




# GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding

Virginie Marques<sup>1,2</sup>  | Tristan Milhau<sup>3</sup> | Camille Albouy<sup>4</sup>  | Tony Dejean<sup>3</sup> |  
Stéphanie Manel<sup>2</sup> | David Mouillot<sup>1,5</sup> | Jean-Baptiste Juhel<sup>1</sup> 

<sup>1</sup>MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Montpellier, France

<sup>2</sup>CEFE, EPHE, CNRS, UM, UPV, IRD, PSL Research University, Montpellier, France

<sup>3</sup>SPYGEN, Le Bourget-du-Lac, France

<sup>4</sup>IFREMER, Unité Ecologie et Modèles pour l'Halieutique, Nantes cedex 3, Nantes, France

<sup>5</sup>Australian Research Council Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, Qld, Australia

## Correspondence

Virginie Marques, MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Montpellier, France.  
Email: virginie.marques01@gmail.com

## Funding information

ANR-BiodivERsA; ANR-BiodivERsA, Grant/Award Number: RESERVEBENEFIT

Editor: Trishna Dutta

## Abstract

**Aim:** Environmental DNA metabarcoding has recently emerged as a non-invasive tool for aquatic biodiversity inventories, frequently surpassing traditional methods for detecting a wide range of taxa in most habitats. The major limitation currently impairing the large-scale application of eDNA-based inventories is the lack of species sequences available in public genetic databases. Unfortunately, these gaps are still unknown spatially and taxonomically, hindering targeted future sequencing efforts.

**Innovation:** We propose GAPeDNA, a user-friendly web interface that provides a global overview of genetic database completeness for a given taxon across space and conservation status. As an application, we synthesized data from regional checklists for marine and freshwater fishes along with their IUCN conservation status to provide global maps of species coverage using the European Nucleotide Archive public reference database for 19 metabarcoding primers. This tool automatizes the scanning of gaps in these databases to guide future sequencing efforts and support the deployment of eDNA inventories at larger scale. This tool is flexible and can be expanded to other taxa and primers upon data availability.

**Main conclusions:** Using our global fish case study, we show that gaps increase towards the tropics where species diversity and the number of threatened species are the highest. It highlights priority areas for fish sequencing like the Congo, the Mekong and the Mississippi freshwater basins which host more than 60 non-sequenced threatened fish species. For marine fishes, the Caribbean and East Africa host up to 42 non-sequenced threatened species. By presenting the global genetic database completeness for several primers on any taxa and building an open-access, updatable and flexible tool, GAPeDNA appears as a valuable contribution to support any kind of eDNA metabarcoding study.

## KEYWORDS

environmental DNA, genetic markers, IUCN, marine and freshwater fish, non-indigenous species, reference database, shiny, threatened species

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. *Diversity and Distributions* published by John Wiley & Sons Ltd

## 1 | INTRODUCTION

Aquatic ecosystems are increasingly impacted by human activities, threatening their biodiversity and causing major disruptions in their functioning (Cinner et al., 2016; Link & Watson, 2019; Reid et al., 2019). Marine systems are under severe defaunation with numerous local species extinctions (McCauley et al., 2015) and also experiencing the highest rates of biodiversity changes under the combined effects of climate change and direct human impacts (Blowes et al., 2019). Freshwater ecosystems are even more at risk, with fishes being among the most threatened vertebrates due to habitat degradation or exotic species introduction (Collen et al., 2014). In this context, efficient non-invasive methods are urgently needed to accurately monitor aquatic biodiversity including rare, highly mobile and elusive species in order to set appropriate conservation management.

Among the many ways to survey aquatic biodiversity, eDNA metabarcoding has recently emerged as a promising approach, frequently surpassing traditional inventory methods in detectability potential (Boussarie et al., 2018; Carraro, Hartikainen, Jokela, Bertuzzo, & Rinaldo, 2018; Stat et al., 2019; Valentini et al., 2016). Exogenous DNA released by animals in the environment, through shed skin, mucus or faeces, can be retrieved by filtering water and amplified via polymerase chain reaction (PCR) using universal primers (Ficetola, Miaud, Pompanon, & Taberlet, 2008). High-throughput sequencing of the amplified DNA fragments provides a list of sequences over which corresponding species can be assigned by comparison with available genetic databases like the European Nucleotide Archive (ENA) (Dickie et al., 2018; Kanz et al., 2005).

However, the major limitation currently impairing the large-scale application of eDNA inventories is the incompleteness of species sequences available in public genetic databases, considerably reducing the breadth of detected biodiversity. Historically, eDNA studies have primarily focused on well-known species-poor freshwater systems (Jerde, Wilson, & Dressler, 2019), but recently, eDNA biodiversity inventories have spread all over the globe, across a wide range of ecosystems encompassing less studied and more diverse taxa and habitats (Cilleros et al., 2019; Jerde et al., 2019; Yamamoto et al., 2017). A recent study on European aquatic systems shows that genetic coverage varies widely among taxonomic groups, databases and the level of monitoring (Weigand et al., 2019) with, for example, European freshwater fish lacking genetic coverage on the 12S mitochondrial marker for 64% of the 627 species.

Teleostean fishes represent the largest group of vertebrates with more than 32,000 species ("www.fishbase.org,") and a total biomass estimated at 0.7 Gt (Bar-On, Phillips, & Milo, 2018). They represent the most extensively studied taxonomic group using eDNA with up to 60% of the publications on vertebrates (Tsuji, Takahara, Doi, Shibata, & Yamanaka, 2019) and play a significant role in carbon cycling (Wilson et al., 2009) and food security (Hicks et al., 2019). Despite their cultural, commercial and ecological importance, fish populations are increasingly depleted or threatened due to overfishing (Anticamara, Watson, Gelchu, & Pauly, 2011) and habitat alterations (Collen et al., 2014). Surprisingly, the extent to which genetic

reference databases cover fish biodiversity for the most widely used metabarcoding primers is unknown, while it ultimately determines the amount and the composition of species potential revealed by eDNA surveys. This kind of information is currently available, albeit scattered across different databases, but we still lack a tool facilitating the assessment and visualization of genetic species coverage for a given region, a given taxon and a given primer pair.

Here, we filled this gap by developing a user-friendly, flexible and interactive web interface linking reference genetic databases to regional species lists. Using regional freshwater and marine fish checklists, we assessed geographical variations in species diversity coverage versus gap for different metabarcoding primer pairs. Then, we highlighted the geographical bias in genetic coverage and disparities according to the native and conservation status of species (IUCN), providing valuable recommendations for future eDNA investigations at global scale.

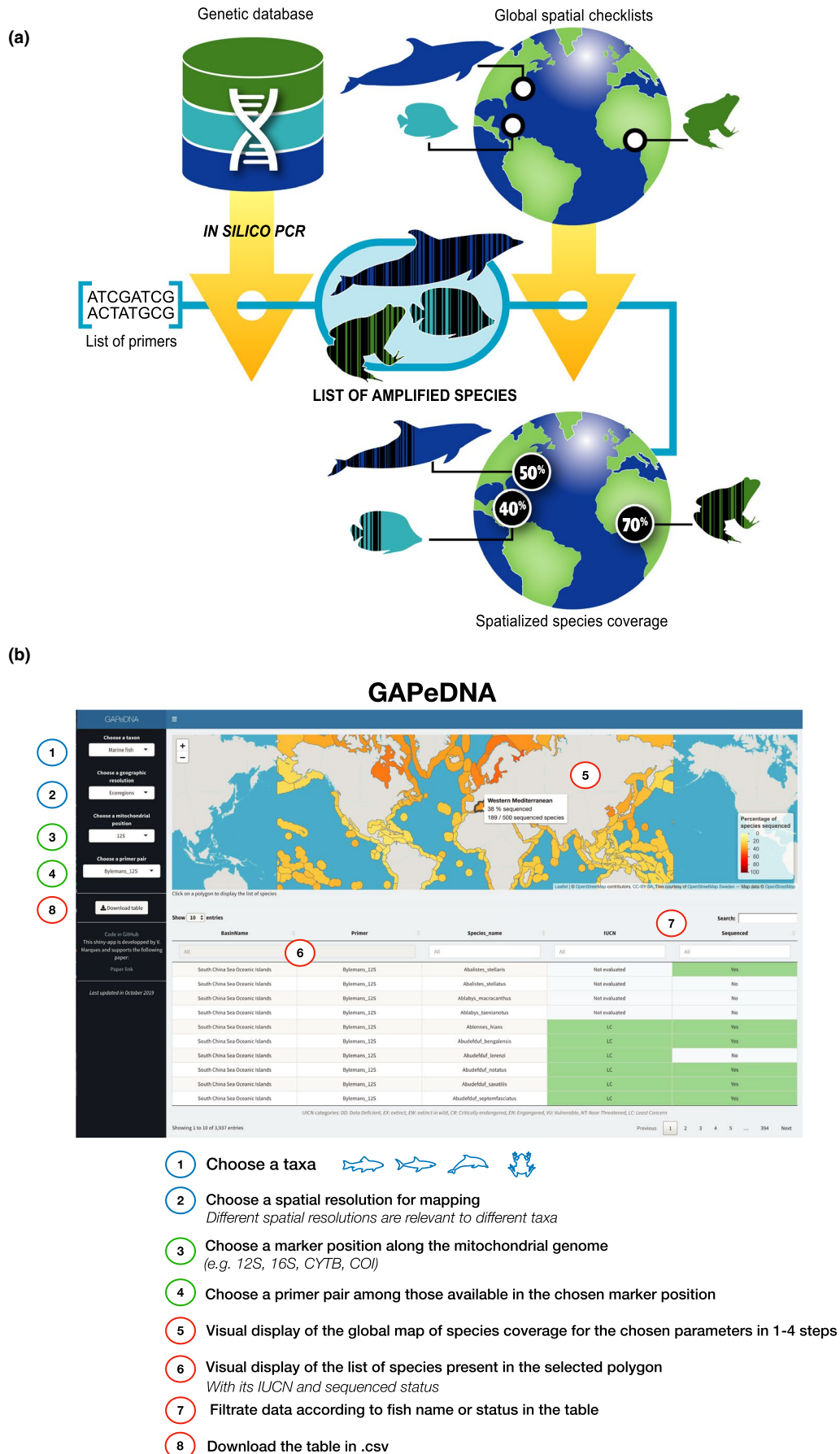
## 2 | METHODS

### 2.1 | Interactive web interface: GAPeDNA

To facilitate the global assessment and visualization of regional gaps in genetic databases for environmental DNA metabarcoding, we developed a user-friendly interactive web interface called GAPeDNA (<https://shiny.cefe.cnrs.fr/GAPeDNA/>, Figure 1), using the shiny R package (Chang, Cheng, Allaire, Xie, & McPherson, 2019). This interface allows researchers and stakeholders to easily locate gaps in the reference genetic databases at global scale for a selection of fish metabarcoding primers. A virtual PCR using the selected primers is performed on a selected online genetic database. The list of the amplified species is then compared to a spatialized checklist to generate the percentage of species referenced in each spatial unit or area (e.g. basins and ecoregions for freshwater and marine fishes, respectively) (Figure 1a). This percentage is then displayed with an interactive global map in GAPeDNA. This interface is flexible and can display results for several primer pairs per taxon and several spatial units, and allows the user to choose between several options (Figure 1b). We present the application for fish, but users are encouraged to suggest new taxa, which requires to have (a) at least one primer pair targeting the taxa using metabarcoding and (b) globally georeferenced species checklists. It also allows to visualize which species are actually sequenced for a given primer when clicking on the area of interest, under which conservation status (i.e. IUCN category) these species are, and extract this information as a comma-separated values (CSV) file. Users can thus quickly grasp information regarding sequencing priorities depending on their research interest.

### 2.2 | Genetic sequence database and genetic coverage by markers

To illustrate the distribution of species coverage, we used the European Nucleotide Archive (ENA) (Kanz et al., 2005) (release 138,



**FIGURE 1** Illustration of the process for generating map and data in the GAPeDNA web application (a) and details on the interface (b). User's spatial choices are in blue and green, genetic choices are in green, and visual displays are in red

downloaded in January 2019) as the genetic reference database for fish species. This database was formatted using *obiconvert* from the OBITOOLS toolkit (Boyer et al., 2016) to run in silico PCRs (i.e. virtual PCR based on primer affinity to sequences). Yet, primer sequences need to be present within the sequence fragment deposited online to be detectable using this in silico approach.

An extensive literature search was conducted to identify the most commonly used primer pairs targeting fish for metabarcoding on ISI Web of Science with the following keywords: "fish" AND "metabarcoding" AND "primer" AND "environmental DNA". We discarded primer pairs not primarily targeting fish, only targeting a restricted group of fish or containing errors. Following this filtering, we retained 23 primer pairs from 18 papers (Table S1), from five regions in the mitochondrial genome (hereafter referred as markers), namely 12S, 16S, 18S, COI and CytB. All primer pairs were used individually to run in silico PCRs using *ecoPCR* from OBITOOLS (Boyer et al., 2016), with three mismatches allowed. All species amplified by each primer pair were compared to the regional fish checklists of both marine and freshwater environments, to obtain the percentage of species coverage by spatial unit and by primer. Fish names obtained from GenBank were checked and updated using FishBase as the sole reference. We further discarded four primer pairs with low performance (global fish coverage < 0.05%) to avoid bias when comparing markers (Table S2), so we proceeded with a total of 19 primer pairs on 4 markers, as the only primer pair located on the 18S rDNA marker was discarded. The successful virtual amplification of a species by a primer pair is conditional to (a) species presence in the public genetic database and (b) the primer ability to amplify the sequence. Hence, primer pairs lacking universality for fish sequence amplification show an overall low coverage, even if located on a genetic marker with a larger sequence coverage in online database, as they are unable to amplify those due to primer specificity.

### 2.3 | Global species checklists and status

The checklist for freshwater fish was extracted from a global-scale database of fish diversity at the basin scale (Tedesco et al., 2017). The authors reviewed a large body of information from 1,436 distinct sources over 3,119 drainage basins, covering more than 80% of Earth surface and comprising 14,953 fish species, so 90% of all freshwater fishes were recorded in FishBase ([www.fishbase.org](http://www.fishbase.org)). Although all biogeographic realms are well represented, some

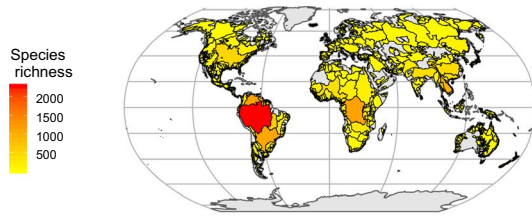
regional gaps remain in the database due to the scarcity of information or the probable low number of freshwater taxonomists in some regions like South-East Asia. The global diversity of marine fishes was assembled using OBIS (OBIS Ocean Biogeographic Information and System, (n.d.)) and regional checklists (Albouy et al., 2019; Pellissier, Heine, Rosauer, & Albouy, 2018), including manual verification to remove taxonomic classification errors. It contains available occurrence data for all marine teleost and agnathan fishes, so a total of 14,202 species representing 82% of all marine fish species were recorded in FishBase. The original spatial resolution was a 1° grid for all marine environments. For visualization and interpretation purposes, this grid was then coerced at two supplementary biogeographic spatial scales according to Marine Ecoregions (Spalding et al., 2007) (a) at the province scale, with 62 distinct units, and (b) at the ecoregion scale, with 232 distinct units. Latitudes and longitudes were computed as the centroid of each polygon at the finest resolution for both environments using the R package *sf* (Pebesma, 2016), and land areas were removed using polygons from Natural Earth Data (<https://www.naturalearthdata.com/>). Areas were calculated using the Mollweide equal-area projection and presented in figures using the Robinson projection.

For freshwater environments, a species is considered as non-indigenous in a given basin only if this species is able to complete its entire life cycle and harbours self-sustaining populations in that basin (Tedesco et al., 2017). A species is considered as indigenous when never occurring as non-indigenous in any basin following the original data (Tedesco et al., 2017). We acknowledge that some of the species classified as indigenous may have been introduced in another basin but have still not been identified, detected or been referred as such into global databases. However, our dataset represents currently the most recent and precise data on non-indigenous freshwater species at the global scale (Tedesco et al., 2017). For marine systems, we used the information supplied in FishBase and only considered species flagged as "introduced," excluding species categorized as "questionable" or "non-settled."

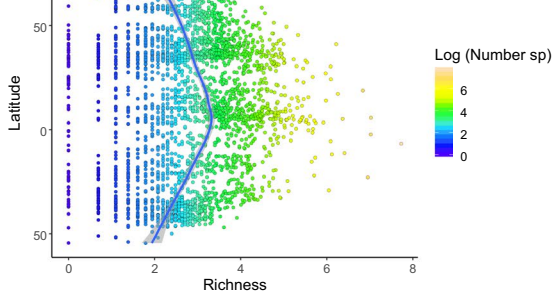
Regarding the conservation status of species, we retrieved data from the redlist R package (Chamberlain, n.d.) to assign each species from both freshwater and marine environments into an IUCN Red List category. The abbreviation "DD" represents Data Deficient, "LC" represents Least Concern, and all Threatened or Near-Threatened categories were grouped under the "Threatened & NT" status. We excluded species identified as "EX" for Extinct and "EW" for Extinct in the Wild. Where no data were available, we assigned the value "NA."

**FIGURE 2** Global and latitudinal distributions of freshwater fish species richness on log scale (a, b), coverage by online genetic database for the Miya primer pair targeting the 12S mitochondrial rDNA region (c, d), the Kocher primer targeting the cytochrome B mitochondrial rDNA region (e, f), the DiBattista primer targeting the 16S rDNA region (g, h) and the Ward f2 primer targeting the COI mitochondrial region (i, j). The number of species along latitude (b) is  $\log_{10}$ -scaled and obtained from the finest resolution, here by basin. Global latitudinal patterns of all primer pairs are given in Figures S5 and S6, and the global distribution maps are reproducible and interactive using the web application (<https://shiny.cefe.cnrs.fr/GAPeDNA/>). Primers were chosen to represent the most used primer pair for each genetic marker in fish eDNA studies (Tsuji et al., 2019)

(a)

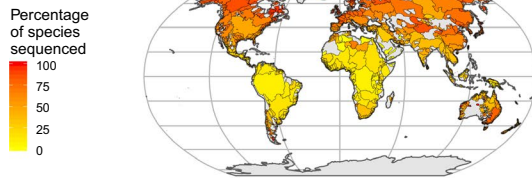


(b)

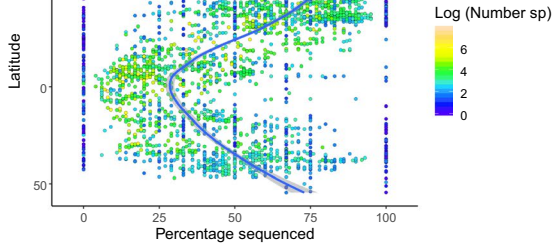


(c)

Miya\_12S

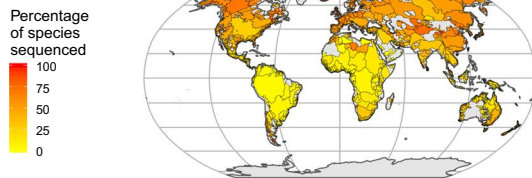


(d)

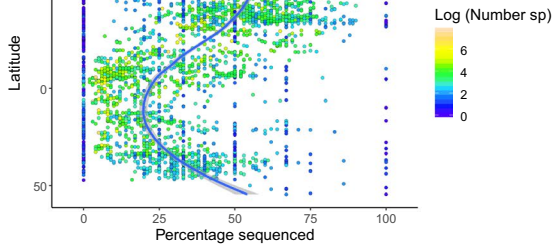


(e)

Kocher\_CYTB

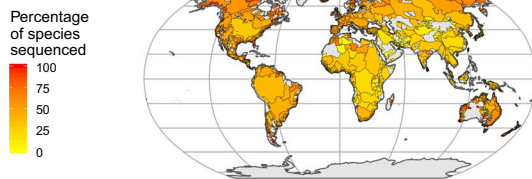


(f)

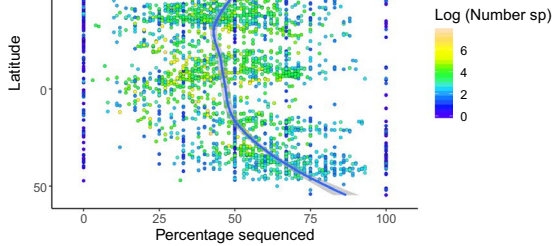


(g)

DiBattista\_16S

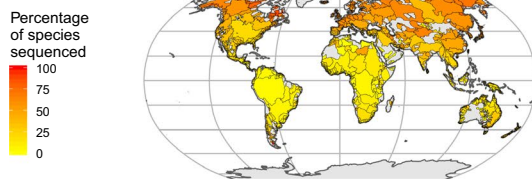


(h)

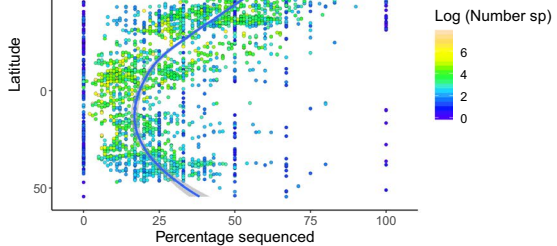


(i)

Ward\_COI\_f1



(j)



### 3 | RESULTS

#### 3.1 | Global distribution of genetic database completeness and gaps

The 3,119 freshwater drainage basins, located across all continents (except the poles), largely varied in terms of surface, from 2 km<sup>2</sup> to 5,888,417 km<sup>2</sup> (Amazon) with a mean of 31,996 km<sup>2</sup> ( $SD = 209,732$  km<sup>2</sup>). Their species richness ranged from 1 to 2,273 with a mean of 33 species ( $SD = 71$ ), with an increasing number of species towards the equator following the classical latitudinal gradient (Figure 2a,b). Across the 232 marine ecoregions, species richness also greatly increased towards the equator, from 14 species (East Antarctic) to 3,937 species (South China Sea Oceanic Islands; Figure 4a,b). Marine ecoregion area varied from 19,000 km<sup>2</sup> (Puget Trough, Northern America) to 2,647,573 km<sup>2</sup> (Hawaii) with a mean of 588,862 km<sup>2</sup> ( $SD = 460,459$  km<sup>2</sup>), and no correlation between area and fish species richness was observed (Figure S1).

Global coverage of fish species in GenBank largely varied according to both the marker position along the mitochondrial genome and among primers for a given position (Figures 2 and 3), with a global coverage for freshwater species ranging between 7% for COI Ward and 26% for 16S McInnes, and a coverage for marine species between 4% for Thomsen Cytb cb and 30% for Shaw 16S (Table S2). For a given primer pair, species coverage also greatly varied along the latitudinal gradient, with a U-shaped relationship peaking in high absolute latitudes for most of the primers in freshwater systems. For example, the 16S McInnes primer pair had a mean coverage of 89% between 48° and 52° latitude (84 basins) and only 40% between -2° and 2° latitude (54 basins). This contrast was also marked for primers targeting the 12S mitochondrial rDNA region. For example, the 12S Miya primer pair covered 83% of the fish checklist in high latitudes (between 48° and 52°), but only 23% close to the equator (between -2° and 2° latitude, Figure 2d). The Cytb from Thomsen 2cbl and 2deg (Figures S5 and S6) covered, respectively, 13% and 18% of the fish checklists, but showed no geographical gradient.

In marine ecosystems, the latitudinal gradient in species coverage was less pronounced with several primer pairs showing a steady decrease in coverage with decreasing latitude (Figure 3). Tropical fish assemblages along the equator were less sequenced than northern temperate assemblages, but were generally more sequenced than in negative latitude ecoregions towards the south pole, as opposed to freshwater systems. Only the 12S Bylemans primer pair, covering 13% of marine fishes, showed no geographical pattern (Figure S6).

#### 3.2 | Genetic coverage of native versus. Non-indigenous species

Environmental DNA can be used to track non-indigenous species in ecosystems. However, only the primers located on the 12S and 16S had a mean species coverage superior to 50% for all 605 identified non-indigenous freshwater fishes (Figure 4a). For the primers on the COI and Cytb, less than half of all non-indigenous fishes were amplified and sequenced. Only two primers, both on the 16S, had a coverage for more than 60% of non-indigenous species, while none had a coverage above 57% for the 12S primers. However, these species still had an overall larger coverage in databases compared to native species, the maximum for native species being 31% for a 16S marker and 15% or 19% for 12S and Cytb markers, respectively.

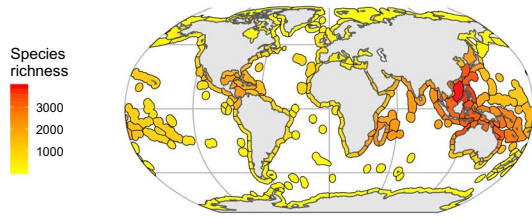
For the marine fishes, we identified 196 species as non-indigenous in at least one region of the marine realm, two times less than the 605 species identified in freshwater. However, global patterns of coverage were similar (Figure 4b), albeit with a wider coverage of marine non-indigenous species compared to their freshwater counterparts (maximum 12S coverage of 69% versus 57%). Overall, for both categories, non-indigenous species were more sequenced than indigenous species, but 20% to 80% of fish species remain to be sequenced depending on the genetic marker.

#### 3.3 | Genetic coverage of fish species with different IUCN conservation status

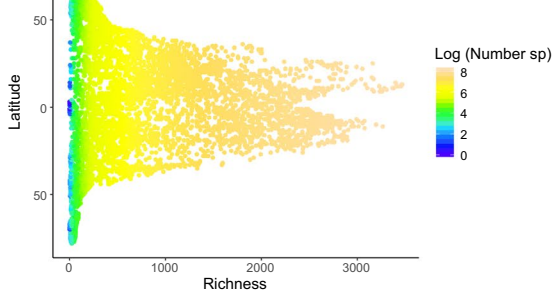
Most of freshwater fish species were not evaluated (NA, 45.9% of total) or Least Concern (LC, 33.2%). However, 1,758 species (11.7%) were classified as threatened by including Vulnerable (VU), ENdangered (EN) and CRitically endangered (CR) species or Near Threatened (NT) categories of the IUCN Red List ([www.iucnredlist.org](http://www.iucnredlist.org), Figure S2). The genetic database coverage of fish species according to their IUCN status showed consistent patterns for all markers (Figure 5a and 5c). Species classified as Least Concern (LC) were always more represented in genetics databases compared to non-evaluated (NA), data deficient species (DD) (Figure S3) or threatened species (T & NT). Freshwater basins where the most threatened species remain to be sequenced using the 12S Miya primer pair were mainly located around the equator with 79 species in the Congo Basin and 63 species in the Mekong Basin or in the Northern Hemisphere with a maximum of 72 species in the Mississippi Basin (Figure 5b). These basins also host the highest number of threatened species, independent of reference filling (Figure S4).

**FIGURE 3** Global and latitudinal distributions of marine fish species richness (a, b), coverage of online genetic database for the Miya primer pair targeting the 12S mitochondrial rDNA region (c, d), the Kocher primer targeting the cytochrome B mitochondrial rDNA region (e, f), the DiBattista primer targeting the 16S rDNA region (g, h) and the Ward f2 primer targeting the COI mitochondrial region (i, j). The number of species along latitude (b) is log-scaled and obtained from the finest resolution, here by a 1° grid. Global latitudinal patterns of all primer pairs are given in Figures S5 and S6, and the global distribution maps are reproducible and interactive using the web application (<https://shiny.cefe.cnrs.fr/GAPeDNA/>). Primers were chosen to represent the most used primer pair for each genetic marker in fish eDNA studies (Tsuji et al., 2019)

(a)

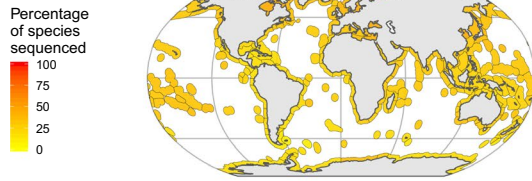


(b)

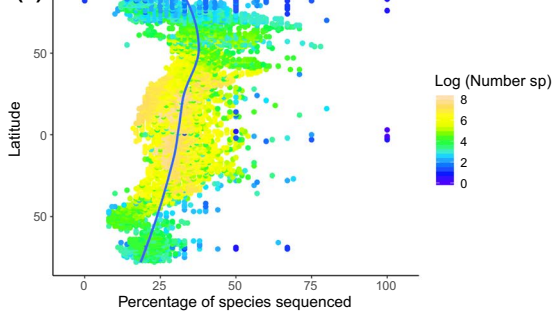


(c)

Miya\_12S

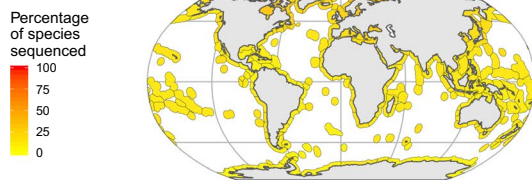


(d)

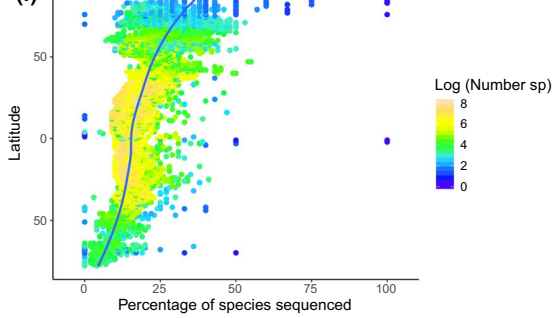


(e)

Kocher\_CYTB

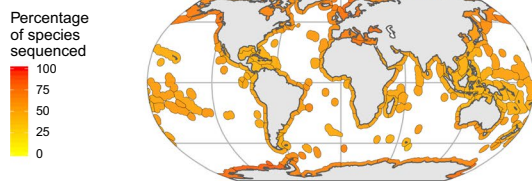


(f)

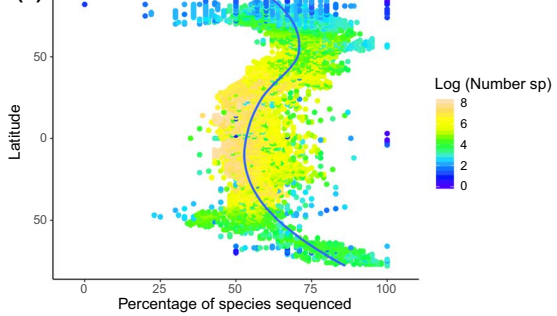


(g)

DiBattista\_16S

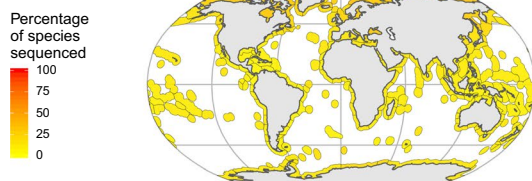


(h)

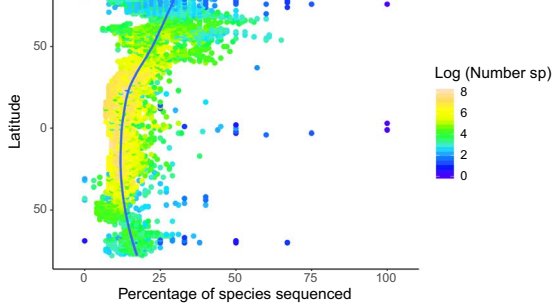


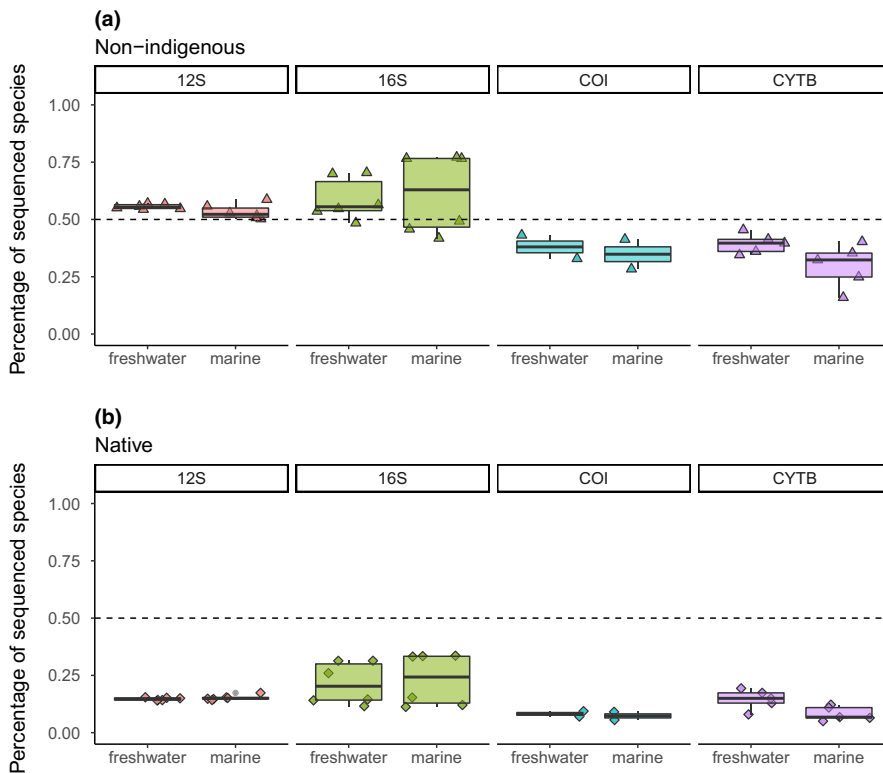
(i)

Ward\_COI\_f1



(j)





**FIGURE 4** Percentage of species coverage (a) in marine systems for non-indigenous (196) and native species (12,290) and (b) in freshwater for non-indigenous (605) and native species (14,348) depending on the marker primer position. Each triangle represents a primer pair

In marine environments, 3.5% of all species were classified under an IUCN Red List status compared to 11.7% in freshwater systems (Figure S2), and around the same proportion of fishes were unevaluated or data deficient (49% versus 55% for freshwater). Genetic coverage was systematically higher for threatened species compared to Least Concern (LC) species, albeit never exceeding 50% for any primer or ecoregion (Figure 5). Species listed as LC consistently had a higher coverage than unevaluated or data deficient species (Figure S3). Marine ecoregions hosting the most threatened species remaining to be sequenced using the 12S Miya primers were also located around the equator, particularly in the Caribbean with a maximum of 42 species in the south-western Caribbean ecoregion or in the Eastern Coast of Africa with a maximum of 32 species in the Delagoa ecoregion (Figure 5d).

## 4 | DISCUSSION

### 4.1 | Genetic markers and primer selection

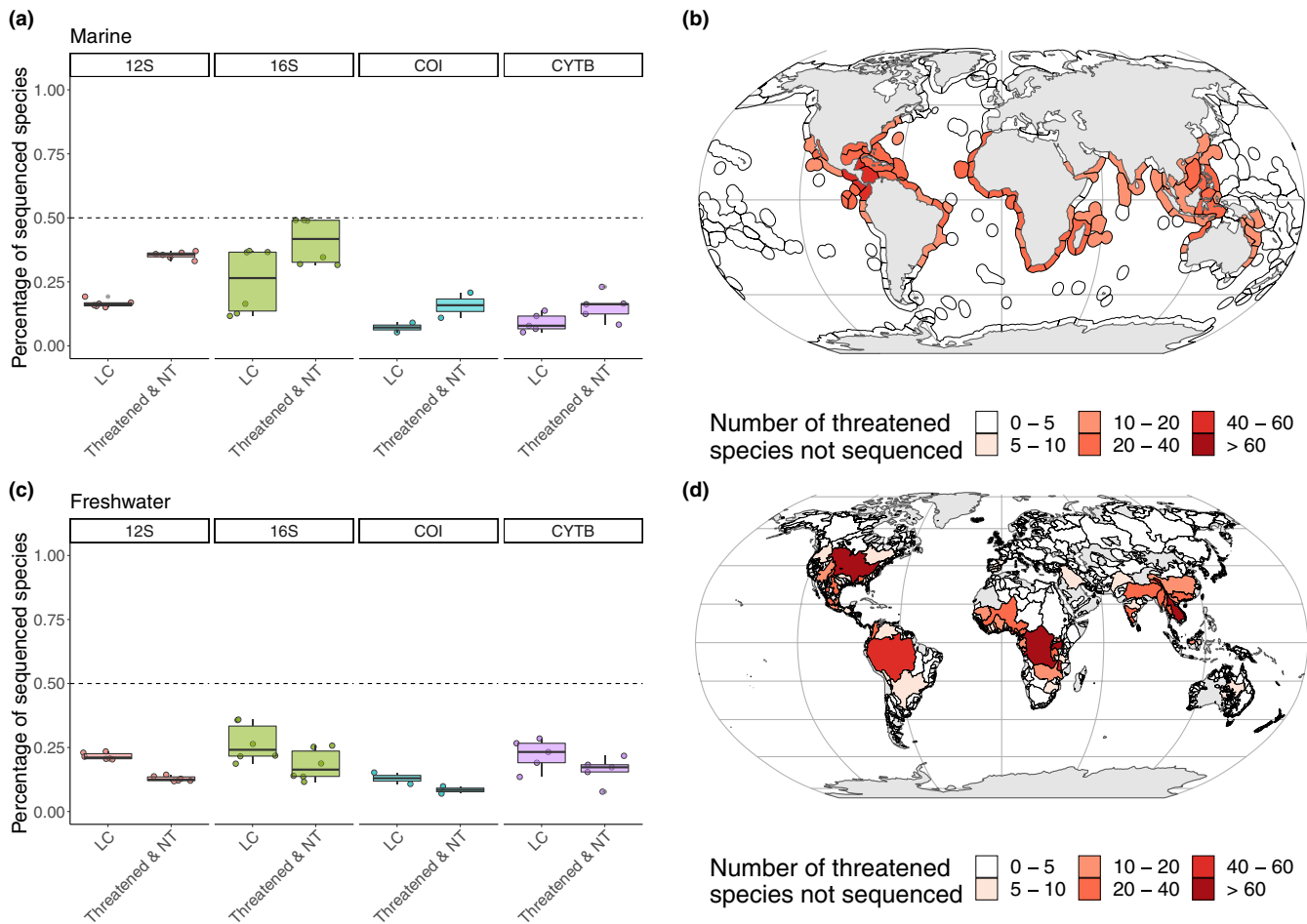
eDNA is currently limited by the scarcity of species sequences available in online public genetic databases. We provide here a spatialized global assessment of fish sequence coverage and gaps in databases, using published eDNA primers, and displayed on an online, semi-automated and flexible application called GAPeDNA. Our study considers all existing markers and most primers capable of theoretically amplifying fish species by *in silico* PCR, regardless of their performance, avoiding a bias in the choice of a genetic marker or primer pair. The marker and primer selection must be motivated

by their efficiency to detect the targeted taxa owing to their specificity and sensitivity. A general consensus is emerging in fish eDNA studies towards the use of 12S primers (Collins et al., 2019; Weigand et al., 2019). Primers located on the 12S mitochondrial region have been recognized as the best to specifically amplify fishes, unlike COI primers which lack specificity, resulting in low fish detectability (Valentini et al., 2016). Unfortunately, we show that the 12S still has a very low species completeness in genetic databases, with strong spatial disparities. With the goal to sequence a maximum of species, it is crucial to reach a consensus in the genetic marker selection to join efforts towards a globally coordinated sampling strategy for this genetic marker. Once species gaps in the 12S sequences will be almost filled, it would pave the way to install eDNA metabarcoding as a robust and standard monitoring and inventory tool, capable of fish identification to the species level in every location.

### 4.2 | Mapping species coverage gaps to improve eDNA monitoring

The global diversity of both freshwater and marine fishes is not well covered in public genetic databases. Globally, we show a higher coverage around high latitude in the Northern Hemisphere consistent across the genetic markers and primers while tropical areas, which host more species, have more species gaps in public sequence databases (Figures 2 and 3). For freshwater fishes, the genetic species coverage exhibits a clear U-shaped pattern for almost all markers along the latitudinal diversity gradient (Hillebrand, 2004) (Figure 2), with a minimum percentage of sequenced species around the





**FIGURE 5** Percentage of coverage according to two IUCN categories: Least Concern (LC) or Threatened and Near Threatened (NT) for all primer pairs and global gap in threatened species not sequenced illustrated for the Miya 12S primers in (a, b) marine systems and (c, d) freshwater systems. Each dot represents one primer pair. The threatened category includes the categories Vulnerable (VU), Endangered (E) and CRitically endangered (CR). The categories Not Evaluated (NA) and Data Deficient (DD) were not represented. All the categories are displayed in Figure S3

equator. For marine fishes, species coverage declines with declining latitude, and the minimum percentage of species sequenced is around the low latitudes of the Southern Hemisphere where marine fish diversity is the lowest.

The location of the gaps may drive the future sampling efforts required to fill them. Tropical environments are under-represented in public sequence databases and will require a costly, time-consuming and globally coordinated efforts to both describe and sequence the numerous species left to be discovered, as well as sequence the numerous species already described (Juhel et al., 2020; Pinheiro, Moreau, Daly, & Rocha, 2019). Environmental DNA is settling as an efficient inventory tool that can overcome hurdles encountered when sampling in tropical ecosystems. In many large water bodies, such as the Mekong or the Amazon, water turbidity prevents visual census leaving the eDNA the only non-invasive monitoring method (Cilleros et al., 2019; Yamamoto et al., 2017). The need to fill species gaps is urgent in these environments as they are experiencing major turnover in species identities with unknown consequences on ecosystem functioning and resilience (Magurran et al., 2018).

Tropical marine ecosystems are biodiversity hotspots, particularly the Coral Triangle (Barlow et al., 2018; Myers, Mittermeier, Mittermeier, Da Fonseca, & Kent, 2000). Tropical countries also tend to have a high dependency to fish resources (Andrello et al., 2017; Barange et al., 2014), stressing the importance of securing a sustainable exploitation of fish which requires monitoring assessments and correct evaluations of biodiversity as these both aspects are intimately linked (Duffy, Lefcheck, Stuart-Smith, Navarrete, & Edgar, 2016; Lefcheck et al., 2019). For instance, crypto-benthic fishes (<5cm) have been recently shown to contribute massively to coral reef functioning (Brandl et al., 2019), particularly by feeding fish consumed by humans, but they are still poorly inventoried. Tropical countries are also projected to undergo among the most severe environmental impacts related to human population expansion and climate change (Barlow et al., 2018), highlighting the importance of conducting ecological studies and setting appropriate conservation programs. For instance, mesophotic reefs (30–150 metres depth) are still poorly known while they potentially host very different species assemblages that can be also affected by climate

change (Lesser, Slattery, Laverick, Macartney, & Bridge, 2019; Rocha et al., 2018). Their exploration will require new eDNA-based protocols (fish sampling for reference database and water filtering) that must complement visual surveys that remain limited at this depth. Yet, there is a clear publication bias with the most diverse ecosystems being the least studied in ecology (Hickisch et al., 2019). So, the efforts to achieve genetic database completeness are massive but necessary in such highly diverse environments in order to tackle major conservation challenges like the protection of vulnerable but still poorly described biodiversity.

### 4.3 | Environmental DNA metabarcoding to monitor non-indigenous species

Among the numerous threats that all aquatic environments are currently facing lies non-indigenous species, which have the potential to disrupt entire ecosystems when declared as invasive (Albins & Hixon, 2013; Bax, Williamson, Agüero, Gonzalez, & Geeves, 2003; Clavero & García-Berthou, 2005). For example, the Nile perch (*Lates niloticus*), introduced in the 1950s in the Lake Victoria, drove around half of the hundreds of native Cichlid fish species to extinction through predation and competition (McGee et al., 2015; Witte et al., 1992). As traditional methods struggle to detect those species at an early stage of installation, eDNA offers an important potential for early detection below the traditional detection threshold (Dougherty et al., 2016; Hunter et al., 2015). Yet, a successful detection of species introduction relies on database completeness for those species. We show that, even among fish species identified as non-indigenous in freshwater ecosystems, up to 30% are currently missing in the best curated 16S database (Figure 4). For the genetic marker 12S, a maximum of 55% of non-indigenous species are sequenced per basin, twice as much as native species. It was expected that more non-indigenous species would be genetically referenced compared to native ones since referencing species occurrence outside their native range necessarily assumes their observation and a large proportion of introductions being intentional for recreational fishing (Leprieur, Beauchard, Blanchet, Oberdorff, & Brosse, 2008), making tissue for genetic sequencing easily available. We highlight here that despite a higher coverage for non-indigenous species (Figure 4), the potential of eDNA to detect invasion events and provide early warning signals is still limited while crucial for mitigating deleterious effects (Vander Zanden, Hansen, Higgins, & Kornis, 2010).

### 4.4 | Sequencing threatened species to support their monitoring

Environmental DNA has a great potential in biodiversity conservation, addressing the constraints of detecting elusive or low-abundant species missed by traditional surveys. The proportions of threatened species estimated by the IUCN Red List (11% of freshwater and 3%

of marine fishes) are likely underestimated as 48% of fish species are unevaluated while 7 to 9% are Data Deficient (Figure S2). Although the fate of Data Deficient species remains largely unexplored, they form the category with the least coverage in public genetic databases and are estimated to hide a large proportion of already threatened species (Bland, Collen, Orme, & Bielby, 2015). Even among threatened species, less than 50% have referenced sequences across all genetic markers, and surprisingly, their coverage is lower than Least Concern species for freshwater fishes. This can be due to the high number of threatened freshwater fishes, mainly located in hard-to-explore tropical regions (Collen et al., 2014).

Most threatened freshwater fishes live in large tropical basins such as the Congo, the Mekong or the Amazon (Figure 5). However, the Mississippi Basin, although located in a well-developed and science-leading country, the United States, where conservation measures and monitoring programs are well established, hosts 72 threatened species that are not sequenced for a 12S primer pair. So, efforts to complement genetic reference databases must be widespread and are not only related to the level of species richness or economic development, as often assumed.

### 4.5 | Interactive online application to support eDNA metabarcoding studies

We developed the user-friendly web app interface GAPeDNA to synthesize this large amount of information and make it easily accessible, even without any coding skills. It allows users to select a taxonomic group (at the moment, only freshwater and marine fish are available), the spatial unit or area, the genetic markers of interest and the corresponding primers to evaluate their global spatialized species coverage in public genetic databases, and have access to the corresponding list of species per spatial unit and status (IUCN). This permits the assessment of species remaining to be sequenced for a given spatial zone and sets priorities for sequencing. Although this study is focused on fish as an example, any new taxa can be added to GAPeDNA, providing necessary information is given: 1) primers suited for metabarcoding and 2) global spatialized species checklists. This can thus expand the reach and potential of this tool within the metabarcoding scientific community and managers using eDNA for ecological surveys.

As the adoption of eDNA metabarcoding as a standard and robust monitoring approach worldwide depends on its ability to identify organisms at the species level, we hope that our tool and its potential as demonstrated by the fish example included here will encourage researchers, managers, foundations and institutions to work towards a joint effort for a global sequencing effort targeting taxa of interest to enhance eDNA metabarcoding inventories.

### ACKNOWLEDGEMENTS

We thank Cyril Bernard for assistance with the deployment of the application on the online server, and Pierre Lopez for the illustration in Figure 1a.

## CONFLICT OF INTEREST

TM and TD are research scientists in a private company, specialized in the use of eDNA for species detection.

## PEER REVIEW

The peer review history for this article is available at <https://publons.com/publon/10.1111/ddi.13142>.

## DATA AVAILABILITY STATEMENT

Source code and data for the app are available on GitHub: <https://github.com/virginiemarques/GAPeDNA>. The web-app application is accessible on both GitHub and CEFE online server: <https://shiny.cefe.cnrs.fr/GAPeDNA/>.

## ORCID

Virginie Marques  <https://orcid.org/0000-0002-5142-4191>

Camille Albouy  <https://orcid.org/0000-0003-1629-2389>

Jean-Baptiste Juhel  <https://orcid.org/0000-0003-2627-394X>

## REFERENCES

- Albins, M. A., & Hixon, M. A. (2013). Worst case scenario: Potential long-term effects of invasive predatory lionfish (*Pterois volitans*) on Atlantic and Caribbean coral-reef communities. *Environmental Biology of Fishes*, 96(10–11), 1151–1157. <https://doi.org/10.1007/s10641-011-9795-1>
- Albouy, C., Archambault, P., Appeltans, W., Araújo, M. B., Beauchesne, D., Cazelles, K., ... Gravel, D. (2019). The marine fish food web is globally connected. *Nature Ecology & Evolution*, 3(8), 1153–1161. <https://doi.org/10.1038/s41559-019-0950-y>
- Andrello, M., Guilhaumon, F., Albouy, C., Parravicini, V., Scholtens, J., Verley, P., ... Mouillot, D. (2017). Global mismatch between fishing dependency and larval supply from marine reserves. *Nature Communications*, 8(1), 16039. <https://doi.org/10.1038/ncomm516039>
- Anticamara, J. A., Watson, R., Gelchu, A., & Pauly, D. (2011). Global fishing effort (1950–2010): Trends, gaps, and implications. *Fisheries Research*, 107(1–3), 131–136. <https://doi.org/10.1016/j.fishres.2010.10.016>
- Barange, M., Merino, G., Blanchard, J. L., Scholtens, J., Harle, J., Allison, E. H., ... Jennings, S. (2014). Impacts of climate change on marine ecosystem production in societies dependent on fisheries. *Nature Climate Change*, 4(3), 211–216. <https://doi.org/10.1038/nclimate2119>
- Barlow, J., França, F., Gardner, T. A., Hicks, C. C., Lennox, G. D., Berenguer, E., ... Graham, N. A. J. (2018). The future of hyperdiverse tropical ecosystems. *Nature*, 559(7715), 517–526. <https://doi.org/10.1038/s41586-018-0301-1>
- Bar-On, Y. M., Phillips, R., & Milo, R. (2018). The biomass distribution on Earth. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25), 6506–6511. <https://doi.org/10.1073/pnas.1711842115>
- Bax, N., Williamson, A., Aguero, M., Gonzalez, E., & Geeves, W. (2003). Marine invasive alien species: A threat to global biodiversity. *Marine Policy*, 27(4), 313–323. [https://doi.org/10.1016/S0308-597X\(03\)00041-1](https://doi.org/10.1016/S0308-597X(03)00041-1)
- Bland, L. M., Collen, B., Orme, C. D. L., & Bielby, J. (2015). Predicting the conservation status of data-deficient species. *Conservation Biology*, 29(1), 250–259. <https://doi.org/10.1111/cobi.12372>
- Blowes, S. A., Supp, S. R., Antão, L. H., Bates, A., Bruelheide, H., Chase, J. M., ... Dornelas, M. (2019). The geography of biodiversity change in marine and terrestrial assemblages. *Science*, 366(6463), 339–345. <https://doi.org/10.1126/science.aaw1620>
- Boussarie, G., Bakker, J., Wangensteen, O. S., Mariani, S., Bonnin, L., Juhel, J.-B., ... Mouillot, D. (2018). Environmental DNA illuminates the dark diversity of sharks. *Science Advances*, 4(5), eaap9661. <https://doi.org/10.1126/sciadv.aap9661>
- Boyer, F., Mercier, C., Bonin, A., Bras, Y. L., Taberlet, P., & Coissac, E. (2016). OBITOOLS: A UNIX-inspired software package for DNA metabarcoding. *Molecular Ecology Resources*, 16(4), 176–182. <https://doi.org/10.1111/1755-0998.12428>
- Brandl, S. J., Tornabene, L., Goatley, C. H. R., Casey, J. M., Morais, R. A., Côté, I. M., ... Bellwood, D. R. (2019). Demographic dynamics of the smallest marine vertebrates fuel coral reef ecosystem functioning. *Science*, 364(6446), 1189–1192. <https://doi.org/10.1126/science.aav3384>
- Carraro, L., Hartikainen, H., Jokela, J., Bertuzzo, E., & Rinaldo, A. (2018). Estimating species distribution and abundance in river networks using environmental DNA. *Proceedings of the National Academy of Sciences of the United States of America*, 115(46), 11724–11729. <https://doi.org/10.1073/pnas.1813843115>
- Chamberlain, S. (2018). *rredlist: 'IUCN' Red List Client. R package version 0.5.0*. <https://CRAN.R-project.org/package=rredlist>
- Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2019). *shiny: Web Application Framework for R. R package version 1.3.2*. <https://CRAN.R-project.org/package=shiny>
- Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., ... 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. <https://doi.org/10.1111/1755-0998.12900>
- Cinner, J. E., Huchery, C., MacNeil, M. A., Graham, N. A. J., McClanahan, T. R., Maina, J., ... Mouillot, D. (2016). Bright spots among the world's coral reefs. *Nature*, 535(7612), 416–419. <https://doi.org/10.1038/nature18607>
- Clavero, M., & Garcia-Berthou, E. (2005). Invasive species are a leading cause of animal extinctions. *Trends in Ecology and Evolution*, 20(3), 110. <https://doi.org/10.1016/j.tree.2005.01.003>
- Collen, B., Whitton, F., Dyer, E. E., Baillie, J. E. M., Cumberlidge, N., Darwall, W. R. T., ... Böhm, M. (2014). Global patterns of freshwater species diversity, threat and endemism. *Global Ecology and Biogeography*, 23(1), 40–51. <https://doi.org/10.1111/geb.12096>
- Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., ... Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, 10(11), 1985–2001. <https://doi.org/10.1111/2041-210X.1>
- Dickie, I. A., Boyer, S., Buckley, H. L., Duncan, R. P., Gardner, P. P., Hogg, I. D., ... Weaver, L. (2018). Towards robust and repeatable sampling methods in eDNA-based studies. *Molecular Ecology Resources*, 18(5), 940–952. <https://doi.org/10.1111/1755-0998.12907>
- Dougherty, M. M., Larson, E. R., Renshaw, M. A., Gantz, C. A., Egan, S. P., Erickson, D. M., & Lodge, D. M. (2016). Environmental DNA (eDNA) detects the invasive rusty crayfish *Orconectes rusticus* at low abundances. *Journal of Applied Ecology*, 53(3), 722–732. <https://doi.org/10.1111/1365-2664.12621>
- Duffy, J. E., Lefcheck, J. S., Stuart-Smith, R. D., Navarrete, S. A., & Edgar, G. J. (2016). Biodiversity enhances reef fish biomass and resistance to climate change. *Proceedings of the National Academy of Sciences of the United States of America*, 113(22), 6230–6235. <https://doi.org/10.1073/pnas.1524465113>
- Ficetola, G. F., Miaud, C., Pompanon, F., & Taberlet, P. (2008). Species detection using environmental DNA from water samples. *Biology Letters*, 4(4), 423–425. <https://doi.org/10.1098/rsbl.2008.0118>
- Hickisch, R., Hodgetts, T., Johnson, P. J., Sillero-Zubiri, C., Tockner, K., & Macdonald, D. W. (2019). Effects of publication bias on

- conservation planning. *Conservation Biology*, 33(5), 1151–1163. <https://doi.org/10.1111/cobi.13326>
- Hicks, C. C., Cohen, P. J., Graham, N. A. J., Nash, K. L., Allison, E. H., D'Lima, C., ... MacNeil, M. A. (2019). Harnessing global fisheries to tackle micronutrient deficiencies. *Nature*, 574(7776), 95–98. <https://doi.org/10.1038/s41586-019-1592-6>
- Hillebrand, H. (2004). On the generality of the latitudinal diversity gradient. *The American Naturalist*, 163(2), 192–211. <https://doi.org/10.1086/381004>
- Hunter, M. E., Oyler-McCance, S. J., Dorazio, R. M., Fike, J. A., Smith, B. J., Hunter, C. T., ... Hart, K. M. (2015). Environmental DNA (eDNA) sampling improves occurrence and detection estimates of invasive Burmese pythons. *PLoS ONE*, 10(4), 1–17. <https://doi.org/10.1371/journal.pone.0121655>
- Jerde, C. L., Wilson, E. A., & Dressler, T. L. (2019). Measuring global fish species richness with eDNA metabarcoding. *Molecular Ecology Resources*, 19(1), 19–22. <https://doi.org/10.1111/1755-0998.12929>
- Juhel, J.-B., Utama, R. S., Marques, V., Vimono, I. B., Sugeha, H. Y., Kadarusman, ... Hocdé, R. (2020). Accumulation curves of environmental DNA sequences predict coastal fish diversity in the coral triangle. *Proceedings of the Royal Society B: Biological Sciences*, 287(1930), 20200248.
- Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., ... Apweiler, R. (2005). The EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, 33(Database issue), D29–D33. <https://doi.org/10.1093/nar/gki098>
- Lefcheck, J. S., Innes-Gold, A. A., Brandl, S. J., Steneck, R. S., Torres, R. E., & Rasher, D. B. (2019). Tropical fish diversity enhances coral reef functioning across multiple scales. *Science Advances*, 5(3), eaav6420. <http://dx.doi.org/10.1126/sciadv.aav6420>
- Leprieur, F., Beauchard, O., Blanchet, S., Oberdorff, T., & Brosse, S. (2008). Fish invasions in the world's river systems: When natural processes are blurred by human activities. *PLoS Biology*, 6(2), 404–410. <https://doi.org/10.1371/journal.pbio.0060028>
- Lesser, M. P., Slattery, M., Laverick, J. H., Macartney, K. J., & Bridge, T. C. (2019). Global community breaks at 60 m on mesophotic coral reefs. *Global Ecology and Biogeography*, 28(10), 1403–1416. <https://doi.org/10.1111/geb.12940>
- Link, J. S., & Watson, R. A. (2019). Global ecosystem overfishing: Clear delineation within real limits to production. *Science Advances*, 5(6), 1–12. <https://doi.org/10.1126/sciadv.aav0474>
- Magurran, A. E., Deacon, A. E., Moyes, F., Shimadzu, H., Dornelas, M., Phillip, D. A. T., & Ramnarine, I. W. (2018). Divergent biodiversity change within ecosystems. *Proceedings of the National Academy of Sciences of the United States of America*, 115(8), 1843–1847. <https://doi.org/10.1073/pnas.1712594115>
- McCauley, D. J., Pinsky, M. L., Palumbi, S. R., Estes, J. A., Joyce, F. H., & Warner, R. R. (2015). Marine defaunation: Animal loss in the global ocean. *Science*, 347(6219), 247–254. <https://doi.org/10.1126/science.1255641>
- McGee, M. D., Borstein, S. R., Neches, R. Y., Buescher, H. H., Seehausen, O., & Wainwright, P. C. (2015). A pharyngeal jaw evolutionary innovation facilitated extinction in Lake Victoria cichlids. *Science*, 350(6264), 1077–1079. <https://doi.org/10.1126/science.aab0800>
- Myers, N., Mittermeier, R., Mittermeier, C., Da Fonseca, G., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403, 853–858. <https://doi.org/10.1038/35002501>
- OBIS (2020). Ocean Biodiversity Information System. Intergovernmental Oceanographic Commission of UNESCO. Retrieved February 2, 2018. Retrieved from <https://obis.org/>.
- Pebesma, E. (2016). GeoSPARQL (Perry and Herring, 2012), and open source libraries that empower the open source geospatial software landscape including GDAL (Warmerdam, 2008), GEOS (GEOS Development Team, 2017), and liblwgeom (a PostGIS component). *The R Journal*, 10(1), 439–446.
- Pellissier, L., Heine, C., Rosauer, D. F., & Albouy, C. (2018). Are global hotspots of endemic richness shaped by plate tectonics? *Biological Journal of the Linnean Society*, 123(1), 247–261. <https://doi.org/10.1093/biolinnean/blx125>
- Pinheiro, H. T., Moreau, S., Daly, M., & Rocha, L. A. (2019). Will DNA barcoding meet taxonomic needs? *Science*, 365(6456), 873–875.
- Reid, A. J., Carlson, A. K., Creed, I. F., Eliason, E. J., Gell, P. A., Johnson, P. T. J., ... Cooke, S. J. (2019). Emerging threats and persistent conservation challenges for freshwater biodiversity. *Biological Reviews*, 94(3), 849–873. <https://doi.org/10.1111/brv.12480>
- Rocha, L. A., Pinheiro, H. T., Shepherd, B., Papastamatiou, Y. P., Luiz, O. J., Pyle, R. L., & Bongaerts, P. (2018). Mesophotic coral ecosystems are threatened and ecologically distinct from shallow water reefs. *Science*, 284, 281–284. <https://doi.org/10.1126/science.aaq1614>
- Spalding, M. D., Fox, H. E., Allen, G. R., Davidson, N., Ferdaña, Z. A., Finlayson, M., ... Robertson, J. (2007). Marine ecoregions of the World: A bioregionalization of coastal and shelf areas. *BioScience*, 57(7), 573. <https://doi.org/10.1641/B570707>
- Stat, M., John, J., DiBattista, J. D., Newman, S. J., Bunce, M., & Harvey, E. S. (2019). Combined use of eDNA metabarcoding and video surveillance for the assessment of fish biodiversity. *Conservation Biology*, 33(1), 196–205. <https://doi.org/10.1111/cobi.13183>
- Tedesco, P. A., Beauchard, O., Bigorne, R., Blanchet, S., Buisson, L., Conti, L., ... Oberdorff, T. (2017). Data descriptor: A global database on freshwater fish species occurrence in drainage basins. *Scientific Data*, 4, 1–6. <https://doi.org/10.1038/sdata.2017.141>
- Tsuji, S., Takahara, T., Doi, H., Shibata, N., & Yamanaka, H. (2019). The detection of aquatic macroorganisms using environmental DNA analysis—A review of methods for collection, extraction, and detection. *Environmental DNA*, 1(2), 99–108. <http://dx.doi.org/10.1002/edn3.21>
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., ... Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4), 929–942. <https://doi.org/10.1111/mec.13428>
- Vander Zanden, M. J., Hansen, G. J. A., Higgins, S. N., & Kornis, M. S. (2010). A pound of prevention, plus a pound of cure: Early detection and eradication of invasive species in the Laurentian Great Lakes. *Journal of Great Lakes Research*, 36(1), 199–205. <https://doi.org/10.1016/j.jglr.2009.11.002>
- Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., ... Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *Science of the Total Environment*, 678, 499–524. <https://doi.org/10.1016/j.scitotenv.2019.04.247>
- Wilson, R. W., Millero, F. J., Taylor, J. R., Walsh, P. J., Christensen, V., Jennings, S., & Grosell, M. (2009). Contribution of fish to the marine inorganic carbon cycle. *Science*, 323, 359–362. <https://doi.org/10.1126/science.1157972>
- Witte, F., Goldschmidt, T., Wanink, J., van Oijen, M., Goudswaard, K., Witte-Maas, E., & Bouton, N. (1992). The destruction of an endemic species flock: Quantitative data on the decline of the haplochromine cichlids of Lake Victoria. *Environmental Biology of Fishes*, 34(1), 1–28. <https://doi.org/10.1007/BF00004782>
- Yamamoto, S., Masuda, R., Sato, Y., Sado, T., Araki, H., Kondoh, M., Miya, M. (2017). Environmental DNA metabarcoding reveals local fish communities in a species-rich coastal sea. *Scientific Reports*, 7, 40368. <https://doi.org/10.1038/srep40368>

**BIOSKETCHES**

**Virginie Marques** is a PhD student at the University of Montpellier (France). She is interested in reef fish ecology and applying environmental DNA metabarcoding to study the distribution of aquatic vertebrates. She is part of the MEGAFUNA Consortium, supported by Monaco Explorations (<https://www.monacoexplorations.org/en/>) and SpyGen company (<http://www.spygen.com/fr/>), which ambitions to provide a global view of fish biodiversity across the oceans using emerging non-destructive technological tools (eDNA, remote videos and artificial intelligence).

Authors' contribution: V.M., J.B.J. and D.M. designed the research; T.M. and C.A. provided the data; V.M., J.B.J. and D.M. analysed the data and interpreted the results; V.M. and J.B.J. wrote the initial draft and designed the figures; V.M., J.B.J., S.M., T.D. and D.M. wrote and corrected the paper; all authors approved the final draft; and V.M. designed and created the interactive application.

**SUPPORTING INFORMATION**

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

**How to cite this article:** Marques V, Milhau T, Albouy C, et al. GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Divers Distrib.* 2020;00:1–13. <https://doi.org/10.1111/ddi.13142>