ECOGRAPHY

Research

Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA metabarcoding sequences

Virginie Marques, Pierre-Édouard Guérin, Mathieu Rocle, Alice Valentini, Stéphanie Manel, David Mouillot and Tony Dejean

V. Marques (https://orcid.org/0000-0002-5142-4191) ✉ (virginie.marques01@gmail.com) and D. Mouillot, MARBEC, Univ. de Montpellier, CNRS, Ifremer, IRD, Montpellier, France. – P.-É. Guérin, S. Manel and VM, CEFE, Univ. Montpellier, CNRS, EPHE-PSL Univ., IRD, Univ. Paul Valéry Montpellier, Montpellier, France. – M. Rocle, Compagnie Nationale du Rhône, Direction de l'Ingénierie, Lyon, France. – A. Valentini and T. Dejean, SPYGEN, Le Bourget-du-Lac, France.

Ecography
43: 1–12, 2020
doi: 10.1111/ecog.05049

Subject Editor: Simon Creer
Editor-in-Chief: Miguel Araújo
Accepted 15 July 2020

Human activities impact all ecosystems on Earth, which urges scientists to better understand biodiversity changes across temporal and spatial scales. Environmental DNA (eDNA) metabarcoding is a promising non-invasive method to assess species composition in a wide range of ecosystems. Yet, this method requires the completeness of a reference database, i.e. a list of DNA sequences attached to each species of the regional pool, which is rarely met. As an alternative, molecular operational taxonomic units (MOTUs) can be extracted as clusters of sequences. However, the extent to which the diversity of MOTUs can predict the diversity of species across spatial scales is unknown. Here, we used 196 samples along the Rhone river (France) for which the reference database is complete to assess whether a blind eDNA approach can reliably predict the ground-truth number of species at different spatial scales. Using the 12S rDNA teleo primer, we curated and clustered 60 million sequences into MOTUs using a new assembled bioinformatic pipeline. We show that stringent quality filters were necessary to remove artefact noise, notably MOTUs present in a single PCR replicate, which represented 55% of MOTUs (103). Post-clustering cleaning also removed 19 additional erroneous MOTUs and only discarded one truly present species. We then show that the diversity of retained fish MOTUs accurately predicted the local (α, $r = 0.98$) and regional (γ) ground-truth species diversity (67 MOTUs versus 63 species), but also the species dissimilarity between samples (β-diversity, $r = 0.98$). This work paves the way towards extending the use of eDNA metabarcoding in community ecology and biogeography despite major gaps in genetic reference databases.

Keywords: 12S primer, α-β-δ-diversity, clustering, metabarcoding, MOTUs, reference database

NORDIC SOCIETY OIKOS

www.ecography.org

Introduction

In the new era of the Anthropocene, most ecosystems are experiencing severe human impacts and environmental changes with major consequences on species diversity (McCauley et al. 2015, Hughes et al. 2017, Isbell et al. 2019). Nevertheless, the ongoing reorganization of biodiversity is still poorly quantified and understood (but see Blowes et al. 2019) for two major reasons. First, the losses or gains of species are scale dependent with complex results emerging at the local or regional spatial scale (Vellend et al. 2013, Dornelas et al. 2014). For instance, several studies show that local species diversity is on average constant over time (Dornelas et al. 2014, Magurran et al. 2018), even under human impacts, while other studies report alarming species losses regionally or globally (Galetti et al. 2014, Doherty et al. 2016, Finderup Nielsen et al. 2019). Thus, any biodiversity monitoring should be spatially explicit (McGill et al. 2015) with three major components 1) local or α-diversity for the number of species within a given site, 2) spatial variation or β-diversity in species composition among sites and 3) regional or γ-diversity for the number of species within a geographical area containing all sites (Whittaker 1972). Second, biases and gaps in biodiversity inventories prevent accurate and comparable assessments across space and time (Hortal et al. 2015). This is particularly problematic when species are rare, small, cryptic or elusive or when ecosystems are either species-rich like in the tropics or hardly accessible like the deep sea (Mora et al. 2008, Menegotto and Rangel 2018). Hence, there is an urgent need for standardized and accurate biodiversity monitoring methods across spatial scales allowing reliable inter-study comparisons.

The metabarcoding of environmental DNA (eDNA) has the potential to fill this gap as it has been shown to surpass most traditional methods in species detection for both terrestrial and aquatic ecosystems (Bohmann et al. 2014, Valentini et al. 2016, Ruppert et al. 2019, Sales et al. 2020). Indeed, all organisms shed cells containing DNA in their environment, as intra or extra-cellular material, and can be retrieved for up to a few days (Dejean et al. 2011, Collins et al. 2018, Harrison et al. 2019). Amplification and high-throughput eDNA sequencing followed by bioinformatic analyses produce a list of sequences with the ultimate goal to assess species composition in a given site. This bioinformatic step requires the completeness of a reference database, i.e. a list of sequences attached to each species in the regional pool, to accurately assign each eDNA sequence to a given species. Yet, reference databases are often incomplete (Weigand et al. 2019). An estimated 91% of eukaryotic species inhabiting the ocean are yet to be described (Mora et al. 2011a) while only 13% of all described Teleostean fish species are referenced in public reference databases like the European Nucleotide Archive (ENA) (Leinonen et al. 2011) for the 12S ribosomal DNA fragment amplified by the teleo primers (Valentini et al. 2016), limiting the extent of species diversity revealed by eDNA metabarcoding.

Currently, completing reference databases would require massive sampling and sequencing efforts since many species still remain undiscovered due to their intrinsic nature (rare, small or elusive) or their unexplored habitat (e.g. deep sea) (Menegotto and Rangel 2018). Moreover, polymerase chain reaction (PCR) and sequencing generate numerous errors, overestimating the true number of species by several orders of magnitude (Edgar and Flyvbjerg 2015, Flynn et al. 2015). Thus, accurate methods able to assess biodiversity without complete reference databases while considering PCR and sequencing errors are urgently needed.

The microbial field pioneered methodological advances to infer biological diversity without a complete reference by clustering similar sequences into molecular operational taxonomic units (MOTUs) (Huse et al. 2010). However, these approaches focus mainly on fungi or unicellular organisms where the concept of species remains challenging (Pawlowski et al. 2018, Lladó Fernández et al. 2019). Even if clustering-based analyses are increasingly used in eDNA studies targeting vertebrates (Andruszkiewicz et al. 2017, Bakker et al. 2017, Closek et al. 2019, Sales et al. 2019), using the diversity of MOTUs as a reliable proxy for species diversity has yet not been evaluated. For instance, Closek et al. (2019) reported a large overestimation with more than 1300 MOTUs for 92 fish taxa only in the Californian Current upwelling ecosystem. The extent to which the metabarcoding of vertebrate eDNA can provide a reliable blind estimation of species diversity across spatial scales is unknown.

Here we evaluate how clusters of vertebrate eDNA sequences can predict species diversity across spatial scales. More precisely, we quantify how MOTUs can accurately predict local (α) and regional (γ) species diversity but also composition species dissimilarity between samples (β-diversity). For this, we focused on teleost fishes which are highly vulnerable to anthropogenic threats (Mora et al. 2011b) and represent the main group of vertebrates studied with eDNA (Tsuji et al. 2019). First, we highlight the geographic and taxonomic gaps in the reference database for the 12S mtDNA fragment, which is known to perform well with the teleo primer (Collins et al. 2019) designed by Valentini et al. (2016). Then, we assemble a metabarcoding bioinformatic pipeline based on sequence clustering using SWARM (Mahé et al. 2015), post-clustering using LULU (Frøslev et al. 2017) and stringent quality filters to analyze eDNA sequences from 196 samples along 500 km of the Rhône river (France). From the composition of MOTUs in each sample, we estimate α-, β- and γ-diversity and compare them to their analogs obtained with ground-truth assignment of all sequences using the complete reference database without clustering. Finally, we discuss strengths and weaknesses of this approach based on eDNA sequence clustering to assess taxonomic diversity across spatial scales, even when lacking exhaustive reference databases.

Material and methods

Global taxonomic and spatial gap analysis for fish

Recent fish metabarcoding studies indicate that primers located on the 12S ribosomal rRNA locus (12S rDNA)

perform better (i.e. detect more species, with less bias and more specific amplification) than primers based on alternatives loci (Ribosomal locus 16S, the cytochrome c oxidase I gene (COI)) (Collins et al. 2019, Weigand et al. 2019). Although the COI gene and associated primers might cover a larger proportion of fish species in the reference database and have a higher interspecific variability, their lack of suitable conserved region complicates the definition of taxa-specific primers. COI primers exhibit a clear lack of consistency across replicates, have a low specificity leading to a low amplification of target organisms with often less than 5% of cleaned reads assigned to fish (Collins et al. 2019) resulting in a low detectability power (Deagle et al. 2014, Bylemans et al. 2018, Collins et al. 2019). Among the fish eDNA 12S markers, we selected the teleo marker (forward primer-ACACCGCCCGTCACTCT, reverse primer-CTTCCGGTACACTTACCATG) (Valentini et al. 2016) given its high ability to detect fish species even in highly diverse ecosystems (Civade et al. 2016, Valentini et al. 2016, Bylemans et al. 2018, Pont et al. 2018, Cantera et al. 2019, Cilleros et al. 2019).

We first assessed the global taxonomic coverage of the teleo primers by performing in silico PCR using ecoPCR (Boyer et al. 2016) on the entire public database ENA (Leinonen et al. 2011) (release 138, January 2019). To build our reference database, we allowed a maximum of three mismatches and compared the results with the complete fish taxonomy from FishBase (Froese and Pauly 2019). For the spatial analysis, we extracted freshwater fish checklists of all drainage basins from the most recent and comprehensive data at the global scale (Tedesco et al. 2017), covering about 80% of inland waters. We obtained marine checklists from OBIS (OBIS Ocean Biogeographic Information System) at 1° resolution (Albouy et al. 2019), and used them to estimate fish composition within marine ecoregions globally (Spalding et al. 2007).

eDNA sampling and sequencing

We downloaded the sequence data from a previous study by Pont et al. (2018). The complete dataset encompasses 196 eDNA samples collected along 500 km of the Rhone River (France, Supplementary material Appendix 1 Fig. A1), corresponding to 103 distinct sites with field replicates (between 1 and 4 samples per site) in 2016. Among those, the original study used only 118 samples corresponding to 59 sites, but all samples were collected and processed in parallel. For each sample, 30 l of freshwater water were filtered, extracted, amplified and sequenced (Pont et al. 2018).

Clustering methods

Accurately delineating 'true' biological sequences from PCR and sequencing noise has been an ongoing challenge since the emergence of next generation sequencing (NGS) technologies. Clustering sequences into molecular operational taxonomic units (MOTUs) or defining exact sequence variants (ESVs) as proxies for species is a common practice in the prokaryote microbial field but also to study unicellular eukaryotes or fungi (Huse et al. 2010, Schmidt et al. 2013, Zimmermann et al. 2015, Callahan et al. 2017) and more recently eDNA of vertebrates (Closek et al. 2019, Sales et al. 2019).

While clustering has been historically limited to the creation of MOTUs based on a fixed similarity threshold, usually 97% (Stackebrandt and Goebel 2008, Edgar 2018), it poorly generalizes across markers or biological models (Edgar and Flyvbjerg 2015, Mahé et al. 2015, Nguyen et al. 2015, Callahan et al. 2017). As an alternative, new methods generate either ESV like the divisive amplicon denoising algorithm (DADA2) (Callahan et al. 2016) or MOTUs from de novo clustering algorithms based on sequence distribution and abundance to correct errors, like SWARM (Mahé et al. 2015). SWARM is an agglomerative unsupervised de novo single-linkage-clustering algorithm, building networks to define MOTUs based on sequence proximity and relative abundance (Mahé et al. 2015). While a threshold-based algorithm simply groups sequences together according to a fixed value, SWARM forms chains linking sequences based on their similarity and analyses the pattern to optimally break the network and delineate MOTUs (Mahé et al. 2014, 2015). So, the 'true' sequence is expected to be the most abundant while less abundant but close sequences are considered as erroneous as they are more likely to accumulate errors. This process avoids the dependence on a fixed value, which is not recommended in eDNA metabarcoding with short barcodes where only one mismatch can imply a different species (Miya et al. 2015).

Pipelines workflow

We based our analysis on two different pipelines: one where each unique sequence is independently assigned to a given species (called the Species pipeline) and the other one which clusters sequences into MOTUs using the SWARM algorithm (called the MOTU pipeline). In the Species pipeline, a complete reference database is required to assign a taxa to each sequence. In the MOTU pipeline, each MOTU also requires a taxonomic assignment but the completeness of the reference database is not required, as a partially complete reference database is sufficient to exclude MOTUs representing non-specific amplification, in our case, all non-fish taxa.

First, pre-processing steps were common for both pipelines (Fig. 1). Reads were assembled using VSEARCH (Rognes et al. 2016), demultiplexed at the PCR replicate level and primers trimmed using CUTADAPT (Martin 2011) adapted from an existing metabarcoding pipeline (<https://github.com/frederic-mahe/swarm/wiki/Fred's-metabarcoding-pipeline>). No mismatches were allowed in tags for demultiplexing while sequences containing ambiguous nucleotides were discarded. Two additional steps were applied in the pre-processing for the MOTU pipeline. First unsupervised clustering was performed with SWARM, using a minimum distance of one nucleotide between each MOTU (d = 1), as one mismatch can separate two distinct species with

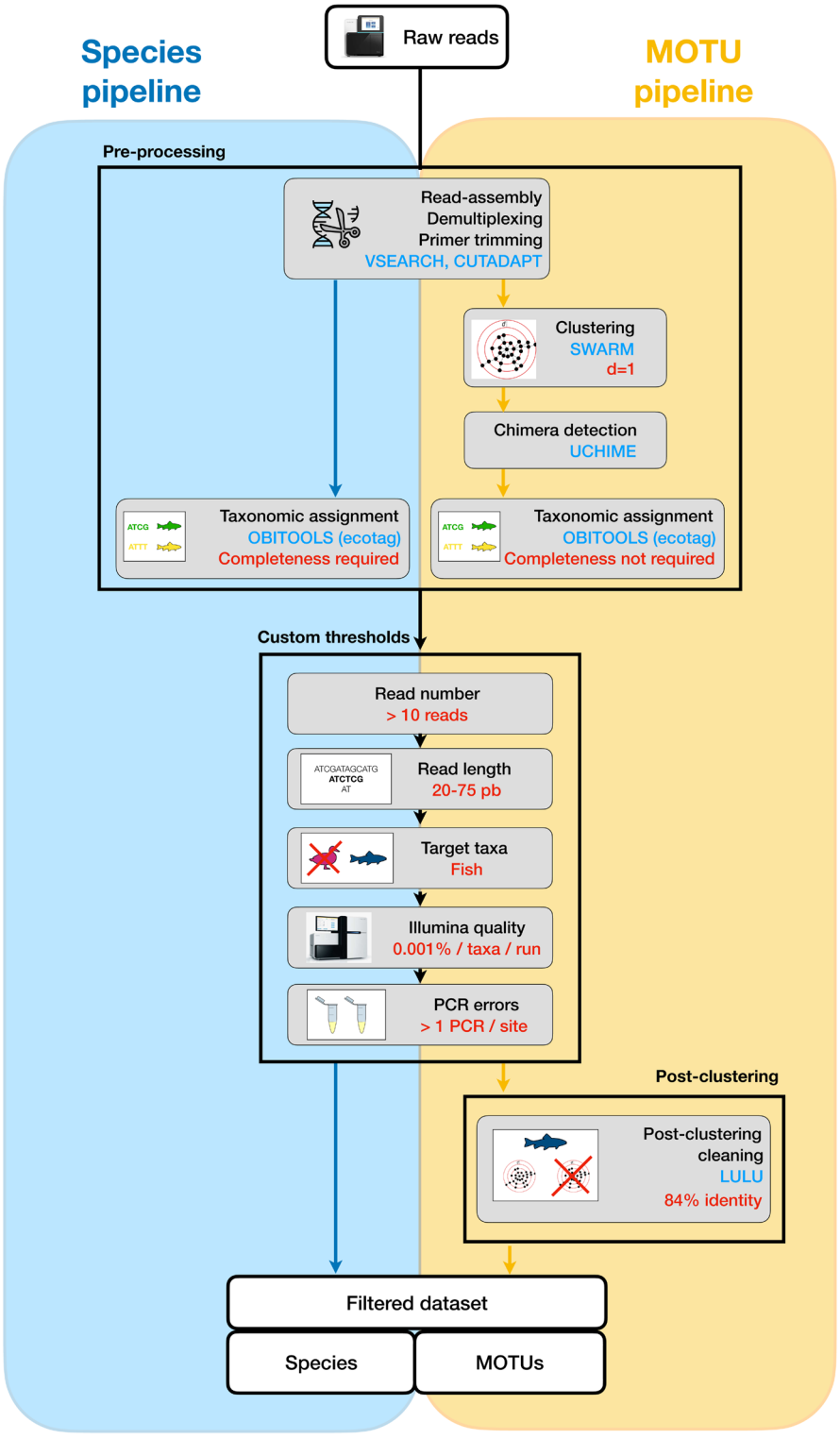

Figure 1. Illustration of the entire pipeline with three main steps: pre-processing, clustering, application of thresholds and post-clustering. Programs used are in blue and thresholds or requirements in red. Blue lines represent the classical alternative paths for the ground-truth method (Species pipeline), i.e. with the complete reference database and no clustering, whereas yellow lines represent the MOTU-based pipeline (MOTU pipeline), while black lines represent shared steps.

our primer. Taxonomic assignments of all unique sequences or MOTUs were then performed by ecotag, a lowest common ancestor (LCA) algorithm from the Obitools toolkit relying on the National Center for Biotechnology Information (NCBI) phylogeny tree (Boyer et al. 2016). Then, a set of custom and already published thresholds were applied on unique sequences for both the Species and MOTU pipelines (Fig. 1) (Valentini et al. 2016). All sequences or MOTUs with less than 10 reads, too short (< 20 bp), too long (> 75 bp) (Valentini et al. 2016) or not assigned to a fish phylum were discarded.

Each site usually has 2 samples as field replicates (except for 13 sites where the number of samples ranges from 1 to 4), and each sample has 12 PCR replicates, so most sites are represented by 24 individual PCRs (range: 12–48 PCRs replicates). For each site, we discarded all MOTUs or sequences present in only one PCR replicate (Civade et al. 2016). To avoid tag-jump noise (Schnell et al. 2015), all sequences with an abundance frequency of less than 0.001 per taxon/MOTU and per library were discarded. For the MOTU pipeline only, we then used the LULU algorithm (Frøslev et al. 2017) to clean MOTUs identified as erroneous based on sequence identity between MOTUs, abundances and patterns of co-occurrence. We used the blastn command line with the megablast algorithm to create the file matching all pairwise MOTUs to infer their similarity percentage. Then, to apply LULU, we used the 84% identity threshold (Frøslev et al. 2017) but also ran a sensitivity analysis with changes in the main parameters, i.e. the cross influence of identity threshold percentage and co-occurrence percentage (Supplementary material Appendix 1 Fig. A3).

Taxonomic assignments

For both pipelines, taxa assignments were performed on both our local database, exhaustive for resident species of the regional pool, and ENA (release 138, January 2019). For the Species pipeline, associating the local database with ENA (Leinonen et al. 2011) detected 24 extra species, among which 12 matches at 98% to our local database but at 100% in a public database to a foreign species (Supplementary material Appendix 1 Table A1). Those foreign species were unlikely to be present in the river, and most likely resulted from PCR or sequencing errors of local species randomly matching with foreign species. To avoid artificially inflating regional diversity from incorrect assignments, we only considered ENA assignments when our local database performed poorly (< 98% similarity). Among the 12 remaining species detected only by ENA and matching at < 98% to our local database, all were marine species from the Mediterranean Sea but 11 have records indicating a tolerance for brackish water while 6 were clearly known to enter estuaries (Supplementary material Appendix 1 Table A2). Most of those were also commonly consumed by humans, and DNA could have been transported into the river from sewage waters. Those extra species were hence kept for further analyses as they were unlikely to be errors generated at the PCR or sequencing step and they unlikely represent a methodological artefact.

Before analysis, assignments from ecotag were corrected to be more stringent as the algorithm can sometimes validate genus or family-level assignments to sequences with low similarity, which we chose to not trust blindly. This is due to the functioning of the ecotag algorithm (Boyer et al. 2016) and can happen in clades with a low species coverage in the reference database. We decided to add a level of standardization and only validate assignments at the species level for sequences matching at > 98% similarity, at 96–98% for the genus level, at 90–96% for the family level and at less than 90% similarity for the order or higher level for all sequences matching following a pilot study on phylogenetic signal for this marker (Supplementary material Appendix 1 Fig. A2).

Controlling taxonomic redundancy

When a sequence has a low percentage of similarity (< 98%), it can correspond to 1) a species absent from databases, 2) noise from PCR/sequencing errors from actual sequenced species or 3) rare but strong intra-specific variation at this non-coding locus which is prone to rapid mutations or insertions (Leinonen et al. 2011, Valentini et al. 2016). A common NGS metabarcoding issue is that for one species sequence matching at 100%, it can generate several noise variants matching at less than 100% (Frøslev et al. 2017). Hence, when counting the total number of taxa to infer the level of diversity, there is always a clear overestimation. For example, one *Salmo trutta* sequence with 100% similarity to a reference database would likely be accompanied by sequences matching at 97%, assigned at the *Salmo* genus and 95% assigned at the Salmonidae family. Where one species is present, the total taxa count can be three. To correct the number of taxa while being conservative, we created an estimated species count based on taxonomic correction for redundancy. A genus, family or order assignment can only be kept if there is no species already belonging to that rank, otherwise it would be more likely to be an error since the genetic databases are exhaustive for local resident species, the rest representing only a minority of rare sequences.

To evaluate the performance of LULU in the MOTU pipeline, we grouped taxa following this logic up to the family level. If a MOTU is assigned to a family for which a species representative is also detected, we assumed an error for this species and taxonomic redundancy. If a MOTU is assigned to an order only, it was not considered to represent an additional species.

Diversity comparison across scales

To assess the performance of our MOTU-based approach we calculated regional (γ) diversity, local (α) or sample diversity and dissimilarity between samples (β) with each pipeline. For the Species pipeline we retained all sequences matching at > 98% similarity cleaned for taxonomic redundancy to

count the number of distinct species. For the MOTU pipeline, we retained all MOTUs assigned to a fish taxa regardless of their similarity percentage. We used the software R ver. 3.6, where sample or α-diversity was computed as richness, i.e. plain species count. β diversity was computed using the Sorensen index, with the beta.temp and beta.multi functions from betapart package (Baselga and Orme 2012). A low value of dissimilarity between samples indicates similar communities, on a scale from 0 (identical) to 1 (totally dissimilar so no species or MOTU is common). We used the Mantel correlation test for pairwise sample comparisons.

Results

Global gaps in fish reference databases

Our analysis reveals that only 4243 out of 33 124 teleostean fish species (13%) are sequenced in the region amplified using the teleo primers, for both marine and freshwater environments (Fig. 2a). At higher taxonomic rank, we show that 38% of genera have at least one representative species sequenced for the 12S on the teleo fragment, this percentage reaching up to 80% for families. Next, we highlight a strong spatial heterogeneity between marine and freshwater environments but also among freshwater basins and marine ecoregions (Fig. 2b–c). For freshwater ecosystems, the proportion of fish species being referenced for the 12S fragment ranges from 0 to 100%, with tropical basins having an overall lower coverage than their temperate counterparts, except for Oceania where the proportion of sequenced species is among the highest. South America and Africa have by far the lowest coverage among all continents. For marine ecosystems, disparities are less pronounced but coverage varies between 10 and 53%. Ecoregions in Europe and Northern America have the highest coverage whereas tropical and southern ecoregions are the least covered.

γ-diversity assessment after filtering and clustering processes

In the 196 samples along the Rhone river, we obtain 60 689 053 reads of 299 225 distinct sequences with a mean of 309 617 reads per sample prior to any filtering (Table 1). First, we analyzed the eDNA metabarcoding data with the complete reference database (local database and ENA combined) with the Species pipeline (Fig. 1). We detect a total of 63 fish species (Table 1). Our new assembled MOTU pipeline applied on the same raw dataset identifies 67 MOTUs out of which 61 (91%) could be subsequently identified at the species level, i.e. matching at least at 98% of similarity with a species in the reference database.

We find that 98% of unique sequences and 96% of unique MOTUs correspond to either low abundant (< 10 reads) or non-fish species, so represent artefacts, noise or unspecific amplifications (Table 1), while only accounting for 12.5% and 4.4% of total reads, respectively. Sequence length filtering has a low influence, removing only 1 MOTU and no species. While removing only 0.004% of the total read count, our PCR filter removing all reads found in only one PCR replicate per site eliminates 45 MOTUs assigned to species (from 108 to 63) among which only 4 are possibly resident to the area (Supplementary material Appendix 1 Table A3). All other eliminated taxa are absent in the river and likely result from errors, contaminations from sewage waters or methodological artefacts. This PCR replicate filter also discards more than half of the detected MOTUs (86 out of 189, Table 1) representing mainly taxonomic redundancy and low-quality reads.

Following the PCR replicate filtering step, only 50 out of 86 MOTUs are represented by one taxon (Fig. 3), revealing either redundancy, with several MOTUs corresponding to the same taxa, or a lack of identification at the species level for the 36 remaining MOTUs. The application of LULU decreases the total number of MOTUs from 86 to 67 (Fig. 3). In particular, the number of taxa represented by more than 1 MOTU decreases from 15 (up to 6 MOTUs per taxa) to 8 after cleaning with LULU. Following this step, the lost MOTU representing a real taxa corresponds to a complex of two cyprinid fish species (*Ctenopharyngodon idella* and *Hypophthalmichthys molitrix*) for which teleo marker is not resolutive at the species level.

Finally, the regional pool (γ-diversity) of our fish Rhone dataset is comprised of 67 MOTUs among which 61 can be assigned to a species with 98% similarity while the ground-truth value is 63 fish species using the Species pipeline (Table 1).

Estimates of α and β species diversity using MOTUs

For each sample, we calculated the local (α-) diversity obtained by each pipeline so in terms of species and MOTUs. Overall the correlation between the number of MOTUs and the number of species is high and significant ($r = 0.98$; $p < 0.001$; Fig. 4a). The mean difference in local diversity across samples between the two pipelines is of 1.02 (SD = 1.5) with the MOTU-based approach underestimating true α-diversity. The maximum difference in local diversity is 5 (Fig. 4a), meaning that for one sample five less MOTUs are detected compared to the number of species identified with the reference database.

Since a similar value of α-diversity detected by the two pipelines does not necessary imply the same community composition, we performed a dissimilarity analysis (β-diversity) between samples pairs for both methods using the Sorensen index. We detect a high and significant correlation ($r = 0.98$, $p < 0.001$, Fig. 4b) between pairwise sample dissimilarity estimated with the Species and MOTU pipelines. We highlight no over or underestimation of dissimilarity by one pipeline compared to the other. Overall, the MOTU pipeline generates lower dissimilarity for 71% of pairs of samples compared to the Species pipeline but in 95% of all cases, the inferred level of dissimilarity has less than 0.1 difference between the two pipelines.

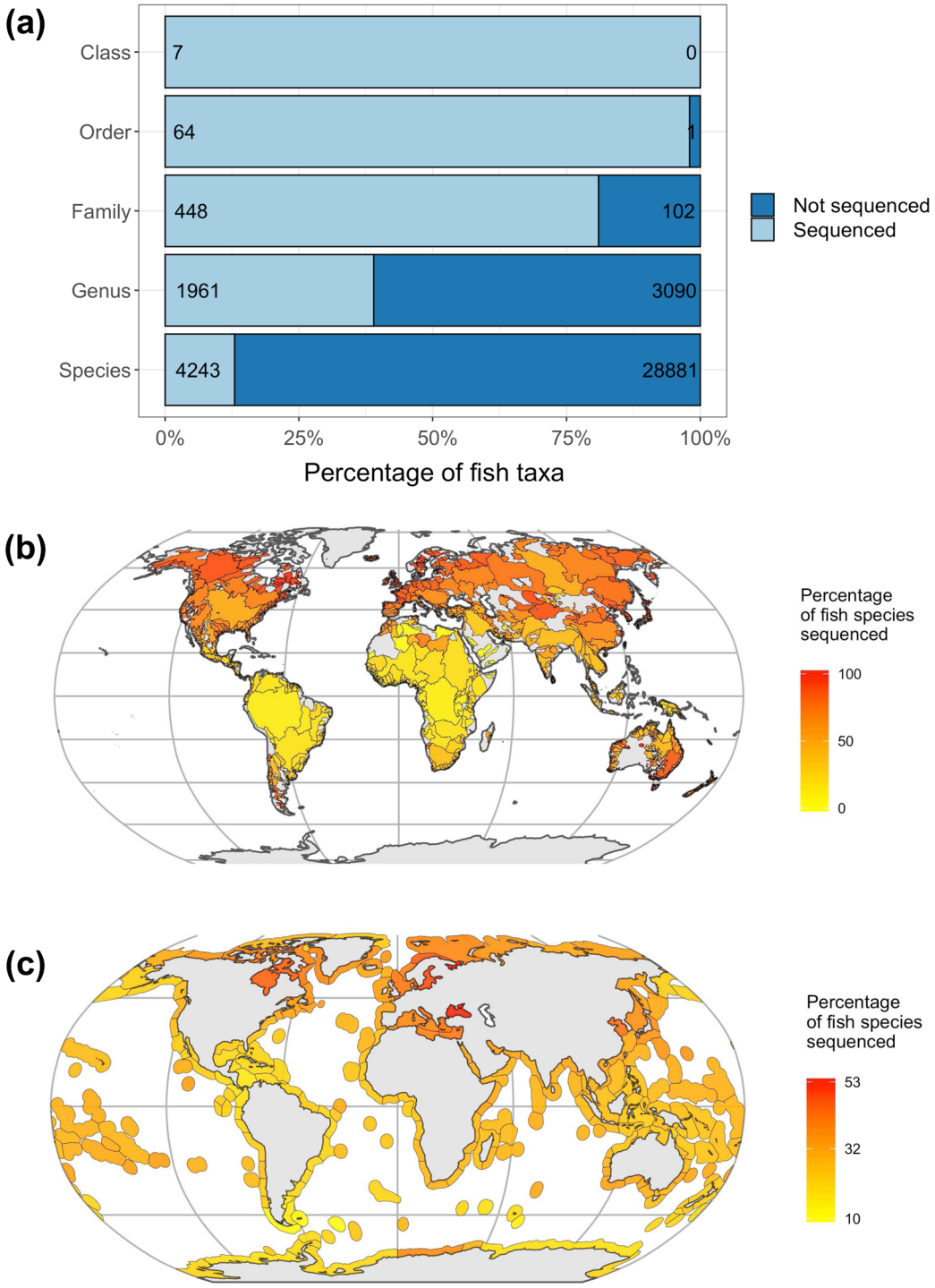

Figure 2. Percentage of sequenced freshwater and marine fish species using the teleo primer per taxonomic level (a), per freshwater basin (b) and per marine ecoregion (c).

Discussion

While eDNA metabarcoding represents a promising tool for scaling-up biodiversity inventories (Berry et al. 2019, Ruppert et al. 2019), its strong dependence on genetic reference databases limits its application in many regions of the world, as well as for some taxonomic groups or some habitats (Weigand et al. 2019). Indeed, even diverse yet well-studied ecosystems such as coral reefs do not have exhaustive genetic references for most lineages and the majority of commonly used primers in eDNA metabarcoding (DiBattista et al. 2017, West et al. 2020).

Table 1. Numbers (#) of reads, sequences, species and MOTUs identified and retained at each step of our Species or MOTUs pipelines (Fig. 1) with # Species representing the number of taxa corrected for taxonomic redundancy (see Methods). Details for each step are presented in Methods and Fig. 1.

	Species pipeline			MOTU pipeline	
Steps	# Reads	# Sequences	# Species	# Reads	# MOTUs
No filter	60 689 053	299 225	399	60 684 944	5375
> 10 reads	55 655 419	7819	227	60 593 926	568
Fish taxa	53 253 228	6424	108	57 988 700	190
Length filter	53 253 170	6422	108	57 988 623	189
> 1 PCR/site	53 021 739	6121	63	57 759 482	86
LULU	–	–	–	57 736 566	67

Some reference-free tools exist, but their application remains mostly limited to unicellular or fungi organisms, with different aims and constraints compared to eDNA studies targeting vertebrates. Moreover, such tools do not provide plausible diversity levels for most applications on vertebrate eDNA (Andruszkiewicz et al. 2017, Closek et al. 2019, Siegenthaler et al. 2019). Further, a proper testing of whether those approaches provide reliable diversity estimates is lacking (Pedrós-Alió 2006, Huse et al. 2010, Lladó Fernández et al. 2019) beyond controlled mock communities (Frøslev et al. 2017, Alberdi et al. 2018). In our study, we assembled a set of bioinformatic tools to generate fish MOTUs and assess the level of diversity across spatial scales based on the use of eDNA metabarcoding, using a well-known river system as a case study.

We show that, at the regional level, our MOTU-based pipeline provides a comparable estimate of species diversity with 67 MOTUs when 63 species are detected. However, some MOTUs represent either errors or unreferenced species, and 8 species remain undetected due to clustering and stringent filtering. Such weakness arises as many species have close sequences to each other and co-occur. So, it remains impossible for any algorithm to distinguish close species from errors. This dilemma – distinguishing rare MOTUs from errors – is inherent to clustering techniques (Huse et al. 2010, Frøslev et al. 2017, Pawlowski et al. 2018). Despite numerous attempts to solve this issue, there is still a trade-off between allowing false positives and creating false negatives (Reeder and Knight 2009). Among the MOTUs representing taxonomic redundancy, at least 3 taxa (*Gobio gobio*, *Alosa* sp.,

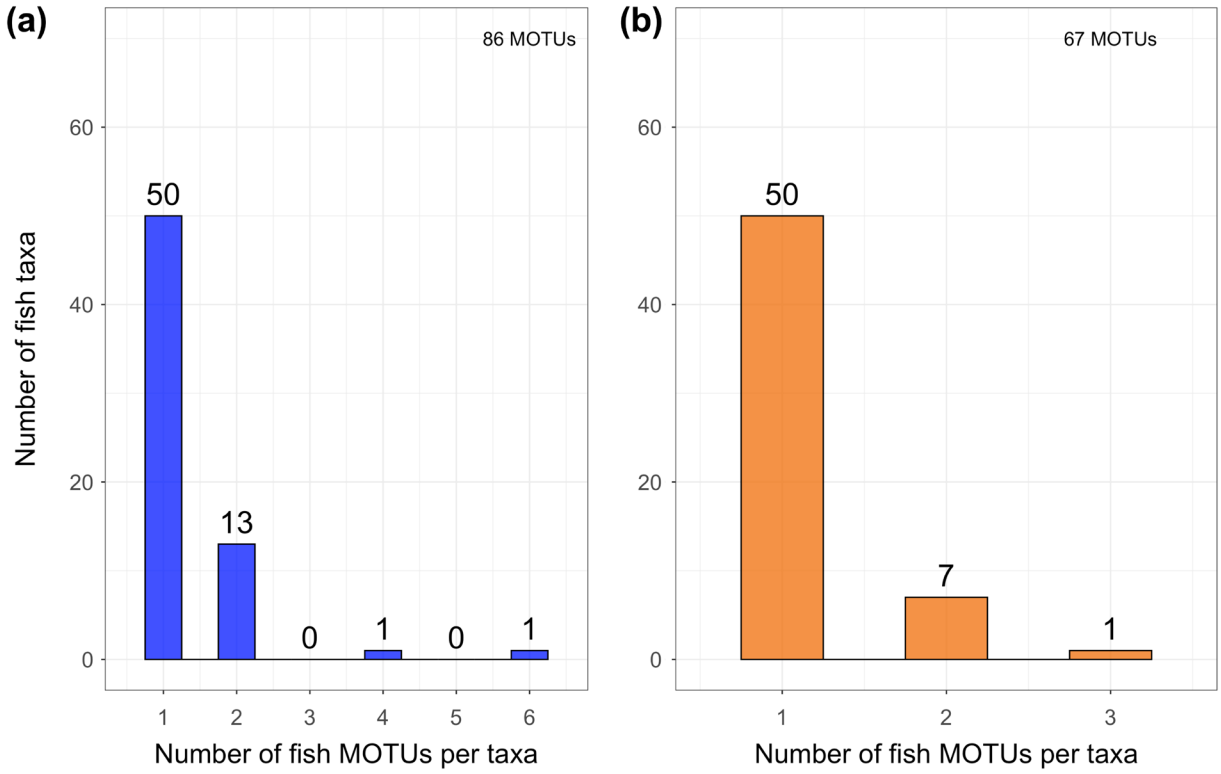

Figure 3. Distribution of the number of MOTUs per fish taxa (a) before LULU cleaning and (b) after LULU cleaning for taxonomic redundancy (Frøslev et al. 2017).

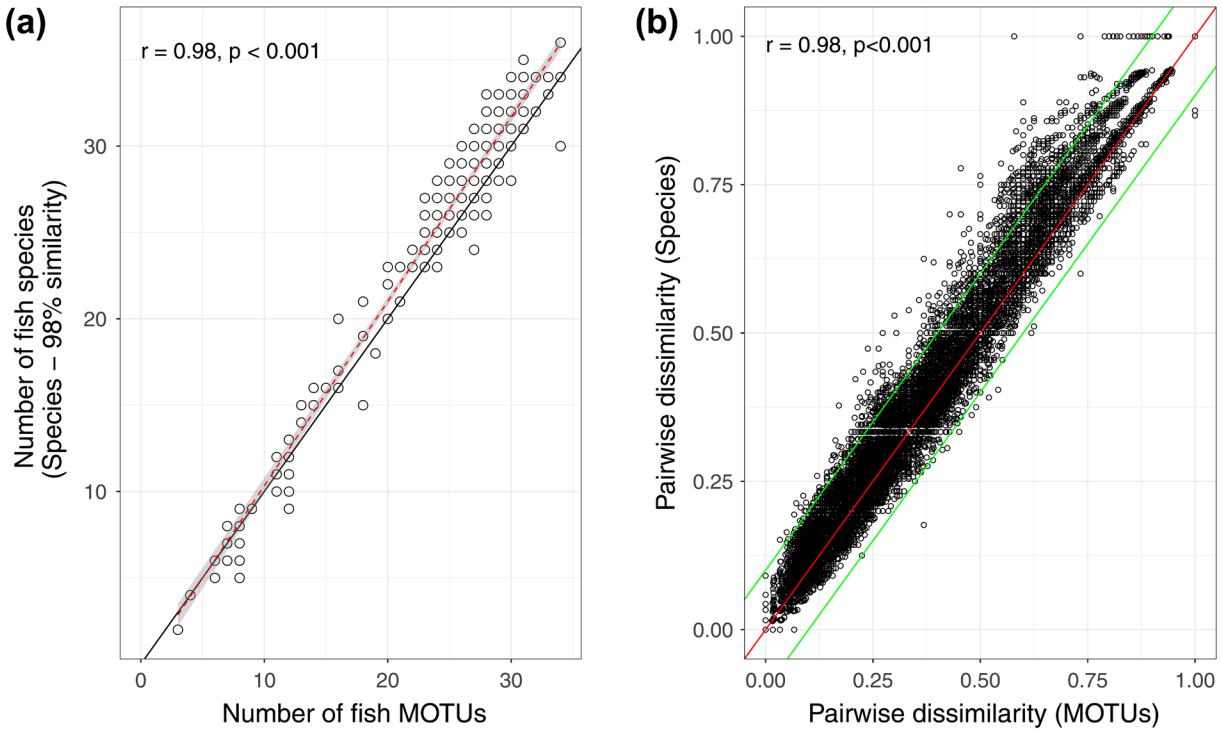

Figure 4. The Pearson linear correlation shows the strength of the relationship between the number of species and the number of MOTUs identified with our two pipelines for each sample (a). The black line represents the identity slope and the red line represents the linear regression between the number of species and that of MOTUs. (b) The Mantel correlation shows the relationship between the Species and the MOTU pipeline for pairwise sample dissimilarity. Each dot represents the β-diversity value for a pair of samples estimated by either one of our pipelines (Species versus MOTUs), red line represents the identity slope and green lines represent respectively the +0.1 and −0.1 limits.

Phoxinus phoxinus) are known to hybridize (Alexandrino et al. 2006) or are under taxonomic revision with the potential existence of multiple species displaying genetic variations (Kottelat and Persat 2005, Collin and Fumagalli 2011) while for one species (*Dicentrarchus labrax*), genetic public databases (Sayers et al. 2019) highlight a marked intra-specific variability.

Sequencing and PCR errors are common in metabarcoding datasets (Siegwald et al. 2017), but as eDNA barcodes are usually short to enhance detectability (Bohmann et al. 2014, Deiner et al. 2017), one mismatch generated randomly can easily correspond to a distinct but closely related species. This poses the risk of false-positive detection, like in the present study, where several foreign species were detected (Supplementary material Appendix 1 Table A2). Yet, none of the false positive species detected with the Species pipeline were retained as a MOTU after the clustering process, highlighting the strength of our clustering approach to clean false positive errors when they likely arise from PCR and sequencing errors. When using a classical metabarcoding pipeline without a stringent cleaning or clustering step to infer diversity from short sequences, those false positive species might remain in the global pool of detected species which would require special care to flag and exclude such errors (i.e. manual alignment of sequences and verification of species geographical distribution). We also show the extent to which SWARM is able to assign the correct sequence as the representative of each MOTU, since 61 (out of 67) MOTUs perfectly match to a species from the reference database. Our results stress the importance to combine post-clustering filters based on PCR replicates and a cleaning algorithm to remove spurious MOTUs.

Since our MOTU-based pipeline slightly overestimates regional diversity with 67 MOTUs obtained compared to 63 species identified, a key question is how it can impact local diversity assessment. We found a slight tendency for MOTUs to underestimate species richness, with less than 2 MOTUs of difference compared to the number of species for most samples. The underestimation of diversity stemming from missed species (8 species so 13% of the regional pool) is not totally compensated by the overestimation caused by taxonomic redundancy in the regional pool. Further, no outliers were identified over all 196 samples. We also show that most of mentioned pitfalls do not impact patterns of dissimilarity at the community scale, as results are similar whether they are based on blind MOTUs or species identification. In summary, the assessment of local diversity is nearly not impacted by the absence of a complete reference database, both estimates are highly correlated (98%) with a mean difference of one species between pipelines.

While these results are valid using the teleo marker (12S rDNA, ~60 bp long), we could not validate our pipeline using other primer sets due to time and financial constraints. This pipeline can still be applied to other markers, but it would require a marker with a similar level of taxonomic specificity and limited intra-specific variation, to avoid an over-estimation of taxonomic diversity due to haplotype diversity. An application with another primer would require more investigation to test if threshold adjustments are necessary to match its specificities (i.e. PCR replicates number, minimum number of reads, LULU parameters, minimum distance in SWARM clustering). We suggest the design of a small pilot study in a well-known system to validate its blind predictive power before larger-scale applications.

We show that our approach using MOTUs delivers robust estimates of species diversity at the three geographic scales, unlocking new potential for biodiversity monitoring through eDNA. With more than 75% of fish families potentially detectable, our approach can go beyond the simple delineation of sequences within clusters when further assigning taxonomy to our MOTUs. In particular, the use of assignment algorithms such as the Lowest Common Ancestor (LCA) algorithm (Boyer et al. 2016, Gao et al. 2017) is well suited for taxonomic assignment in eDNA studies with incomplete reference database. We can then estimate the potential number of species per family when the sequence coverage within families is sufficient for such assessment. While a family assignment has limitations, ecological characteristics are generally well conserved for species within a given family (Brandl et al. 2018) and allow relevant metrics of ecological analyses to be computed at this scale. As the minimum coverage within family necessary for robust detection using LCA is likely to vary across taxa and goes beyond the scope of this study, a complete coverage is not requested and our approach can provide an accurate estimate of species diversity within family for ecological studies. Yet, we highlight some limitations when it comes to conservation policies for which unnamed MOTUs will not be satisfying. As conservation programs usually focus on few taxa which are mostly rare, threatened, invasive or emblematic (Pimm et al. 2018, Enquist et al. 2019, Hannah et al. 2020), achieving the complete sequencing of those target species is urgent but realistic in the near future, as opposed to the sequencing of most vertebrates. The current filling of global DNA databases is sufficient for our approach to work globally and across scales. Diversity indices derived from this method are shown to be reliable at α, β and γ scales to infer similar ecological conclusions as those based on classical species identification.

Conclusion

While it has widely been reported that molecular biodiversity inventories outperform classical inventories (videos, acoustic) in the open environment (Thomsen et al. 2016, Boussarie et al. 2018), we demonstrate that, in the absence of a complete genetic reference database, a bioinformatic pipeline using Molecular Operational Taxonomic Units is able to provide robust estimates of species diversity across spatial scales. Even if some species cannot being distinguished after the clustering step, a common issue due to genetic proximity between close taxa (Fahner et al. 2016), the geographic biodiversity patterns are highly similar to those obtained with a species-based method. As false negatives are inherent to any inventory method in ecology (Field et al. 2007) and while false positives are rarer but to avoid at all cost (Chambert et al. 2015), we suggest a precautionary approach where some 'true' observations could be lost in order to reduce the risk of false observations. Given the current state of genetic database coverage, a species-based eDNA approach is only achievable in freshwater ecosystems located in the Northern hemisphere, where the coverage exceeds 50% of fish species (Fig. 2). For all other ecosystems, our study is the proof of concept demonstrating that, given an appropriate primer set as well as filtering and cleaning processes, MOTUs can be used to accurately assess the level of biodiversity at all scales: local, turnover and regional. We thus advocate the need to focus sequencing efforts in priority towards 1) families with no genetic coverage so presently virtually undetectable with our approach and 2) conservation-important like invasive species or IUCN Red List species for which unassigned MOTUs cannot substitute. This work paves the way towards extending the use of eDNA in community ecology and biogeography even for poorly known ecosystems or lineages, and install eDNA as a standard monitoring tool (Jarman et al. 2018). It also reinforces its initial goal of versatility and high comparability to monitor any kind of ecosystem and compare communities across wide environmental gradients.

Data and code availability

The Species (<https://gitlab.mbb.univ-montp2.fr/edna/bash_105vsearch_ecotag>) and MOTU pipelines (<https://gitlab.mbb.univ-montp2.fr/edna/bash_swarm>) are freely accessible in Gitlab. All sequencing data is already available on Dryad: <https://doi.org/10.5061/dryad.t4n42rr> (Pont et al. 2019).

Acknowledgements – We thank SPYGEN and CNR team for contributing to the field work and/or the laboratory analysis, and Franck Pressiat from the CNR for valuable comments on the manuscript.
Funding – Funding for this work was provided by the 'Compagnie Nationale du Rhône' (CNR) and SPYGEN.
Author contributions – VM, TD, DM, MR and SM contributed to the study design, PEG and VM wrote the bioinformatic pipeline, TD and MR conducted the fieldwork, VM and AV performed the analysis, and all authors contributed towards writing, reviewing and editing the article.
Conflicts of interest – MR is a research engineers of a French electricity generation companies, AV and TD are research scientists in a private company, specialized on the use of eDNA for species detection.

References

Alberdi, A. et al. 2018. Scrutinizing key steps for reliable metabarcoding of environmental samples. – Methods Ecol. Evol. 9: 134–147.

Albouy, C. et al. 2019. The marine fish food web is globally connected. – Nat. Ecol. Evol. 3: 1153–1161.

Alexandrino, P. et al. 2006. Interspecific differentiation and intraspecific substructure in two closely related clupeids with extensive hybridization, *Alosa alosa* and *Alosa fallax*. – J. Fish Biol. 69: 242–259.

Andruszkiewicz, E. A. et al. 2017. Biomonitoring of marine vertebrates in Monterey Bay using eDNA metabarcoding. – PLoS One 12: e0176343.

Bakker, J. et al. 2017. Environmental DNA reveals tropical shark diversity in contrasting levels of anthropogenic impact. – Sci. Rep. 7: 1–11.

Baselga, A. and Orme, C. D. L. 2012. Betapart: an R package for the study of beta diversity. – Methods Ecol. Evol. 3: 808–812.

Berry, T. E. et al. 2019. Marine environmental DNA biomonitoring reveals seasonal patterns in biodiversity and identifies ecosystem responses to anomalous climatic events. – PLoS Genet. 15: e1007943.

Blowes, S. A. et al. 2019. The geography of biodiversity change in marine and terrestrial assemblages. – Science 366: 339–345.

Bohmann, K. et al. 2014. Environmental DNA for wildlife biology and biodiversity monitoring. – Trends Ecol. Evol. 29: 358–367.

Boussarie, G. et al. 2018. Environmental DNA illuminates the dark diversity of sharks. – Sci. Adv. 4: eaap9661.

Boyer, F. et al. 2016. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. – Mol. Ecol. Resour. 16: 176–182.

Brandl, S. J. et al. 2018. The hidden half: ecology and evolution of cryptobenthic fishes on coral reefs. – Biol. Rev. 93: 1846–1873.

Bylemans, J. et al. 2018. Toward an ecoregion scale evaluation of eDNA metabarcoding primers: a case study for the freshwater fish biodiversity of the Murray–Darling Basin (Australia). – Ecol. Evol. 8: 8697–8712.

Callahan, B. J. et al. 2016. DADA2: high resolution sample inference from Illumina amplicon data. – Nat. Methods 13: 581–583.

Callahan, B. J. et al. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. – ISME J. 11: 2639–2643.

Cantera, I. et al. 2019. Optimizing environmental DNA sampling effort for fish inventories in tropical streams and rivers. – Sci. Rep. 9: 1–11.

Chambert, T. et al. 2015. Modeling false positive detections in species occurrence data under different study designs. – Ecology 96: 332–339.

Cilleros, K. et al. 2019. Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): a test with Guianese freshwater fishes. – Mol. Ecol. Resour. 19: 27–46.

Civade, R. et al. 2016. Spatial representativeness of environmental DNA metabarcoding signal for fish biodiversity assessment in a natural freshwater system. – PLoS One 11: e0157366.

Closek, C. J. et al. 2019. Marine vertebrate biodiversity and distribution within the central California Current using environmental DNA (eDNA) metabarcoding and ecosystem surveys. – Front. Mar. Sci. 6: 732.

Collin, H. and Fumagalli, L. 2011. Evidence for morphological and adaptive genetic divergence between lake and stream habitats in European minnows (*Phoxinus phoxinus*, Cyprinidae). – Mol. Ecol. 20: 4490–4502.

Collins, R. A. et al. 2018. Persistence of environmental DNA in marine systems. – Commun. Biol. 1: 185.

Collins, R. A. et al. 2019. Non-specific amplification compromises environmental DNA metabarcoding with COI. – Methods Ecol. Evol. 10: 1985–2001.

Deagle, B. E. et al. 2014. DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. – Biol. Lett. 10: 20140562.

Deiner, K. et al. 2017. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. – Mol. Ecol. 26: 5872–5895.

Dejean, T. et al. 2011. Persistence of environmental DNA in freshwater ecosystems. – PLoS One 6: e23398.

DiBattista, J. D. et al. 2017. Assessing the utility of eDNA as a tool to survey reef-fish communities in the Red Sea. – Coral Reefs 36: 1245–1252.

Doherty, T. S. et al. 2016. Invasive predators and global biodiversity loss. – Proc. Natl Acad. Sci. USA 113: 11261–11265.

Dornelas, M. et al. 2014. Assemblage time series reveal biodiversity change but not systematic loss. – Science 344: 296–299.

Edgar, R. C. 2018. Updating the 97% identity threshold for 16S ribosomal RNA OTUs. – Bioinformatics 34: 2371–2375.

Edgar, R. C. and Flyvbjerg, H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. – Bioinformatics 31: 3476–3482.

Enquist, B. J. et al. 2019. The commonness of rarity: global and future distribution of rarity across land plants. – Sci. Adv. 5: 1–14.

Fahner, N. A. et al. 2016. Large-scale monitoring of plants through environmental DNA metabarcoding of soil: recovery, resolution and annotation of four DNA markers. – PLoS One 11: e0157505.

Field, S. A. et al. 2007. Improving precision and reducing bias in biological surveys: estimating false-negative error rates. – Ecol. Appl. 13: 1790–1801.

Finderup Nielsen, T. et al. 2019. More is less: net gain in species richness, but biotic homogenization over 140 years. – Ecol. Lett. 22: 1650–1657.

Flynn, J. M. et al. 2015. Toward accurate molecular identification of species in complex environmental samples: testing the performance of sequence filtering and clustering methods. – Ecol. Evol. 5: 2252–2266.

Froese, R. and Pauly, D. 2019. Fishbase. – <www.fishbase.org>.

Frøslev, G. T. et al. 2017. Algorithm for post-clustering curation of DNA amplicon data yields reliable biodiversity estimates. – Nat. Commun. 8: 1188.

Galetti, M. et al. 2014. Defaunation in the Anthropocene. – Science 345: 401–406.

Gao, X. et al. 2017. A Bayesian taxonomic classification method for 16S rRNA gene sequences with improved species-level accuracy. – BMC Bioinform. 18: 1–10.

Hannah, L. et al. 2020. 30% land conservation and climate action reduces tropical extinction risk by more than 50%. – Ecography 43: 943–953.

Harrison, J. B. et al. 2019. Predicting the fate of eDNA in the environment and implications for studying biodiversity. – Proc. R. Soc. B 286: 20191409.

Hortal, J. et al. 2015. Seven shortfalls that beset large-scale knowledge of biodiversity. – Annu. Rev. Ecol. Evol. Syst. 46: 523–549.

Hughes, T. P. et al. 2017. Coral reefs in the Anthropocene. – Nature 546: 82–90.

Huse, S. M. et al. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. – Environ. Microbiol. 12: 1889–1898.
Isbell, F. et al. 2019. Deficits of biodiversity and productivity linger a century after agricultural abandonment. – Nat. Ecol. Evol. 3: 1533–1538.
Jarman, S. N. et al. 2018. The value of environmental DNA biobanking for long-term biomonitoring. – Nat. Ecol. Evol. 2: 1192–1193.
Kottelat, M. and Persat, H. 2005. The genus *Gobio* in France, with redescription of *G. gobio* and description of two new species (Teleostei: Cyprinidae). – Cybium 29: 211–234.
Leinonen, R. et al. 2011. The European nucleotide archive. – Nucleic Acids Res. 39: 44–47.
Lladó Fernández, S. et al. 2019. The concept of operational taxonomic units revisited: genomes of bacteria that are regarded as closely related are often highly dissimilar. – Folia Microbiol. 64: 19–23.
Magurran, A. E. et al. 2018. Divergent biodiversity change within ecosystems. – Proc. Natl Acad. Sci. USA 115: 1843–1847.
Mahé, F. et al. 2014. Swarm: robust and fast clustering method for amplicon-based studies. – PeerJ 2: e593.
Mahé, F. et al. 2015. Swarm v2: highly-scalable and high-resolution amplicon clustering. – PeerJ 3: e1420.
Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. – EMBnet.journal 17: 10.
McCauley, D. J. et al. 2015. Marine defaunation: animal loss in the global ocean. – Science 347: 247–254.
McGill, B. J. et al. 2015. Fifteen forms of biodiversity trend in the anthropocene. – Trends Ecol. Evol. 30: 104–113.
Menegotto, A. and Rangel, T. F. 2018. Mapping knowledge gaps in marine diversity reveals a latitudinal gradient of missing species richness. – Nat. Commun. 9: 4713.
Miya, M. et al. 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. – R. Soc. Open Sci. 2: 150088.
Mora, C. et al. 2008. The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. – Proc. R. Soc. B 275: 149–155.
Mora, C. et al. 2011a. How many species are there on earth and in the ocean? – PLoS Biol. 9: e1001127.
Mora, C. et al. 2011b. Global human footprint on the linkage between biodiversity and ecosystem functioning in reef fishes. – PLoS Biol. 9: e1000606.
Nguyen, N.-P. et al. 2015. A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. – Biofilms Microbiomes 1: 10–13.
Pawlowski, J. et al. 2018. The future of biotic indices in the ecogenomic era: integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. – Sci. Total Environ. 637–638: 1295–1310.
Pedrós-Alió, C. 2006. Marine microbial diversity: can it be determined? – Trends Microbiol. 14: 257–263.
Pimm, S. L. et al. 2018. How to protect half of earth to ensure it protects sufficient biodiversity. – Sci. Adv. 4: 1–9.
Pont, D. et al. 2018. Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. – Sci. Rep. 8: 1–13.
Pont, D. et al. 2019. Data from: Environmental DNA reveals quantitative patterns of fish biodiversity in large rivers despite its downstream transportation. – Dryad, Dataset, <https://doi.org/10.5061/dryad.t4n42rr>.
Reeder, J. and Knight, R. 2009. The 'rare biosphere': a reality check. – Nat. Methods 6: 636–637.
Rognes, T. et al. 2016. VSEARCH: a versatile open source tool for metagenomics. – PeerJ 4: e2584.
Ruppert, K. M. et al. 2019. Past, present and future perspectives of environmental DNA (eDNA) metabarcoding: a systematic review in methods, monitoring and applications of global eDNA. – Global Ecol. Conserv. 17: e00547.
Sales, N. G. et al. 2019. Influence of preservation methods, sample medium and sampling time on eDNA recovery in a neotropical river. – Environ. DNA 1: 119–130.
Sales, N. G. et al. 2020. Fishing for mammals: landscape-level monitoring of terrestrial and semi-aquatic communities using eDNA from riverine systems. – J. Appl. Ecol. 57 707–716.
Sayers, E. W. et al. 2019. GenBank. – Nucleic Acids Res. 47: D94–D99.
Schmidt, P. A. et al. 2013. Illumina metabarcoding of a soil fungal community. – Soil Biol. Biochem. 65: 128–132.
Schnell, I. B. et al. 2015. Tag jumps illuminated – reducing sequence-to-sample misidentifications in metabarcoding studies. – Mol. Ecol. Resour. 15: 1289–1303.
Siegenthaler, A. et al. 2019. Metabarcoding of shrimp stomach content: harnessing a natural sampler for fish biodiversity monitoring. – Mol. Ecol. Resour. 19: 206–220.
Siegwald, L. et al. 2017. Assessment of common and emerging bioinformatics pipelines for targeted metagenomics. – PLoS One 12: e0169563.
Spalding, M. D. et al. 2007. Marine ecoregions of the world: a bioregionalization of coastal and shelf areas. – Bioscience 57: 573.
Stackebrandt, E. and Goebel, B. M. 2008. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. – Int. J. Syst. Evol. Microbiol. 44: 846–849.
Tedesco, P. A. et al. 2017. Data Descriptor: a global database on freshwater fish species occurrence in drainage basins. – Sci. Data 4: 1–6.
Thomsen, P. F. et al. 2016. Environmental DNA from seawater samples correlate with trawl catches of subarctic, deepwater fishes. – PLoS One 11: e0165252.
Tsuji, S. et al. 2019. The detection of aquatic macroorganisms using environmental DNA analysis – a review of methods for collection, extraction and detection. – Environ. DNA 1: 99–108.
Valentini, A. et al. 2016. Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. – Mol. Ecol. 25: 929–942.
Vellend, M. et al. 2013. Global meta-analysis reveals no net change in local-scale plant biodiversity over time. – Proc. Natl Acad. Sci. USA 110: 19456–19459.
Weigand, H. et al. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: gap-analysis and recommendations for future work. – Sci. Total Environ. 678: 499–524.
West, K. M. et al. 2020. eDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical island ecosystem. – Mol. Ecol. 29: 1069–1086.
Whittaker, R. H. 1972. Evolution and measurement of species diversity. – Taxon 21: 213–251.
Zimmermann, J. et al. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. – Mol. Ecol. Resour. 15: 526–542.

Supplementary material (available online as Appendix ecog-05049 at <www.ecography.org/appendix/ecog-05049>). Appendix 1.