Received: 26 October 2022 | Revised: 8 May 2023 | Accepted: 9 May 2023

DOI: 10.1002/edn3.433

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

WILEY

METHOD

An automated workflow to assess completeness and curate GenBank for environmental DNA metabarcoding: The marine fish assemblage as case study

Cristina Claver | Oriol Canals | Leire G. de Amézaga | Iñaki Mendibil | Naiara Rodriguez-Ezpeleta

AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Sukarrieta, Bizkaia, Spain

Correspondence
Cristina Claver and Naiara Rodriguez-Ezpeleta, AZTI, Marine Research, Basque Research and Technology Alliance (BRTA), Sukarrieta, Bizkaia, Spain.
Email: cclaver@azti.es and nrodriguez@azti.es

Funding information
Eusko Jaurlaritza, Grant/Award Number: GENGES project; H2020 European Institute of Innovation and Technology, Grant/Award Number: 817806; Hezkuntza, Hizkuntza Politika Eta Kultura Saila, Eusko Jaurlaritza, Grant/Award Number: Predoctoral

Abstract

To successfully implement environmental DNA-based (eDNA) diversity monitoring, the completeness and accuracy of reference databases used for taxonomic assignment of eDNA sequences are among the challenges to be tackled. Here, we have developed a workflow that evaluates the current status of GenBank for marine fishes. For a given combination of species and barcodes, a gap analysis is performed and potentially erroneous sequences are identified. Our gap analysis based on the four most used genes (cytochrome c oxidase subunit 1, 12S rRNA, 16S rRNA, and cytochrome b) for fish eDNA metabarcoding found that COI, the universal choice for metazoans, is the gene covering the highest number of Northeast Atlantic marine fishes (70%), while 12S rRNA, the preferred region for fish-targeting studies, only covers about 50% of the species. The presence of too close and too distant barcode sequences as expected by their taxonomic classification confirms the existence of erroneous sequences in GenBank that our workflow can detect and eliminate. Comparing taxonomic assignments of real marine eDNA samples with raw and clean reference databases for the most used 12S rRNA barcodes (*teleo* and *MiFish*), we confirmed that both barcodes perform differently and demonstrated that the application of the database cleaning workflow can result in drastic changes in community composition. Besides providing a tool for reference database curation, this study confirms the need to increase 12S rRNA reference sequences for European marine fishes and evidences the dangers of taxonomic assignments by directly querying GenBank. We have developed a workflow that evaluates the current status of GenBank for marine fishes. For a given combination of species and barcodes, a gap analysis is performed and potentially erroneous sequences are identified.

KEYWORDS
12S rRNA, environmental DNA, metabarcoding, *MiFish*, reference database, *teleo*

1 | INTRODUCTION

Environmental DNA (eDNA) metabarcoding studies are often based on public reference databases on whose accuracy and completeness lies the reliability of taxonomic assignment (Richardson et al., 2018; Virgilio et al., 2010). Some public databases have filtering options and analysis tools available for quality controls, such as the trackability of the voucher specimens, but are focused on specific regions (e.g., BOLD (Ratnasingham & Hebert, 2007) mostly covers the COI gene). This severely limits their use when targeting taxa for which the best-performing primers are located in other regions, such as ribosomal genes. The most complete reference database is GenBank (Benson et al., 2012), but it acts as a mere sequence repository and its unverified submission process often leads to misannotated sequences (Steinegger & Salzberg, 2020). Although the reliability of GenBank for a range of DNA-based monitoring applications has been praised (Leray et al., 2019), it has also been contested (Locatelli et al., 2020). If the number of expected species in the study area is modest, sequences of the species of interest can be downloaded and manually curated to remove misannotated sequences, and barcoding of the missing species in public repositories can be performed to complete the reference database (Collins et al., 2021; Thomsen et al., 2016; West et al., 2021). On the other hand, when the number of species expected in a region is very large (e.g., fishes in a large marine area), manual curation of the database is unviable (Leray et al., 2020) and incompleteness of the database is expected (Weigand et al., 2019). Incompleteness of reference databases can lead to false-negative detection, leading to, for example, failure in detecting alien species (Klymus et al., 2017); inaccuracy of reference databases can lead to false-positive detections, resulting, for example, in incorrectly reporting species presence (Port et al., 2016).

Reference database completeness and accuracy are especially relevant for marine fishes, with about 20.000 described species (WoRMS Editorial Board, 2022), which makes manual curation and completion of databases difficult. eDNA metabarcoding is being increasingly applied to their study during the last years (Fraija-Fernández et al., 2020; Tsuji et al., 2019) and has arisen as a promising, alternative tool for monitoring this important resource (Gilbey et al., 2021). Fish eDNA metabarcoding studies have been conducted using a variety of barcodes, the most common ones being those based on the mitochondrial cytochrome b (cytb), small (12S) and large (16S) subunit ribosomal RNA (rRNA), and cytochrome c oxidase subunit 1 (COI) genes (Zhang et al., 2020). From these, the COI-based barcodes are considered standard for animal metabarcoding studies (Leray et al., 2013; Vrijenhoek, 1994), have been sequenced for a broad range of European marine fishes (Weigand et al., 2019), and curated reference databases are available (Oliveira et al., 2016). However, because eDNA extracted from water samples contains traces of many abundant organisms other than fish, the use of COI metazoan universal primers results in over-amplification of non-target taxa (Collins et al., 2019; Fraija-Fernández et al., 2020), and fish-specific primers have been developed, mostly based on the 12S rRNA gene, such as those amplifying the *teleo* (Valentini et al., 2016) or *MiFish* (Miya et al., 2015) regions

Studies using marine water eDNA metabarcoding to assess fish diversity based on 12S rRNA perform taxonomic assignment in a variety of ways. Some authors assign taxonomy by directly querying GenBank (e.g., Lamy et al., 2021; Sato et al., 2021; Zhou et al., 2022), which might result in erroneous assignments due to the presence of problematic records (Li et al., 2018). Others rely on filtered versions of GenBank containing only the target barcode of the sequences from the species of interest. This filtering can be done either using the information in the record definition (Arranz et al., 2020; Barco et al., 2022; Gold et al., 2021; Iwasaki et al., 2013; Machida et al., 2017; Mariani et al., 2021; Russo et al., 2021) or based on similarity searches (Heller et al., 2018; Leray et al., 2022). Yet, although the use of these filtered versions is popular in marine fish eDNA metabarcoding studies (Kawato et al., 2021; Kume et al., 2021; Nguyen et al., 2020; Oka et al., 2021; Polanco et al., 2021), the methods used to extract the sequences are not meant to remove potential contaminations other than those identifiable through their labelling (e.g., Arranz et al., 2020). Finally, other studies attempt to reduce potentially erroneous sequences by visual inspection of phylogenetic trees (e.g. Canals et al., 2021; Collins et al., 2019, 2021; Fraija-Fernández et al., 2020), but this approach is tedious, not viable for databases composed by a large number of species, and limited by the low phylogenetic resolution of short barcodes (Polanco et al., 2021; Zhang et al., 2020). Thus, more dynamic screening tools are necessary to overcome reference database quality issues and meet the high expectations concerning global biodiversity eDNA monitoring. Here, to assist future marine fish eDNA metabarcoding studies, we have developed an automated workflow to (i) perform a gap analysis of GenBank for a list of species of interest, (ii) create a reference database of specific barcodes for the species of interest, and (iii) detect and eliminate the most obvious spurious sequences. As a study case, we have applied this workflow to the fish inhabiting the European Marine Regions. We have assessed the gaps for COI and 12S rRNA-based barcodes, generated a curated reference database for the most widely used (i.e., *teleo* and *MiFish*) regions from the 12S rRNA gene, and compared the performance of the taxonomic assignment using the reference database before and after database curation on marine eDNA samples. Finally, we contribute to the reference database completeness by barcoding the 12S rRNA sequence of 21 different fish species. This newly developed workflow, which can be applied to any mitochondrial barcode and set of species, and results derived from it constitute a step ahead for increasing the completeness and accuracy of reference databases for marine fish eDNA metabarcoding studies. This, together with additional barcoding efforts to populate reference databases, is a major milestone for making fish eDNA biomonitoring reliable and trustworthy.

2 | MATERIALS AND METHODS

A summary of the procedures followed is presented in Figure 1 and all the scripts used are available on GitHub (https://github.com/rodriguez-ezpeleta/NEA_fish_DB).

2.1 | Fish checklist assembly and reference sequence retrieval

The list of fish species present in the northeast Atlantic and adjacent seas was assembled from FishBase (Froese & Pauly, 2022) by retrieving the species occurring in the European Marine Regions (Baltic Sea, Barents Sea, Black Sea, Canary Current, Celtic Biscay Shelf, Faroe Plateau, Greenland Sea, Iberian Coastal, Iceland Shelf and Sea, North Sea, Norwegian Sea, and Mediterranean Sea) using the R package rfishbase (Boettiger et al., 2012), and taxonomy was extracted from World Register of Marine Species (WoRMS Editorial Board, 2022). All mitochondrial gene records available in GenBank for the species in the reference list were identified using eUtils (Sayers, 2008) and were assigned as belonging to one of the most common genes used in metazoan metabarcoding surveys, that is, cytochrome oxidase I (COI), cytochrome b (cytb), 12S rRNA (12S), and 16S rRNA (16S), based on their definition or, those with ambiguous definition, based on BLAST searches (Altschul et al., 1990) against complete COI, 12S, 16S, and/or cytb sequences, respectively; matches were considered if query sequences had ≥60% sequence similarity with the complete sequences.

2.2 | Barcoding gap analysis

The barcoding gap analysis was carried out for two barcodes of the two most widely used genes in fish metabarcoding studies: *mlCOI* (Leray et al., 2013) and *folCOI* (Vrijenhoek, 1994) from COI, and *teleo* (Valentini et al., 2016) and *MiFish* (Miya et al., 2015) from 12S rRNA. First, sequences from all COI and 12S rRNA records identified above were downloaded from GenBank. Using *mothur* (Schloss et al., 2009), these sequences were aligned against reference alignments of COI and 12S rRNA reference sequences (previously aligned with MAFFT (Katoh & Standley, 2013)) and trimmed to the 12S rRNA and COI regions. Then, complete *folCOI*, *mlCOI*, *teleo*, and *MiFish* barcode regions were identified using *cutadapt* (Martin, 2011), and partial sequences covering at least 90% of the barcode region were identified using *mothur* with the complete barcode regions as template. Both the complete and partial barcodes were kept for the barcoding gap analysis. Similarity matrices were calculated based on sequence similarity scores obtained by all-against-all BLAST analysis. Similarity value distributions were visualized in heatmaps for six different taxonomic categories: intraspecific (SP), intra-genus (GE), intra-family (FA), intra-order (OR), intra-class (CL), and intra-phylum (PH), and classified into five levels so that Level 1 comprises the range of intraspecific similarity values (excluding outliers) and Levels 2–5 comprise values resulting from dividing uniformly the range of values between the minimum similarity value and the lowest value of Level 1. Pairs with no BLAST hits among them because of not enough coverage or too distant were reported as "No dist."

2.3 | Automated curation of reference databases

To identify potentially erroneous sequences in the database, a series of rules were developed according to how sequences clustered

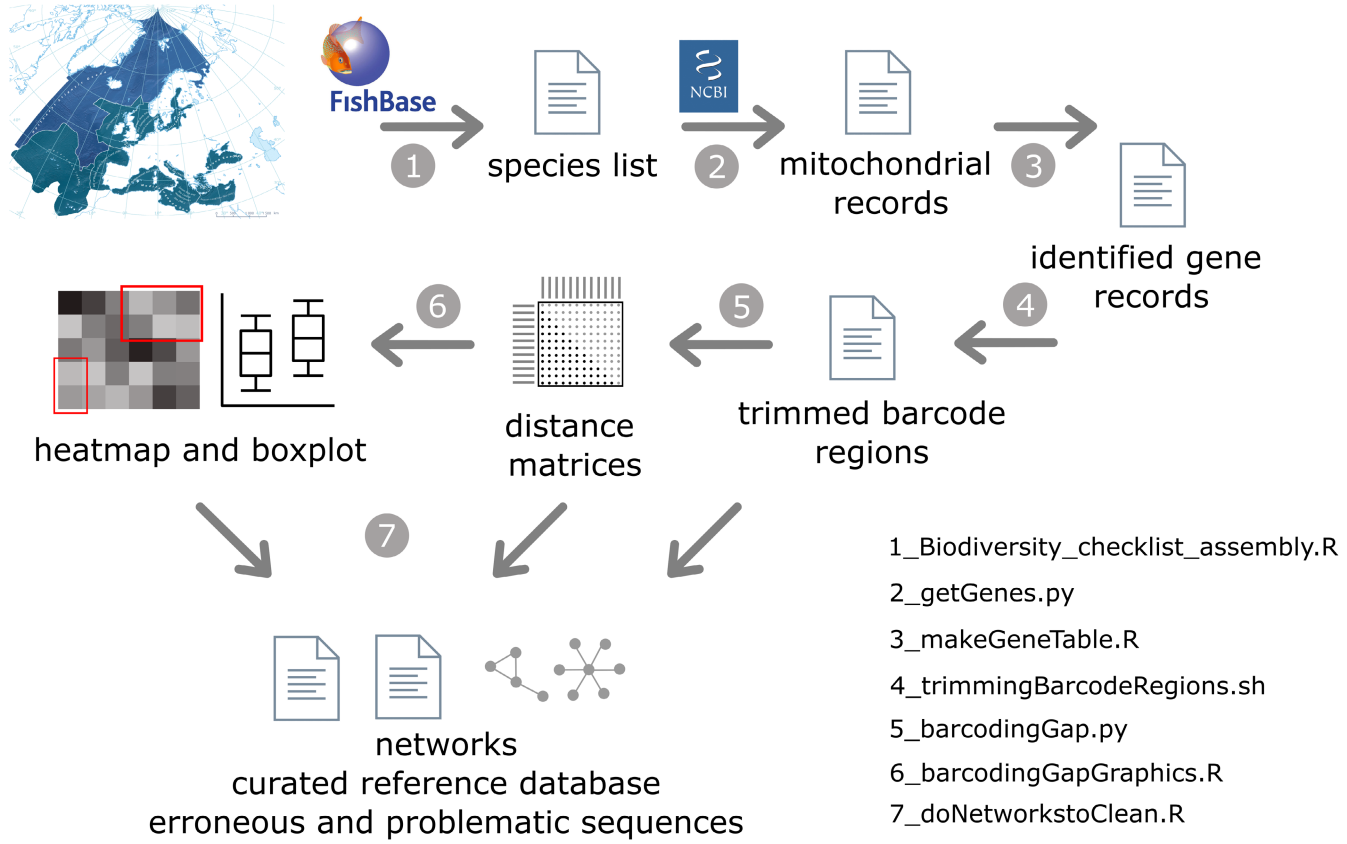

FIGURE 1 Schematic view of the workflow developed in this study. Numbers correspond to specific scripts.

within a given combination of level and taxonomic category, focusing on the squares far from the diagonal in heatmaps, which represent sequences that are too similar or too different given their taxonomy. We focused on the extremes, selecting levels 3 to 5 for intraspecific relationships (i.e., too distant pairs) and levels 1 and 2 for intra-phylum, intra-class, and intra-order interspecific relationships (i.e., too similar pairs). For each chosen taxonomic classification and level, independent networks were created, and the decision tree method developed (Figure S1) was applied to tag sequences as correct, erroneous, or problematic on the basics of how they clustered in a network of "too similar" or "too different" sequences, respectively. Using this decision tree, sequences more similar to sequences of other classes or orders than to sequences of the same species, genus, or family are tagged as erroneous if there is enough information to conclude which of the sequences is potentially erroneous within the network and tagged as problematic when the information is not enough to resolve it. Networks that are too complicated to resolve by the decision tree can be visually inspected, combined with other evidence such as blast searches or phylogenetic trees, and manually tag the sequences as correct, erroneous, or problematic. Erroneous-tagged sequences are removed from the database and erroneous and problematic-tagged sequences are compiled in two independent lists including the reason for their classification as erroneous or problematic. Finally, a curated database is created and outputted in *fasta* and *tax* formats, which are the files required for the posterior use of the database for taxonomic classification.

2.4 | Amplicon data generation, bioinformatic processing and analysis

We analyzed marine water samples with the two most used barcodes in fish eDNA metabarcoding studies (i.e., *teleo* and *MiFish*). For that aim, 30 5-L water samples were collected at different locations, time, and depths in the Bay of Biscay (Figure S2). Water filtering, DNA extraction, and amplification with the *teleo* primer pair (Valentini et al., 2016) were performed as described in Fraija-Fernández et al. (2020). An approximate DNA volume of 100 μL was extracted from each sample. The concentration of the extractions was calculated, and the DNA concentration of the samples was homogenized to 5 ng/μl by diluting samples exceeding the desired concentration. For both primer pairs, three replicate PCR amplifications were done per sample in a final volume of 20 μL, including 10 μL of KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 0.4 μL of each amplification primer (final concentration of 0.2 μM), 7.2 μL of Milli-Q water, and 2 μL (10 ng) template DNA. The thermocycling profile for PCR amplification with *MiFish* primer pair (Miya et al., 2015) included 3 min at 95°C; 35 cycles of 20 s at 98°C, 15 s at 60°C, and 15 s at 72°C; and finally, 5 min at 72°C. Replicate PCR products were combined and purified using AMPure XP beads (Beckman Coulter) following manufacturer's instructions and used as templates for the generation of 12×8 dual-indexed amplicons in the second PCR following the "16S Metagenomic Sequence Library Preparation" protocol (Illumina) using the Nextera XT Index Kit (Illumina). PCR-negative controls resulted in no visible amplification in agarose gels. Multiplexed PCR products were purified using the AMPure XP beads, quantified using Quant-iT dsDNA HS assay kit using a Qubit® 2.0 Fluorometer (Life Technologies), and adjusted to 4 nM. Then, 5 μL of each sample were pooled, checked for size and concentration using the Agilent 2100 bioanalyzer (Agilent Technologies), sequenced using the 2×300 paired-end protocol on the Illumina MiSeq platform (Illumina), and demultiplexed based on their barcode sequences. The quality of demultiplexed reads was verified with FASTQC (Andrews, 2010). Primer pairs were removed using *cutadapt* (Martin, 2011), allowing a maximum error rate of 20%. Reads longer than 30 nucleotides and containing the two primer sequences were kept and merged using pear (Zhang et al., 2014) with a minimum overlap of 10 nucleotides for *MiFish* and 20 nucleotides for *teleo*. Pairs with average quality lower than 33 Phred score were removed with *Trimmomatic* (Bolger et al., 2014) and those reads shorter than 60 and 140 nucleotides for *teleo* and *MiFish*, respectively, not covering the target region or containing ambiguous positions were discarded using *mothur*. Potential chimeras were detected based on UCHIME (Edgar et al., 2011) and removed. Taxonomy was assigned to unique reads using the Bayesian classifier method (Wang et al., 2007) implemented in *mothur* (*cutoff*=60) using the *teleo* and *MiFish* databases before and after the automated curation process. Only reads assigned to species level were considered in subsequent steps. Ordination of communities (considering only shared species between both barcodes) was carried out using non-metric multidimensional scaling (NMDS; *metaMDS* function, vegan package version 4.1.1 (Oksanen et al., 2013)) analyses based on Bray–Curtis dissimilarities (*vegdist* function, vegan package). ANOSIM (analysis of similarity; Clarke (1993)) was used to test if samples were grouped according to the factor barcode (*anosim* function, vegan package).

2.5 | Generation of 12S rRNA sequences

Fin and muscle tissue samples from morphologically identified specimens (Table S1) were obtained during the CSIC SUMMER-2020 survey in the Western Mediterranean Sea (Balearic Islands, Alboran Sea, Gulf of Cadiz, and Atlantic Ocean) and from fishing vessels landing in the port of Ondarroa (Basque Country, Spain). For each sample, genomic DNA was extracted from muscle tissue or fin using the Wizard Genomic DNA Purification kit (Promega) following manufacturer's instructions for "Isolating Genomic DNA from Tissue Culture Cells and Animal Tissue." Extracted DNA was resuspended in Milli-Q water and its concentration was determined with NANODROP (Thermo Scientific™). The extracted DNA was then amplified using the *MarineFish* primer pair (Jin et al., 2013), a 900- to 1100-bp-long 12S rRNA region covering both *teleo* and *MiFish* regions, by mixing 10 μL of 2X PCR Master Mix (Fisher Scientific), 0.4 μL of each primer, 2 μL DNA template (1–20 ng), and 7.2 μL of nuclease-free water, and using the following amplification conditions: 95°C for 3 min; 35 cycles of denaturation at 95°C for 30 s, annealing at 56°C for 30 s, and extension at 72°C for 75 s; and final extension at 72°C for 10 min. The PCR products were migrated in a 2% agarose gel in TAE buffer

and purified using ILUSTRA EXOSTAR1-Step (Cytiva) following manufacturer's conditions and sent for Sanger sequencing. Forward and reverse sequences were merged and SeqTrace software (Stucky, 2012) was used for quality filtering (minimum confidence score of 30). Sequences were submitted to GenBank (accession numbers available in Table S1) and added to the above-generated *teleo* and *MiFish* reference databases.

3 | RESULTS

3.1 | Assessment of database completeness for the most used fish eDNA metabarcoding markers

The list of Northeast Atlantic and Mediterranean marine fishes compiled included 1791 species: 1603 Actinopterygii, 174 Elasmobranchii, 8 Holocephali, 4 Petromyzonti, and 2 Myxini (Table S2). In total, 1277, 1067, 1047, and 898 fish species have COI, 12S, 16S, and cytb gene records available, respectively, including 42,115 COI, 27,546 cytb, 8542 16S, and 6820 12S sequences (Figure 2a). The COI gene is the one with the highest number of sequences and species coverage (70%), and cytb, despite having the second highest number of sequences, is the one with the lowest species coverage (50%). This is due to a high number of cytb records belonging to a small number of species (e.g., Atlantic cod *Gadus morhua*, European anchovy *Engraulis encrasicolus*, or milkfish *Chanos chanos*). 12S and 16S rRNA exhibit similar species coverage values (about 60%). The COI-based barcodes (*folCOI*, *mlCOI*) have the highest species coverage (>70%), whereas 12S rRNA barcodes cover between 40% (*teleo*) and 48% (*MiFish*) of the species (Figure 2b,c). To increase the 12 rRNA-based barcode coverage, we have sequenced the *teleo* and *MiFish* regions of 21 species, from which 5 and 16 had none or only one of the barcodes available at the time of submission (**Table S1**).

3.2 | Using the barcoding gap principle for potential error detection in reference databases

Distance matrices resulted in more than half a billion sequence pair comparisons for both COI barcodes and about 8 and 5 million pair comparisons for *MiFish* and *teleo*, respectively. In all barcodes, the average pairwise similarity decreases as sequences belong to more distant taxonomic categories (Figure 3a), but an unexpected number of outliers representing low similarity in pairs of sequences belonging to the same species and high similarity in sequences belonging to taxonomically distant species are noticeable. The categorization of distance ranges in levels (Table S3) revealed the number of pairs that do not behave as expected according to the barcoding gap principle,

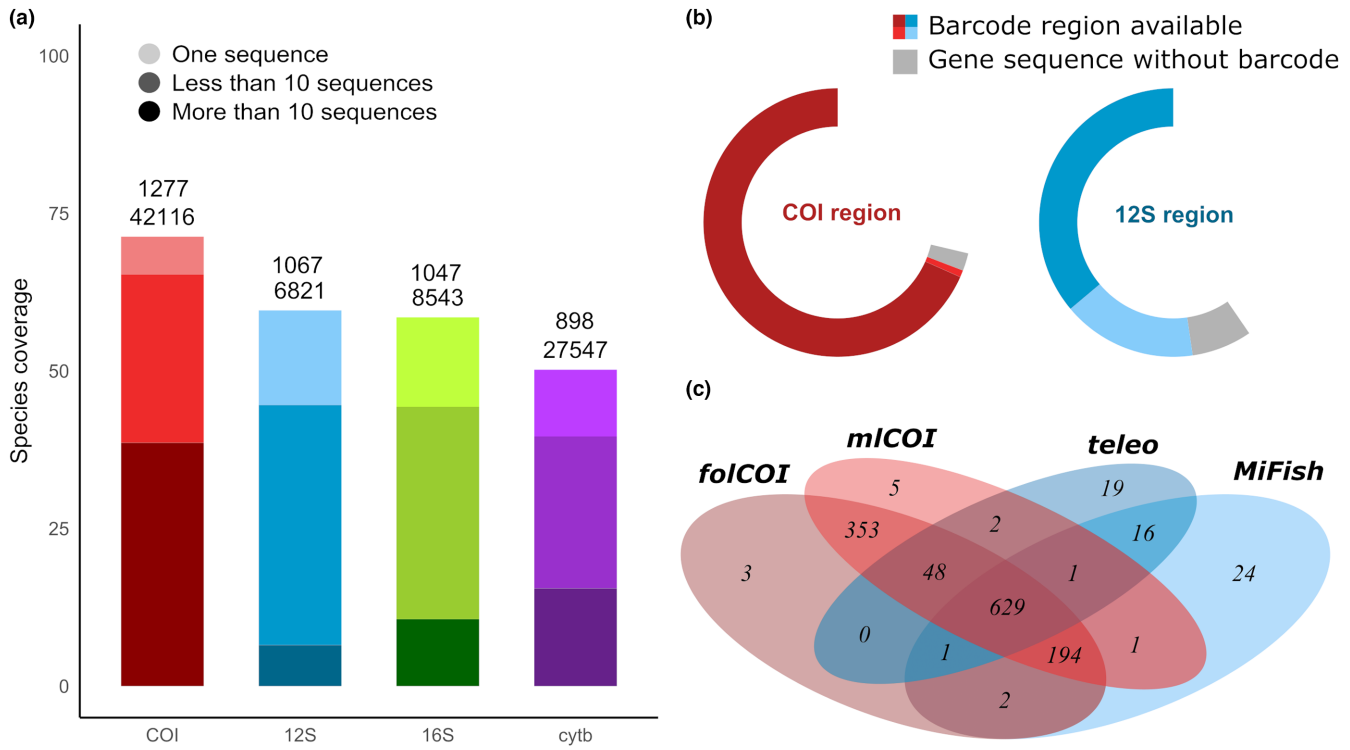

FIGURE 2 Reference database gap analysis. (a) Cumulative coverage (%) of European marine fishes for each gene. Numbers on bars indicate the number of species for which there are sequences available (above) and the total number of sequences available in GenBank (below) for each gene. (b) Barcode availability for COI and 12S gene markers. Dark red and dark blue represent portion of species with both barcodes available (i.e., *mlCOI* and *folCOI* for COI and MiFish and teleo for 12S). Light red and light blue represent portion of species with only one of the two barcodes available (i.e., *mlCOI* or *folCOI* for COI and MiFish or teleo for 12S). (c) Venn diagram showing the number of species with available references for *mlCOI*, *folCOI*, *teleo*, and *MiFish* barcodes.

that is, those that are too similar but belong to different species, or those that are too distant while being taxonomically close (at the most top-right and bottom-left squares of Figure 3b).

3.3 | Diagnosis and flagging of sequences by automatic screening

Our decision tree approach (see Methods; Figure S1) applied to networks within these pairs identified potentially erroneous and problematic sequences. Notably, our approach detected spurious sequences for both 12S rRNA barcodes (summarized in Tables S4 and S5). Broadly, three types of networks were distinguished according to their clustering structure. The first type consists of networks formed by a central sequence (both for intra- and interspecific relationships), where that central sequence is evaluated; for example, one central sequence of *Alburnus alburnus* was more similar to *Phoxinus phoxinus* sequences than to other sequences of *A. alburnus* for both *teleo* and *MiFish* (Figure 4a,b), so it was classified as potentially erroneous, and one central sequence of *Carcharodon carcharias* was identical to *Cetorhinus maximus* sequences (Figure 4c,d) while showing low similarity to other *C. carcharias* sequences and was also tagged as potentially erroneous. The second type consists of networks with no central sequence with only two species, where all sequences of the network are analyzed one by one; for example, we found *Engraulis encrasicolus* sequences being not only more similar than expected to *Istiophorus albicans* sequences but also very similar to each other (including those not represented in the network) and located within expected intraspecific distances (Figure 4e), so they were labeled as correct. Although for *I. albicans* there were no other sequences in the database, *I. albicans* sequences were more similar to *Engraulis* sequences than to other sequences from the *Istiophorus* genus, being thus, the *I. albicans* sequences labeled as erroneous. The third type consists of networks formed by more than two species and no central sequence, which cannot be analyzed automatically and require manual inspection; for example, a network with two non-gadoid sequences (belonging to *Argyropelecus gigas* and *Crystallogobius linearis*) that are identical to many *Gadidae* sequences (Figure 4f), where *C. linearis* sequence would be classified as erroneous for being more similar to sequences of the *Gadidae* family than to other sequences of the same species but the sequence of *A. gigas* would be classified as problematic due to lack of information to compare within the database because there are no more sequences for *A. gigas*, and no intra-genus relationships are available.

3.4 | Performance evaluation of raw and curated reference databases

For the 30 samples included in this study, we obtained 2,274,886 and 1,462,841 *MiFish* and *teleo* reads, respectively (Tables S6 and S7), from which ≈90% were assigned to the species level. For *teleo*, the Atlantic sailfish *Istiophorus albicans* represented 35% of the reads when using the raw database. Because *I. albicans* sequences were labeled as erroneous by our automated workflow in the *teleo* database, reads previously assigned to sailfish using the raw database were classified as anchovy with the clean database, leading to more coherent results across barcodes (Figure 5). A total of 94 species were identified in the study, from which 28 were detected by both barcodes (Figure 6). Although the relative abundance of some species (e.g., *Sprattus sprattus* and *Trachurus trachurus*) was substantially different in several samples, a positive correlation between the relative read abundance of most of the species detected by both barcodes is observed (Figure 7a). A total of 66 species were only detected by one of the barcodes, but they represented a very small percent of reads (<1%). Finally, the barcode used was not supported to be the main factor determining the fish community composition of the samples (Figure 7b) (ANOSIM test, R: 0.047, p-value: 0.0161).

4 | DISCUSSION

4.1 | Database completeness

Marine fishes constitute an important resource globally (FAO, F., 2012), whose management scale monitoring is costly and time consuming with traditional methods. Thus, eDNA metabarcoding has arisen as a promising, alternative tool applied in an increasing number of studies, including invasive species detection (Sepulveda et al., 2020), migration pattern discovery (Thalinger et al., 2019), or behavior assessment (Canals et al., 2021). In this context, the availability of curated and complete databases will be foremost for the uptake of eDNA-based approaches in fisheries monitoring. From the most used genes for fish metabarcoding, we confirm that COI-based barcodes considered standard for metazoans (Hebert et al., 2003) are the most abundant in GenBank (Porter & Hajibabaei, 2018). Yet, the non-fish taxa amplification in marine eDNA water samples (Collins et al., 2019; Fraija-Fernández et al., 2020) makes 12S rRNA-based barcodes more suitable (McClenaghan et al., 2020; Zhang et al., 2020) even with less species coverage in reference databases (Collins et al., 2019). We highlight that the 12S rRNA gene, although being the most used region for fishes, is only sequenced for half of the fish species inhabiting European marine waters, with the actual number of species available for specific barcodes (i.e., *teleo* and *MiFish*) even lower. This is due to *MiFish* and *teleo* barcodes being non-overlapping, so the existence of a 12S rRNA sequence for a given species does not imply that both regions are covered.

In our analyses, based on marine water samples, the difference in completeness of *MiFish* and *teleo* databases does not result in major differences in the overall community, since the most abundant species in our samples are present in both databases. While 70% of the species were detected by only one barcode, they represented a very small read abundance, likely reflecting rare DNA (Kelly et al., 2017; Stat et al., 2019). Despite ongoing efforts to increase the coverage of reference databases, sequences of key species are lacking, and

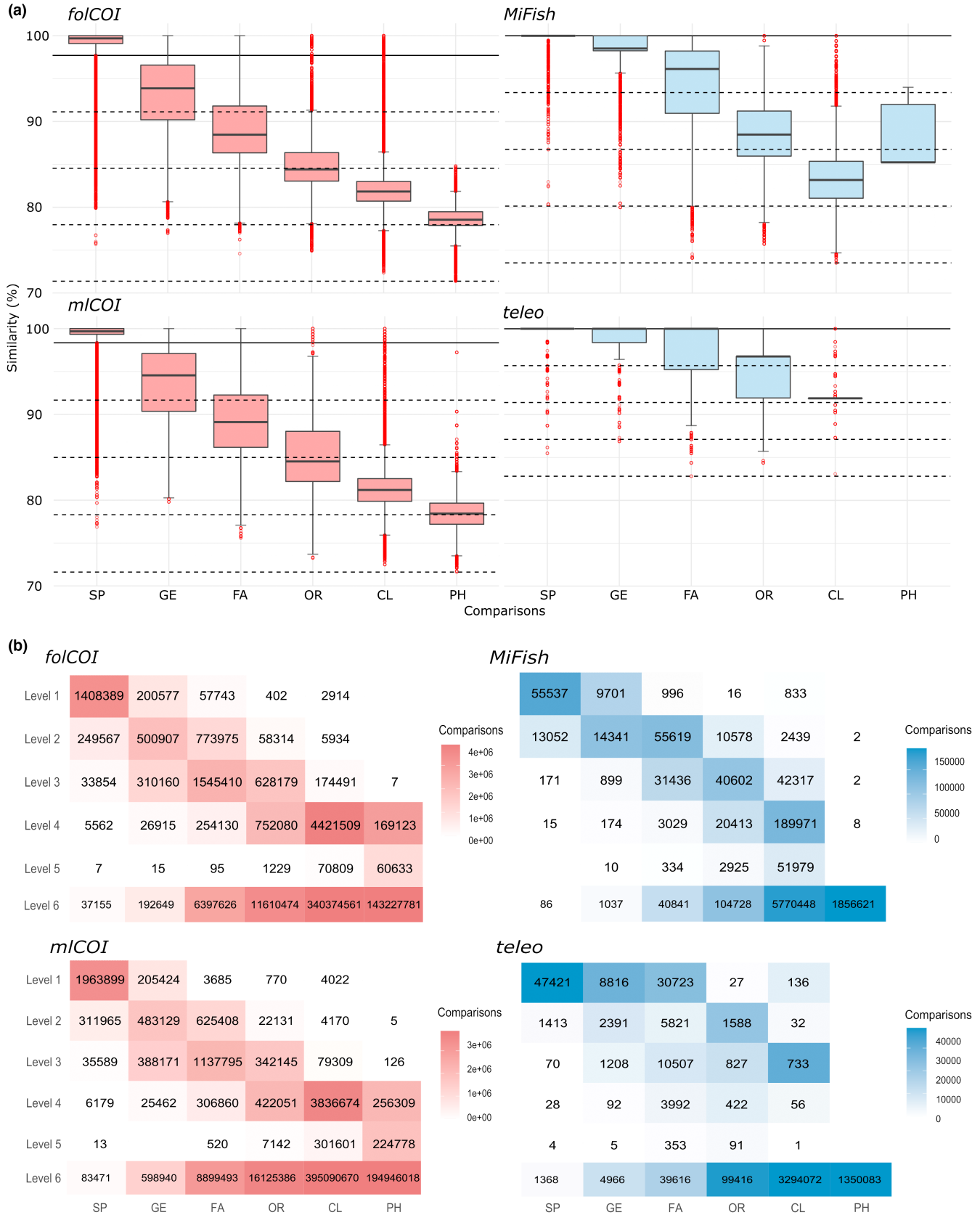

FIGURE 3 Barcoding gap analysis. (a) Boxplots depicting sequence pair distances (%) for intraspecific (SP) and interspecific divergences at different taxonomic levels (GE, FA, OR, CL, and PH). Levels are indicated with dashed horizontal lines (note that the values vary for each barcode). Outliers are represented with red dots. (b) Heatmaps representing number of pairs within different divergence levels.

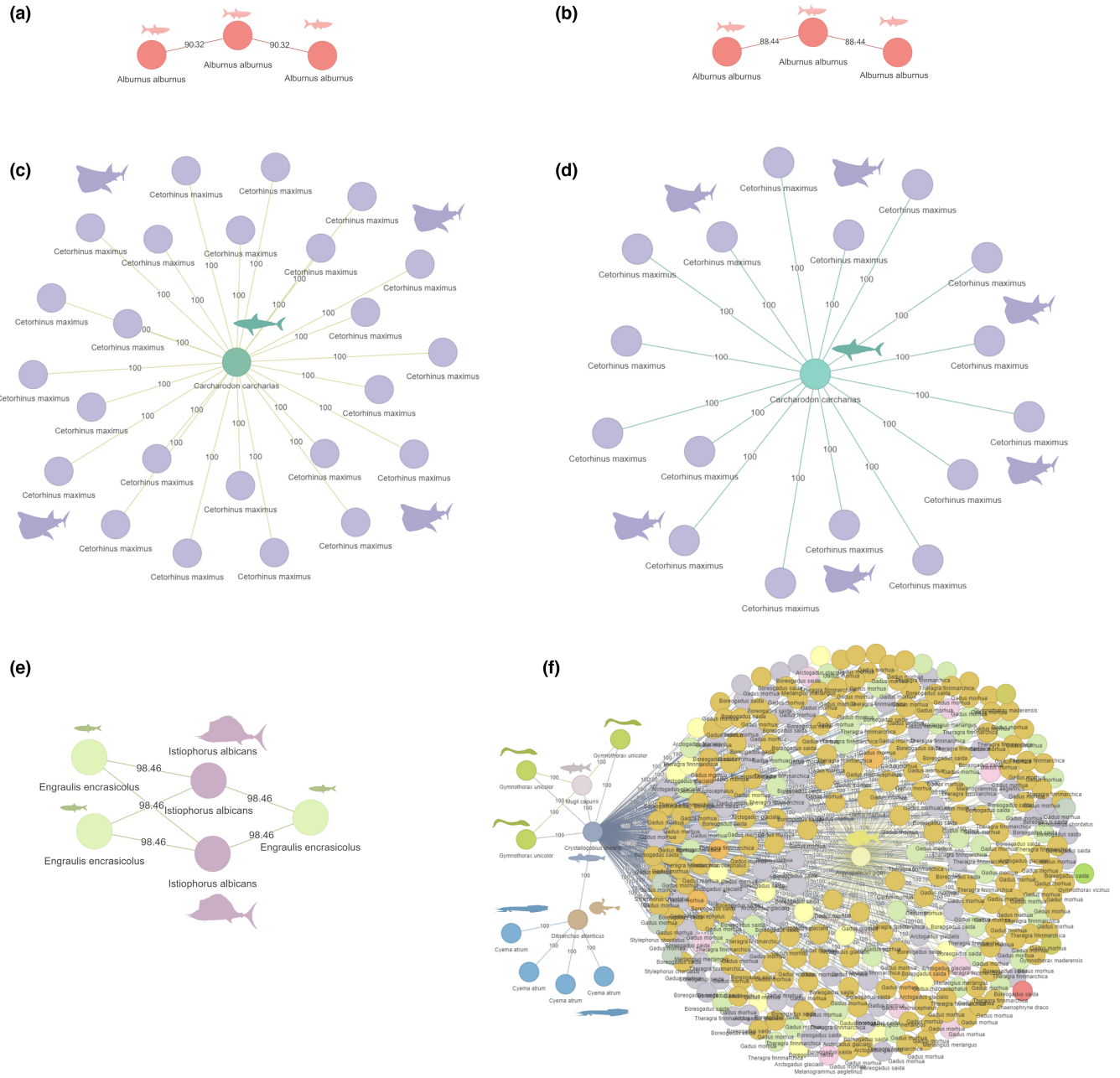

FIGURE 4 Examples of developed networks for *teleo* (left) and *MiFish* (right) databases. (a, b) Intraspecific distance analysis of the common bleak *Alburnus alburnus* sequences of (a) *MiFish* region in level 3 and (b) *teleo* region in level 4. (c, d) Interspecific (intra-order level) distance analysis of one sequence of the great white shark *Carcharodon carcharias* and sequences of the barking shark *Cetorhinus maximus* in level 1 of (c) *MiFish* and (d) *teleo* sequences. (e) Complex structure network formed by two species. (f) Complex multispecies structure network of interspecific (intra-class level) comparisons in Level 1.

shifting fish eDNA metabarcoding studies to remote areas with less known diversity, such as the deep sea, or specific applications, such as invasive species detection, increases its relevance. For instance, because in this study we have identified as erroneous the only two available records of *teleo* region for *Istiophorus albicans*, it turns out to be one more missing species in the database. To contribute to completing the 12S rRNA barcode reference database, required for present and future eDNA-based fish monitoring, we have barcoded both missing and poorly represented species, including deep sea and commercial fishes.

4.2 | Database accuracy

Public reference databases function as open sources of information where researchers submit their sequence data, enhancing reproducibility and transparency (Deiner et al., 2017; Leray et al., 2020). However, the free and open submission process is a "double-edged sword" because it leads to unverified record accumulation (Porter & Hajibabaei, 2018), some of which result in misannotated sequences (Steinegger & Salzberg, 2020). Contaminant amplification and data entry error cases in GenBank have been reported previously (Leray et al., 2019). Also,

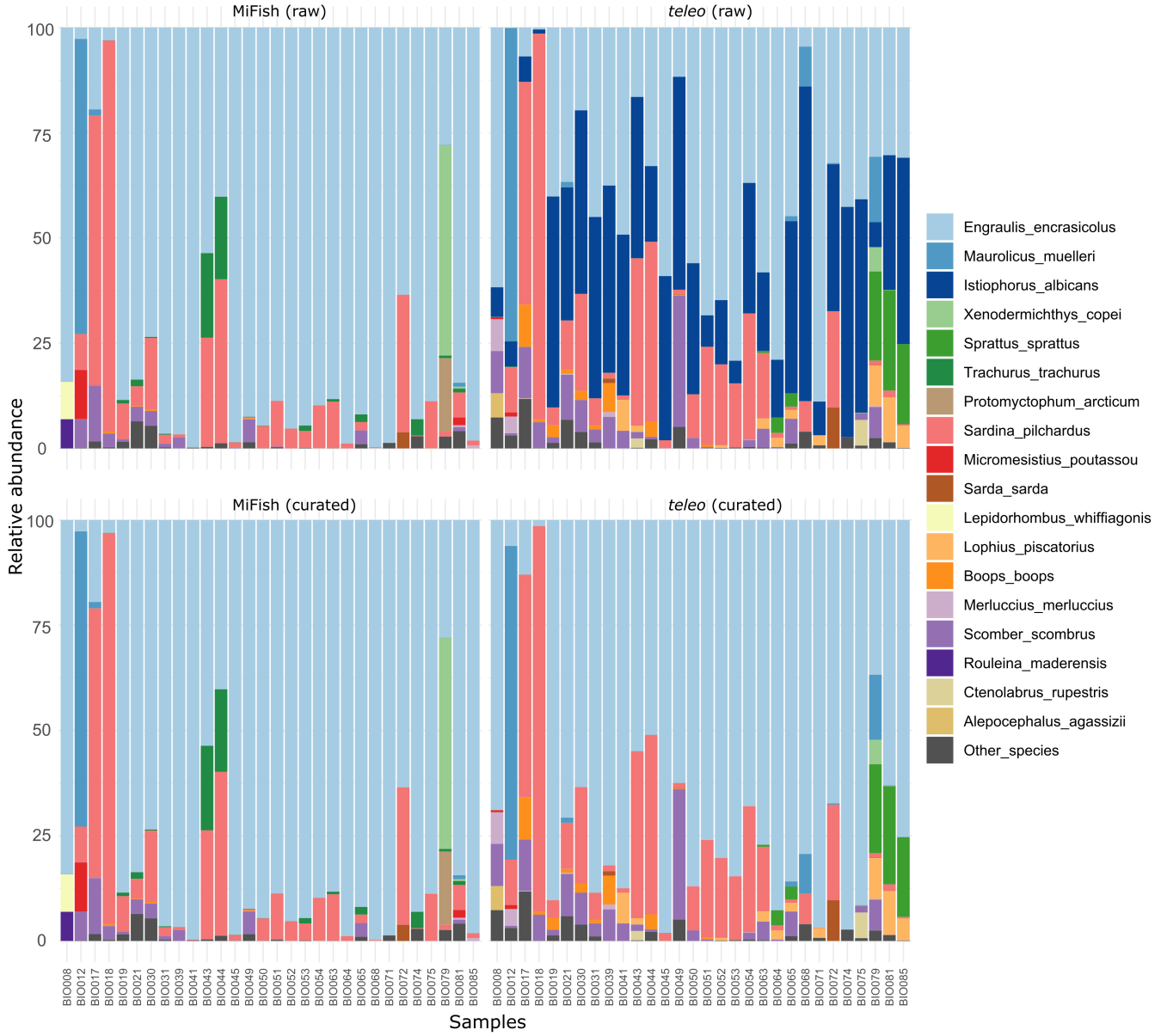

FIGURE 5 Barplots showing the read relative abundance for the 17 most abundant species using *MiFish* (left) and *teleo* (right) barcodes, performing the taxonomic assignment against raw (top) and curated (bottom) reference databases. Less abundant species are merged into "Other_species."

misidentifications of sampled specimens can occur, especially when referring to species with morphological similarities (Lyon et al., 2018), rare species, or individuals derived from fishery vessels (Figueiredo et al., 2020; Kirsch et al., 2018). This can be due to lack of taxonomists (Buyck, 1999) or to rapid classification onboard based on the most likely species (FAO, 2004). Even if the voucher specimen is correctly identified, additional issues can occur downstream of the sample processing and analysis. For instance, contaminant DNA of *Homo sapiens* (Kryukov & Imanishi, 2016), bacteria (Strong et al., 2014), or species in previously extracted samples could result in erroneously labeled sequences in the database or the formation of chimeric sequences (Haas et al., 2011). Effort is being made to identify incorrect records (Bucklin et al., 2021; Leray et al., 2019), but their removal takes time because errors are not always reported, much less corrected. An example of the magnitude of the consequences is given in this work with the misassignment of *Engraulis encrassicolus* sequences to *Istiophorus albicans* (Figure 5). *E. encrassicolus* is the most abundant small pelagic in the Bay of Biscay (Uriarte et al., 1996), whereas *I. albicans*, instead, is rare in the region (ICCAT, 2019). Because they correspond to very different consumer levels in the food web and their commercial importance is different in the study area, the raw database-derived results would have led to a wrong interpretation of ecological and economical relevance; the use of our barcoding gap-based error detection method has allowed to identify and solve the issue.

The so-called barcoding gap relies on the principle that the larger the difference between intraspecific and interspecific genetic distances, the more accurate the taxonomic classification (Hebert et al., 2004). For fishes, the barcoding gap has been examined for

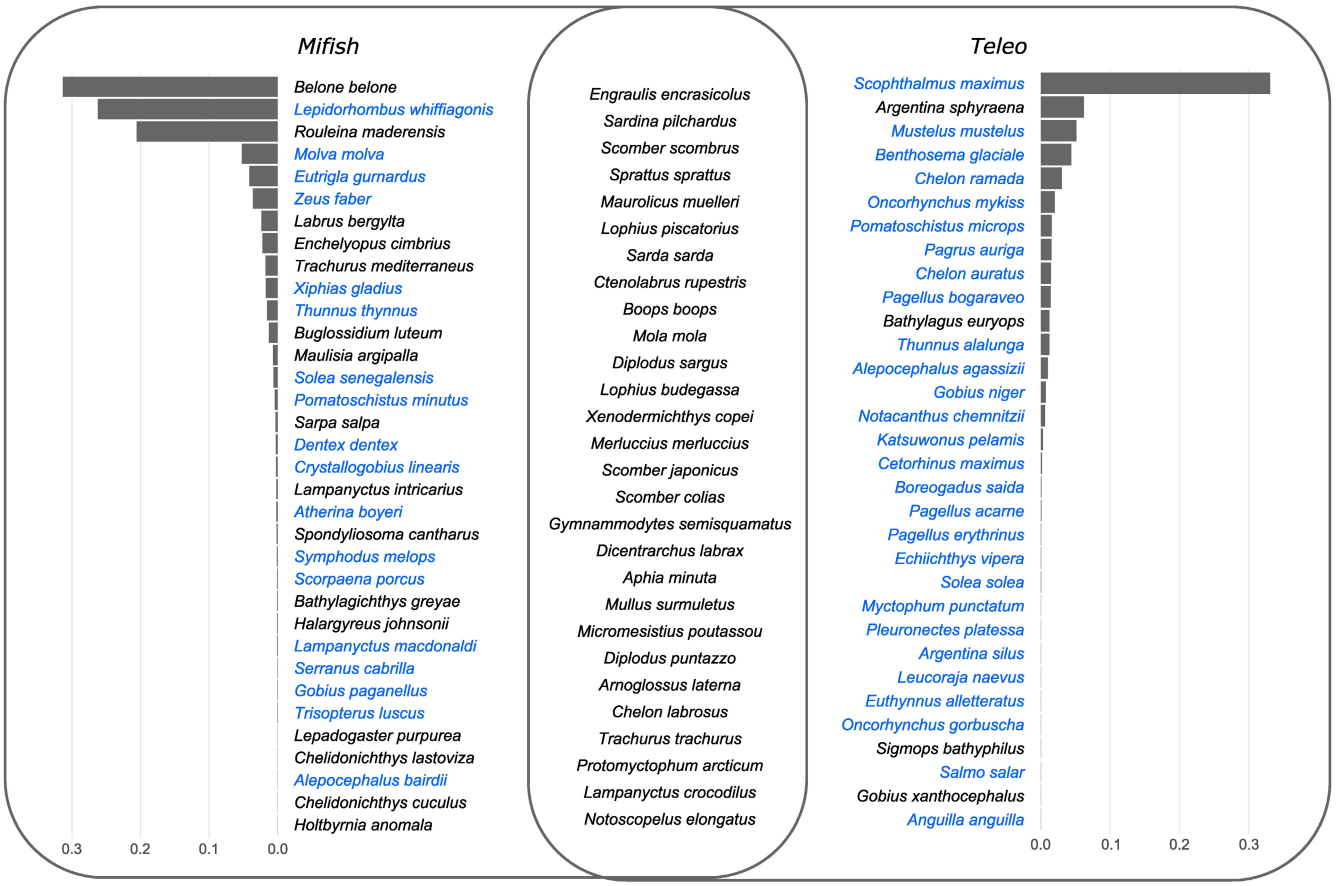

FIGURE 6 Venn diagram representing the species identified in this study. Bar plots indicate the relative abundance (%) of species detected with only one barcode. Species highlighted in blue also have representative sequences in the reference database of the other barcode but were not detected.

large (500–900 bp) mitochondrial regions (Cawthorn et al., 2012; Li et al., 2018), but eDNA metabarcoding usually relies on shorter DNA fragments (≈ 60–170 bp), for which the barcoding gap requires further examination. Here, we examine the barcoding gap for short barcodes widely used for fish and found an excess number of pairs of sequences that do not follow the barcoding gap principle (Figure 3). For these outlier pairs, defining a strategy to identify the erroneous sequence should be feasible considering the high taxonomic distance between them. Cases closer to the diagonal of the heatmap are more complex and make it especially challenging to identify whether sequences are truly erroneous or whether natural reasons make the pair be out of the diagonal. For example, low taxonomic discrimination by the 12S gene has been reported within fish genera (e.g., Sebastes, *Anarchias*) and families (e.g., *Gadidae*, *Cyprinidae*, *Istiophoridae*) (Gold et al., 2021; Johnstone et al., 2007; Thomsen et al., 2016), which could make sequences appear more similar than expected according to taxonomy. Similarly, biological phenomena such as inter-specific introgression could make sequences from the same species appear more distant than expected and species from different species closer than expected (Viñas & Tudela, 2009). These challenging cases are not limited to the accuracy of the reference database but to the chosen barcode region, and different primer combinations are a promising solution to tackle them (Ficetola & Taberlet, 2023).

4.3 | Toward an automated database curation procedure

Our workflow performs a quick screening to detect erroneously labeled sequences and flag problematic ones; additionally, it provides the networks for the sequences that did not result in a clear diagnostic due to the complexity of the distance relationships so that they can be manually inspected. Thus, this workflow is a significant step in automatically improving GenBank-based reference databases for diverse taxa. Unlike other steps in the bioinformatics processing of sequencing data, there is a notable lack of homogeneity in the reference database curation for taxonomic assignment between similar studies. The most accurate approach for manual curation is the use of phylogenetic trees, which allows detailed inspection for erroneous sequence detection (; Collins et al., 2019; Leray et al., 2019). However, manual inspection of phylogenetic trees is not viable for large databases and has limitations for short and unequal-length sequences. Here, we explore an alternative solution and propose a workflow for spurious sequence detection based on network analysis. Briefly, the approach considers spurious sequences that are more similar to sequences from other species than to sequences of their own (labeled) species. We have focused on 12S rRNA barcodes as being the preferred region for fish eDNA metabarcoding, yet the

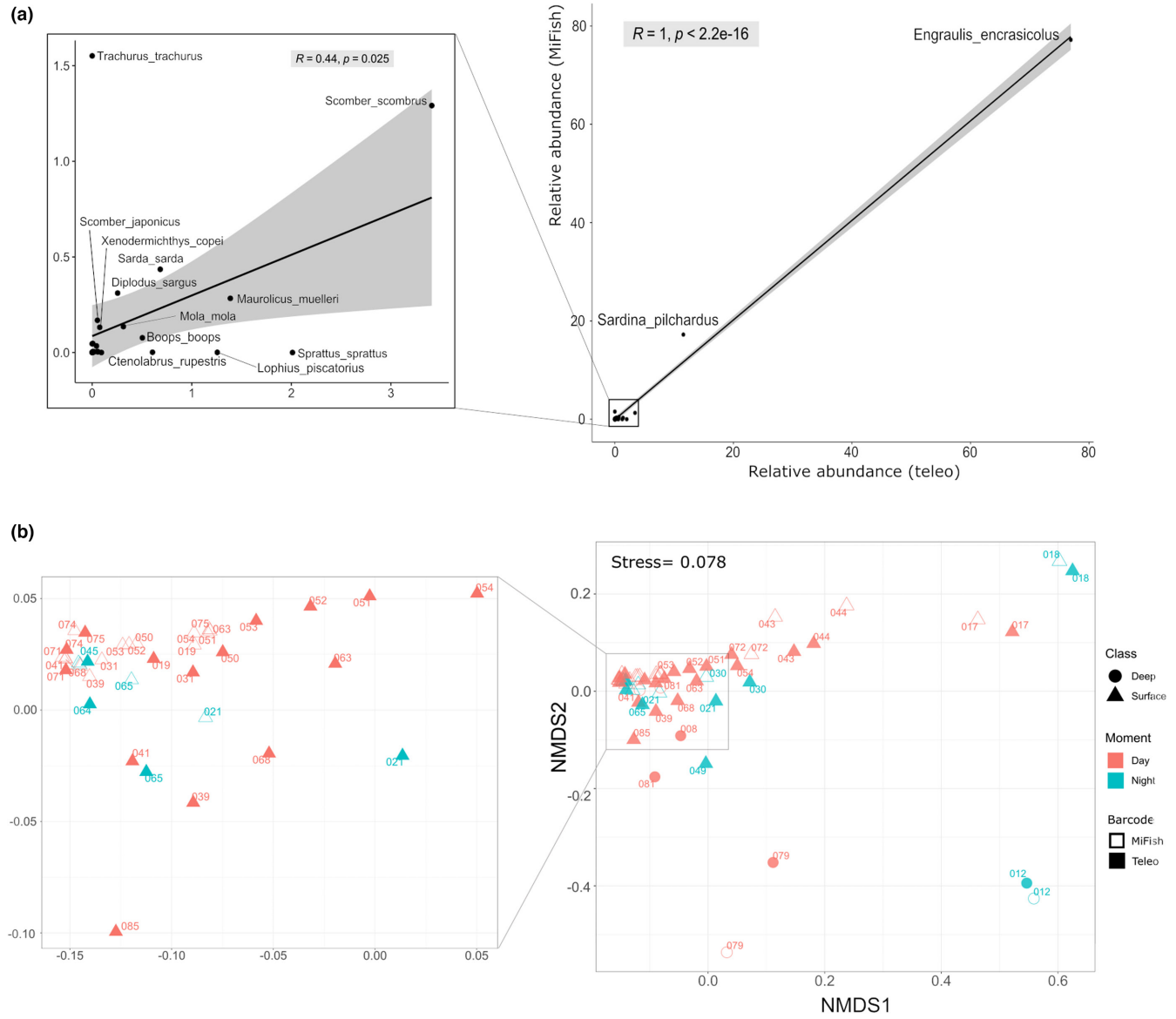

FIGURE 7 (a) Relationship between relative abundance of reads of shared species between *MiFish* and *teleo* barcodes. Shaded area represents the 95% confidence interval of the linear regression. (b) Non-metric multidimensional scaling (NMDS) for shared species using the two markers.

method can be used for any barcode. We acknowledge that the tool assumes a linear relationship between similarity and taxonomic relatedness, which is not always fulfilled by real sequences. Yet, this assumption ensures the detection of errors in the extreme cases, and, moreover, the tool allows to modify taxonomic ranges and distance levels to be included in the analyses so that less extreme cases can also be inspected. Thus, this workflow not only allows to retrieve the barcode sequences corresponding to a given list of species but performs a first screening of spurious sequences, allowing one to eliminate, flag, or further inspect them. A limiting factor of the method is related to the poor representation of some species in the reference databases due to the existence of barcoded species with only one record available because verifying single records is complex with any screening method. Distance matrices rely on the confidence of close records, which will rarely be mislabeled. Although being an unlikely scenario, it is a limitation to take into consideration, especially when single records are abundant such as in our study.

Noteworthy, our method was able to detect a particularly challenging but existing problem in genetic databases: the chimeric sequences. For example, although present in both the *teleo* and *MiFish* reference databases, *I. albicans* sequences were only labeled as erroneous in the *teleo* database. In the *MiFish* database, *I. albicans* sequences were more similar to other *Istiophorus* sequences than to sequences belonging to the genus *Engraulis*. This can be explained by the formation of chimeras between the target species and other species during the barcoding process, either in the PCR or assembly steps (Haas et al., 2011). The fact that no reads were classified as *I. albicans* with the *MiFish* raw database supports the chimeric structure of the sequence, being some regions truly from *I. albicans* and others from *Engraulis*. In line with the above, it is noteworthy to remark that database curation substantially

changed the taxonomic assignment of *teleo* reads, which highlights the importance of caution and critical reasoning when analyzing metabarcoding data to avoid wrong interpretations or misunderstandings. Although minor differences were observed between the taxonomic assignments of *MiFish* reads using raw and curated versions of the database, potential erroneous sequences belonging to species not detected in the study were also identified, which may be problematic in studies involving other fish assemblages. To guarantee reliable fish eDNA metabarcoding applications, such as management-scale diversity monitoring, the suitability and quality of reference databases need to be considered. The completeness and accuracy evaluation are to become good practices in the field with the use of tools such as the pipeline developed in this study.

AUTHOR CONTRIBUTIONS

C.C., O.C., and N.R-E. conceived the study. I.M. carried out the fieldwork. I.M. and L.G.A. performed the laboratory work. C.C. and N.R-E developed code. C.C. wrote the first draft of the manuscript and O.C. and N.R-E critically contributed with revisions. All authors gave the final approval for publication.

ACKNOWLEDGMENTS

The authors are grateful to the crew of R/V Ramón Margalef for their help during sampling, especially to María Santos, and to Iker Pereda for his help in code debugging. This work has been funded by the European Union's Horizon 2020 research and innovation program (project SUMMER with grant agreement No. 817806), the Department of Environment, Planning, Agriculture and Fisheries (Basque Government) through the project GENGES, and the Department of Education (Basque Government) through a predoctoral grant to Cristina Claver.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Raw sequencing reads and associated metadata are available on the NCBI SRA (BioProject PRJNA894161). Developed scripts and corresponding output files are available at GitHub (https://github.com/rodriguez-ezpeleta/NEA_fish_DB).

ORCID

Cristina Claver https://orcid.org/0000-0003-4071-8976
Naiara Rodriguez-Ezpeleta https://orcid.org/0000-0001-6735-6755

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham Bioinformatics.

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), 1–8.

Barco, A., Kullmann, B., Knebelsberger, T., Sarrazin, V., Kuhs, V., Kreutle, A., Pusch, C., & Thiel, R. (2022). Detection of fish species from marine protected areas of the North Sea using environmental DNA. *Journal of Fish Biology*, *101*, 722–727.

Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2012). GenBank. *Nucleic Acids Research*, *41*(D1), D36–D42.

Boettiger, C., Lang, D. T., & Wainwright, P. (2012). Rfishbase: Exploring, manipulating and visualizing FishBase data from R. *Journal of Fish Biology*, *81*(6), 2030–2039.

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120.

Bucklin, A., Peijnenburg, K. T., et al. (2021). Toward a global reference database of COI barcodes for marine zooplankton. *Marine Biology*, *168*(6), 1–26.

Buyck, B. (1999). Taxonomists are an endangered species in Europe. *Nature*, *401*(6751), 321.

Canals, O., Mendibil, I., Santos, M., Irigoien, X., & Rodríguez-Ezpeleta, N. (2021). Vertical stratification of environmental DNA in the open ocean captures ecological patterns and behavior of deep-sea fishes. *Limnology and Oceanography Letters*, *6*(6), 339–347.

Cawthorn, D.-M., Steinman, H. A., & Witthuhn, R. C. (2012). Evaluation of the 16S and 12S rRNA genes as universal markers for the identification of commercial fish species in South Africa. *Gene*, *491*(1), 40–48.

Clarke, K. R. (1993). Non-parametric multivariate analyses of changes in community structure. *Australian Journal of Ecology*, *18*(1), 117–143.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *10*(11), 1985–2001.

Collins, R. A., Trauzzi, G., Maltby, K. M., Gibson, T. I., Ratcliffe, F. C., Hallam, J., Rainbird, S., Maclaine, J., Henderson, P. A., Sims, D. W., Mariani, S., & Genner, M. J. (2021). Meta-fish-lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes. *Journal of Fish Biology*, *99*(4), 1446–1454.

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.

Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C., & Knight, R. (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics*, *27*(16), 2194–2200.

FAO. (2004). *Implementation issues associated with listing commercially exploited aquatic species on CITES appendices*. FAO Fisheries Report. Food & Agriculture Org.

FAO. (2012). *The state of world fisheries and aquaculture*. Opportunities and challenges.

Ficetola, G. F., & Taberlet, P. (2023). Towards exhaustive community ecology via DNA metabarcoding. *Molecular Ecology*, 1–10.

Figueiredo, I., Maia, C., Lagarto, N., & Serra-Pereira, B. (2020). Bycatch estimation of Rajiformes in multispecies and multigear fisheries. *Fisheries Research*, *232*, 105727.

Fraija-Fernández, N., Bouquieaux, M. C., et al. (2020). Marine water environmental DNA metabarcoding provides a comprehensive fish diversity assessment and reveals spatial patterns in a large oceanic area. *Ecology and Evolution*, *10*(14), 7560–7584.

Froese, R., & Pauly, D. (2022). "FishBase from , version (06/2022)". www.fishbase.org

Gilbey, J., Carvalho, G., Castilho, R., Coscia, I., Coulson, M. W., Dahle, G., Derycke, S., Francisco, S. M., Helyar, S. J., Johansen, T., Junge, C., Layton, K. K. S., Martinsohn, J., Matejusova, I., Robalo, J. I., Rodríguez-Ezpeleta, N., Silva, G., Strammer, I., Vasemägi, A., & Volckaert, F. A. M. (2021). Life in a drop: Sampling environmental

DNA for marine fishery management and ecosystem monitoring. *Marine Policy*, *124*, 104331.

Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., Walker, H. J., Jr., Burton, R. S., Kacev, D., Martz, L. D., & Barber, P. H. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resources*, *21*(7), 2546–2564.

Haas, B. J., Gevers, D., Earl, A. M., Feldgarden, M., Ward, D. V., Giannoukos, G., Ciulla, D., Tabbaa, D., Highlander, S. K., Sodergren, E., Methé, B., DeSantis, T. Z., The Human Microbiome Consortium, Petrosino, J. F., Knight, R., & Birren, B. W. (2011). Chimeric 16S rRNA sequence formation and detection in sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, *21*(3), 494–504.

Hebert, P. D., Ratnasingham, S., & De Waard, J. R. (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.

Hebert, P. D. N., Stoeckle, M. Y., et al. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, *2*(10), e312.

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). A database of metazoan cytochrome c oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, *5*(1), 1–7.

ICCAT. (2019). International Comission for the conservation of Atlantic tunas. In *Report of the stranding Comitee on research and statistics (SCRS)*. https://www.iccat.int/Documents/Meetings/Docs/2019/REPORTS/2019_SCRS_ENG.pdf

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.

Jin, X., Zhao, S., et al. (2013). Universal primers to amplify the complete mitochondrial 12S rRNA gene in marine fish species. *Genetics and Molecular Research*, *12*(4), 4575–4578.

Johnstone, K., Marshall, H., et al. (2007). Biodiversity genomics for species at risk: Patterns of DNA sequence variation within and among complete mitochondrial genomes of three species of wolffish (Anarhichas spp.). *Canadian Journal of Zoology*, *85*(2), 151–158.

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.

Kawato, M., Yoshida, T., Miya, M., Tsuchida, S., Nagano, Y., Nomura, M., Yabuki, A., Fujiwara, Y., & Fujikura, K. (2021). Optimization of environmental DNA extraction and amplification methods for metabarcoding of deep-sea fish. *MethodsX*, *8*, 101238.

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.

Kirsch, J. E., Day, J. L., Peterson, J. T., & Fullerton, D. K. (2018). Fish misidentification and potential implications to monitoring within the San Francisco estuary, California. *Journal of Fish and Wildlife Management*, *9*(2), 467–485.

Klymus, K. E., Marshall, N. T., & Stepien, C. A. (2017). Environmental DNA (eDNA) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. *PLoS One*, *12*(5), e0177643.

Kryukov, K., & Imanishi, T. (2016). Human contamination in public genome assemblies. *PLoS One*, *11*(9), e0162424.

Kume, M., Lavergne, E., Ahn, H., Terashima, Y., Kadowaki, K., Ye, F., Kameyama, S., Kai, Y., Henmi, Y., Yamashita, Y., & Kasai, A. (2021). Factors structuring estuarine and coastal fish communities across Japan using environmental DNA metabarcoding. *Ecological Indicators*, *121*, 107216.

Lamy, T., Pitz, K. J., Chavez, F. P., Yorke, C. E., & Miller, R. J. (2021). Environmental DNA reveals the fine-grained and hierarchical spatial structure of kelp forest fish communities. *Scientific Reports*, *11*(1), 1–13.

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences*, *116*(45), 22651–22656.

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2020). Reply to Locatelli *et al*: Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group. *Proceedings of the National Academy of Sciences*, *117*(51), 32213–32214.

Leray, M., Knowlton, N., & Machida, R. J. (2022). MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. In *MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences*. DNA.

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*(1), 34.

Li, X., Shen, X., Chen, X., Xiang, D., Murphy, R. W., & Shen, Y. (2018). Detection of potential problematic Cytb gene sequences of fishes in GenBank. *Frontiers in Genetics*, *9*, 30.

Locatelli, N. S., McIntyre, P. B., et al. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences*, *117*(51), 32211–32212.

Lyon, J. P., Tonkin, Z., Moloney, P. D., Todd, C., & Nicol, S. (2018). Conservation implications of angler misidentification of an endangered fish. *Aquatic Conservation: Marine and Freshwater Ecosystems*, *28*(6), 1396–1402.

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*(1), 1–7.

Mariani, S., Fernandez, C., et al. (2021). Shark and ray diversity, abundance and temporal variation around an Indian Ocean Island, inferred by eDNA metabarcoding. *Conservation Science and Practice*, *3*(6), e407.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. Journal*, *17*(1), 10–12.

McClenaghan, B., Fahner, N., Cote, D., Chawarski, J., McCarthy, A., Rajabi, H., Singer, G., & Hajibabaei, M. (2020). Harnessing the power of eDNA metabarcoding for the detection of deep-sea fishes. *PLoS One*, *15*(11), e0236540.

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.

Nguyen, B. N., Shen, E. W., Seemann, J., Correa, A. M. S., O'Donnell, J. L., Altieri, A. H., Knowlton, N., Crandall, K. A., Egan, S. P., McMillan, W. O., & Leray, M. (2020). Environmental DNA survey captures patterns of fish and invertebrate diversity across a tropical seascape. *Scientific Reports*, *10*(1), 1–14.

Oka, S.-i., Doi, H., Miyamoto, K., Hanahara, N., Sado, T., Miya, M., et al. (2021). Environmental DNA metabarcoding for biodiversity monitoring of a highly diverse tropical fish community in a coral reef lagoon: Estimation of species richness and detection of habitat segregation. *Environmental DNA*, *3*(1), 55–69.

Oksanen, J., Blanchet, F. G., et al. (2013). Package 'vegan'. *Community Ecology Package, Version*, *2*(9), 1–295.

Oliveira, L. M., Knebelsberger, T., Landi, M., Soares, P., Raupach, M. J., & Costa, F. O. (2016). Assembling and auditing a comprehensive DNA

barcode reference library for European marine fishes. *Journal of Fish Biology*, *89*(6), 2741–2754.
Andrea Polanco, F., Richards, E., Flück, B., Valentini, A., Altermatt, F., Brosse, S., Walser, J.-C., Eme, D., Marques, V., Manel, S., Albouy, C., Dejean, T., & Pellissier, L. (2021). Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. *Environmental DNA*, *3*(6), 1113–1127.
Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, *25*(2), 527–541.
Porter, T. M., & Hajibabaei, M. (2018). Over 2.5 million COI sequences in GenBank and growing. *PLoS One*, *13*(9), e0200177.
Ratnasingham, S., & Hebert, P. D. (2007). BOLD: The barcode of life data system. *Molecular Ecology Notes*, *7*(3), 355–364.
Richardson, R. T., Bengtsson-Palme, J., Gardiner, M. M., & Johnson, R. M. (2018). A reference cytochrome c oxidase subunit I database curated for hierarchical classification of arthropod metabarcoding data. *PeerJ*, *6*, e5126.
Russo, T., Maiello, G., Talarico, L., Baillie, C., Colosimo, G., D'Andrea, L., di Maio, F., Fiorentino, F., Franceschini, S., Garofalo, G., Scannella, D., Cataudella, S., & Mariani, S. (2021). All is fish that comes to the net: Metabarcoding for rapid fisheries catch assessment. *Ecological Applications*, *31*(2), e02273.
Sato, M., Inoue, N., Nambu, R., Furuichi, N., Imaizumi, T., & Ushio, M. (2021). Quantitative assessment of multiple fish species around artificial reefs combining environmental DNA metabarcoding and acoustic survey. *Scientific Reports*, *11*(1), 1–14.
Sayers, E. (2008). "E-utilities Quick Start." from 2018 Oct 24.
Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., & Weber, C. F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09
Sepulveda, A. J., Nelson, N. M., Jerde, C. L., & Luikart, G. (2020). Are environmental DNA methods ready for aquatic invasive species management? *Trends in Ecology & Evolution*, *35*(8), 668–678.
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.
Steinegger, M., & Salzberg, S. L. (2020). Terminating contamination: Large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biology*, *21*(1), 1–12.
Strong, M. J., Xu, G., Morici, L., Splinter Bon-Durant, S., Baddoo, M., Lin, Z., Fewell, C., Taylor, C. M., & Flemington, E. K. (2014). Microbial contamination in next generation sequencing: Implications for sequence-based analysis of clinical samples. *PLoS Pathogens*, *10*(11), e1004437.
Stucky, B. J. (2012). SeqTrace: A graphical tool for rapidly processing DNA sequencing chromatograms. *Journal of Biomolecular Techniques: JBT*, *23*(3), 90.
Thalinger, B., Wolf, E., Traugott, M., & Wanzenböck, J. (2019). Monitoring spawning migrations of potamodromous fish species via eDNA. *Scientific Reports*, *9*(1), 1–11.
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.
Tsuji, S., Takahara, T., Doi, H., Shibata, N., & Yamanaka, H. (2019). The detection of aquatic macroorganisms using environmental DNA analysis—A review of methods for collection, extraction, and detection. *Environmental DNA*, *1*(2), 99–108.
Uriarte, A., et al. (1996). Bay of Biscay and Ibero Atlantic anchovy populations and their fisheries. *Scientia Marina*, *60*, 237–255.
Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., … Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*(4), 929–942.
Viñas, J., & Tudela, S. (2009). A validated methodology for genetic identification of tuna species (genus Thunnus). *PLoS One*, *4*(10), e7606.
Virgilio, M., Backeljau, T., Nevado, B., & de Meyer, M. (2010). Comparative performances of DNA barcoding across insect orders. *BMC Bioinformatics*, *11*(1), 1–10.
Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299.
Wang, Q., Garrity, G. M., Tiedje, J. M., & Cole, J. R. (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, *73*(16), 5261–5267.
Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, S. A., Bouchez, A., Borja, A., Čiamporová-Zaťovičová, Z., Ferreira, S., … Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, *678*, 499–524.
West, K., Travers, M. J., Stat, M., Harvey, E. S., Richards, Z. T., DiBattista, J. D., Newman, S. J., Harry, A., Skepper, C. L., Heydenrych, M., & Bunce, M. (2021). Large-scale eDNA metabarcoding survey reveals marine biogeographic break and transitions over tropical North-Western Australia. *Diversity and Distributions*, *27*(10), 1942–1957.
WoRMS Editorial Board. (2022). World Register of Marine Species. https://www.marinespecies.org at VLIZ https://doi.org/10.14284/170
Zhang, J., Kobert, K., Flouri, T., & Stamatakis, A. (2014). PEAR: A fast and accurate Illumina paired-end reAd mergeR. *Bioinformatics*, *30*(5), 614–620.
Zhang, S., Zhao, J., & Yao, M. (2020). A comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution*, *11*(12), 1609–1625.
Zhou, S., Fan, C., Xia, H., Zhang, J., Yang, W., Ji, D., Wang, L., Chen, L., & Liu, N. (2022). "combined use of eDNA metabarcoding and bottom trawling for the assessment of fish biodiversity in the Zhoushan Sea." Frontiers in marine science: 2056.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Claver, C., Canals, O., de Amézaga, L. G., Mendibil, I., & Rodriguez-Ezpeleta, N. (2023). An automated workflow to assess completeness and curate GenBank for environmental DNA metabarcoding: The marine fish assemblage as case study. *Environmental DNA*, *5*, 634–647. https://doi.org/10.1002/edn3.433