luke R. Thompson, Ocean Chemistry and Ecosystems Division, Atlantic Oceanographic and Meteorological Laboratory, National Oceanic and Atmospheric Administration, Miami, Florida, USA.

Funding information
National Oceanic and Atmospheric Administration, Grant/Award Number: NA06OAR4320264 06111039 and NA20OAR4320472

Abstract

Fish eDNA metabarcoding is a nonintrusive, time- and cost-effective way to detect biodiversity, with potential applications in ecosystem-based management and fisheries assessment. Nevertheless, fish-specific eDNA resources are currently not well-developed, and many fish species are not yet sequenced in reference databases. We developed Mitohelper, a Python-based command line tool to annotate and align fish mitochondrial sequences available in the existing MitoFish database. Mitohelper improves MitoFish annotations by adding gene names and additional taxonomic classifications. Using these improved reference datasets, Mitohelper's *getrecord* function searches MitoFish for available mitochondrial (cytochrome *c* oxidase, 12S rRNA, and others) gene sequences against a user-provided list of fish taxonomic names. Mitohelper's *getalignment* function aligns (often partial) mitochondrial gene sequences to a user-specified full-length reference sequence for the assessment and visualization of overlapping sequencing regions. To facilitate the development of taxonomic classifiers, we combined Mitohelper's 12S rRNA dataset with SILVA's 16S rRNA and 18S rRNA datasets in a QIIME 2-compatible format. By providing valuable information on taxonomic and gene region coverage of currently available fish mitochondrial data, Mitohelper's functions promote informed experimental design that can guide sequencing and analysis strategies. In summary, Mitohelper improves the breadth and functionality of eDNA data resources, as well as the accuracy of taxonomic classification. Mitohelper and its reference datasets are updated approximately monthly and available at https://github.com/aomlomics/mitohelper.

KEYWORDS
computational biology, ecology

1 | INTRODUCTION

Aquatic ecosystems support a rich diversity of micro- and macroorganisms across the kingdoms of life. These organisms interact with each other and their environments, forming complex networks that underpin ecosystem functions beneficial to humans (Zinger et al., 2012). These functions include climate regulation, primary productivity, biogeochemical cycling, and resource provisioning (Zinger et al., 2012). Understanding species diversity and interactions in aquatic ecosystems is thus a critical component of ecosystem-based management (Palumbi et al., 2009), including fisheries assessment (Townsend et al., 2019). The emergence and increasing availability of high-throughput sequencing technology have garnered extensive interest in the use of environmental

DNA (eDNA) metabarcoding as a non-invasive, time-effective, and cost-effective method to detect aquatic biodiversity across multiple trophic levels (Deiner et al., 2017). Particularly for fishes, eDNA metabarcoding has the potential to detect greater diversity than traditional surveys (e.g., net tows or bottom trawls) with higher power, especially for elusive fish species and over large spatial scales (Castañeda et al., 2020; Deiner et al., 2017). In a typical eDNA metabarcoding workflow, one or more marker genes are amplified from community DNA, then sequenced with high-throughput methods to assess species diversity (Deiner et al., 2017). Marker genes commonly used for fish eDNA metabarcoding are mitochondrial-encoded, including cytochrome *c* oxidase subunit I (COI) (Bakker et al., 2017; Leray et al., 2013; Simpfendorfer et al., 2016; B. C. Stoeckle, Beggel, et al., 2017), cytochrome *b* (Hanfling et al., 2016; Jo et al., 2019; Lacoursière-Roussel et al., 2016; Murakami et al., 2019), and various ribosomal RNAs (rRNAs). 16S rRNA and 18S rRNA have been used in fish eDNA metabarcoding studies (Bessey et al., 2020; Stat et al., 2019) and broader surveys across eukaryotic groups (Berry et al., 2019; Djurhuus et al., 2020; Holman et al., 2019; Kelly et al., 2017; Waraniak et al., 2019). In recent years, 12S rRNA is increasingly used in fish eDNA metabarcoding (Andruszkiewicz et al., 2017; Kelly et al., 2014; Lafferty et al., 2020; Miya et al., 2015; Stoeckle et al., 2017, 2020; Thomsen et al., 2016; Yamamoto et al., 2017). 12S rRNA-based primers offer similar taxonomic resolution as COI-based primers, but with higher specificity to fishes due to lower primer-template mismatches and lower PCR amplification bias (Bylemans et al., 2018).

Taxonomic assignment is a crucial step in assessing eDNA community composition, involving the comparison of processed read sequences to known sequences of interest that are taxonomically curated within reference databases (Deiner et al., 2017). Specialized sequence databases include the Barcode of Life Data System (BOLD), which contains mostly COI sequences from the Animalia, Plantae, Fungi, and Protista kingdoms of life (Ratnasingham & Hebert, 2007). Specialized rRNA gene repositories, such as the SILVA rRNA database (Quast et al., 2013) and the Ribosomal Database Project (RDP) database (Cole et al., 2009) are also frequently used for bacteria, archaea, fungal, and eukaryote taxonomic assignment. The National Center for Biotechnology Information (NCBI) maintains generalized nucleotide sequence databases like GenBank (including basic local alignment search tool (BLAST) nucleotide databases like nt) and the more-curated RefSeq database, as well as specialized subrepositories like the Organelle Genome Resources and the ribosomal RNA BLAST databases (Sayers et al., 2019). Nevertheless, NCBI data are not quality-controlled to the level required for precise taxonomic analysis and sequences can be inaccurately annotated, leading to taxonomic misclassification (Iwasaki et al., 2013; Machida et al., 2017; Wangensteen & Turon, 2017). Furthermore, curated fish systematics dataset like the classification scheme in *Fishes of the World* (Nelson et al., 2016), as well as sequence datasets like BOLD and NCBI, are not readily compatible with existing taxonomic assignment algorithms (Machida et al., 2017). This has motivated efforts to develop pipelines for data curation and formatting (Curd et al., 2019; Machida et al., 2017; Wangensteen & Turon, 2017) and sequence data integration across databases (Arranz et al., 2020).

Despite the wealth of available sequence data, databases focusing on aquatic organisms are less comprehensive than terrestrial organisms. Aquatic organisms are often under-represented or absent from reference databases (Iwasaki et al., 2013; Wangensteen & Turon, 2017). For fish-related eDNA research, fish-specific sequence databases can expedite taxonomic assignment, facilitate customized annotations, and inform taxonomic coverage of available sequence data (Iwasaki et al., 2013; Wangensteen & Turon, 2017). The MitoFish database (Iwasaki et al., 2013; Sato et al., 2018) is currently the only public fish-specific mitochondrial resource. MitoFish (Iwasaki et al., 2013; Sato et al., 2018) contains high-quality reannotated complete mitogenome data, as well as complete and partial mitogenome data downloaded approximately monthly from NCBI RefSeq and GenBank (Sayers et al., 2019). MitoFish also offers the MitoAnnotator and MiFish pipelines for mitogenome annotation and eDNA analyses (Iwasaki et al., 2013; Sato et al., 2018). Nonetheless, only a small portion (~ 0.4% of sequence data and ~ 8% of species) of MitoFish data is annotated, while remaining records are FASTA-formatted and annotated with only their accession numbers and taxonomic (from species to order) information. Parsing MitoFish sequence data to identify gene names and taxonomic classifications can be time- and resource-consuming, necessitating the development of helper tools that expand the annotations available for these fish mitochondrial sequences.

We developed Mitohelper, a Python-based command line tool to annotate MitoFish data and facilitate experimental design, alignment visualization, and reference sequence analysis in fish-specific eDNA studies. Mitohelper's *getrecord* function identifies available mitochondrial (all genes, COI only, or 12S rRNA only) sequence records from a user-submitted list of fish taxonomic names. A separate *getalignment* function aligns a set of input mitochondrial sequences with a user-provided reference sequence to identify matching gene regions. Mitohelper's reference datasets integrate (a) MitoFish data (accession number and sequence); (b) NCBI data (gene definition and taxonomic classification); and 3) fish systematics data from *Fishes of the World* in tab-separated format files and are publicly available in our GitHub repository (https://github.com/aomlomics/mitohelper). Our repository also includes preformatted QIIME-compatible ribosomal RNA sequence and taxonomy databases (12S rRNA gene alone or 12S rRNA + 16S rRNA + 18S rRNA genes) that can be used for taxonomic assignment in fish eDNA analyses.

2 | MATERIALS AND METHODS

Complete and partial fish mitogenome sequences were downloaded from monthly MitoFish (Iwasaki et al., 2013) releases available from http://mitofish.aori.u-tokyo.ac.jp on the 2nd of each month. Mitogenome accession numbers were matched with NCBI accession

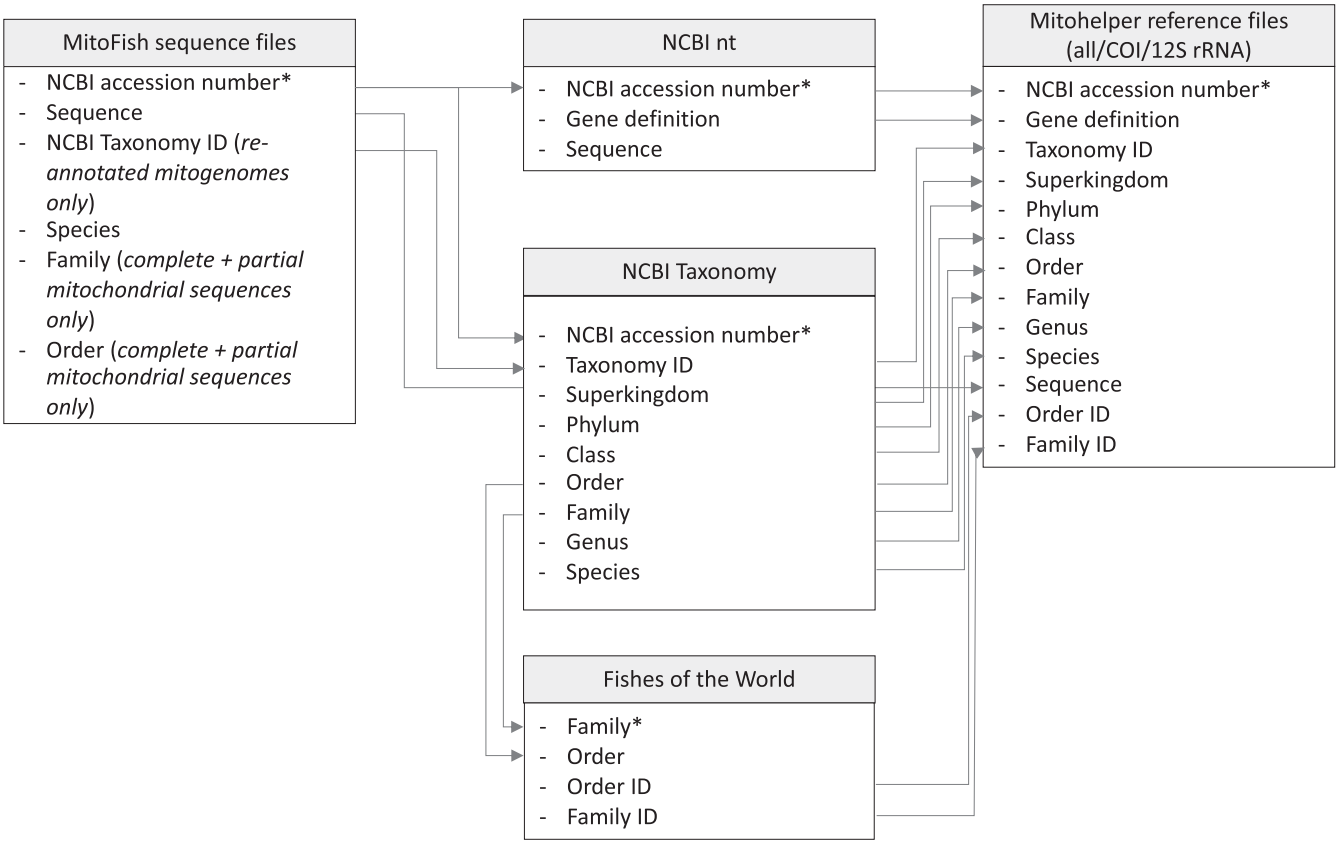

FIGURE 1 Data schema illustrating how MitoFish (Iwasaki et al., 2013), NCBI (Sayers et al., 2019), and *Fishes of the World* (Nelson et al., 2016) data were integrated to generate the Mitohelper reference files. Each box represents a dataset and entries in that dataset. * denotes the primary key of each dataset

numbers and gene definitions from the nt blast database (Sayers et al., 2019), using Bash and Python scripts described in the developer wiki page (https://github.com/aomlomics/mitohelper/wiki; Figure 1). From the mitogenome entries, corresponding taxonomic information was retrieved from the NCBI taxonomy database using the R package taxonomizr (https://www.rdocumentation.org/packages/taxonomizr; Figure 1). Entries with missing taxonomic information were manually reviewed and fixed. Order and family numbers (IDs) from the fifth edition of *Fishes of the World* (Nelson et al., 2016) were downloaded from https://sites.google.com/site/fotw5th/ and matched with order and family names in the dataset using R's merge function. The COI gene dataset was extracted from the merged mitogenome records using the keywords (case-insensitive) "(COI)," "CO1," "COX1," "(COXI)," "COI," "cytochrome *c* oxidase subunit I," "cytochrome *c* oxidase subunit 1," "complete genome," "complete mito," and "cytochrome oxidase subunit I." Cytochrome *b* records were filtered out from this dataset using the keywords (case-insensitive) "cytochrome *b*" and "*cytb*." The 12S rRNA gene dataset was extracted from the merged mitogenome records using the keywords (case-insensitive) "12S ribo," "12S rRNA," "12S small," "complete genome," and "complete mito." Records that did not contain these keywords but contained the more general keyword "small subunit" were manually reviewed and included in the dataset if they were described as 12S rRNA records in the corresponding literature, or classified to be 12S rRNA by NCBI's web blastn search (Ye et al., 2006). Records that did not contain these keywords but were manually verified to be COI or 12S rRNA gene sequences from literature, such as partial mitogenome records and those described in fish 12S rRNA metabarcoding studies, for example, (Stoeckle et al., 2018), were included in the dataset.

Mitohelper functions were written and tested using Python v3.6.10 (Sanner, 1999). The Python package click v7.1.2 (https://click.palletsprojects.com) was used to build the command line interface that contains two main commands, *getrecord* and *getalignment*, which are run from the command line following the base command *mitohelper.py*. *getrecord* uses case-insensitive regular expression searches to retrieve sequence records, FASTA-formatted sequences (optional), and QIIME-formatted hierarchical taxonomy output (optional) from a user-provided fish taxonomy list. *getalignment* provides an option to run a user-specified blastn algorithm (traditional blastn, blastn-short, megablast, or discontiguous dc-megablast) to locally align sequences in an input FASTA file with a user-specified reference sequence. *getalignment* outputs the blastn results in tab-separated format (-outfmt 7). For each input sequence, *getalignment* extracts the local alignment positions with the highest alignment bit score to the reference sequence from the tab-separated table. If more than one alignment shares the highest bit score, the script will combine the hits and report the earliest start position and latest

end position. The alignment positions are then parsed by the Python package pandas v0.25.3 (McKinney, 2010) and visualized with the Python package seaborn v0.10.1 (Waskom & team, s. d., 2020), which is based on matplotlib v3.3.0 (Hunter, 2007).

To create QIIME-compatible rRNA datasets, the 12S rRNA gene reference dataset was split into taxonomy and sequence tables, then imported into QIIME 2 v2020.6 (Bolyen et al., 2019). Sequence data were filtered using the QIIME 2 plugin RESCRIPt (REference Sequence annotation and CuRatIon Pipeline) (Bolyen et al., 2019). Sequences with ≥ 5 ambiguous bases and homopolymers ≥ 8 bp long were removed from the dataset. Remaining sequences and taxonomies were dereplicated in uniq mode to retain identical sequences with different taxonomies. The length-filtering step was skipped because amplicons generated using MiFish primers (Sato et al., 2018) are generally short (~170 bp). To create the combined 12S rRNA + 16S rRNA + 18S rRNA datasets, dereplicated 12S rRNA sequences and taxonomies were exported from QIIME 2. RESCRIPt-processed full-length 16S rRNA and 18S rRNA gene data from the SILVA database v138 release (Quast et al., 2013) was downloaded from https://docs.qiime2.org/2020.11/data-resources/ and exported from the QIIME 2 environment. 12S rRNA, 16S rRNA, and 18S rRNA gene and taxonomy data were then concatenated and imported back into QIIME 2 (Bolyen et al., 2019).

3 | RESULTS

As of January 2021, there are 668,366 records in MitoFish v3.63 (Iwasaki et al., 2013). Based on our analysis, these species belonged to 89 known taxonomic orders, 567 families, 4,495 genera, and 37,910 species. 95% of the records (n = 627,768) were from species belonging to Actinopteri, the largest taxonomic class of fish (Figure 2). Mitohelper's data processing pipeline identified ~ 34% (n = 224,712) of the MitoFish records to be COI and ~ 6% (n = 37,314) to be 12S rRNA. These records were used to create reference datasets used for Mitohelper's functions. Of the 567 fish families in the MitoFish dataset, 13 were absent in the COI subset and 24 were absent in the 12S rRNA subset (Table 1). Six families commonly absent in the COI and 12S rRNA subsets included Cynodontidae, Hypnidae, Leptocharidae, Normanichthyidae, Notocheiridae, and

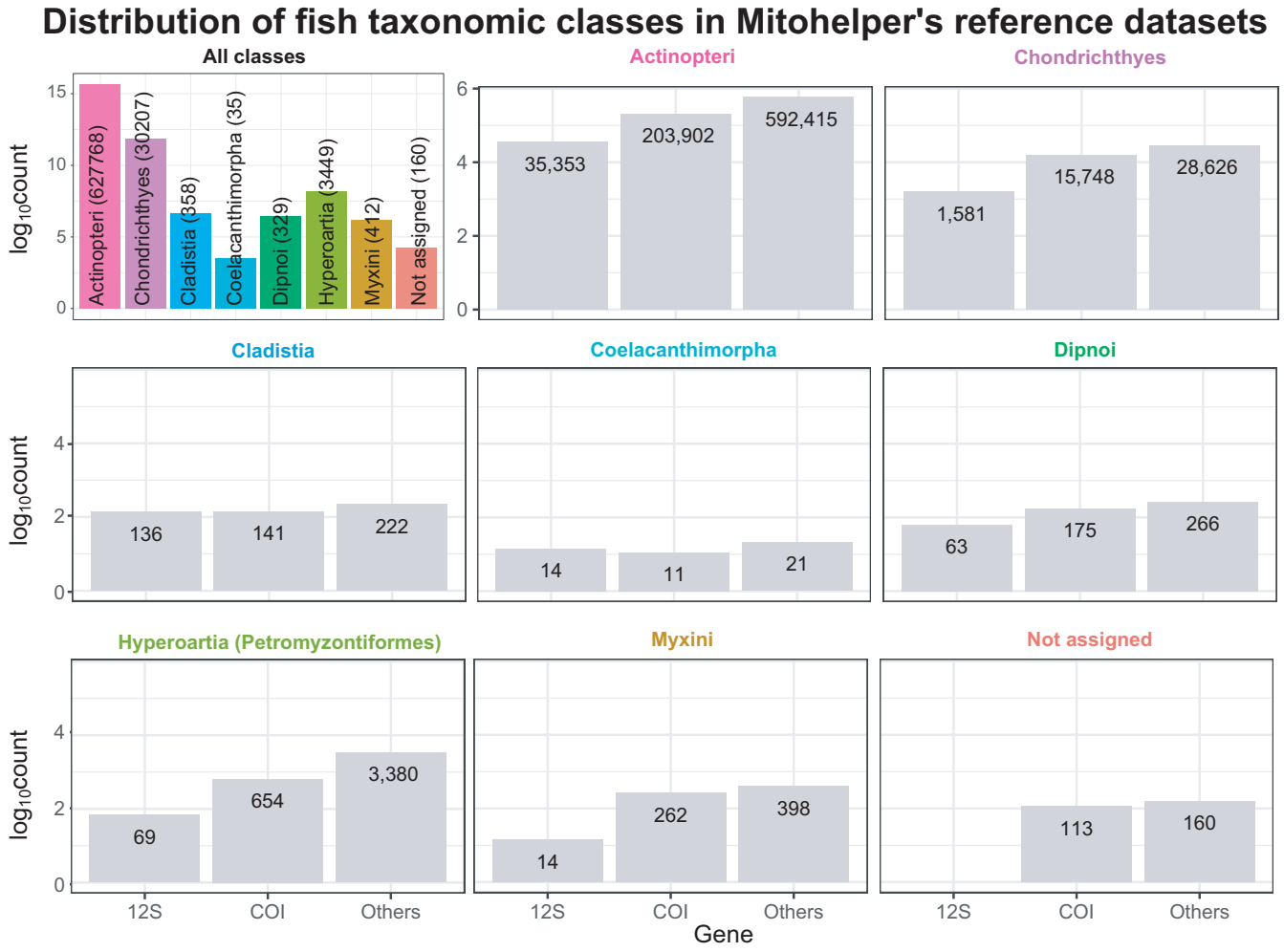

FIGURE 2 Distribution of taxonomic classes in Mitohelper's datasets, including 12S rRNA, COI (cytochrome *c* oxidase subunit I), and other gene records. The number of records (y-axis) from each taxonomic class is $\log_{10}$-transformed. Actual count values are shown on each bar [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 1 Fish families present in MitoFish database, but not in COI and/or 12S rRNA subsets of the database

Fish families missing from COI subset	Fish families missing from 12S rRNA subset
Aenigmachannidae	Akysidae
Anchariidae	Aphyonidae
Bathysauroididae	Bathysauropsidae
Bembropidae	Brachaeluridae
Cynodontidae	Congiopodidae
Hypnidae	**Cynodontidae**
Leptochariidae	Distichodontidae
Normanichthyidae	Gibberichthyidae
Notocheiridae	**Hypnidae**
Parascylliidae	Kryptoglanidae
Radiicephalidae	**Leptochariidae**
Tarumaniidae	Milyeringidae
Thalasseleotrididae	**Normanichthyidae**
	Notocheiridae
	Pantanodontidae
	Parascorpididae
	Pataecidae
	Plectrogeniidae
	Schilbidae
	Scoloplacidae
	Stephanoberycidae (genome assembly GCA_900312575 available on NCBI)
	Thalasseleotrididae
	Zanclorhynchidae
	Zanobatidae

Note: Families absent from both subsets are highlighted in bold.

Thalasseleotrididae (Table 1). The COI subset contained records from 25,305 species, while the 12S rRNA subset contained records from 11,961 species.

getrecord uses case-insensitive regular expression searches to retrieve sequence records from a user-provided fish taxonomy list (Figure 3). *getrecord* supports seven main taxonomic levels, including species (L7), genus (L6), family (L5), order (L4), and phylum (L3). Users can submit a plain text file containing fish taxonomic names (at a taxonomic rank at or below the specified rank being searched), specify the reference database, and indicate the taxonomic rank to search all queries against (Figure 3). For example, given a list of fish species names, a user can extract 12S rRNA sequence records ("-d mitofish.12S. Dec2020.tsv") matching the family/families of these fish species ("-l 5"). *getrecord* outputs matching records in tab-separated values (TSV) format with the following headers: Query, Accession, Gene definition, taxid (NCBI Taxonomy ID), Superkingdom, Phylum, Class, Order, Family, Genus, Species, Sequence, OrderID (from *Fishes of the World*), and FamilyID (from *Fishes of the World*; Figure 3). Accession and sequence data of retrieved records can also be written into an output FASTA file using the optional "--fasta" switch (Figure 3). With the optional "--taxout" switch, *getrecord* can also write the accession numbers and taxonomic information of retrieved records to a hierarchical taxonomy file that can be directly imported into QIIME 2 (Bolyen et al., 2019).

getalignment uses local sequence similarity search results to retrieve the alignment positions of a set of input sequences relative to a user-specified reference sequence (Figure 3). The input to *getalignment* can be a FASTA-formatted nucleotide sequence file generated using Mitohelper's *getrecord* or other methods (Figure 3). Because alignment positions are relative, the user will also have to prepare a separate FASTA file (specified using the "-r/--reference-sequence" option), which contains a single nucleotide reference sequence for all the input sequences to be searched against (Figure 3). Example reference sequence files, including full-length 12S and 18S rRNA gene sequences from *Danio rerio* (zebrafish) and 16S rRNA sequence from *Escherichia coli* strain K-12 substr. MG1655, are provided in the "testdata" folder of the Mitohelper GitHub repository. For the sequence similarity search, the user can select one of the following blastn tasks: blastn, blastn-short, megablast, or dc-megablast, using the "blastn-task" option (Figure 3). Because of differences in word size and scoring parameters, the number of hits and maximum bit scores produced by various blastn tasks is different (see Figure 4 for example). Traditional blastn reports hits with at least 11 nucleotide exact matches (word size = 11), while blastn-short (word size = 7) is optimized for similarity searches involving short nucleotide sequences (<50 bp) (BLAST® Command Line Applications User Manual, 2008). On the other hand, traditional megablast (word size = 28) is optimized for detecting highly identical sequences, such as those from the same species or closely related species, while dc-megablast (word size = 11 with nonconsecutive base matching) is optimized for more divergent sequences, such as those from more distantly related species (BLAST® Command Line Applications User Manual, 2008).

getalignment also accepts a tab-separated blastn output file (generated using the BLAST application's -outfmt 6 and -oufmt 7 flags) as input (Figure 3). In this case, *getalignment* expects the reference sequence to be the first query sequence and the subject sequence. Hence, the "-r/--reference-sequence" option is not needed, and "blastn-task" should be set to "none." *getalignment* produces a tabular output listing the accession numbers of sequences producing significant alignments, their start and end positions relative to the reference sequence, and the highest bit score of the alignment (Figure 3). An output PDF file that graphically summarizes these alignment positions and scores in a line plot similar to NCBI's web blast output is also provided (Figure 3).

4 | DISCUSSION

We developed a Python-based search tool, Mitohelper, to facilitate preliminary stages of fish eDNA research. The *getrecord* command

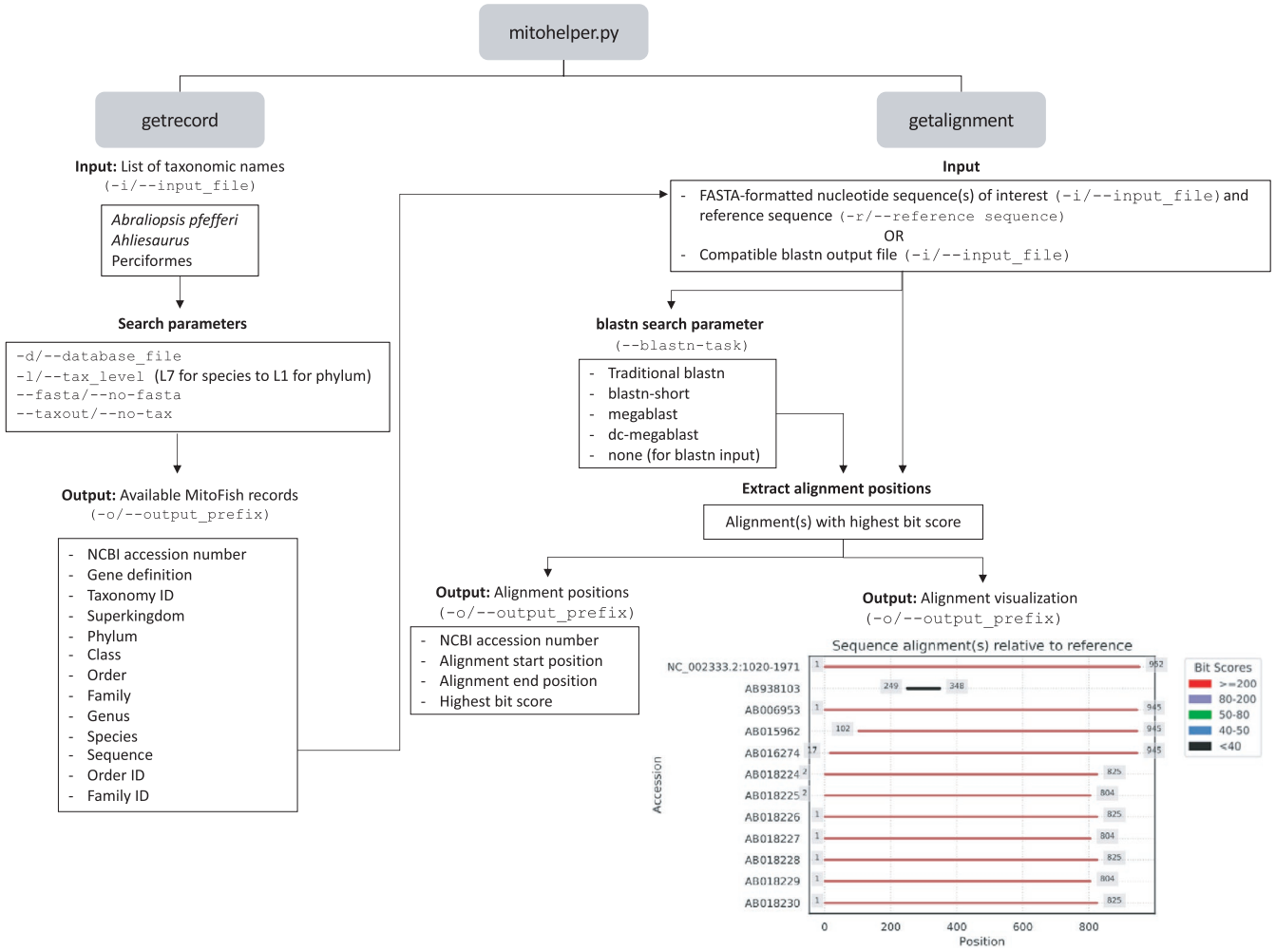

FIGURE 3 Overview of Mitohelper's functions (*getrecord* and *getalignment*) and outputs [Colour figure can be viewed at wileyonlinelibrary.com]

is useful for researchers interested in surveying the presence/absence of mitochondrial reference sequences for specific fish taxa, while the *getalignment* command enables the visualization and examination of sequenced mitochondrial gene region(s). These functions are important for experimental design in fish eDNA studies because not all mitochondrial regions from all fish species have been sequenced to date. Since COI has traditionally been used for animal barcoding (Hebert et al., 2003), full-length COI sequences are available for more fish species than other marker gene sequences, such as the 12S rRNA gene. Compared to 12S rRNA sequence records, there are about six times as many COI records and twice as many fish species represented in MitoFish (Iwasaki et al., 2013; Sato et al., 2018). Furthermore, because different 12S rRNA sequencing primers used in eDNA studies target different regions of the gene, partial sequences publicly available are specific only to the sequenced region. By providing insights on the taxonomic and gene region coverage of MitoFish (Iwasaki et al., 2013; Sato et al., 2018), Mitohelper guides research decisions pertaining to the sequencing of additional voucher specimens, the choice of sequencing primers, and the choice of downstream analysis methods.

MitoFish (Iwasaki et al., 2013; Sato et al., 2018), compiled from NCBI RefSeq and GenBank (Sayers et al., 2019), represents one of the most comprehensive mitochondrial fish sequence resource in terms of species coverage. With the Mitohelper reference dataset, we also created QIIME-compatible rRNA datasets that can be useful for developing taxonomic classifiers, based on either 12S rRNA gene sequences, or 12S rRNA gene sequences + 16S rRNA gene sequences + 18S rRNA gene sequences. The inclusion of 16S rRNA and 18S rRNA sequences to 12S rRNA gene sequence classifier data can improve the detection of contaminant rRNA sequences, such as bacterial 16S rRNA sequences concurrently amplified in metazoan 12S rRNA studies (Machida et al., 2012). Similar to previous observations (Arranz et al., 2020), we note that data unique to other resources, such as BOLD (Ratnasingham & Hebert, 2007), can be missing in the MitoFish database. Mitohelper extracts COI and 12S rRNA datasets predominantly using keyword searches, which are affected by metadata accuracy and annotation quality (Arranz et al., 2020). During the development of Mitohelper, we manually added sequence records without typical keywords associated with mitochondrial marker gene records, although our list of atypically annotated records may not be exhaustive. Sequence-based

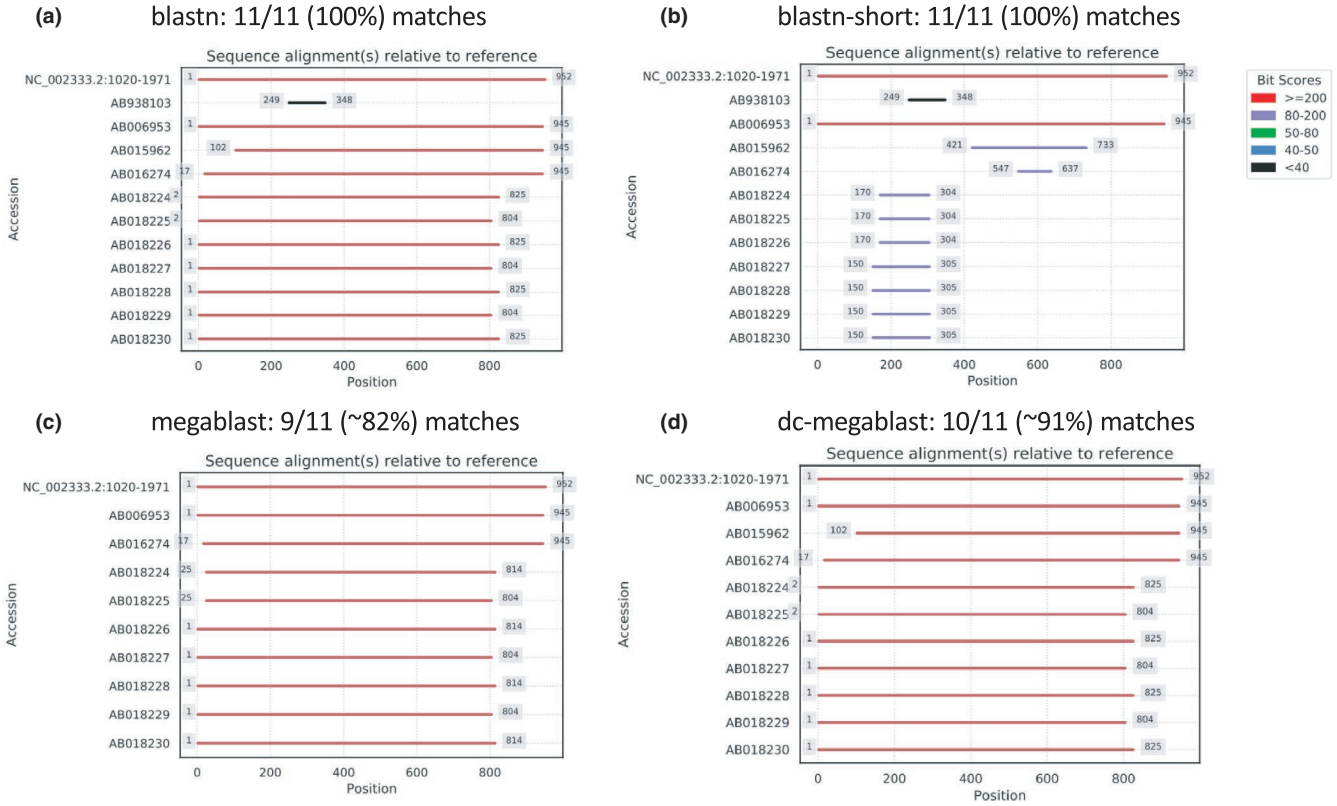

FIGURE 4 Examples of Mitohelper's *getalignment* graphical output generated using (a) traditional blastn, (b) blastn-short, (c) megablast, and (d) discontiguous (dc)-megablast. Alignment positions for each sequence are indicated in gray boxes, and line colors correspond to the alignment bit score values. These alignments were based on the 12S rRNA gene, using the sequence from *Danio rerio* (zebrafish; NC_002333) as reference [Colour figure can be viewed at wileyonlinelibrary.com]

approaches for generating reference databases, such as Creating Reference libraries Using eXisting tools (CRUX) (Curd et al., 2019) circumvents this limitation, albeit with a trade-off between sensitivity and computing time.

While Mitohelper's *getrecord* function looks up mitochondrial gene information from the reference datasets, the *getalignment* function retrieves alignment positions from input sequence(s) aligned to a reference sequence. Unlike the prokaryotic 16S rRNA marker gene, where primer and sequence positions across different species are unanimously labeled according to their corresponding positions on the *E. coli* 16S rRNA gene (Lane et al., 1985), a reference species does not yet exist for eukaryotic and fish eDNA studies. Although Mitohelper allows users to select their own reference sequence for alignments, to have a consistent frame of reference, we recommend the use of *Danio rerio* (zebrafish) as a reference species for fish eDNA studies. This is because *D. rerio* is a widely used model organism for experimental, especially developmental, studies with a fully sequenced genome (Howe et al., 2013). Despite the lack of a reference species, structurally annotated reference alignments have been created for protein-coding, tRNA, rRNA, and noncoding genes from the mitogenomes of 250 fish species (Satoh et al., 2016). These reference alignments, along with a universal reference species, will promote standardization of primer names and sequence positions, thus improving existing annotations for fish marker gene sequences.

In summary, Mitohelper is a tool aimed at improving fish mitochondrial sequence annotations and at evaluating taxonomy and gene region coverage of currently available sequences. As fish eDNA research continues to grow, there is an impetus to improve the capabilities of eDNA-specific data resources and analysis tools. We also emphasize the importance of simultaneous barcoding and metabarcoding efforts to increase the sequence and taxonomic coverage of fish reference databases for accurate taxonomic classification and diversity predictions. Future fish eDNA studies will benefit from comprehensive reference data resources, as well as standardized annotation and analysis pipelines, which facilitates scientific discovery, reproducibility, and comparability.

ACKNOWLEDGMENTS

This research was carried out [in part] under the auspices of the Cooperative Institute for Marine and Atmospheric Studies (CIMAS), a Cooperative Institute of the University of Miami, and the National Oceanic and Atmospheric Administration (NOAA), cooperative agreement # NA20OAR4320472. This work was also supported by award NA06OAR4320264 06111039 to the Northern Gulf Institute (NGI) at Mississippi State University from NOAA's Office

of Oceanic and Atmospheric Research (OAR), U. S. Department of Commerce.

AUTHOR CONTRIBUTIONS

SJL wrote and tested the software tool, curated the databases, analyzed and interpreted output from the software tool, and wrote the manuscript; LRT conceived of the software tool, provided guidance to its creation, and edited the manuscript.

DATA AVAILABILITY STATEMENT

Mitohelper is publicly available on GitHub at https://github.com/aomlomics/mitohelper. Raw data used to build the Mitohelper repository were downloaded from MitoFish, NCBI, and the *Fishes of the World* homepage (URLs in Materials and Methods section). The data processing workflow (including all scripts) and Mitohelper's algorithm are documented in the developer wiki (https://github.com/aomlomics/mitohelper/wiki). Reference datasets, including Mitohelper-compatible and QIIME-compatible datasets, are available at http://doi.org/10.5281/zenodo.4458571.

ORCID

Shen Jean Lim https://orcid.org/0000-0003-4578-5318
Luke R. Thompson https://orcid.org/0000-0002-3911-1280

REFERENCES

Andruszkiewicz, E. A., Sassoubre, L. M., & Boehm, A. B. (2017). Persistence of marine fish environmental DNA and the influence of sunlight. *PLoS One*, *12*(9), e0185043. https://doi.org/10.1371/journal.pone.0185043

Arranz, V., Pearman, W. S., Aguirre, J. D., & Liggins, L. (2020). MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. *Scientific Data*, *7*(1), 209. https://doi.org/10.1038/s41597-020-0549-9

Bakker, J., Wangensteen, O. S., Chapman, D. D., Boussarie, G., Buddo, D., Guttridge, T. L., Hertler, H., Mouillot, D., Vigliola, L., & Mariani, S. (2017). Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. *Scientific Reports*, *7*(1), 16886. https://doi.org/10.1038/s41598-017-17150-2

Berry, T. E., Saunders, B. J., Coghlan, M. L., Stat, M., Jarman, S., Richardson, A. J., Davies, C. H., Berry, O., Harvey, E. S., & Bunce, M. (2019). Marine environmental DNA biomonitoring reveals seasonal patterns in biodiversity and identifies ecosystem responses to anomalous climatic events. *PLOS Genetics*, *15*(2), e1007943. https://doi.org/10.1371/journal.pgen.1007943

Bessey, C., Jarman, S. N., Berry, O., Olsen, Y. S., Bunce, M., Simpson, T., Power, M., McLaughlin, J., Edgar, G. J., & Keesing, J. (2020). Maximizing fish detection with eDNA metabarcoding. *Environmental DNA*, *2*(4), 493–504. https://doi.org/10.1002/edn3.74

BLAST® Command Line Applications User Manual (2008). *National Center for Biotechnology Information (US)*. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK279684

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

Bylemans, J., Gleeson, D. M., Hardy, C. M., & Furlan, E. (2018). Toward an ecoregion scale evaluation of eDNA metabarcoding primers: A case study for the freshwater fish biodiversity of the Murray-Darling Basin (Australia). *Ecology and Evolution*, *8*(17), 8697–8712. https://doi.org/10.1002/ece3.4387

Castañeda, R. A., Van Nynatten, A., Crookes, S., Ellender, B. R., Heath, D. D., MacIsaac, H. J., Mandrak, N. E., & Weyl, O. L. F. (2020). Detecting native freshwater fishes using novel non-invasive methods. *Frontiers in Environmental. Science*, *8*(29), https://doi.org/10.3389/fenvs.2020.00029

Cole, J. R., Wang, Q., Cardenas, E., Fish, J., Chai, B., Farris, R. J., Kulam-Syed-Mohideen, A. S., McGarrell, D. M., Marsh, T., Garrity, G. M., & Tiedje, J. M. (2009). The Ribosomal Database Project: Improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, *37*, D141–145. https://doi.org/10.1093/nar/gkn879

Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., & Meyer, R. S. (2019). Anacapa toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210x.13214

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

Djurhuus, A., Closek, C. J., Kelly, R. P., Pitz, K. J., Michisaki, R. P., Starks, H. A., Walz, K. R., Andruszkiewicz, E. A., Olesin, E., Hubbard, K., Montes, E., Otis, D., Muller-Karger, F. E., Chavez, F. P., Boehm, A. B., & Breitbart, M. (2020). Environmental DNA reveals seasonal shifts and potential interactions in a marine community. *Nature Communications*, *11*(1), 254. https://doi.org/10.1038/s41467-019-14105-1

Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., & Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*(13), 3101–3119. https://doi.org/10.1111/mec.13660

Hebert, P. D. N., Ratnasingham, S., & Waard, J. R. D. (2003). Barcoding animal life: Cytochrome *c* oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(suppl_1), S96–S99. https://doi.org/10.1098/rsbl.2003.0025

Holman, L. E., de Bruyn, M., Creer, S., Carvalho, G., Robidart, J., & Rius, M. (2019). Detection of introduced and resident marine species using environmental DNA metabarcoding of sediment and water. *Scientific Reports*, *9*(1), 11559. https://doi.org/10.1038/s41598-019-47899-7

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., McLaren, K., Matthews, L., McLaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., … Stemple, D. L. (2013). The zebrafish reference genome sequence and its relationship to the human genome. *Nature*, *496*(7446), 498–503. https://doi.org/10.1038/nature12111

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., & Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, *30*(11), 2531–2540. https://doi.org/10.1093/molbev/mst141

Jo, T., Murakami, H., Yamamoto, S., Masuda, R., & Minamoto, T. (2019). Effect of water temperature and fish biomass on environmental DNA

shedding, degradation, and size distribution. *Ecology and Evolution*, *9*(3), 1135–1146. https://doi.org/10.1002/ece3.4802

Kelly, R. P., Closek, C. J., O'Donnell, J. L., Kralj, J. E., Shelton, A. O., & Samhouri, J. F. (2017). Genetic and manual survey methods yield different and complementary views of an ecosystem. *Frontiers in Marine. Science*, *3*(283), https://doi.org/10.3389/fmars.2016.00283

Kelly, R. P., Port, J. A., Yamahara, K. M., & Crowder, L. B. (2014). Using environmental DNA to census marine fishes in a large mesocosm. *PLoS One*, *9*(1), e86175. https://doi.org/10.1371/journal.pone.0086175

Lacoursière-Roussel, A., Rosabal, M., & Bernatchez, L. (2016). Estimating fish abundance and biomass from eDNA concentrations: Variability among capture methods and environmental conditions. *Molecular Ecology Resources*, *16*(6), 1401–1414. https://doi.org/10.1111/1755-0998.12522

Lafferty, K. D., Garcia-Vedrenne, A. E., McLaughlin, J. P., Childress, J. N., Morse, M. F., & Jerde, C. L. (2020). At Palmyra Atoll, the fish-community environmental DNA signal changes across habitats but not with tides. *Journal of Fish Biology*, 1–11. https://doi.org/10.1111/jfb.14403

Lane, D. J., Pace, B., Olsen, G. J., Stahl, D. A., Sogin, M. L., & Pace, N. R. (1985). Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences*, *82*(20), 6955–6959. https://doi.org/10.1073/pnas.82.20.6955

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

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

Machida, R. J., Leray, M., Ho, S. L., & Knowlton, N. (2017). Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*, 170027. https://doi.org/10.1038/sdata.2017.27

McKinney, W. (2010). *Data structures for statistical computing in python*,. In S. M. van der Walt, Jarrod (Ed.), Proceedings of the 9th Python in Science Conference.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7), 150088. https://doi.org/10.1098/rsos.150088

Murakami, H., Yoon, S., Kasai, A., Minamoto, T., Yamamoto, S., Sakata, M. K., Horiuchi, T., Sawada, H., Kondoh, M., Yamashita, Y., & Masuda, R. (2019). Dispersion and degradation of environmental DNA from caged fish in a marine environment. *Fisheries Science*, *85*(2), 327–337. https://doi.org/10.1007/s12562-018-1282-6

Nelson, J. S., Grande, T. C., & Wilson, M. V. H. (2016). *Fishes of the World* (5th ed.). John Wiley & Sons Inc..

Palumbi, S. R., Sandifer, P. A., Allan, J. D., Beck, M. W., Fautin, D. G., Fogarty, M. J., Halpern, B. S., Incze, L. S., Leong, J.-A., Norse, E., Stachowicz, J. J., & Wall, D. H. (2009). Managing for ocean biodiversity to sustain marine ecosystem services. *Frontiers in Ecology and the Environment*, *7*(4), 204–211. https://doi.org/10.1890/070135

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Research*, *41*, D590–596. https://doi.org/10.1093/nar/gks1219

Ratnasingham, S., & Hebert, P. D. (2007). bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355-364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Sanner, M. F. (1999). Python: A programming language for software integration and development. *Journal of Molecular Graphics and Modelling*, *17*(1), 57–61.

Sato, Y., Miya, M., Fukunaga, T., Sado, T., & Iwasaki, W. (2018). MitoFish and MiFish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, *35*(6), 1553–1555. https://doi.org/10.1093/molbev/msy074

Satoh, T. P., Miya, M., Mabuchi, K., & Nishida, M. (2016). Structure and variation of the mitochondrial genome of fishes. *BMC Genomics*, *17*(1), 719. https://doi.org/10.1186/s12864-016-3054-y

Sayers, E. W., Agarwala, R., Bolton, E. E., Brister, J. R., Canese, K., Clark, K., Connor, R., Fiorini, N., Funk, K., Hefferon, T., Holmes, J. B., Kim, S., Kimchi, A., Kitts, P. A., Lathrop, S., Lu, Z., Madden, T. L., Marchler-Bauer, A., Phan, L., … Ostell, J. (2019). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *47*(D1), D23–D28. https://doi.org/10.1093/nar/gky1069

Simpfendorfer, C. A., Kyne, P. M., Noble, T. H., Goldsbury, J., Basiita, R. K., Lindsay, R., Shields, A., Perry, C., & Jerry, D. R. (2016). Environmental DNA detects critically endangered largetooth sawfish in the wild. *Endangered Species Research*, *30*, 109–116. https://doi.org/10.3354/esr00731

Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., & Harvey, E. S. (2019). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. *Conservation Biology*, *33*(1), 196–205. https://doi.org/10.1111/cobi.13183

Stoeckle, B. C., Beggel, S., Cerwenka, A. F., Motivans, E., Kuehn, R., & Geist, J. (2017). A systematic approach to evaluate the influence of environmental conditions on eDNA detection success in aquatic ecosystems. *PLoS One*, *12*(12), e0189119. https://doi.org/10.1371/journal.pone.0189119

Stoeckle, M. Y., Das Mishu, M., & Charlop-Powers, Z. (2018). GoFish: A versatile nested PCR strategy for environmental DNA assays for marine vertebrates. *PLoS One*, *13*(12), e0198717. https://doi.org/10.1371/journal.pone.0198717

Stoeckle, M. Y., Das Mishu, M., & Charlop-Powers, Z. (2020). Improved environmental DNA reference library detects overlooked marine fishes in New Jersey, United States. *Frontiers in Marine Science*, *7*(226), https://doi.org/10.3389/fmars.2020.00226

Stoeckle, M. Y., Soboleva, L., & Charlop-Powers, Z. (2017). Aquatic environmental DNA detects seasonal fish abundance and habitat preference in an urban estuary. *PLoS One*, *12*(4), e0175186. https://doi.org/10.1371/journal.pone.0175186

Thomsen, P. F., Møller, P. R., Sigsgaard, E. E., Knudsen, S. W., Jørgensen, O. A., & Willerslev, E. (2016). Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. *PLoS One*, *11*(11), e0165252. https://doi.org/10.1371/journal.pone.0165252

Townsend, H., Harvey, C. J., deReynier, Y., Davis, D., Zador, S. G., Gaichas, S., Weijerman, M., Hazen, E. L., & Kaplan, I. C. (2019). Progress on implementing ecosystem-based fisheries management in the United States through the use of ecosystem models and analysis. *Frontiers in Marine. Science*, *6*, 641, https://doi.org/10.3389/fmars.2019.00641

Wangensteen, O. S., & Turon, X. (2017). Metabarcoding techniques for assessing biodiversity of marine animal forests. In S. Rossi, L. Bramanti, A. Gori, & C. Orejas (Eds.), *Marine animal forests: The ecology of benthic biodiversity hotspots* (pp. 445–473). Springer International Publishing.

Waraniak, J. M., Marsh, T. L., & Scribner, K. T. (2019). 18S rRNA metabarcoding diet analysis of a predatory fish community across seasonal changes in prey availability. *Ecology and Evolution*, *9*(3), 1410–1430. https://doi.org/10.1002/ece3.4857

Waskom, M., & Team, s. d. (2020). *mwaskom/seaborn*. Zenodo. doi:10.5281/zenodo.592845

Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Minamoto, T., & Miya, M. (2017). Environmental DNA metabarcoding

reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, *7*, 40368. https://doi.org/10.1038/srep40368

Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: improvements for better sequence analysis. *Nucleic Acids Research*, *34*, W6–W9. https://doi.org/10.1093/nar/gkl164

Zinger, L., Gobet, A., & Pommier, T. (2012). Two decades of describing the unseen majority of aquatic microbial diversity. *Molecular Ecology*, *21*(8), 1878–1896. https://doi.org/10.1111/j.1365-294X.2011.05362.x

How to cite this article: Jean Lim S, Thompson LR. Mitohelper: A mitochondrial reference sequence analysis tool for fish eDNA studies. *Environmental DNA*. 2021;3:706–715. https://doi.org/10.1002/edn3.187