ORIGINAL ARTICLE

DNA metabarcoding of mock communities highlights potential biases when assessing Neotropical fish diversity

Heron Oliveira Hilário[1] | Izabela Santos Mendes[1,2] | Naiara Guimarães Sales[3] | Daniel Cardoso Carvalho[1,2]

[1]Conservation Genetics Lab, Postgraduate Program in Vertebrate Biology, Pontifical Catholic University of Minas Gerais, PUC Minas, Belo Horizonte, Brazil

[2]Postgraduate Program in Genetics, Institute of Biological Sciences, Federal University of Minas Gerais, Belo Horizonte, Brazil

[3]School of Science, Engineering and Environment, University of Salford, Salford, UK

Correspondence
Daniel Cardoso Carvalho, Conservation Genetics Lab, Postgraduate Program in Vertebrate Biology, Pontifical Catholic University of Minas Gerais, PUC Minas, Belo Horizonte, Brazil.
Email: carvalho.lgc@gmail.com

Naiara Guimarães Sales, School of Science, Engineering and Environment, University of Salford, Salford, UK.
Email: n.guimaraessales@salford.ac.uk

Abstract

Despite the increasing popularity of DNA metabarcoding in the assessment of aquatic ecosystems using fish eDNA or ichthyoplankton, challenges have hampered its broader application in the Neotropical freshwaters. Using five mock communities composed of fish species from two Neotropical River basins, we evaluated the influence of DNA concentration and choice of mitochondrial 12S molecular markers (MiFish, NeoFish, and Teleo) on species detection and relative read abundance (RRA) using DNA meta-barcoding. Of the three 12S markers analyzed, only MiFish detected all species from all mock communities. The performance of a taxonomy-free approach using ASV/MOTUs was not as precise as assigning DNA reads to species using a curated 12S library that includes approximately 100 fish species since more than one ASV/MOTU was observed for the same specimen. Thus, here we showcase the importance of a custom reference database to allow precise assignment of Neotropical fish species in metabarcoding studies and that the RRA is dependent on community composition, marker, and DNA concentration. We highlight the importance of controlled experiments using known species communities before large investments are made in assessing biodiversity using non-invasive methods that apply DNA metabarcoding.

KEYWORDS
12S gene, ASV, biodiversity, bioinformatics, freshwater fish, MOTU

1 | INTRODUCTION

Metabarcoding has been recognized as a powerful tool for non-invasive biomonitoring and management of biodiversity (Boivin-Delisle et al., 2020; Deiner et al., 2017), including megadiverse realms such as the Amazonian ichthyofauna (Cantera et al., 2022).

The use of universal primers and high-throughput sequencing have enabled the assessment of biodiversity and community composition for target taxa from many types of environmental samples, unraveling patterns linked to environmental variables or human impacts (Deiner et al., 2017; Elbrecht & Leese, 2015; Hänfling et al., 2016; Riaz et al., 2011).

Heron Oliveira Hilário and Izabela Santos Mendes should be considered joint first author.

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

HILÁRIO ET AL.

Despite the increasing popularity of this method in the assessment of aquatic ecosystems using fish eDNA or ichthyoplankton pools (Carvalho, 2022; Cilleros et al., 2019; Sales et al., 2019; Sales et al., 2021, Silva et al., 2023), the vast majority of fish metabarcoding studies are still concentrated in temperate regions, in well-characterized and reasonably accessible environments (Lawson Handley et al., 2019; McDevitt et al., 2019). Considering the highly diverse Neotropical ichthyofauna, appropriate implementation of metabarcoding biomonitoring is hindered by a lack of knowledge of the local biodiversity (e.g., undescribed or cryptic species), primer biases, and the incompleteness of reference databases (Jackman et al., 2021; Sales, Mariani, et al., 2018; Sato et al., 2017). Additionally, the multiple bioinformatic pipelines used for sequence processing and taxonomic identification (Cilleros et al., 2019; Jackman et al., 2021; Sales et al., 2019) may also influence the ultimate resolution of species detected in a given sample (Majaneva et al., 2018).

Since its adoption as the marker of choice for DNA barcoding of animal species, a vast amount of sequence, taxonomic, and metadata information has been amassed on the cytochrome oxidase subunit 1 (cox1 or COI), especially through the efforts of the "Consortium for the Barcode of Life" (Porter & Hajibabaei, 2018). Despite this knowledge, the internal low conservation of COI markers hampers the development of shorter mini-barcodes, and usually requires primer cocktails to enable balanced polymerase chain reaction (PCR) amplification of all target fish species, which also results in side amplification of bacteria and other non-target taxa (Collins et al., 2019). Ideally, DNA markers used for metabarcoding must also correspond to a highly polymorphic locus and also facilitate the design of highly specific primers (Deagle et al., 2014). Concerning fish metabarcoding, the 12S rRNA mitochondrial gene has emerged as an alternative marker to COI, leading to the development of many primers targeting this region (Zhang et al., 2020). For instance, Miya et al. designed the MiFish markers for fish eDNA amplification based on 880 mitochondrial genomes of mostly marine fishes (2015), while Valentini et al. (2016) developed the Teleo marker from 86 fish species—64 European freshwater and 22 marines. Together, these two primer sets represented the marker of choice in over 52% of recent studies involving fish eDNA (Xiong et al., 2022). More recently, aiming to investigate Neotropical fish biodiversity, Milan et al. (2020) developed the NeoFish marker based on the 12S region of 132 fishes of the Neotropical São Francisco River Basin. These three markers target the 12S gene in adjacent regions and were initially considered able to contain enough genetic variation to allow the differentiation of most fish species. However, Mifish and Teleo have already been demonstrated to fail in the detection of some fish taxa (Jackman et al., 2021; Zhang et al., 2013) and, to this date, no comprehensive and comparative evaluation of their amplification or taxonomic classification for fish from megadiverse Neotropical rivers has been conducted, which hinders informed decisions regarding their suitability for different study systems.

Regarding the processing of sequencing data and taxonomic assignment, eDNA studies of the Neotropics have mostly relied on clustering techniques based on molecular operational taxonomic units (MOTUs), with the application of a 97% sequence similarity threshold for species identification (e.g., Cilleros et al., 2019; Sales et al., 2021). More recently, studies have started to demonstrate the need of including distinct approaches (i.e., ZOTUs and ASVs) to describe the recovered eDNA diversity (Dal Pont et al., 2021). Clustering methods merge a set of sequences into meaningful biological identities, which are hence considered as a proxy for species. Alternatively, the amplicon sequence variants (ASVs) (or ESVs, exact sequence variants) have been proposed to provide a finer resolution, allowing the detection of single-nucleotide differences over the sequenced gene region (Callahan et al., 2017). Sometimes interpreted as intraspecific variability, which can be a desirable output of eDNA, ASVs may provide additional ecological signals that can be extrapolated from when considering species (García-García et al., 2019). This sequence diversity might provide an increased resolution but could also be prone to harbor sequencing errors (e.g., errors originating from PCR amplification and sequencing artifacts). Recently, the importance of combining both denoising and clustering procedures has been raised by Antich et al. (2021) when considering the standard barcode fragment of COI. Yet, to our knowledge, an evaluation of applying both methods using mock communities to uncover the real advantages of each approach has still not been applied to other markers (i.e., markers targeting the 12S gene region) targeting the Neotropical megadiverse fish communities. The strengths and weaknesses of the molecular markers used for eDNA metabarcoding can be revealed through the investigation of mock communities which enables the identification not only of species that failed to be amplified but also the extent to which biased amplification occurs.

In this study, we aim to better understand the use of metabarcoding in the detection of Neotropical fish diversity by evaluating the influence of the following features of sample composition, using fish mock communities: (1) how the proportion of input DNA and (2) molecular marker choice influences the detection of species in mock communities and (3) to compare the performance of taxonomy-free approach using ASV/MOTUs and assignment of DNA reads to species using a curated library (Figure 1). Moreover, a robust database for the 12S region was built to enable metabarcoding studies of eastern shield Neotropical species that can be used to evaluate primers' complementarity on species detection and to investigate preferentially amplified species. Our results highlighted the importance of controlled experiments using known species communities and the availability of a robust reference database for metabarcoding studies, especially in megadiverse realms.

2 | MATERIAL AND METHODS

2.1 | Building a local 12S reference database

A custom reference database was developed using previously molecularly identified specimens using DNA barcoding and traditional taxonomy (de Carvalho et al., 2011; Pugedo et al., 2016) deposited at the Conservation Genetics Lab collection (PUC Minas University).
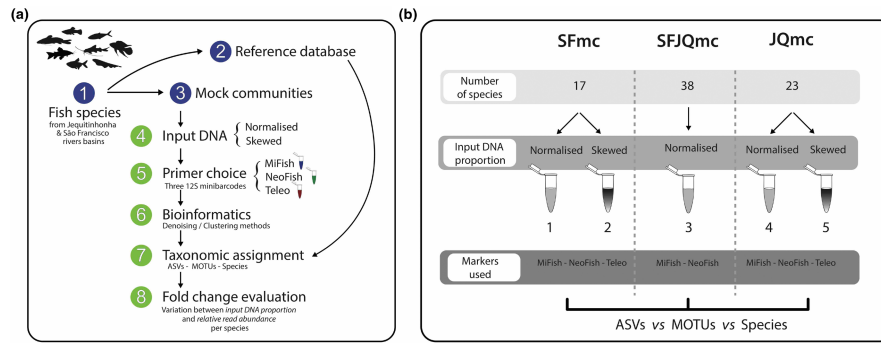
FIGURE 1 Study framework (a) and description of analyzed mock communities including different species composition and DNA input, markers evaluated, and proxies used for biodiversity assessment (b). All mock communities built using species from the São Francisco River Basins (SFRB) and Jequitinhonha River Basin (JQRB). (1) Normalized São Francisco River mock community (SFmc) comprising 23 species from the SFRB with the same concentration of DNA (10 ng/μL) and (2) SFmc skewed using different DNA concentrations. (3) São Francisco and Jequitinhonha combined mock community (SFJQmc) built with 38 unique species from the normalized Jequitinhonha River mock community (JQmc) JQmc and SFmc. (4) normalized Jequitinhonha River mock community (JQmc) composed of 23 species from the JQRB built using equal concentrations of DNA. (5) JQmc skewed: mock community composed of 23 species from the JQRB built with skewed concentrations of DNA.

The 12S reference library was built using a DNA fragment of 600 bp obtained by Sanger DNA sequencing as described in Milan et al., 2020. The targeted 12S fragment includes the region used by three 12S markers: (1) NeoFish (~190 bp, Milan et al., 2020); (2) MiFish (~170 bp, Miya et al., 2015); and (3) Teleo (~70 bp, Valentini et al., 2016).

The *Conservation Genetics Lab 12S Sequence Database* (*LGC 12Sdb* hereafter) comprises 187 sequences representing 99 fish species (Table S1) collected from the Jequitinhonha (JQ) and São Francisco (SF) River Basins. Considering the Jequitinhonha River Basin, 40 species are represented by 53 sequences obtained in this study (complete description of the fish collection, sample storage, and processing available in Data S1). From the São Francisco River Basin, 70 fish species are represented by 134 sequences (corresponding to sequences obtained in this study and combined with data retrieved from Milan et al., 2020). The 12S sequences were deposited into GenBank and are available under the accession numbers: OR198493–OR198679.

2.2 | Assemblage of mock communities

All mock communities were constructed using known ratios of input DNA, including DNA from fish specimens also included in the aforementioned reference database (Figure 1). A total of five mock communities were analyzed, including two composed of 17 species collected from the Jequitinhonha River Basin (JQRB), namely JQmc (Table S2), of which one was constructed using equal concentrations of DNA (JQmc normalized) and the other one based on non-normalized or skewed concentrations of DNA (JQmc skewed). Two other mock communities were constructed comprising 23 species from the São Francisco River Basin (SFRB) (SFmc, Table S2), also including one normalized (SFmc normalized) and one skewed version (SFmc skewed). A fifth mock community contained 38 unique species used to build the SFmc and JQmc combined and was normalized (SFJQmc, Table S2).

DNA extraction was carried out using samples of the same tissues used for the construction of the LGC 12Sdb, with the Wizard® Genomic DNA Purification Kit (Promega). DNA was extracted from fish fin clips for all samples included in the mock communities to ensure consistency and avoid potential biases related to mitochondrial/genomic DNA ratio. DNA integrity was assessed using 2% agarose gel electrophoresis, and DNA concentration was determined using *Qubit™ 1x dsDNA HS* assay (Thermo Fisher). Normalized mock communities contained all DNA extracts pooled at a concentration of 10 ng/μL DNA per analyzed species, whereas skewed mock communities contained the original DNA concentration obtained from the tissue DNA extractions (corresponding to between 0.26% and 51.58% of the total DNA, Table S2).

2.3 | Amplification, library preparation, and sequencing

The JQmc was amplified with all three markers (NeoFish, MiFish, and Teleo) (Table S3); SFmc and SFJQmc were both amplified only with NeoFish and MiFish primers because they have similar amplicon sizes and are suitable for combining in a single HTS Illumina run. Each mock community was amplified independently in three replicates for each marker. Each sample/primer combination was first amplified in 25 μL PCR triplicates using the high-fidelity AmpliTaq Gold (Applied Biosystems). To account for possible tag jumping and contamination, one positive control, not included in any of the analyzed mock communities, was used including a marine cartilaginous fish DNA (*Prionace glauca*).

Twelve libraries were prepared using the three markers, NeoFish, MiFish, and Teleo. Markers were designed with 5′ Illumina library construction adaptors, and a two-step PCR procedure was

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

HILÁRIO ET AL.

used for library preparation using the *Nextera XT kit* (Illumina) following the manufacturer's instructions. In the first step, a fragment of the mitochondrial 12S rRNA gene was amplified using the MiFish, NeoFish, and Teleo markers. The first PCR reaction was performed on a 25 μL final volume, using AmpliTaq Gold, with 2 μL DNA. The thermal cycles for this step were as follows: initial denaturation at 95°C for 2 min, followed by 35 cycles of denaturation at 95°C for 1 min, annealing at 55°C (Teleo), 60°C (MiFish), and 60°C (NeoFish) for 30 s, and elongation at 72°C for 1 min, followed by final elongation at 72°C for 7 min. PCR products were purified using *Agencourt AMpure* beads (Beckman Coulter) according to the manufacturer's instructions and used as templates for the second PCR. The Illumina sequencing adaptors plus the 8-bp identifier indices were added in the subsequent PCR using a forward and reverse fusion primer (Table S4). The second PCR reaction was performed on a 25 μL final volume, containing 10 μL 2.51×KAPA HiFi HotStart ReadyMix (KAPA Biosystems), 5 μL of each primer, and 5 μL of the first PCR products. The thermal cycles of the second PCR were as follows: initial denaturation at 95°C for 3 min, followed by eight cycles of denaturation at 98°C for 20 s, annealing and elongation combined at 72°C for 15 s, with a final elongation at 72°C for 5 min. DNA concentration was estimated using the *Qubit 1x dsDNA HS* assay kit and a Qubit fluorometer (Thermo Fisher Scientific). Libraries were pooled in equimolar concentrations and sequenced on the MiniSeq platform (Illumina) using the Mid-output Kit and 2×150 bp paired-end sequencing following the manufacturer's instructions.

2.4 | Bioinformatic analyses

All bioinformatics analyses were conducted in R v.4.1.1 (R Core Team, 2021), using a pipeline based on DADA2 v1.16.0 (Callahan et al., 2016) and phyloseq v1.34.0 (McMurdie & Holmes, 2013). R scripts are fully available on Github (https://github.com/heronoh/fish_eDNA). Raw sequence data are available at https://zenodo.org/.

In brief, reads were submitted to the removal of undetermined bases and quality filtering (Q-scores ≥ 30). Only reads containing the expected primer sequence corresponding to each sample were kept for downstream analysis. Error correction, read-pair merging, and chimaera identification and removal were performed using the default settings of DADA2 functions. No length truncation was performed since primer removal automatically clips uninformative regions, but the resulting ASVs beyond the expected amplicon length range for each marker were discarded (NeoFish: >185 & <200; MiFish: >165 & <180; Teleo: >60 & <75). Subsequently, identified ASVs were clustered into MOTUs using SWARM v3.1.0 (Mahé et al., 2014, 2015), using the *fastidious* option and *d = 1*. Taxonomic assignments were conducted using the DADA2 RDP classifier and the *LGC 12Sdb*. Additionally, all ASVs were searched against local NCBInr using an automated BLASTn 2.10.1+ function with minimum similarity and minimum coverage (-perc_identity 95 and -qcov_hsp_perc 95). After taxonomic assignment and in order to curate and correct partial assignments and unexpected species, phylogenetic

trees were built for each marker using the correspondent ASVs and the marker amplicon region for all database species. Sequences were aligned using MUSCLE, and a maximum-likelihood tree was generated using MEGAX.

The relative read abundance (RRA) was obtained for each ASV, dividing the number of copies of the ASVs by the sum of the number of copies of all ASVs in the sample. A test of significance was conducted using an analysis of similarity (ANOSIM) to compare primer's depiction of both normalized and skewed versions of the same mock community, and the relationship between RRA and input DNA concentration for each sample was further assessed through Pearson's correlation coefficient. Linear regression models between the proportion of input DNA and RRA were generated for each primer, considering each species separately, to evaluate biases on amplification (Tables S5A–C). It was not possible to evaluate the difference between normalized and skewed mock communities since the narrow distribution of the input DNA on the normalized versions prevented reliable regressions.

We then used the actual proportion of the corresponding species' DNA in the respective mock community to calculate a *fold change* between DNA input and the resulting RRA for each species, using the formula *(frac{num}{denom}) if (num > denom), and as (frac{−denom}{num}) if (num < denom)*, and the resulting values were used to generate bar plots and boxplots.

The total counts of ASVs, MOTUs, and correctly identified species were used to compare diversity estimates for each mock community and primer. A *fold change* between −1 and 1 corresponds to a 1:1 relative proportion of input DNA and sequence counts on samples. A *fold change* <−1 indicates a lower abundance of the species among the sequences than its actual abundance in the mock community input DNA. *Fold-change* values >1 indicate that the species is more represented on the sequences. For example, a *fold change* of 10 indicates that the species was 10 times more abundant on the ASVs than it was on the input DNA. An optimal primer pair would recover fold-change values close to zero, with a minimum variation between taxa. It is important to highlight that all the markers we targeted here occur in multiple copies in each cell which can skew read abundances (Krehenwinkel et al., 2017). Hence, fold-change values of zero will rarely be found, even in the absence of amplification bias. However, as amplification bias amplifies exponentially with PCR cycle number, its effect on fold change should be more severe than simple copy number variation.

3 | RESULTS

3.1 | 12S reference database

The updated 12S reference database (*LGC 12Sdb*) is comprised of 187 specimens from the São Francisco (SF) and Jequitinhonha (JQ) River Basins, with sequences ranging from 568 to 721 bp (663 bp on average), representing 73 genera, 28 families, 9 orders, and 99 species; 15 species of which are shared between the basins and are

represented by different specimens from each drainage. Considering the known ichthyofauna, the species diversity of the database corresponds to approximately 75% for the JQ and 46% for the SF River Basins. The database encompasses the amplicon regions of the three 12S markers NeoFish, MiFish, and Teleo, but, for the latter two, some species are partially represented after trimming low-quality bases on the ends of the alignment.

Concerning the species used in the mock communities, the database covers 100% of the NeoFish amplicon, an average of 86% of the MiFish amplicon total length (min: 69.6%, max: 97.24%, on the 5′ end), and 88.89% of the Teleo amplicon (min: 81.49%, max: 97.23%, on the 3′ end) (Figure S1). These reference sequences, despite their length, allowed the correct identification of almost all species included in the mock communities to the species level. The exceptions were for two species pairs: *Prochilodus argenteus* and *P. hartii*, and *Hoplias brasiliensis* and *H. intermedius*, which shared identical amplicon sequences for all three analyzed markers despite complete amplicon sequencing. Therefore, due to this hindrance in the taxonomic resolution, each species pair was considered as a single entity (*P. argenteus/P. hartii* and *H. brasiliensis/H. intermedius*) leading to a total of 38 molecular identifiable species added to the mock communities. Thus, for all further analyses (primer performance; RRA vs. input DNA, and others), each of these species' pairs were considered as a single entity.

3.2 | Data quality control and processing

A total of 2.66 million raw reads were obtained for the five mock communities (1.34 million—NeoFish, 0.92 million—MiFish, and 0.39 million—Teleo). After raw data processing to remove undetermined bases and primers, read merging, chimaera removal, and error correction, a total of 1.57 million sequences were retained for the three analyzed markers: 0.60 million for NeoFish (45%), 0.65 million for MiFish (70%), and 0.31 million for Teleo (79%). Sequencing read depth per mock community was 174,016 reads on average, with a minimum of 150,286 reads. Read counts along all pipeline steps are detailed in Table S6.

3.3 | Primers' performance on species detection

Of the three 12S markers analyzed, only MiFish was able to detect all species from all mock communities. The NeoFish marker was able to detect 34 of 38 species in the SFJQmc (89.5%), 20 of 23 in the SFmc (86.96%), and 16 of 17 in the JQmc (94.11%) (Figure 2). NeoFish presented a poor performance on the Pseudopimelodidae, Pimelodidae and Heptapteridae families, failing to detect four species *Imparfinis minutus*, *Microglanis leptostriatus*, *Pimelodus pohli*, and *Steindachneridion amblyurum* in all mock communities. Teleo detected 16 of 17 species (94.11%) in the normalized JQmc, failing to detect *Prochilodus costatus*; and 15 of 17 species in the skewed JQmc, missing *P. costatus* and *A. lacustris*.

Concerning ASVs, for the 38 species present in the SFJQmc, MiFish recovered 48 ASVs while NeoFish retrieved 42 ASVs (Figure 2). Considering the 23 species in the SFmc, MiFish recovered 29 ASVs in both normalized and skewed mock communities, while NeoFish recovered 26 and 24 ASVs in the normalized and skewed versions, respectively (Figure 2). For the normalized and skewed JQmc, both MiFish and Teleo retrieved 17 ASVs, and NeoFish recovered 33 ASVs in the normalized, but only 20 ASVs in the skewed version (Figure 2).

P. maculatus, from SFmc, was the species with the highest number of ASVs generated with MiFish, with three in the normalized version and four in the non-normalized. For NeoFish, *Tetragonopterus chalceus* presented four to five ASVs, and *M. garmani* had one to six ASVs recovered. For the Teleo marker, *Hypostomus nigrolineatus* and *Delturus brevis* had two ASVs each (Table S5A–C).

3.4 | Fold variation between input DNA and RRA

MiFish detected some species with mean fold change lower than −10× (e.g., *Microglanis leptostriatus*, *Pterygoplichthys etentaculatus*, *Pamphorichtys hollandi*, and *Myleus micans*). The species with the lowest fold-change value were *Myleus micans* (< −50×) and *Pamphorichthys hollandi* (< −200×). In contrast, *Moenkhausia costae* had a fold increase of 13.9×, and *Pimelodus maculatus* had an increase of ~5×. For *Gymnotus carapo*, fold change varied from −1.12× to 9×; *Pterygoplichthys etentaculatus*, between −7.5× and −12.2×; and *Tetragonopterus chalceus*, between −1.16× and 40×; *Tetragonopterus chalceus* also had a great variation, ranging from −2.7× to 49.9× (Figure 3).

The Teleo marker detected *Astyanax lacustris* in the normalized JQmc with an extremely low RRA (0.0000246%; fold change < −500×) but was completely missed on the non-normalized pool. Other underrepresented species for this marker were as follows: *Gymnotus carapo* (~130× fold decrease), *Delturus brevis* (~10× fold decrease), *Rhamdia quelen* (~7.5× fold decrease), and *Hypostomus nigrolineatus* (~5× fold decrease). On the other hand, *Megaleporinus garmani* was overrepresented, with a fold increase of ~7× (Figure 3).

The three primers showed similar fold-change profiles for some species, such as *R. quelen*, *W. maculata*, *T. galeatus*, *H. nigrolineatus*, *D. brevis*, and *H. steindachneri*, which were all under amplified. Conversely, *M. garmani* and *C. gilbert* were overamplified by all primers. The species missed by NeoFish presented mild negative fold variation with MiFish, between −2 and −10. The only species not detected by Teleo (*P. costatus*) was not under- or overamplified by MiFish and NeoFish.

3.5 | Correlation between input DNA and RRA

Comparisons between input DNA and correspondent RRA showed a significant correlation only for the skewed JQmc amplified by NeoFish, indicating a moderate correlation, with $r^2 = 0.419$
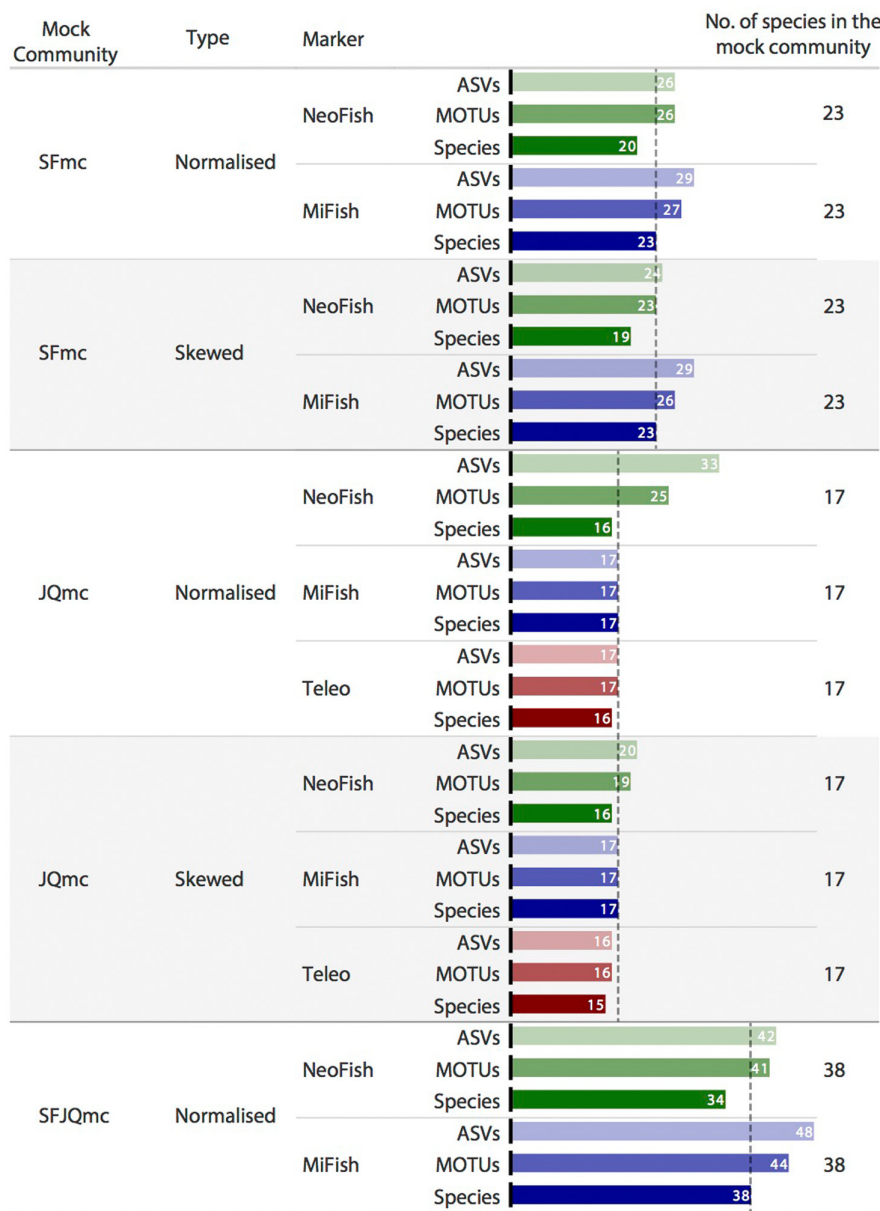
FIGURE 2 Numbers of ASVs, MOTUs, and assigned species for all mock communities and markers. Bars represent the values for each category. Markers are identified by colors: NeoFish—green, MiFish—blue, and Teleo—red. The dotted lines highlight the number of species included in each mock community. Number of species DNA added in each mock community: SFJQmc=38; SFmc=23; and JQmc=17.

(p-value=0.0123; Figure 4). For the other markers and mock communities, no significant correlations were obtained, but MiFish had similar regression patterns for SFmc and JQmc, presenting the closest profiles for regression lines. NeoFish had different profiles for the mock communities, with a less steep regression line for the SFmc. When mock communities were evaluated in combination (i.e., SFJQmc), no significant correlation was observed (Table S7).

4 | DISCUSSION

Despite a sharp increase in DNA metabarcoding studies, their potential to reveal biodiversity patterns in the Neotropics remains largely unexplored (Zinger et al., 2020). Here, we present a 12S fish reference database for approximately 100 fish species that allows a broader range of species assignment using non-invasive molecular

methods. We showcased the influence of community composition, proportion of input DNA, molecular marker, and taxonomy-free versus species assignment caveats on species detection using mock communities.

Although all three 12S molecular markers (Mifish, Neofish, and Teleo) allowed amplification and detection of most species from the mock communities, MiFish had the best performance and was the only marker able to identify most species (Figure 2). These findings are in agreement with Polanco et al. (2021), who compared the performance of MiFish and Teleo using *in silico* PCR, and environmental samples from a tropical and temperate origin. They reported that both markers had dropouts and recommended the use of multiple primers to increase species detection probability using combinations that allow amplification of all target species. However, Neotropical fish biodiversity is far more complex than the mock communities used in the present work. For instance, the SF River Basin hosts
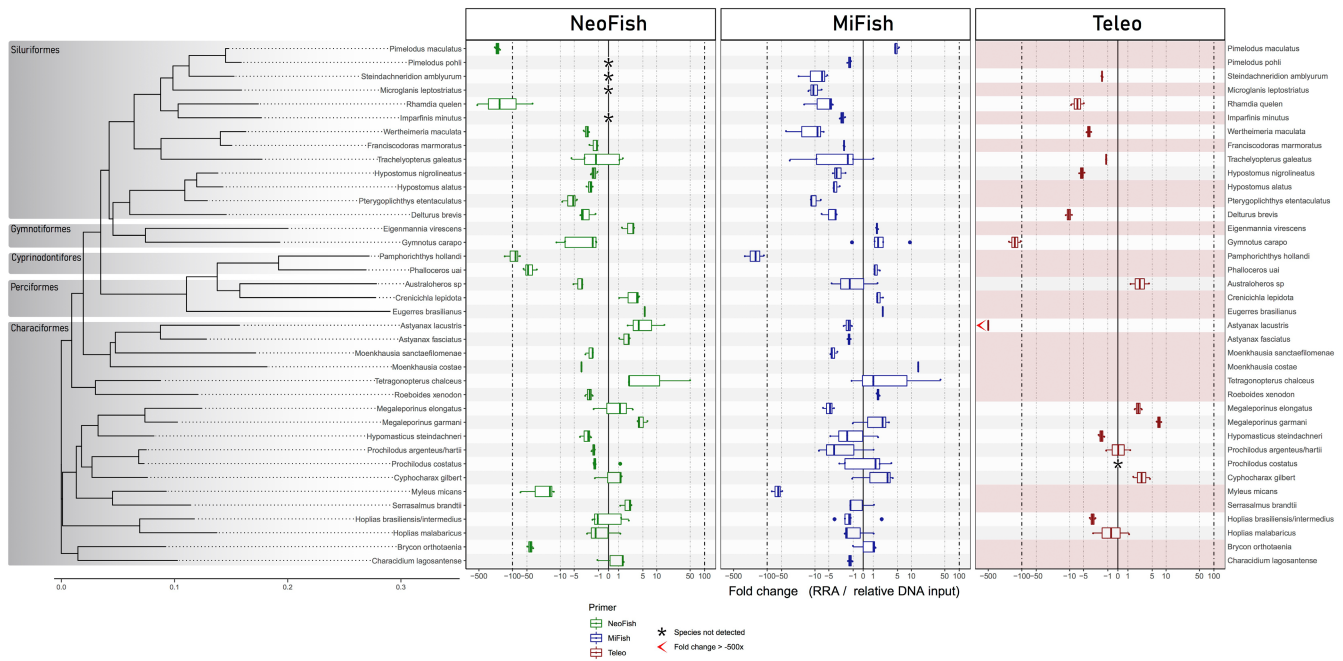
FIGURE 3 Fold variation between input DNA and relative read abundance (RRA) for SFmc, JQmc, and SFJQmc species. Bars represent the range of the fold change between the input DNA and RRA for each species and marker. Undetected species are marked by *. Light red spaces indicate species absent in the mock communities amplified with Teleo.
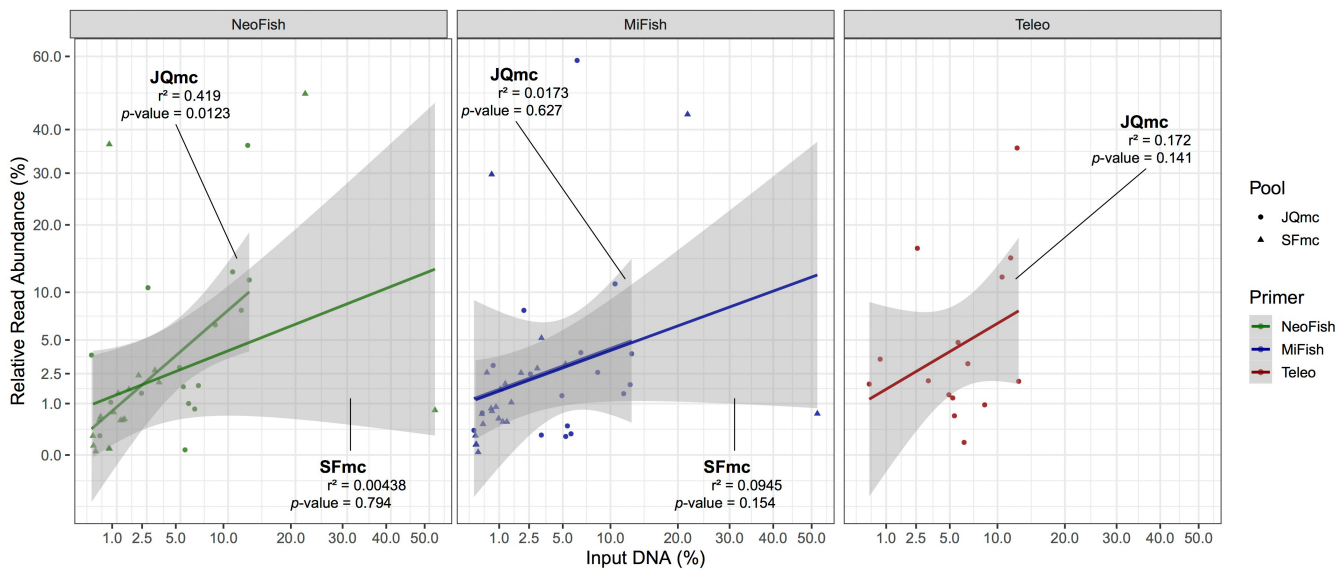

FIGURE 4 Correlations between relative read abundance (RRA) and correspondent input DNA added in each mock community. Each dot represents one species in a mock community. Markers are identified by colors: NeoFish—green, MiFish—blue, and Teleo—red.

at least 300 fish species (Barbosa et al., 2017). Therefore, it is important to consider that the identification outcomes can differ from different species assemblages and, in order to investigate less characterized and diverse environments such as the Neotropical rivers, the use of multiple and complementary markers should be considered, especially in preliminary eDNA studies where target biodiversity is not known.

None of the 12S short amplicons used in this study presented enough molecular resolution to discriminate between the congeners species *Prochilodus argenteus/hartii* and *Hoplias brasiliensis/intermedius*. However, it is known that these genera have issues concerning their taxonomy (Rosso et al., 2018), and hybrids exist (Sales, Pessali, et al., 2018). Moreover, Santos et al. (2021) reported that the ribosomal 12S and 16S genes have the lowest sequence variation in the whole mitochondrial genome of *Prochilodus* species. Therefore, other loci could be evaluated in the future to further improve these species' resolution issues, if necessary, its discrimination.

8 | **WILEY**— Environmental DNA
Dedicated to the study and use of environmental DNA for basic and applied sciences
Open Access

HILÁRIO ET AL.

The NeoFish marker, recently developed to detect Neotropical fish (Milan et al., 2020), did not perform well to detect Siluriformes species, especially from the Heptapteridae family. These species share a divergent base at the 3′ end of the forward primer that may explain such poor performance (Table S8). In order to overcome this limitation and with the use of our custom database, we propose a simple redesign by displacing the NeoFish forward primer with a single base toward the 5′ (NeoFishF_v2: 5′CCGCCGTCGCAAGCTTACCC-3′). It is known that even universal primers designed using large databases such as MiFish often require different versions to be able to recover and identify all species (Miya et al., 2015). For instance, when Miya et al. (2019) observed that MiFish primer did not amplify DNA of 10 species of sea sculpins from four genera *Pseudoblennius*, *Furcina*, *Ocynectes*, and *Vellitor* (Scorpaeniformes: Cottidae), which are common reef-associated fish species along the Pacific coast of central Japan, the authors suggested the use of optimized primers to accommodate the sequence variation of these taxa. Furthermore, a series of later experiments provided reasonable results that more accurately reflected the regional fish community compositions. The effectiveness of such a modification in the NeoFish marker should be tested in a future study.

Using mock communities with different input DNA proportions, we could simulate real eDNA sampling scenarios where rare target species may present low DNA concentration and thus may be proportionally under-amplified. In the present study, both NeoFish and Teleo detected less species in the skewed versions of the mock communities, retrieving one species less than the normalized versions. These species were present at lower than 0.6% of the input DNA in the mock skewed communities (*Pimelodus maculatus*, with 0.58% on skewed SFmc, not detected by NeoFish, and *Astyanax lacustris*, with 0.26% on skewed JQmc, not detected by Teleo). Therefore, the use of deeper sequencing, with more reads per sample, is advised to enable rare species detection, as proposed by Van der Loos and Nijland (2021), who also highlighted the importance of mock communities as positive controls to enable the evaluation of the quantitative signal and to define correction factors.

Despite some species presenting similar amplification profiles and fold changes by all three markers, other species behaved differently for each molecular marker, being under- or overrepresented on the yielded reads, or even not amplified, a pattern known as "PCR dropouts" (Miya et al., 2020). This contrasting amplification profile between different markers is known to result from many factors, but primer annealing leading to biased amplification would be a plausible explanation. However, quality of the tissue sample, tissue age, and preservation of samples would also affect the ratio of mitochondrial/genomic DNA. Even though we have used the same type/age of tissue, fixed with ethanol, the mitochondrial/genomic DNA ratio could affect RRA fold variation. Nonetheless, the effect on RRA fold variation would be similar for all three markers, allowing us to control for bias using our multimarker metabarcoding approach.

This is the first study to estimate fold changes from input DNA to RRA for fish species using mock communities of different species compositions. Our results demonstrated that fold changes from input DNA to RRA are dependent on the primer/species pair. Understanding the primer interaction with the species, both in silico through mismatch detection and PCR and in vitro through fold-change evaluation, could enable calibration of field experiments with the application of corrections for species/primers with known under- and over-amplification behavior. Similarly, Duke and Burton (2020) and Sard et al. (2019) found generalized differences between primer pairs and their ability to detect species in a mock community. Future studies could tease apart the effects of amplification bias from other variables that may impact detection thresholds. In fact, McLaren et al. (2019) developed a model for correcting RRA of samples by the use of reference mock communities of known composition and pointed out that RRA outcome of a given species depends not only on its amplification profile by the chosen marker but also on the amplification profiles of the other species in the mixture.

Both ASV and MOTU measurements slightly overestimated richness, with values higher than the original number of species present in the mock communities, since a single specimen yielded multiple ASVs/MOTUs. Thus, if ASVs or MOTUs were used as a proxy for species this could lead to an overestimation of species richness. Therefore, the use of taxonomy-free metabarcoding, which has been proposed in the investigation of highly diverse and poorly characterized species assemblies (Mächler et al., 2021), as an alternative tool to enable ecological assessments of megadiverse Neotropical fish fauna, should be used with caution. The sequence diversity represented on these proxies may arise from many sources, including intraspecific marker diversity as well as errors incorporated in PCR and sequencing steps (Tsuji et al., 2020). Moreover, in our study, the mock community from the same basin amplified with the same marker in normalized and skewed configurations yielded different numbers of ASVs and MOTUs but had identical species assigned. Furthermore, this discrepancy was more pronounced in the mock community with the highest number of species in its composition.

The cut-off necessary to remove putative contaminations and/or tag jumping (RRA ≤ 0.02%) resulted in the loss of three species for the NeoFish marker and one for both MiFish and Teleo (Tables S5A–C). Although the use of minimum cut-off values for RRA has been proposed to mitigate contaminations within eDNA metabarcoding studies (Shirazi et al., 2021), this practice is subjective and can result in the removal of species that are present in the environment.

Thus, here we showcased the importance of a custom reference database to allow precise assignment of Neotropical fish species in metabarcoding studies, and that the RRA is dependent on community composition, marker, and DNA concentration. Moreover, since species concentration influences amplification and the number of ASVs/MOTUs recovered for a single species, studies focusing on estimating species abundance using metabarcoding must consider these methodological caveats. The poor performance of NeoFish marker to identify some target taxa highlights the importance of controlled experiments using known species communities before large investments are made in assessing biodiversity using non-invasive methods such as environmental DNA.

WILEY | **9**

CONFLICT OF INTEREST STATEMENT

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in Zenodo at http://doi.org/10.5281/zenodo.7601113 (https://zenodo.org/record/7601113#.Y9xF0nbMK00).

BENEFIT-SHARING STATEMENT

Benefits from this research accrue from the sharing of our data and results on public databases as described above.

ORCID

Heron Oliveira Hilário https://orcid.org/0000-0002-2548-1053
Naiara Guimarães Sales https://orcid.org/0000-0002-2922-3561
Daniel Cardoso Carvalho https://orcid.org/0000-0003-4517-5069

REFERENCES

Antich, A., Palacin, C., Wangensteen, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, *22*, 1–24.

Barbosa, J. M., Soares, E. C., Cintra, I. H. A., Hermann, M., & Araújo, A. R. R. (2017). Perfil da ictiofauna da bacia do rio São Francisco. *Acta of Fisheries and Aquatic Resources*, *5*(1), 70–90.

Boivin-Delisle, D., Laporte, M., Burton, F., Dion, R., Normandeau, E., & Bernatchez, L. (2020). Using environmental DNA for biomonitoring of freshwater fish communities: Comparison with established gill-net surveys in a boreal hydroelectric impoundment. *Environmental DNA*, *3*(1), 105–120. https://doi.org/10.1002/edn3.135

Callahan, B. J., McMurdie, P. J., & Holmes, S. P. (2017). Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. *The ISME Journal*, *11*(12), 2639–2643.

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581–583.

Cantera, I., Coutant, O., Jézéquel, C., Decotte, J. B., Dejean, T., Iribar, A., Vigouroux, R., Valentini, A., Murienne, J., & Brosse, S. (2022). Low level of anthropization linked to harsh vertebrate biodiversity declines in Amazonia. *Nat Commun*, *13*(1), 3290. https://doi.org/10.1038/s41467-022-30842-2

Carvalho, D. C. (2022). Ichthyoplankton DNA metabarcoding: Challenges and perspectives. *Molecular Ecology*, *31*(6), 1612–1614.

Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., Iribar, A., Taberlet, P., Vigouroux, R., & Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources*, *19*(1), 27–46.

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.

Dal Pont, G., Duarte Ritter, C., Agostinis, A. O., Stica, P. V., Horodesky, A., Cozer, N., Balsanelli, E., Netto, O. S. M., Henn, C., Ostrensky, A., & Pie, M. R. (2021). Monitoring fish communities through environmental DNA metabarcoding in the fish pass system of the second largest hydropower plant in the world. *Sci Rep*, *11*(1), 23167. https://doi.org/10.1038/s41598-021-02593-5

de Carvalho, D. C., Oliveira, D. A., Pompeu, P. S., Leal, C. G., Oliveira, C., & Hanner, R. (2011). Deep barcode divergence in Brazilian freshwater fishes: The case of the São Francisco River basin. *Mitochondrial DNA*, *22*(sup1), 80–86.

Deagle, B. E., Jarman, S. N., Coissac, E., Pompanon, F., & Taberlet, P. (2014). DNA metabarcoding and the cytochrome c oxidase subunit I marker: Not a perfect match. *Biol Lett*, *10*(9), 20140562. https://doi.org/10.1098/rsbl.2014.0562

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de 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.

Duke, E. M., & Burton, R. S. (2020). Efficacy of metabarcoding for identification of fish eggs evaluated with mock communities. *Ecology and Evolution*, *10*(7), 3463–3476.

Elbrecht, V., & Leese, F. (2015). Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. *PLoS One*, *10*(7), e0130324. https://doi.org/10.1371/journal.pone.0130324

García-García, N., Tamames, J., Linz, A. M., Pedrós-Alió, C., & Puente-Sánchez, F. (2019). Microdiversity ensures the maintenance of functional microbial communities under changing environmental conditions. *The ISME Journal*, *13*(12), 2969–2983.

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

Jackman, J. M., Benvenuto, C., Coscia, I., Oliveira Carvalho, C., Ready, J. S., Boubli, J. P., Magnusson, W. E., McDevitt, A. D., & Guimarães Sales, N. (2021). eDNA in a bottleneck: Obstacles to fish metabarcoding studies in megadiverse freshwater systems. *Environmental DNA*, *3*(4), 837–849.

Krehenwinkel, H., Wolf, M., Lim, J. Y., Rominger, A. J., Simison, W. B., & Gillespie, R. G. (2017). Estimating and mitigating amplification bias

in qualitative and quantitative arthropod metabarcoding. *Scientific Reports*, 7(1), 17668.

Lawson Handley, L., Read, D. S., Winfield, I. J., Kimbell, H., Johnson, H., Li, J., Hahn, C., Blackman, R., Wilcox, R., Donnelly, R., Szitenberg, A., & Hänfling, B. (2019). Temporal and spatial variation in distribution of fish environmental DNA in England's largest lake. *Environmental DNA*, 1(1), 26–39. https://doi.org/10.1002/edn3.5

Mächler, E., Walser, J. C., & Altermatt, F. (2021). Decision-making and best practices for taxonomy-free environmental DNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology*, 30(13), 3326–3339.

Mahé, F., Rognes, T., Quince, C., de Vargas, C., & Dunthorn, M. (2014). Swarm: Robust and fast clustering method for amplicon-based studies. *PeerJ*, 2, e593.

Mahé, F., Rognes, T., Quince, C., De Vargas, C., & Dunthorn, M. (2015). Swarm v2: Highly-scalable and high-resolution amplicon clustering. *PeerJ*, 3, e1420.

Majaneva, M., Diserud, O. H., Eagle, S. H., Boström, E., Hajibabaei, M., & Ekrem, T. (2018). Environmental DNA filtration techniques affect recovered biodiversity. *Scientific Reports*, 8(1), 4682.

McDevitt, A. D., Sales, N. G., Browett, S. S., Sparnenn, A. O., Mariani, S., Wangensteen, O. S., Coscia, I., & Benvenuto, C. (2019). Environmental DNA metabarcoding as an effective and rapid tool for fish monitoring in canals. *J Fish Biol*, 95(2), 679–682. https://doi.org/10.1111/jfb.14053

McLaren, M. R., Willis, A. D., & Callahan, B. J. (2019). Consistent and correctable bias in metagenomic sequencing experiments. *eLife*, 8, e46923.

McMurdie, P. J., & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS One*, 8(4), e61217.

Milan, D. T., Mendes, I. S., Damasceno, J. S., Teixeira, D. F., Sales, N. G., & Carvalho, D. C. (2020). New 12S metabarcoding primers for enhanced neotropical freshwater fish biodiversity assessment. *Scientific Reports*, 10(1), 1–12.

Miya, M., Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding: A high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fisheries Science*, 86(6), 939–970.

Miya, M., Sado, T., & eDNA Methods Standardization Committee. (2019). Multiple species detection using MiFish primers. *Environmental DNA Sampling and Experimental Manual Version*, 2, 55–92.

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.

Polanco, F. A., 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.

Porter, T. M., & Hajibabaei, M. (2018). Automated high throughput animal CO1 metabarcode classification. *Scientific Reports*, 8(1), 4226.

Pugedo, M. L., de Andrade Neto, F. R., Pessali, T. C., Birindelli, J. L. O., & Carvalho, D. C. (2016). Integrative taxonomy supports new candidate fish species in a poorly studied neotropical region: The Jequitinhonha River Basin. *Genetica*, 144(3), 341–349.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Riaz, T., Shehzad, W., Viari, A., Pompanon, F., Taberlet, P., & Coissac, E. (2011). ecoPrimers: inference of new DNA barcode markers from whole genome sequence analysis. *Nucleic Acids Res*, 39(21), e145. https://doi.org/10.1093/nar/gkr732

Rosso, J. J., González-Castro, M., Bogan, S., Cardoso, Y. P., Mabragaña, E., Delpiani, M., & de Astarloa, J. M. D. (2018). *Integrative taxonomy reveals a new species of the Hoplias malabaricus species complex (Teleostei: Erythrinidae)*. Verlag Dr.

Sales, N. G., Mariani, S., Salvador, G. N., Pessali, T. C., & Carvalho, D. C. (2018). Hidden diversity hampers conservation efforts in a highly impacted neotropical river system. *Frontiers in Genetics*, 9, 271.

Sales, N. G., Pessali, T. C., Andrade Neto, F. R., & Carvalho, D. C. (2018). Introgression from non-native species unveils a hidden threat to the migratory Neotropical fish Prochilodus hartii. *Biological Invasions*, 20, 555–566.

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., Deiner, K., Præbel, K., Coscia, I., McDevitt, A. D., & Mariani, S. (2021). Space-time dynamics in monitoring neotropical fish communities using eDNA metabarcoding. *Science of the Total Environment*, 754, 142096.

Sales, N. G., Wangensteen, O. S., Carvalho, D. C., & Mariani, S. (2019). Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. *Environmental DNA*, 1(2), 119–130.

Santos, R. P., Melo, B. F., Yazbeck, G. M., Oliveira, R. S., Hilário, H. O., Prosdocimi, F., & Carvalho, D. C. (2021). Diversification of Prochilodus in the eastern Brazilian shield: Evidence from complete mitochondrial genomes (Teleostei, Prochilodontidae). *Journal of Zoological Systematics and Evolutionary Research*, 59(5), 1053–1063.

Sard, N. M., Herbst, S. J., Nathan, L., Uhrig, G., Kanefsky, J., Robinson, J. D., & Scribner, K. T. (2019). Comparison of fish detections, community diversity, and relative abundance using environmental DNA metabarcoding and traditional gears. *Environmental DNA*, 1(4), 368–384.

Sato, H., Sogo, Y., & Yamanaka, H. (2017). Usefulness and limitations of sample pooling for environmental DNA metabarcoding of freshwater fish communities. *Scientific Reports*, 7(1), 1–12.

Shirazi, S., Meyer, R. S., & Shapiro, B. (2021). Revisiting the effect of PCR replication and sequencing depth on biodiversity metrics in environmental DNA metabarcoding. *Ecology and Evolution*, 11(22), 15766–15779.

Silva, L. E., Domingues, R. R., Sales, N. G., Villela, P. M. S., Silva, C. B., & Hilsdorf, A. W. S. (2023). Amazonian ichthyoplankton assessment via DNA metabarcoding: A baseline for detecting spawning sites of migratory fishes. *Biol Conserv*, 284, 110180. https://doi.org/10.1016/j.biocon.2023.110180

Tsuji, S., Miya, M., Ushio, M., Sato, H., Minamoto, T., & Yamanaka, H. (2020). Evaluating intraspecific genetic diversity using environmental DNA and denoising approach: A case study using tank water. *Environmental DNA*, 2(1), 42–52.

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.

Van der Loos, L. M., & Nijland, R. (2021). Biases in bulk: DNA metabarcoding of marine communities and the methodology involved. *Molecular Ecology*, 30(13), 3270–3288.

Xiong, F., Shu, L., Gan, X., Zeng, H., He, S., & Peng, Z. (2022). Methodology for fish biodiversity monitoring with environmental DNA metabarcoding: The primers, databases and bioinformatic pipelines. *Water Biology and Security*, 1(1), 100007. https://doi.org/10.1016/j.watbs.2022.100007

Zhang, J., Kapli, P., Pavlidis, P., & Stamatakis, A. (2013). A general species delimitation method with applications to phylogenetic placements. *Bioinformatics*, 29(22), 2869–2876.

WILEY | **11**

Zhang, S., Zhao, J., & Yao, M. A. (2020). comprehensive and comparative evaluation of primers for metabarcoding eDNA from fish. *Methods in Ecology and Evolution*, *11*, 1609–1625.

Zinger, L., Donald, J., Brosse, S., Gonzalez, M. A., Iribar, A., Leroy, C., Murienne, J., Orivel, J., Schimann, H., Taberlet, P., & Lopes, C. M. (2020). Advances and prospects of environmental DNA in neotropical rainforests. *Advances in Ecological Research*, *62*, 331–373.

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: Hilário, H. O., Mendes, I. S., Guimarães Sales, N., & Carvalho, D. C. (2023). DNA metabarcoding of mock communities highlights potential biases when assessing Neotropical fish diversity. *Environmental DNA*, *00*, 1–11. https://doi.org/10.1002/edn3.456