Environmental DNA

WILEY

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

METHOD OPEN ACCESS

A Confidence Scoring Procedure for eDNA Metabarcoding Records and Its Application to a Global Marine Fish Dataset

Andrea Polanco F.[1] | Romane Rozanski[2,3] | Virginie Marques[2,3] | Martin Helmkampf[4] | David Mouillot[5,6] | Stéphanie Manel[6,7] | Camille Albouy[2,3] | Oscar Puebla[4,8] | Loïc Pellissier[2,3]

[1]Biodiversa Colombia Foundation, Bogota, Colombia | [2]Department of Environmental Systems Science, Ecosystems and Landscape Evolution, Institute of Terrestrial Ecosystems, ETH Zurich, Zurich, Switzerland | [3]Unit of Land Change Science, Swiss Federal Research Institute WSL, Birmensdorf, Switzerland | [4]Leibniz Centre for Tropical Marine Research (ZMT), Bremen, Germany | [5]MARBEC, Univ Montpellier, CNRS, IFREMER, IRD, Montpellier, France | [6]Institute Universitaire de France, Paris, France | [7]CEFE, Univ Montpellier, CNRS, EPHE-PSL University, IRD, Montpellier, France | [8]Institute for Chemistry and Biology of the Marine Environment (ICBM), Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany

Correspondence: Andrea Polanco F. (andrea.polanco@gmail.com) | Romane Rozanski (romane.rozanski@usys.ethz.ch)

Received: 23 June 2024 | **Revised:** 11 February 2025 | **Accepted:** 17 February 2025

Funding: This work was supported by Eidgenössische Technische Hochschule Zürich.

Keywords: biodiversity | environmental DNA | false positive | interpretation | record

ABSTRACT

Environmental DNA (eDNA) metabarcoding is changing the way biodiversity is surveyed in many types of ecosystems. eDNA surveys are now commonly performed and integrated into biodiversity monitoring programs and public databases. Although it is widely recognized that eDNA records require interpretation in light of taxonomy and biogeography, there remains a range of perceptions about how thoroughly records should be evaluated and which ones should be reported. Here, we present a modular procedure, available as an R script, that uses a set of five steps to assess the confidence of species-level eDNA records by assigning them a score from 0 to 5. This procedure includes evaluations of the known geographic distribution of each taxon, the taxonomic resolution of the marker used, the regional completeness of the reference database, the diversification rate, and the range map of each taxon. We tested the procedure on a large-scale marine fish eDNA dataset (572 samples) covering 15 ecoregions worldwide, from the poles to the tropics, using the *teleo* marker on the mitochondrial 12S ribosomal gene. Our analysis revealed broad variation in the average confidence score of eDNA records among regions, with the highest scores occurring along the European and Eastern Atlantic coasts. Generalized linear models applied to record covariates highlighted the significant influences of latitude and species richness on low confidence scores (<2.5). The polar regions notably displayed high proportions of low confidence scores, probably due to the limited completeness of the regional reference databases and the taxonomic resolution of the *teleo* marker. We conclude that only records with high confidence scores (>2.5) should be integrated into biodiversity databases. The medium (2.5) to relatively low-confidence (<2.5) records correspond to species that require further investigation and may be integrated after inspection to ensure high-quality species records.

1 | Introduction

Environmental DNA (eDNA; Taberlet et al. 2012) metabarcoding is revolutionizing species detection and biodiversity surveys, particularly in aquatic environments (Deiner et al. 2017; Harrison et al. 2019). A major advantage of this approach is its ability to document the occupancy of species without the need to capture, collect, or even observe individuals. eDNA metabarcoding is useful for the detection of cryptic and elusive species (Bessey et al. 2020; Nester et al. 2020; Polanco Fernández et al. 2021) and for the detection of non-indigenous species before they have irreversibly invaded ecosystems (LeBlanc

et al. 2020; Coster et al. 2021). As such, eDNA is emerging as an instrumental tool for the large-scale and long-term monitoring of biodiversity, setting the foundation for many applications (Deiner et al. 2021). For example, eDNA metabarcoding can inform the creation or optimization of reserves (Capurso et al. 2023; Mathon et al. 2024), track spatial and temporal biodiversity dynamics (Sevellec et al. 2021; Polanco F. et al. 2022), and contribute to our understanding of species responses to global change (Bernatchez et al. 2024).

Since its first application in marine environments (Thomsen et al. 2012), eDNA metabarcoding has grown and matured rapidly, leading to its increasing use in both fundamental and applied contexts (Kestel et al. 2022; Thomas et al. 2020). Animals shed material from their tissues into the environment, including genetic material that can be retrieved (as eDNA) from environmental samples, such as water or soil, and provide information on biodiversity (Taberlet et al. 2012). DNA is extracted, amplified using specific markers, and sequenced. Then, DNA sequences obtained from eDNA metabarcoding can be analyzed in the form of molecular operational taxonomic units (MOTUs) or amplicon sequence variants (ASVs), representing groups with sufficient similarity to be considered distinct species (historically 3% divergence; Ryberg 2015). Many applications of eDNA metabarcoding require not only the retrieval of MOTUs or ASVs—hereafter referred to as sequences—but also their taxonomic assignment, ideally at the species level. This is the case, for example, in the detection of invasive or endangered species (Boussarie et al. 2018) and in comparisons between historical monitoring data and eDNA detections (Polanco Fernández et al. 2021). This is the role of the bioinformatic step dedicated to taxonomic assignment, in which the sequences recovered from eDNA samples are compared with those contained in a reference database (Hakimzadeh et al. 2024; Mathon et al. 2021). In most software, the quality and completeness of the reference database are not considered (Salvi et al. 2020), which can lead to identification bias (Marques et al. 2021).

Species are identified by matching eDNA sequences to a reference database with data from sequenced specimens that have been identified by taxonomic experts (Hakimzadeh et al. 2024; Mathon et al. 2021). Polymerase chain reaction (PCR) primers are designed to target and amplify the DNA of specific groups of organisms, e.g., fishes for the *teleo* marker (Valentini et al. 2016). Depending on the completeness of the reference database and the genetic variation within and between species, different eDNA markers differ in their ability to identify species and distinguish among closely related species (Zhang et al. 2020). For most taxonomic groups and geographic regions, however, the database is not complete (Marques et al. 2021; Weingand et al. 2019), causing assignments to be less specific (e.g., at the genus or family level), introducing biases into the resulting inferred species lists, or leading to identification errors where sequences are assigned to an incorrect taxonomy label (Claver et al. 2023; Marques et al. 2021). Additionally, sequences recovered from eDNA are generally short and possibly not sufficient to discriminate among species for some genera (De Jonge et al. 2021). In the context of an incomplete database, a full match to a species can be misleading, as it might also match other species in the same genus that have not yet been sequenced (Chorlton 2024).

Even when the same reference database and marker(s) are used, results may differ based on methods of eDNA sampling (Gogarten et al. 2020; Mas-Carrió et al. 2022) and bioinformatic analysis (Mathon et al. 2021; Macé et al. 2022). Incomplete reference databases and low standardization result in a broad range of perceptions of what eDNA biodiversity records mean and how they should be interpreted (Altermatt et al. 2023). For many applications, this lack of standardization also limits the use of eDNA surveys compared with classical methods that rely on direct observations, such as visual surveys (Mathon et al. 2021). While incomplete sampling and variation in the primer region can result in false negatives (Pinfield et al. 2019), the direct comparison of DNA reads with an incomplete reference database increases the risk of generating false positives (Blackman et al. 2023; Dugal et al. 2022). False positives are problematic because they imply the presence of a species that does not actually occur, and they are difficult to correct since this would require evidence for the absence of the considered species. These errors may persist in the literature and biodiversity databases even if they are never validated, propagating influential biases (Jerde 2021; Rodriguez-Martinez et al. 2022). Considering that records from eDNA metabarcoding are starting to be integrated into species occurrence databases (Andersson et al. 2021), there is an urgent need to develop tools to assess the confidence in their validity.

Since biologists decide whether to accept or reject traditional species records based on specific criteria (Zinger et al. 2019), it should in principle be possible to replicate at least part of this process with a standardized and automated procedure for eDNA sequences. The criteria used to assess the plausibility of species records include the previous records of occurrence of the species in the area or nearby (Zinger et al. 2019; Yang et al. 2024), the clarity of the diagnostic characteristics used to identify the species, and the ability to discriminate among species (Cognato et al. 2020). In the context of eDNA, records can be assessed with a combination of natural and technical factors, including species range (e.g., GBIF 2023), evolutionary history (Bellwood et al. 2017; Siqueira et al. 2020), reference database completeness (Marques et al. 2021; Weingand et al. 2019), and eDNA marker resolution.

Here, we propose a procedure for evaluating the confidence of species-level assignments from eDNA metabarcoding based on five steps with three types of information: species ranges, phylogenies, and reference databases. We leverage a unique global database of marine fish eDNA (Polanco Fernández et al. 2021; Marques et al. 2021; Mathon et al. 2021) based on the *teleo* primer pair (Valentini et al. 2016) to investigate the geographic variation in eDNA record confidence and its contributing factors. We consider two types of uncertainties: "known" uncertainty, when sequences do not match any known species and are assigned to a higher taxonomic level (e.g., genus or family), and, more importantly, "hidden" uncertainty, when a sequence is assigned to a species but may be a false positive due to contamination, amplification of degraded DNA, sequencing errors, or misassignments arising from incomplete or incorrect reference databases. Next, we compare the confidence score from the procedure with a subset of the global dataset pertaining to tropical reef fishes collected near Providencia Island and Gayraca Bay in Colombia. We analyze the results in light of the following questions:

1. What fraction of bioinformatic species classifications is assigned a low-confidence score?
2. Does taxonomic confidence vary geographically, and what are the most important factors determining the confidence score?
3. How are confidence scores distributed across species from a dataset, and what is the risk of false positives when this procedure is not applied?

The proposed procedure provides a robust framework for understanding the reliability of eDNA-based species taxonomic assignments, enabling better-informed ecological interpretations and biodiversity assessments.

2 | Material and Methods

2.1 | A Procedure for Scoring the Confidence of eDNA Metabarcoding Records

We developed a framework to label an eDNA metabarcoding record—a sequence assigned to a species—with a confidence score (Figure 1) by implementing five steps. The final score ranges from 0, representing a low-confidence species detection likely to be a false positive, to 5, representing a high-confidence species detection likely to be a true positive. The procedure is implemented using R software and is available as an automated function (available at https://www.polybox.ethz.ch/index.php/s/hd3zE067TWtPcnf, along with an example dataset) that requires four to five input files associated with the five steps described below:

- The first input file (file 1) corresponds to a dataframe containing the name of all the species detected (step 1), their percentages of sequence matching (step 2), their genera (step 3), their diversification rates (step 4), and information about the availability of range maps or occurrence data for the species (step 5; Figure 1).
- The second input file (file 2) is a dataframe containing a regional species list (step 1), including information on the species' genera (step 3), their diversification rates (step 4), and whether they have been sequenced or not (Figure 1).
- The third input file is a list, where each element is a dataframe with information about all the filter coordinates (latitude/longitude) where each species was detected (step 5; Figure 1).
- The fourth and optional fifth input files correspond to the distribution range map (file 4: to be used preferentially) and species occurrence data (file 5: to be used if no range map is available), respectively, for each species detected (step 5).

Due to the substantial amount of data required for all steps, each step is optional and users can provide the corresponding files as needed.

Step 1: Species presence/absence from a regional checklist—The list of species detected with eDNA is compared with a regional species checklist (provided in file 2). In this step, the detected species that are present in the regional checklist are assigned a score of 1, while those not in the checklist receive a score of 0. The regional checklist must be provided by the user. For example, for fish, checklists can be downloaded from GAPeDNA (Marques et al. 2021), an interface that links marker-specific genetic reference databases, generated in silico from the European Molecular Biology Laboratory EMBL (Kanz et al. 2005), to regional species lists at different spatial scales. The data from GAPeDNA are updated regularly by its contributors, and the interface contains information on the sequencing status of each regional species. For the latter, which might be false-positive detections, a list of alternative species from the same genera that are present in the regional checklist is provided as an output from the procedure. An optional parameter allows the user of the procedure to add neighboring ecoregion checklists: in cases where a species is absent from the main ecoregion checklist but present in the neighboring one(s), a score of 0.5 is assigned.

Step 2: Percentage of sequence matching—Each sequence obtained from the bioinformatic procedure (file 1) is assigned a value corresponding to the proportion of sequence matching (between 0 and 1) based on the comparison of the sequence with the one from the reference database used. The match percentage input for each sequence lies between a minimum and a maximum threshold, both chosen by the user (e.g., 97%). The detected species with a 100% match are assigned a score of 1, while those with the minimum match percentage are assigned a score of 0. Species with values between these extremes receive an intermediate score based on a linear function.

Step 3: Gap analysis of the regional reference database—One of the main limitations of eDNA metabarcoding studies is that some species are absent from the reference databases (Schenekar et al. 2020; Altermatt et al. 2023). In this step, the GAPeDNA interface is used to assess the level of database completeness in different ecosystems worldwide. The user extracts the GAPeDNA data (file 2: https://shiny.cefe.cnrs.fr/GAPeDNA/) for one or several areas (province or ecoregion) and primers (e.g., *teleo* or *MiFish*) to assess database completeness at the regional level for each genus. For example, if a genus contains five local species, three of which are sequenced, each detected species from that genus receives a score of three-fifths (0.6). If all species of a given genus are sequenced locally and a sequence is matched to a species, then the detection confidence is considered high, as no other local unsequenced congeneric species are present, and a score of 1 is assigned.

Step 4: Species diversification rate—The diversification rate of fishes refers to the net rate at which new fish species form (speciation) minus the rate at which existing species become extinct (extinction) over a given period. It reflects the balance of these evolutionary processes, shaping the diversity of fish lineages across time and ecosystems (Morlon et al. 2024). The diversification rate varies among taxonomic groups, with some families having a high rate of recent diversification (e.g., Chaetodontidae) and others a lower one (e.g., Scombridae). In eDNA metabarcoding analyses, identifying a sequence to the species level can be problematic if the marker used is not

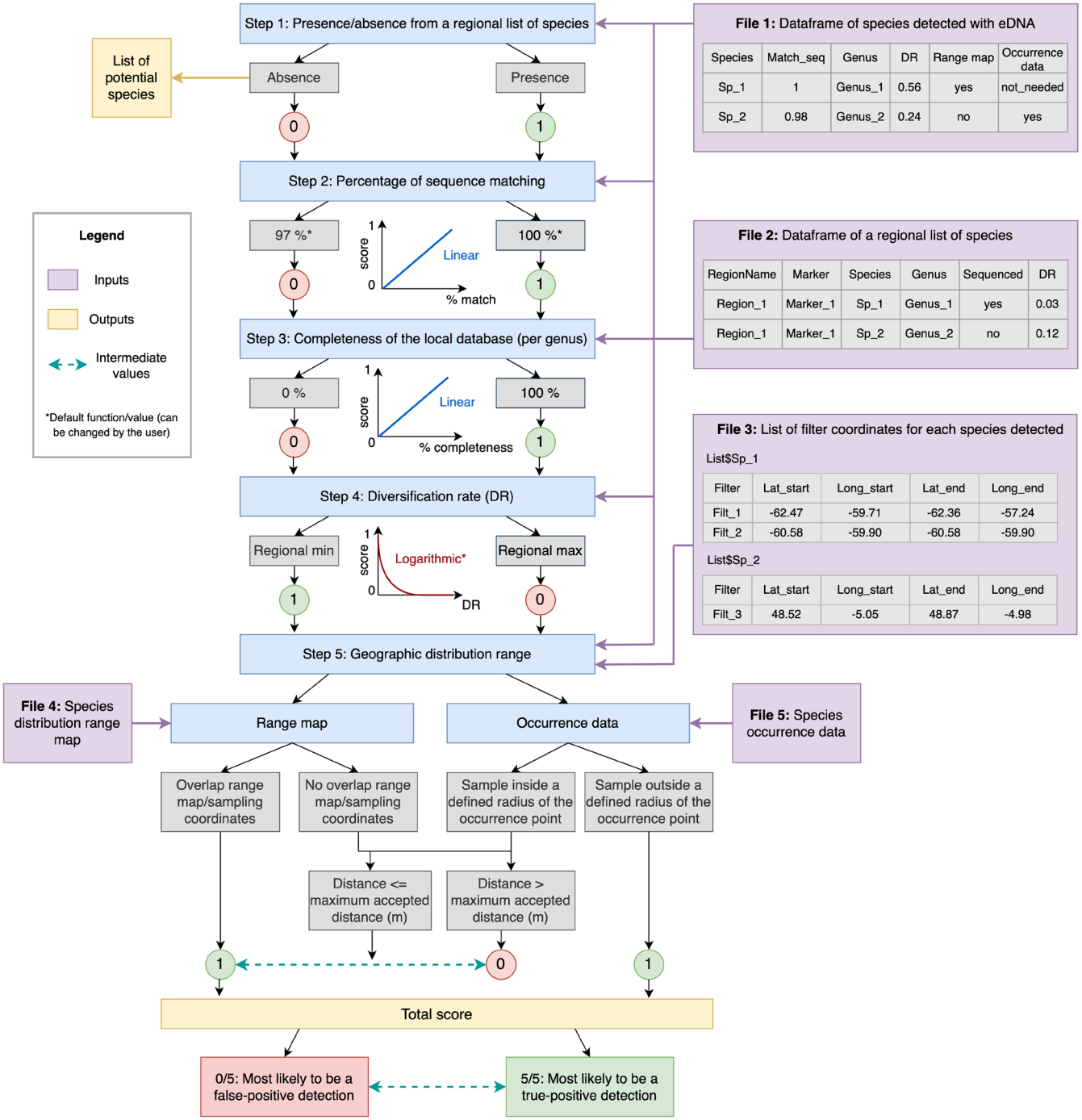

FIGURE 1 | Procedure flow chart summarizing the five steps (1–5) required to assign a score to each species detected with environmental DNA (eDNA) metabarcoding: 1. regional species list, 2. sequence matching, 3. completeness of the reference database, 4. diversification rate, and 5. geographic distribution range.

discriminating enough to distinguish among closely related species in a family with a high rate of recent diversification. To complete this step of the procedure, the user provides the diversification rates of the regional species (file 2; e.g., global fish diversification rates can be downloaded from fishtreeof life.org; Rabosky et al. 2018). The regional species associated with the lowest diversification rate receive a score of 1, while the one with the highest rate is assigned a score of 0. All other species receive intermediate scores based on the fitting of a function (default is logarithmic) that depends on the diversification rate distribution, adapted from species–area relationship (SAR) models implemented with the *SARS* package in R (Matthews et al. 2019; R Core Team 2023).

Step 5: Species geographic range and distribution of occurrences—Comparing eDNA detections of a given species with its distribution range can be useful in the case of incomplete reference databases or non-discriminating markers. In this step, the user provides either a geographic range map of the detected species (file 4) or a spatial occurrence table (file 5), as well as the coordinates of each eDNA sampling location (point or transect; file 3). If a range map is provided, the coordinates of all the eDNA

sampling locations where the species were detected are compared with the species' areas of occurrence on the range map. If at least one eDNA sampling location overlaps with the range map, the species is assigned a score of 1. If none of the sampling locations overlap with the map, distances (in meters) between each sampling location and the nearest polygon of occurrence are computed. If the species is detected only at a distance greater than a maximum threshold (set by the user), it is assigned a score of 0. Species detected at intermediate distances receive a score between 0 and 1 (both excluded) according to a step in distance chosen by the user (e.g., with a maximum threshold of 3 km and a distance step of 1 km, $score_{[0:1000]} = 0.75$; $score_{[1000:2000]} = 0.5$; $score_{[2000:3000]} = 0.25$; $score_{>3000} = 0$). If a table of occurrence is provided, occurrences of the species are compared with the coordinates of all the eDNA sampling locations where the species was detected. If the species was detected in at least one location within a certain radius from the nearest occurrence point (given in meters in the procedure), previously specified by the user, it is assigned a score of 1. For detections at greater distances, the same process as for the range map comparison applies, with a score of 0 assigned if the detection is farther than a defined maximum threshold distance (same threshold as the range map one) and intermediate scores assigned according to the specified distance step.

To generate species geographic ranges (i.e., range maps used in step 5), a range-mapping algorithm that combines an occurrence dataset and convex-hull polygons was applied. The procedure described in Albouy et al. (2019) was followed, using species data sourced from the Ocean Biodiversity Information System OBIS (http://www.iobis.org). A total of 16,238,200 occurrence records were collected from 34,883 OBIS entries. To ensure adequate data quality, data cleaning procedures were applied that involved resolving issues, such as synonyms and misspellings, and identifying rare species, resulting in a set of 11,503,257 occurrences for 11,345 marine fish species. As the OBIS database did not sufficiently represent tropical fish assemblages, the OBIS dataset was merged with a second database that encompasses 6316 coral reef species (Parravicini et al. 2013). Additionally, the analysis was limited to marine species, resulting in a dataset comprising 14,035 fish species. From this pool of species, 840 freshwater and brackish-water species were removed, yielding a final list of 13,195 marine teleost fish species. In addition, distribution maps were reconstructed for each species, defined as the convex polygon surrounding the area where each species was observed (Albouy et al. 2019). The fish range maps were aggregated at a 1° grid resolution for the 13,195 marine fish teleost species (Albouy et al. 2019).

2.2 | Sampling a Global eDNA Dataset for eDNA of Marine Fishes

The procedure was applied to a global eDNA dataset consisting of filtered seawater samples collected at 309 stations in 15 marine regions covering the global ocean from pole to pole (4 polar, 3 temperate, and 8 tropical ecoregions; Mathon et al. 2023). Between 1 and 4 replicates were sampled at each station for a total of 594 eDNA samples. Only samples containing taxa detected at the species level were considered in this study, resulting in 572 samples. Since these data were collected by different collaborators and the sampling method was optimized over time, four different methods were used: (i) collection of 2 L of water in DNA-free sterile plastic bags from a small boat and with closed-circuit rebreather diving (depths 10–40 m), sampling as close as possible to the habitat of benthic fishes (Juhel et al. 2020); (ii) collection of 1 L of water in a sterilized bottle from the surface; (iii) 2-km filtration transect lasting 30 min with two replicates (one on each side of a boat at each station), for a total of 30 L ± 15% of water from just under the surface; (iv) sampling using Niskin bottles. For each sample collected with the first two sampling protocols, the seawater was filtered with sterile Sterivex filter capsules (Merck Millipore; Darmstadt, Germany; pore size 0.22 μm) using disposable sterile syringes. Immediately after filtration, the capsules were filled with a CL1 conservation buffer (SPYGEN, le Bourget du Lac, France) and stored at room temperature. The eDNA filtration device for the other two sampling protocols was composed of an Athena peristaltic pump (Proactive Environmental Products LLC, Bradenton, Florida, USA; nominal flow of 1.0 L min^{-1} ± 15%), a VigiDNA 0.2 μM cross-flow filtration capsule (SPYGEN), and disposable sterile tubing for each filtration capsule. At the end of each filtration, the capsules were filled with 80 mL of CL1 conservation buffer and stored at room temperature. For each sampling campaign, a strict contamination-control protocol was followed (Valentini et al. 2016; Goldberg et al. 2016), which included the use of disposable gloves and single-use filtration equipment.

2.3 | eDNA Extractions and Sequencing

eDNA extractions were performed in a dedicated eDNA laboratory (SPYGEN, Le Bourget du Lac Cedex, France, www.spygen.com) equipped with positive air pressure, UV treatment, and frequent air renewal, following the protocols by Pont et al. (2018) for the VigiDNA capsules and by Juhel et al. (2020) for the Sterivex filter capsules. A teleost-specific 12S mitochondrial rRNA primer pair (*teleo*, forward primer ACACCGCCCGTCACTCT, reverse primer CTTCCGGTACACTTACCATG; Valentini et al. 2016) was used for the amplification of metabarcode sequences. Twelve PCR amplifications per sample were performed in a final volume of 25 μL, with 3 μL of DNA extract as the template. The amplification was performed following the protocol by Pont et al. (2018). The purified PCR products were pooled in equal volumes to achieve a target sequencing depth of 1,000,000 reads per sample. Library preparation and sequencing were performed by Fasteris (Geneva, Switzerland). A total of 45 libraries were prepared using the MetaFast protocol for Illumina sequencing platforms. Paired-end sequencing (2 × 125 bp) was carried out using a HiSeq 2500 sequencer with the HiSeq Rapid Flow Cell v2 and the HiSeq Rapid SBS Kit v2 (Illumina, San Diego, CA, USA), a MiSeq sequencer (2 × 125 bp) with the MiSeq Flow Cell Kit v3 (Illumina), or a NextSeq sequencer (2 × 125 bp) with the NextSeq Mid kit (Illumina), following the manufacturer's instructions. This generated an average of 624,468 sequence reads (paired-end Illumina) per sample.

2.4 | Data Processing

Following sequencing, reads were processed using clustering and post-clustering cleaning to remove errors and estimate

the number of species using MOTUs (Marques et al. 2021). Reads were assembled using *vsearch* ($v = 2$) (Rognes et al. 2016) and were demultiplexed and trimmed with *cutadapt* (v3.4, Cutadapt 2025). Clustering was performed using *Swarm v2* (Mahé et al. 2015) with a *d* value of 1, i.e., a minimum distance of 2 mismatches between clusters.

2.5 | Taxonomic Assignment

Taxonomic assignment of MOTUs was carried out using the lowest common ancestor algorithm *ecotag*, implemented in the OBITools toolkit (Boyer et al. 2016). The European Nucleotide Archive (ENA, Leinonen et al. 2010) was used as a reference database (release 143, March 2020), supplemented by our custom reference database (unpublished) containing approximately 800 sequences. If the sequence matched several taxa with equal percentages of similarity, *ecotag* assigned the sequence to the lowest possible taxonomic level common among all possible matches. Conservative quality filters were then applied: all MOTUs with fewer than 10 reads were discarded, along with those present in only one PCR replicate, to avoid spurious MOTUs originating from a PCR error (Marques et al. 2021). Then, errors generated by index-hopping (MacConaill et al. 2018) were filtered out using a threshold empirically determined per sequencing batch using experimental blanks (combinations of tags not present in the libraries; Taberlet et al. 2018). Tag-jumping (Schnell et al. 2015) was corrected using a threshold of 0.001 of occurrence for a given MOTU within a library. Taxonomic assignments at the species level were accepted as putative species if the percentage of similarity with the reference sequence was 100%, at the genus level if the similarity was 90%–99%, and at the family level if the similarity was 86%–89%. If these criteria were not met, the MOTU was left unassigned. Sequences matching perfectly with more than one species were assigned at the lowest possible level (genus or family). Subsequently, the proportion of detections at four different taxon levels (species, genus, family, and higher levels) was computed for each geographic region. For all taxonomic assignments at the species level, the five-step procedure described above was then run to quantify the confidence in taxonomic assignment. Next, the main factors associated with the generated regional confidence were assessed, along with how these factors vary across regions.

2.6 | Spatial Analysis With Social and Environmental Variables

To explain the spatial variability in taxonomic assignment confidence, the confidence score was related to six non-correlated social and environmental factors computed for France, Norway, Spain, Colombia, Curacao, Indonesia, French Polynesia, and Antarctica. The selected social factors were: (1) population gravity, defined as the human population size divided by the travel time between the sampling station and this population center (summed over a buffer of 500 km around a station; Mathon et al. 2023); (2) marine ecosystem dependency, corresponding to the nutritional, economic, and coastal protection dependence on marine ecosystems at the country scale (Mathon et al. 2023); and (3) gross domestic product (GDP) averaged over the 2020, 2021, and 2022 period (set to 0 for Antarctica due to its lack of a permanent economy). The environmental factors were: (4) species richness; (5) distance to the coast; and (6) the absolute value of the sampling latitude.

2.7 | Statistical Analysis

For each of the 572 samples, the proportion of species with a confidence score $< 2.5/5$ was computed. A generalized linear model (GLM; "stats" R package) was then implemented, with this proportion as the response variable and the social and environmental factors as explanatory variables (with a log transformation applied to population gravity and distance to the coast). To account for the unbalanced number of samples per geographic region, the minimum number of eDNA sample filters (corresponding to 14 filters, collected in Norway) was randomly resampled 100 times for each geographic region. The GLM was then applied to the 100 resulting datasets, using a binomial family associated with a logit regression, suitable for the model proportions. A polynomial term was included in all GLMs to account for potential non-linear relationships with the response variable, and all possible combinations of explanatory variables were tested. To evaluate all models, the average Akaike information criterion corrected for small sample sizes (AICc) and the difference in (delta) AICc between the different models were calculated. The average MacFadden's pseudo R^2 was also calculated, and a Chi-squared test was performed to obtain the *p*-value for each model. Explanatory variables were considered to have a significant influence at $p < 0.05$.

3 | Results

3.1 | Global Distribution of Taxonomic Confidence

In the 594 seawater eDNA samples collected at 309 stations across 15 marine regions, we found a total of 3459 MOTUs, 2701 of which were assigned to at least the family level, and 757 at the species level. Regions with temperate or tropical climates had the highest percentages of MOTUs assigned to the species level (Figure 2; Table S1). The South European Atlantic shelf emerged with 60.4% of MOTU assignments at the species level, followed by the Celtic Sea at 59.7%, and the tropical region of Papua (Indonesia) at 58.9%. In contrast, the Arctic polar regions had lower percentages of MOTUs assigned to the species level, with 9.1% for the North and East Barents Sea and 16.8% for Finnmark (Norway; Figure 2). By applying the five-step procedure to the 572 samples containing MOTUs assigned to the species level with the *teleo* marker, we found that 31% of the species were assigned scores of 4–5, 42% scores of 3–4, 13% scores of 2–3, 5% scores of 1–2, and 0% scores of 0–1 (because all the species had a matching sequence of 100%). Moreover, 9% had no score due to a lack of data for these species in the steps assessed for the confidence of eDNA records. Only 0.3% (2/757 species: *Chanos chanos* and *Megalops atlanticus*) were assigned the highest score of 5. In the Arctic, the low scores were generally associated with

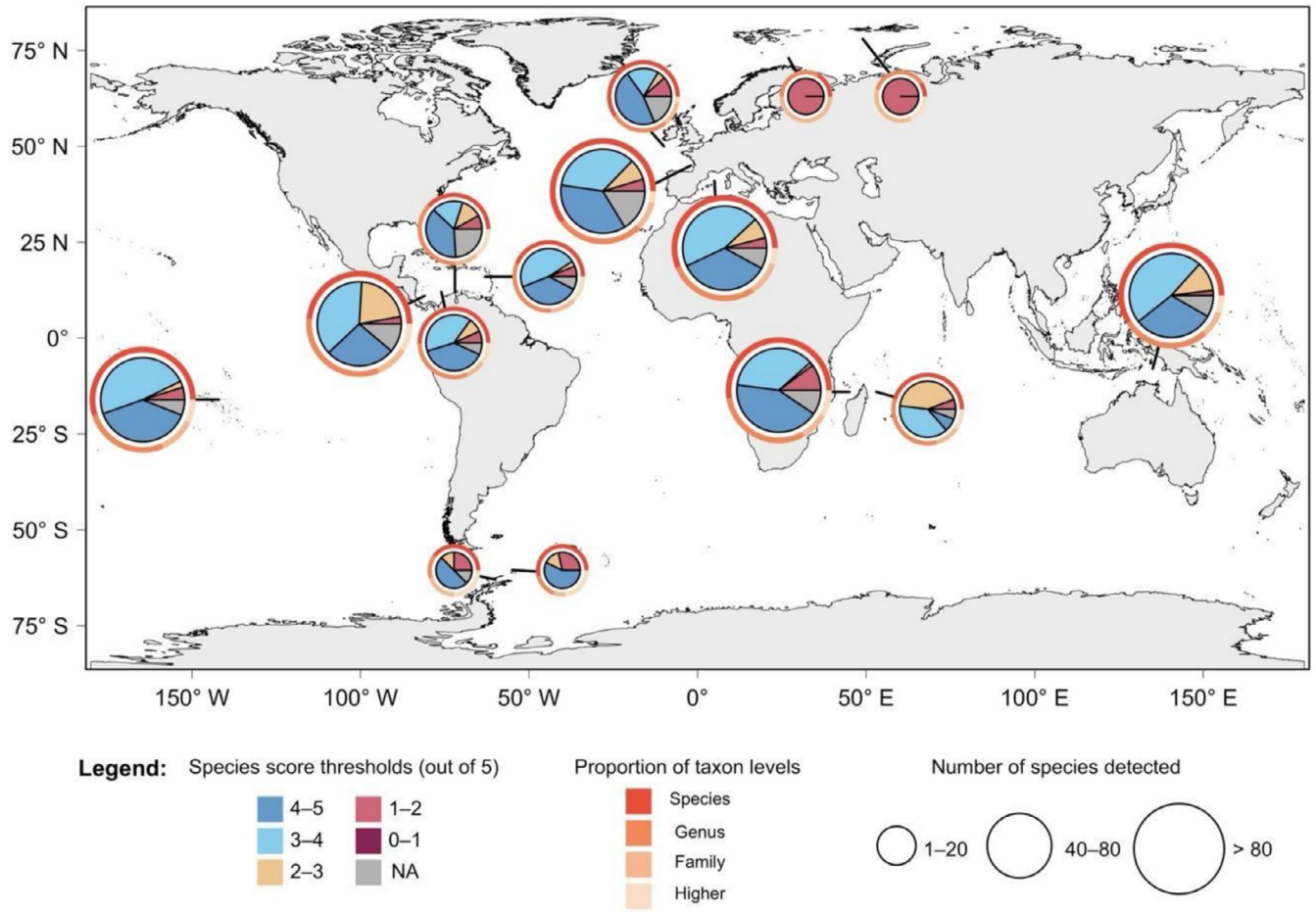

FIGURE 2 | Global map depicting the spatial variation in taxonomic confidence for molecular operational taxonomic units (MOTUs) assigned to the species level (Table S2). The percentages of the different taxonomic levels assigned within each marine geographic region are represented in the ring around the corresponding pie chart. Each pie chart illustrates the percentages of different confidence scores (from 0 to 5) in each region (Species score thresholds), based on the teleo marker. Scores colored in the "NA" category indicate that species information was lacking, such as diversification rates or occurrences, precluding the calculation of a total confidence score. The size of each pie chart corresponds to the number of detected species, with the Arctic and Antarctic having very few species (only two and around ten, respectively), in contrast to the other regions.

poor coverage in the reference database, as many MOTUs were only identified at the level of genus or family.

3.2 | Site-Specific Confidence Scores

Focusing on Providencia Island and Gayraca Bay in the Caribbean, we investigated the eDNA identification confidence in more detail (Figure 3; Tables S2 and S3). Among the 43 species detected near Providencia Island, 4 had a low confidence score below the low confidence threshold of 2.5, while in Gayraca, 2 out of the 32 species fell below this value. Among the low-ranking species, the Pacific red snapper (*Lutjanus peru*) and the spotted rose snapper (*Lutjanus guttatus*) ranked the lowest, with confidence scores of 1.8/5 and 1.9/5, respectively. Among these identifications, five out of six belonged to species found in the tropical eastern Pacific, but with congeneric species present in the Caribbean. The remaining species was the blue hamlet (*Hypoplectrus gemma*), shared between the two locations, which received a score of 2.1/5. The highest scores were assigned to the hogfish (*Lachnolaimus maximus*) near Providencia Island (4.8/5) and the mountain mullet (*Dajaus monticola*) in Gayraca Bay (4.6/5).

3.3 | Spatial Analysis With Social and Environmental Variables

As explanatory variables in the GLMs, we first considered separately the six environmental factors that were only weakly correlated with each other (Figure S1). Three variables—SR (species richness), mean GDP, and latitude—showed a significant association with the proportion of low confidence scores (Figure 4; Figure S2). Regarding species richness, the proportion of low scores decreased as species richness increased (mean pseudo $R^2_{\text{quadraticGLM}} = 0.44$), indicating that a higher species richness was associated with a lower likelihood of potential false-positive detections. Regarding mean GDP, although the pseudo $R^2_{\text{quadraticGLM}}$ value was lower (0.16) than that for species richness, we observed a trend of lower proportions of low confidence scores in areas with higher GDP values. We also found a strong non-linear positive association between latitude and the proportion of low scores (mean pseudo $R^2_{\text{quadraticGLM}} = 0.76$; Figure 4c), with especially high values in the Antarctic and Norway (Arctic) regions. When computing GLMs with all combinations of the explanatory variables, we found that the best model remained the one that included latitude with a quadratic term (Figure 4c;

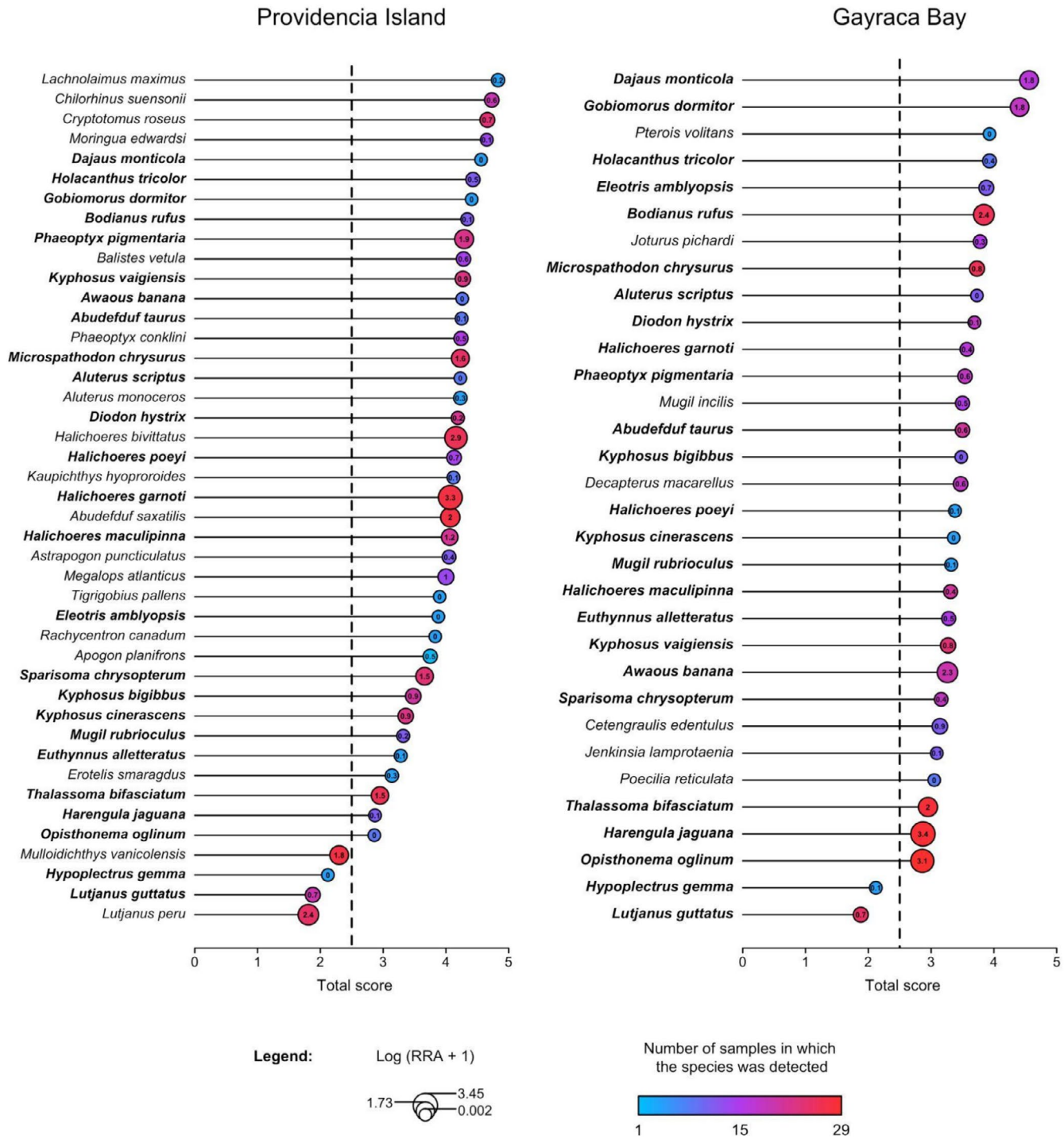

FIGURE 3 | Ranking of the total confidence scores for species from Providencia Island and Gayraca Bay, based on the five steps in the procedure. A high score corresponds to a high confidence in the assignment. Species names in bold are those species common to the two datasets. The color gradient indicates the number of eDNA samples in which each species was detected. The circle size corresponds to the logarithm of the total number of reads, where RRA is relative read abundance (%).

Table S4), indicating that low confidence scores were primarily influenced by the latitudinal geographic location.

4 | Discussion

Addressing the pressing issue of confidence in taxonomic identification from eDNA metabarcoding will enhance the reliability of eDNA data interpretation and bolster its credibility, ensuring more accurate applications in biodiversity surveys. It may thus contribute to the establishment of more robust monitoring programs and databases (Zinger et al. 2019; Takahashi et al. 2023). Here, we propose a procedure for generating confidence scores for eDNA metabarcoding sequences assigned at the species level, based on complementary data including the completeness of the reference database, species

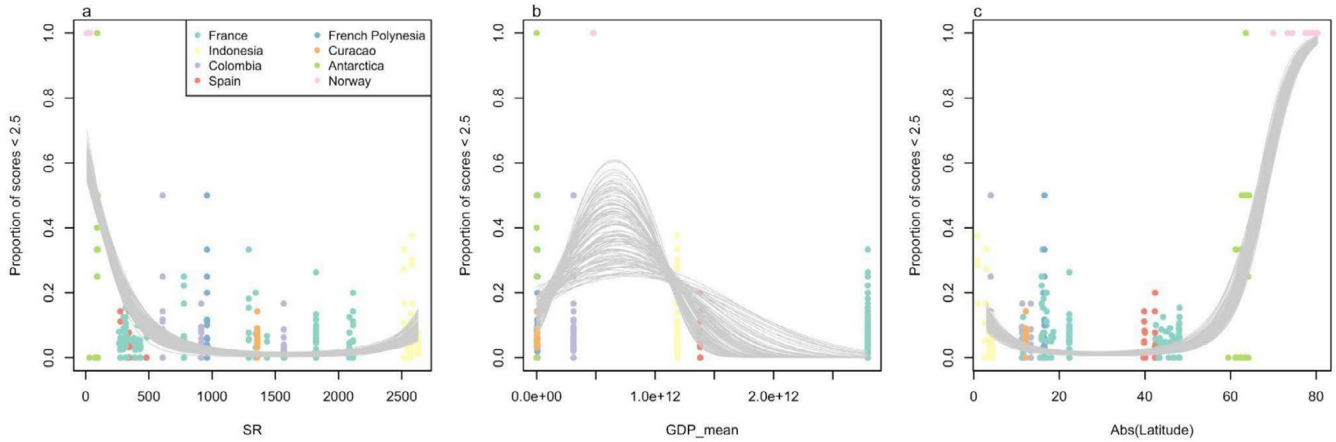

FIGURE 4 | Predictions based on 100 generalized linear models (GLM) for the influence of (a) species richness (SR), (b) mean gross domestic product (GDP), and (c) the absolute value of latitude on the proportion of low confidence scores (< 2.5). In each panel, the gray lines represent the fits of 100 GLMs with a quadratic term. Each dot corresponds to an eDNA sample, with the color corresponding to the geographic region which the sample was collected.

diversification rates, marker resolution, and species geographic range. By applying this procedure to a global marine fish eDNA dataset, we were able to assess how taxonomic confidence varies geographically and identify the main social and environmental variables determining this geographic variation. This application demonstrated that the procedure can reliably quantify confidence in taxonomic assignments. We recommend that metabarcoding practitioners utilize such transparent confidence scoring procedures to characterize species-level detection records. Having this confidence scoring is becoming essential given the increasing number of metabarcoding studies based on different protocols (Klymus et al. 2024). In addition, each step of the procedure is optional and can be removed if the relevant data are not available.

Biases in the taxonomic assignment of eDNA sequences have been attributed to deficiencies in the completeness, reliability, and availability of reference databases (Somervuo et al. 2017; Locatelli et al. 2020; Rodríguez-Ezpeleta et al. 2021; Blackman et al. 2023), conditions that in turn depend on the studied geographic region, the taxonomic group considered, and the targeted genetic region (Keck et al. 2023). At the global scale, there are major geographic differences in the completeness of DNA reference databases (Hillebrand 2004), with some regions and taxonomic groups being well represented and others presenting gaps (Marques et al. 2021). Our study focusing on marine fishes highlights substantial differences in the taxonomic confidence of species-level metabarcoding records across geographic regions. We observed high levels of taxonomic confidence in multiple locations within the European and Eastern Atlantic regions, aligning with the findings of Marques et al. (2021). Furthermore, Marques et al. (2021) reported a latitudinal gradient in species coverage, with lower coverage closer to the tropics. This variability in reference database completeness across regions correlates directly with the levels of taxonomic confidence observed in our study.

While the geographic coverage of the reference database partly determines the confidence of taxonomic assignments, factors associated with the reliability of the eDNA data must also be considered. These include the taxonomic coverage, defined as the percentage of species represented in the reference database out of the total number of species of a genus present in a geographic region; the diversification rate of the taxa in the target group; and the choice of a marker and its resolution (Espinosa-Prieto et al. 2024). Substantial taxa sampling gaps in DNA databases are a widespread issue that may also affect the correct assignment of the sequences to the species level. For marine fauna, the lack of sequences in barcode repositories reaches 80%–94% across metazoan groups (Hestetun et al. 2020). For example, marine fishes, especially commercially important ones in tropical regions like the Caribbean (*Lutjanus* genus) but even those in temperate regions like the Eastern Atlantic (Sparidae family), lack sufficient species sequences in databases (Froese and Pauly 2023). Despite the considerable global occurrence of *Lutjanus*, which comprises 67 species, only 31 species have accessible sequences in NCBI. Similarly, for the Sparidae family, comprising 158 species, only 29 species have sequences available for the 12S *teleo* marker. This lack of reference data hinders taxonomic assignment at the species level.

To better understand the process of taxonomic assignment and to identify the factors contributing to lower confidence, we conducted a detailed analysis of Gayraca Bay and Providencia Island. First, regarding the completeness of DNA barcode datasets, the genus *Hypoplectrus*, which comprises eighteen species (Puebla et al. 2022), is represented by only two 12S sequences in the EMBL reference database associated with the *teleo* marker, both from *Hypoplectrus gemma*. This led some researchers to mistakenly suggest a potential first record for this species in the Southern Greater Caribbean (Polanco Fernández et al. 2021). We nevertheless note that whole-genome data (nuclear and mitochondrial) are publicly available as shotgun resequencing data for > 270 samples from this group (Hench et al. 2019, 2022; Moran et al. 2019; Coulmance et al. 2024). This illustrates that, in the genomic era, lack of coverage for specific markers in reference databases does not necessarily imply a shortage of publicly available data. Furthermore, the genus *Hypoplectrus* exhibits a diversification rate that is among the highest in tropical reef fishes (Siqueira et al. 2020; Hench et al. 2022). As a result, taxonomic assignment at the species level is not possible for this genus

using mitochondrial markers (McCartney et al. 2003; Ramon et al. 2003; Garcia-Machado et al. 2004; Puebla et al. 2022), which highlights the importance of the fourth step of our procedure. Alternative methods, including group-specific primer sets targeting nuclear loci that evolve rapidly (Adams et al. 2019), could potentially be developed. In contrast to the taxonomic assignment of the genus *Hypoplectrus*, the assignment of the hogfish (*Lachnolaimus maximus*), a subtropical species commonly found in the considered Caribbean area (Froese and Pauly 2023; Lieske and Myers 1994), is an example of high taxonomic confidence as assessed by our procedure. The hogfish was assigned the highest score (4.83/5) of all the considered species because it had a 100% sequence match with the reference database, it has a low diversification rate (0.023), it was present in the regional species checklist, and its detection in at least one sampling location overlapped with its range map. We note that *Lachnolaimus* is a monospecific genus, which facilitates the taxonomic assignment of the single species. The same applies to the milkfish (*Chanos chanos*), which was also assigned with high confidence.

Our procedure further allowed us to identify species whose taxonomic assignment was most likely incorrect. For example, the procedure assigned low scores for both the Pacific red snapper (*Lutjanus peru*; 1.81/5) and the spotted rose snapper (1.88/5), revealing probable false-positive detections. Although their sequences matched (100%) with the reference database, these two species are not expected to be present in the Western Atlantic, as they are typically found in the Pacific subtropical region (Froese and Pauly 2023; Robertson and Allen 2015). Both species were absent from the regional checklist, and the eDNA samples in which they were detected were collected away from their nearest area of documented occurrence. These problems in the taxonomic assignment might have arisen from the combination of an incomplete local database for the *Lutjanus* genus (40% of local species sequenced with 12S in the Western Atlantic; Marques et al. 2021) and diversification rates close to or higher than the regional average of 0.19 (0.16 for the spotted rose snapper and 0.22 for the Pacific red snapper; Rabosky et al. 2018). We likely detected *Lutjanus* species that share genetic variation with congeneric species that are only found in the Pacific and have not yet been recorded in the sequence reference database for the Caribbean Sea. By analyzing the above three cases, we illustrate how our procedure helps detect taxonomic assignment problems systematically. Automation of this process is becoming increasingly important as the volume of data obtained with metabarcoding expands rapidly. Automation and standardization enhance data throughput and consistency, reduce human error, and support scalability, all of which are crucial for the effective handling of large-scale datasets and make it possible for researchers to focus more on data interpretation. For species with a confidence score <2.5, we recommend further investigation of their ecology and genetics to assess whether their presence could be possible in the considered area. A species detection with a low confidence score should not be immediately dismissed as a false positive, as it may instead reflect limited information on the species, emphasizing the need for deeper analysis. This warrants additional study, potentially through alternative modeling approaches, to validate results and enhance data accuracy. Additionally, while false positives in eDNA taxonomic assignments are a concern, the presence of eDNA from non-native species in the environment remains possible. This highlights the importance of including both native and invasive species in regional species lists. While our approach requires substantial input information for optimal results, such as reference databases and sequencing data, it offers flexibility by allowing users to select specific steps based on their needs rather than necessitating completion of every step in the procedure. This adaptable structure helps to mitigate data limitations and accommodates varying research goals and resource availability. We recommend that users clearly specify the steps they followed to derive the confidence scores, especially if any steps were omitted.

The inadequacy of species sequences accessible in public genetic databases represents another obstacle hindering the widespread utilization of eDNA inventories on a large scale, which in turn diminishes the scope of biodiversity that can be detected (Marques et al. 2021) or identified correctly. Improving the sequences' availability requires addressing the reasons behind this deficiency of reference data. For example, in the Caribbean tropical region, low information coverage has been reported (Valentini et al. 2016), with only 25% of fish sequences available for the 12 s *teleo* primer, which may be due to a lack of means and resources. Some regions with high biodiversity may be associated with a low capacity for barcoding, and collections may be limited to traditional methods. As a solution, we suggest greater effort in the barcoding of specimens (Beng and Corlett 2020) and more facilitation of means for taxonomists to effectively share this information on the platforms designed for this purpose, because in many cases a lot of biodiversity information remains unshared (De Santana et al. 2021). Existing data standards for genetic sequence data (MIxS and GGBN) can encourage the community to increase the data's suitability for reuse and to avoid situations in which data remain unpublished and inaccessible. Additionally, museum specimens could represent a useful source of sequences to reinforce databases (De Santana et al. 2021).

Ensuring the accuracy of species sequences in public genetic databases presents another large challenge (Robertson 2008). Biodiversity databases, exemplified by GBIF, continually update and refine scientific taxonomic names through essential processes. Accuracy and reliability can be ensured by developing regional reference databases, integrating diagnostic molecular characteristics, and focusing on methodological advancements and standardization in eDNA practices (Abarenkov et al. 2023; Dziedzic et al. 2023; Loeza-Quintana et al. 2020). It is therefore necessary for biodiversity entries from eDNA to be submitted to GBIF with the corresponding sequence. Without this information, it is challenging to assess the confidence of the detection, as done with the procedure presented here. Essentially, the sequence is key to our ability to reanalyze and validate information. GBIF collaborates extensively, engaging in conferences like PISCeS to validate detections and innovate methodologies (Loeza-Quintana et al. 2020). Platforms like GoaT provide access to validated genome-relevant metadata, enhancing project coordination and reliability in taxonomic identification for eDNA sequences (Challis et al. 2023). Efforts to develop such a comprehensive procedure underscore GBIF's commitment to leveraging cutting-edge tools, collaborations, and methodological advancements to harness the potential of eDNA in advancing biodiversity research and conservation efforts.

5 | Conclusions

As more and more species occurrence records from eDNA are deposited in global databases, such as GBIF, a measure of the confidence of these taxonomic assignments would be highly relevant for the reuse of these data for various applications. The implementation of the proposed procedure will enable researchers to identify instances where low confidence necessitates the elevation of taxonomic assignments to a higher taxonomic level. Our procedure is intended to enhance data reliability and align eDNA findings more closely with the actual occurrence of species. The advancement of the use of eDNA technology for various applications relies fundamentally on the taxonomic knowledge of a group and its evolution over time. This accumulation of knowledge over the years has provided the foundation for each of the steps in our procedure for assessing confidence. Consequently, it is essential to promote and support the endeavors of taxonomists, as their diligent work is instrumental in continuously enriching reference databases with meticulously curated data that meet the standards and confidence needed.

Author Contributions

A.P.F., R.R., L.P., C.A., and O.P. jointly designed this study; R.R., V.M., A.P.F., O.P., and M.H. analyzed data. A.P.F., R.R., L.P., C.A., and O.P. produced an initial manuscript draft that was improved through input from the other authors (V.M., M.H., D.M., and S.M.).

Acknowledgments

We thank Eric Coissac and Véronique Helfer for their feedback on an early version of the manuscript. This work was supported by the Monaco Exploration Expedition 2017-2022 funds, the Swiss Polar Institute, the IA-Biodiv ANR project FISH- PREDICT (ANR- 21- AAFI- 0001- 01), the ANR LabCom DiagADNe (ANR-20-LCV1-0008), and the European Union HORIZON EUROPE program ACTNOW (Grant no. 101060072).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The authors have nothing to report.

References

Abarenkov, K., A. F. Andersson, A. Bissett, et al. 2023. *Publishing DNA-Derived Data Through Biodiversity Data Platforms, v1.3*. GBIF Secretariat. https://doi.org/10.35035/Doc-vf1a-nr22.

Adams, C. I., M. Knapp, N. J. Gemmell, et al. 2019. "Beyond Biodiversity: Can Environmental DNA (eDNA) cut It as a Population Genetics Tool?" *Genes* 10, no. 3: 192.

Albouy, C., P. Archambault, W. Appeltans, et al. 2019. "The Marine Fish Food Web Is Globally Connected." *Nature Ecology & Evolution* 3, no. 8: 1153–1161.

Altermatt, F., L. Carraro, M. Antonetti, et al. 2023. "Quantifying Biodiversity Using eDNA From Water Bodies: General Principles and Recommendations for Sampling Designs." *Environmental DNA* 5, no. 4: 671–682.

Andersson, A. F., A. Bissett, A. G. Finstad, et al. 2021. "Publishing DNA-Derived Data Through Biodiversity Data Platforms. v1.0."

Bellwood, D. R., C. H. Goatley, and O. Bellwood. 2017. "The Evolution of Fishes and Corals on Reefs: Form, Function and Interdependence." *Biological Reviews* 92, no. 2: 878–901.

Beng, K. C., and R. T. Corlett. 2020. "Applications of Environmental DNA (eDNA) in Ecology and Conservation: Opportunities, Challenges and Prospects." *Biodiversity and Conservation* 29, no. 7: 2089–2121.

Bernatchez, L., A. L. Ferchaud, C. S. Berger, C. J. Venney, and A. Xuereb. 2024. "Genomics for Monitoring and Understanding Species Responses to Global Climate Change." *Nature Reviews Genetics* 25, no. 3: 165–183.

Bessey, C., S. N. Jarman, O. Berry, et al. 2020. "Maximizing Fish Detection With eDNA Metabarcoding." *Environmental DNA* 2, no. 4: 493–504.

Blackman, R. C., J. C. Walser, L. Rüber, et al. 2023. "General Principles for Assignments of Communities From eDNA: Open Versus Closed Taxonomic Databases." *Environmental DNA* 5, no. 2: 326–342.

Boussarie, G., J. Bakker, O. S. Wangensteen, et al. 2018. "Environmental DNA Illuminates the Dark Diversity of Sharks." *Science Advances* 4, no. 5: eaap9661.

Boyer, F., C. Mercier, A. Bonin, Y. Le Bras, P. Taberlet, and E. Coissac. 2016. "Obitools: A Unix-Inspired Software Package for DNA Metabarcoding." *Molecular Ecology Resources* 16, no. 1: 176–182.

Capurso, G., B. Carroll, and K. A. Stewart. 2023. "Transforming Marine Monitoring: Using eDNA Metabarcoding to Improve the Monitoring of the Mediterranean Marine Protected Areas Network." *Marine Policy* 156: 105–807.

Challis, R., S. Kumar, C. Sotero-Caio, M. Brown, and M. Blaxter. 2023. "Genomes on a Tree (GoaT): A Versatile, Scalable Search Engine for Genomic and Sequencing Project Metadata Across the Eukaryotic Tree of Life." *Wellcome Open Research* 8: 24.

Chorlton, S. D. 2024. "Ten Common Issues With Reference Sequence Databases and How to Mitigate Them." *Frontiers in Bioinformatics* 4, no. 1: 1278228.

Claver, C., O. Canals, L. G. de Amézaga, I. Mendibil, and N. Rodriguez-Ezpeleta. 2023. "An Automated Workflow to Assess Completeness and Curate GenBank for Environmental DNA Metabarcoding: The Marine Fish Assemblage as Case Study." *Environmental DNA* 5, no. 4: 634–647.

Cognato, A. I., G. Sari, S. M. Smith, et al. 2020. "The Essential Role of Taxonomic Expertise in the Creation of DNA Databases for the Identification and Delimitation of Southeast Asian Ambrosia Beetle Species (Curculionidae: Scolytinae: Xyleborini)." *Frontiers in Ecology and Evolution* 8: 27.

Coster, S. S., M. N. Dillon, W. Moore, and G. T. Merovich Jr. 2021. "The Update and Optimization of an eDNA Assay to Detect the Invasive Rusty Crayfish (*Faxonius rusticus*)." *PLoS One* 16, no. 10: e0259084.

Coulmance, F., D. Akkaynak, Y. Le Poul, M. P. Höppner, W. O. McMillan, and O. Puebla. 2024. "Phenotypic and Genomic Dissection of Colour Pattern Variation in a Reef Fish Radiation." *Molecular Ecology* 33, no. 4: e17047.

Cutadapt. 2025. "Cutadapt." https://cutadapt.readthedocs.io/en/v3.4/changes.html#v3-4-2021-03-30.

De Jonge, D. S., V. Merten, T. Bayer, O. Puebla, T. B. Reusch, and H. J. T. Hoving. 2021. "A Novel Metabarcoding Primer Pair for Environmental DNA Analysis of Cephalopoda (Mollusca) Targeting the Nuclear 18S rRNA Region." *Royal Society Open Science* 8, no. 2: 201388.

De Santana, C. D., L. R. Parenti, C. B. Dillman, et al. 2021. "The Critical Role of Natural History Museums in Advancing eDNA for Biodiversity Studies: A Case Study With Amazonian Fishes." *Scientific Reports* 11, no. 1: 18–159.

Deiner, K., H. M. Bik, E. Mächler, et al. 2017. "Environmental DNA Metabarcoding: Transforming How We Survey Animal and Plant Communities." *Molecular Ecology* 26, no. 21: 5872–5895.

Deiner, K., H. Yamanaka, and L. Bernatchez. 2021. "The Future of Biodiversity Monitoring and Conservation Utilizing Environmental DNA." *Environmental DNA* 3, no. 1: 3–7.

Dugal, L., L. Thomas, S. P. Wilkinson, et al. 2022. "Coral Monitoring in Northwest Australia With Environmental DNA Metabarcoding Using a Curated Reference Database for Optimized Detection." *Environmental DNA* 4, no. 1: 63–76.

Dziedzic, E., B. Sidlauskas, R. Cronn, et al. 2023. "Creating, Curating and Evaluating a Mitogenomic Reference Database to Improve Regional Species Identification Using Environmental DNA." *Molecular Ecology Resources* 23, no. 8: 1880–1904.

Espinosa-Prieto, A., L. Hardion, N. Debortoli, and J. N. Beisel. 2024. "Finding the Perfect Pairs: A Matchmaking of Plant Markers and Primers for Multi-Marker eDNA Metabarcoding." *Molecular Ecology Resources* 24: e13937.

Froese, R., and D. Pauly, eds. 2023. *FishBase*. World Wide Web electronic publication. www.fishbase.org, version (02/2023).

Garcia-Machado, E., P. P. Chevalier Monteagudo, and M. Solignac. 2004. "Lack of mtDNA Differentiation Among Hamlets (Hypoplectrus, Serranidae)." *Marine Biology* 144: 147–152.

GBIF: The Global Biodiversity Information Facility. 2023. "What is GBIF?" October 2, 2023. https://www.gbif.org/what-is-gbif.

Gogarten, J. F., C. Hoffmann, M. Arandjelovic, et al. 2020. "Fly-Derived DNA and Camera Traps Are Complementary Tools for Assessing Mammalian Biodiversity." *Environmental DNA* 2, no. 1: 63–76.

Goldberg, C. S., C. R. Turner, K. Deiner, et al. 2016. "Critical Considerations for the Application of Environmental DNA Methods to Detect Aquatic Species." *Methods in Ecology and Evolution* 7, no. 11: 1299–1307.

Hakimzadeh, A., A. Abdala Asbun, D. Albanese, et al. 2024. "A Pile of Procedures: An Overview of the Bioinformatics Software for Metabarcoding Data Analyses." *Molecular Ecology Resources* 24: e13847.

Harrison, J. B., J. M. Sunday, and S. M. Rogers. 2019. "Predicting the Fate of eDNA in the Environment and Implications for Studying Biodiversity." *Proceedings of the Royal Society B: Biological Sciences* 286, no. 1915: 20–191409.

Hench, K., M. Helmkampf, W. O. McMillan, and O. Puebla. 2022. "Rapid Radiation in a Highly Diverse Marine Environment." *Proceedings of the National Academy of Sciences of the United States of America* 119, no. 4: e2020457119.

Hench, K., M. Vargas, M. P. Höppner, W. O. McMillan, and O. Puebla. 2019. "Inter-Chromosomal Coupling Between Vision and Pigmentation Genes During Genomic Divergence." *Nature Ecology & Evolution* 3, no. 4: 657–667.

Hestetun, J. T., E. Bye-Ingebrigtsen, R. H. Nilsson, A. G. Glover, P. O. Johansen, and T. G. Dahlgren. 2020. "Significant Taxon Sampling Gaps in DNA Databases Limit the Operational Use of Marine Macrofauna Metabarcoding." *Marine Biodiversity* 50, no. 5: 70.

Hillebrand, H. 2004. "On the Generality of the Latitudinal Diversity Gradient." *American Naturalist* 163, no. 2: 192–211.

Jerde, C. L. 2021. "Can We Manage Fisheries With the Inherent Uncertainty From eDNA?" *Journal of Fish Biology* 98, no. 2: 341–353.

Juhel, J. B., R. S. Utama, V. Marques, et al. 2020. "Accumulation Curves of Environmental DNA Sequences Predict Coastal Fish Diversity in the Coral Triangle." *Proceedings of the Royal Society B: Biological Sciences* 287, no. 1930: 20200248.

Kanz, C., P. Aldebert, N. Althorpe, et al. 2005. "The EMBL Nucleotide Sequence Database." *Nucleic Acids Research* 33, no. suppl_1: D29–D33.

Keck, F., M. Couton, and F. Altermatt. 2023. "Navigating the Seven Challenges of Taxonomic Reference Databases in Metabarcoding Analyses." *Molecular Ecology Resources* 23, no. 4: 742–755.

Kestel, J. H., D. L. Field, P. W. Bateman, et al. 2022. "Applications of Environmental DNA (eDNA) in Agricultural Systems: Current Uses, Limitations and Future Prospects." *Science of the Total Environment* 847: 157556.

Klymus, K. E., J. D. Baker, C. L. Abbott, et al. 2024. "The MIEM Guidelines: Minimum Information for Reporting of Environmental Metabarcoding Data." *Metabarcoding and Metagenomics* 8: e128689.

LeBlanc, F., V. Belliveau, E. Watson, et al. 2020. "Environmental DNA (eDNA) Detection of Marine Aquatic Invasive Species (AIS) in Eastern Canada Using a Targeted Species-Specific qPCR Procedure." *Management of Biological Invasions* 11, no. 2: 201.

Leinonen, R., R. Akhtar, E. Birney, et al. 2010. "The European Nucleotide Archive." *Nucleic Acids Research* 39, no. suppl_1: D28–D31.

Lieske, E., and R. Myers. 1994. *Collins Pocket Guide. Coral Reef Fishes. Indo-Pacific & Caribbean Including the Red Sea*, 400. Harper Collins Publishers.

Locatelli, N. S., P. B. McIntyre, N. O. Therkildsen, and D. S. Baetscher. 2020. "GenBank's Reliability Is Uncertain for Biodiversity Researchers Seeking Species-Level Assignment for eDNA." *Proceedings of the National Academy of Sciences of the United States of America* 117: 32211–32212.

Loeza-Quintana, T., C. L. Abbott, D. D. Heath, L. Bernatchez, and R. H. Hanner. 2020. "Pathway to Increase Standards and Competency of eDNA Surveys (PISCeS)—Advancing Collaboration and Standardization Efforts in the Field of eDNA." *Environmental DNA* 2, no. 3: 255–260.

MacConaill, L. E., R. T. Burns, A. Nag, et al. 2018. "Unique, Dual-Indexed Sequencing Adapters With UMIs Effectively Eliminate Index Cross-Talk and Significantly Improve Sensitivity of Massively Parallel Sequencing." *BMC Genomics* 19: 1–10.

Macé, B., R. Hocdé, V. Marques, et al. 2022. "Evaluating Bioinformatics Procedures for Population-Level Inference Using Environmental DNA." *Environmental DNA* 4, no. 3: 674–686.

Mahé, F., T. Rognes, C. Quince, C. De Vargas, and M. Dunthorn. 2015. "Swarmv2: Highly-Scalable and High-Resolution Amplicon Clustering." *PeerJ* 3: e1420.

Marques, V., T. Milhau, C. Albouy, et al. 2021. "GAPeDNA: Assessing and Mapping Global Species Gaps in Genetic Databases for eDNA Metabarcoding." *Diversity and Distributions* 27, no. 10: 1880–1892.

Mas-Carrió, E., J. Schneider, B. Nasanbat, et al. 2022. "Assessing Environmental DNA Metabarcoding and Camera Trap Surveys as Complementary Tools for Biomonitoring of Remote Desert Water Bodies." *Environmental DNA* 4, no. 3: 580–595.

Mathon, L., F. Baletaud, A. Lebourges-Dhaussy, et al. 2024. "Three-Dimensional Conservation Planning of Fish Biodiversity Metrics to Achieve the Deep-Sea 30 ×30 Conservation Target." *Conservation Biology* 39: e14368.

Mathon, L., V. Marques, S. Manel, et al. 2023. "The Distribution of Coastal Fish eDNA Sequences in the Anthropocene." *Global Ecology and Biogeography* 32, no. 8: 1336–1352.

Mathon, L., A. Valentini, P. E. Guérin, et al. 2021. "Benchmarking Bioinformatic Tools for Fast and Accurate eDNA Metabarcoding Species Identification." *Molecular Ecology Resources* 21, no. 7: 2565–2579.

Matthews, T. J., K. A. Triantis, R. J. Whittaker, and F. Guilhaumon. 2019. "Sars: An R Package for Fitting, Evaluating and Comparing Species–Area Relationship Models." *Ecography* 42, no. 8: 1446–1455.

McCartney, M. A., J. Acevedo, C. Heredia, et al. 2003. "Genetic Mosaic in a Marine Species Flock." *Molecular Ecology* 12, no. 11: 2963–2973.

Moran, B. M., K. Hench, R. S. Waples, et al. 2019. "The Evolution of Microendemism in a Reef Fish (*Hypoplectrus maya*)." *Molecular Ecology* 28, no. 11: 2872–2885.

Morlon, H., J. Andréoletti, J. Barido-Sottani, et al. 2024. "Phylogenetic Insights Into Diversification." *Annual Review of Ecology, Evolution,*

and Systematics 55, no. 1: 1–21. https://doi.org/10.1146/annurev-ecolsys-102722-020508.

Nester, G. M., M. De Brauwer, A. Koziol, et al. 2020. "Development and Evaluation of Fish eDNA Metabarcoding Assays Facilitate the Detection of Cryptic Seahorse Taxa (Family: Syngnathidae)." *Environmental DNA* 2, no. 4: 614–626.

Parravicini, V., M. Kulbicki, D. R. Bellwood, et al. 2013. "Global Patterns and Predictors of Tropical Reef Fish Species Richness." *Ecography* 36, no. 12: 1254–1262.

Pinfield, R., E. Dillane, A. K. W. Runge, et al. 2019. "False-Negative Detections From Environmental DNA Collected in the Presence of Large Numbers of Killer Whales (*Orcinus orca*)." *Environmental DNA* 1, no. 4: 316–328.

Polanco F., A., C. Waldock, T. Keggin, et al. 2022. "Ecological Indices From Environmental DNA to Contrast Coastal Reefs Under Different Anthropogenic Pressures." *Ecology and Evolution* 12, no. 8: e9212.

Polanco Fernández, A., V. Marques, F. Fopp, et al. 2021. "Comparing Environmental DNA Metabarcoding and Underwater Visual Census to Monitor Tropical Reef Fishes." *Environmental DNA* 3, no. 1: 142–156.

Pont, D., M. Rocle, A. Valentini, et al. 2018. "Environmental DNA Reveals Quantitative Patterns of Fish Biodiversity in Large Rivers Despite Its Downstream Transportation." *Scientific Reports* 8, no. 1: 1–13. https://doi.org/10.1038/s41598-018-28424.

Puebla, O., F. Coulmance, C. J. Estapé, A. M. Estapé, and D. R. R. Robertson. 2022. "A Review of 263 Years of Taxonomic Research on *Hypoplectrus* (Perciformes: Serranidae), With a Redescription of *Hypoplectrus affinis* (Poey, 1861)." *Zootaxa* 5093, no. 2: 101–141.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. https://www.R-project.org/.

Rabosky, D. L., J. Chang, P. F. Cowman, et al. 2018. "An Inverse Latitudinal Gradient in Speciation Rate for Marine Fishes." *Nature* 559, no. 7714: 392–395.

Ramon, M. L., P. S. Lobel, and M. D. Sorenson. 2003. "Lack of Mitochondrial Genetic Structure in Hamlets (*Hypoplectrus* Spp.): Recent Speciation or Ongoing Hybridization?" *Molecular Ecology* 12, no. 11: 2975–2980.

Robertson, D. R. 2008. "Global Biogeographical Data Bases on Marine Fishes: Caveat Emptor." *Diversity and Distributions* 14, no. 6: 891–892.

Robertson, D. R., and G. R. Allen. 2015. *Shorefishes of the Tropical Eastern Pacific*. Smithsonian Tropical Research Institute. Online Information System (02/2023).

Rodríguez-Ezpeleta, N., L. Zinger, A. Kinziger, et al. 2021. "Biodiversity Monitoring Using Environmental DNA." *Molecular Ecology Resources* 21, no. 5: 1405–1409.

Rodriguez-Martinez, S., J. Klaminder, M. A. Morlock, L. Dalén, and D. Y. T. Huang. 2022. "The Topological Nature of Tag Jumping in Environmental DNA Metabarcoding Studies." *Molecular Ecology Resources* 23, no. 3: 621–631.

Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. "VSEARCH: A Versatile Open Source Tool for Metagenomics." *PeerJ* 4: e2584.

Ryberg, M. 2015. "Molecular Operational Taxonomic Units as Approximations of Species in the Light of Evolutionary Models and Empirical Data From Fungi." *Molecular Ecology* 24, no. 23: 5770–5777.

Salvi, D., E. Berrilli, P. D'Alessandro, and M. Biondi. 2020. "Sharpening the DNA Barcoding Tool Through a Posteriori Taxonomic Validation: The Case of *Longitarsus* Flea Beetles (Coleoptera: Chrysomelidae)." *PLoS One* 15, no. 5: e0233573.

Schenekar, T., M. Schletterer, L. A. Lecaudey, and S. J. Weiss. 2020. "Reference Databases, Primer Choice, and Assay Sensitivity for Environmental Metabarcoding: Lessons Learnt From a Re-Evaluation of an eDNA Fish Assessment in the Volga Headwaters." *River Research and Applications* 36, no. 7: 1004–1013.

Schnell, I. B., K. Bohmann, and M. T. P. Gilbert. 2015. "Tag Jumps Illuminated—Reducing Sequence-To-Sample Misidentifications in Metabarcoding Studies." *Molecular Ecology Resources* 15, no. 6: 1289–1303.

Sevellec, M., A. Lacoursière-Roussel, L. Bernatchez, et al. 2021. "Detecting Community Change in Arctic Marine Ecosystems Using the Temporal Dynamics of Environmental DNA." *Environmental DNA* 3, no. 3: 573–590.

Siqueira, A. C., R. A. Morais, D. R. Bellwood, and P. F. Cowman. 2020. "Trophic Innovations Fuel Reef Fish Diversification." *Nature Communications* 11, no. 1: 2669.

Somervuo, P., D. W. Yu, C. C. Xu, et al. 2017. "Quantifying Uncertainty of Taxonomic Placement in DNA Barcoding and Metabarcoding." *Methods in Ecology and Evolution* 8, no. 4: 398–407.

Taberlet, P., A. Bonin, L. Zinger, and E. Coissac. 2018. *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford University Press.

Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann, and E. Willerslev. 2012. "Towards Next-Generation Biodiversity Assessment Using DNA Metabarcoding." *Molecular Ecology* 21, no. 8: 2045–2050.

Takahashi, M., M. Saccò, J. H. Kestel, et al. 2023. "Aquatic Environmental DNA: A Review of the Macro-Organismal Biomonitoring Revolution." *Science of the Total Environment* 873, no. 162: 322.

Thomas, A. C., S. Tank, P. L. Nguyen, J. Ponce, M. Sinnesael, and C. S. Goldberg. 2020. "A System for Rapid eDNA Detection of Aquatic Invasive Species." *Environmental DNA* 2, no. 3: 261–270.

Thomsen, P. F., J. Kielgast, L. L. Iversen, P. R. Møller, M. Rasmussen, and E. Willerslev. 2012. "Detection of a Diverse Marine Fish Fauna Using Environmental DNA From Seawater Samples." *PLoS One* 7, no. 8: e41732.

Valentini, A., P. Taberlet, C. Miaud, et al. 2016. "Next-Generation Monitoring of Aquatic Biodiversity Using Environmental DNA Metabarcoding." *Molecular Ecology* 25, no. 4: 929–942.

Weingand, H., A. J. Beermann, F. Čiampor, et al. 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.

Yang, C. Q., Y. Wang, X. H. Li, et al. 2024. "Environmental Niche Models Improve Species Identification in DNA Barcoding." *Methods in Ecology and Evolution* 15: 2343–2358.

Zhang, S., J. Zhao, and M. Yao. 2020. "A Comprehensive and Comparative Evaluation of Primers for Metabarcoding eDNA From Fish." *Methods in Ecology and Evolution* 11, no. 12: 1609–1625.

Zinger, L., A. Bonin, I. G. Alsos, et al. 2019. "DNA Metabarcoding—Need for Robust Experimental Designs to Draw Sound Ecological Conclusions." *Molecular Ecology* 28, no. 8: 1857–1862.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.