








ORIGINAL ARTICLE

General principles for assignments of communities from eDNA: Open versus closed taxonomic databases

Rosetta C. Blackman^{1,2}  | Jean-Claude Walser³  | Lukas Rüber^{4,5}  |
Jeanine Brantschen^{1,2}  | Soraya Villalba^{4,6} | Jakob Brodersen^{5,6}  | Ole Seehausen^{5,6}  |
Florian Altermatt^{1,2} 

¹Department of Aquatic Ecology, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

²Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zürich, Switzerland

³Genetic Diversity Centre (GDC), Department of Environmental Systems Science (D-USYS), Federal Institute of Technology (ETH), Zürich, Switzerland

⁴Naturhistorisches Museum Bern, Bern, Switzerland

⁵Division of Aquatic Ecology & Evolution, Institute of Ecology & Evolution, University of Bern, Bern, Switzerland

⁶Department of Fish Ecology & Evolution, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Kastanienbaum, Switzerland

Correspondence

Rosetta C. Blackman and Florian Altermatt, Department of Aquatic Ecology, Eawag, Swiss Federal Institute of Aquatic Science and Technology, Überlandstrasse 133, Dübendorf CH-8600, Switzerland.
Email: rosieblackman@gmail.com and florian.altermatt@eawag.ch

Funding information

Bundesamt für Umwelt, Grant/Award Number: 00.5058.PZ/6B1725F08; Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 31003A_173074; University of Zurich Research Priority Programme in Global Change and Biodiversity

Abstract

Metabarcoding of environmental DNA (eDNA) is a powerful tool for describing biodiversity, such as finding keystone species or detecting invasive species in environmental samples. Continuous improvements in the method and the advances in sequencing platforms over the last decade have meant this approach is now widely used in biodiversity sciences and biomonitoring. For its general use, the method hinges on a correct identification of taxa. However, past studies have shown how this crucially depends on important decisions during sampling, sample processing, and subsequent handling of sequencing data. With no clear consensus as to the best practice, particularly the latter has led to varied bioinformatic approaches and recommendations for data preparation and taxonomic identification. In this study, using a large freshwater fish eDNA sequence dataset, we compared the frequently used zero-radius Operational Taxonomic Unit (zOTU) approach of our raw reads and assigned it taxonomically (i) in combination with publicly available reference sequences (open databases) or (ii) with an OSU (Operational Sequence Units) database approach, using a curated database of reference sequences generated from specimen barcoding (closed database). We show both approaches gave comparable results for common species. However, the commonalities between the approaches decreased with read abundance and were thus less reliable and not comparable for rare species. The success of the zOTU approach depended on the suitability, rather than the size, of a reference database. Contrastingly, the OSU approach used reliable DNA sequences and thus often enabled species-level identifications, yet this resolution decreased with the recent phylogenetic age of the species. We show the need to include target group coverage, outgroups and full taxonomic annotation in reference databases to avoid misleading annotations that can occur when using short amplicon sizes as commonly used in eDNA metabarcoding studies. Finally, we make general suggestions to improve the construction and use of reference databases for metabarcoding studies in the future.

Rosetta C. Blackman and Jean-Claude Walser are shared first-authorship.

Ole Seehausen and Florian Altermatt are shared last-authorship.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Environmental DNA* published by John Wiley & Sons Ltd.

1 | INTRODUCTION

Ecological surveying and biodiversity assessment have been a primary goal for ecologists and conservationists aiming to describe and interpret the state and change of biodiversity in an ecosystem. Traditional freshwater monitoring methods developed for target taxa (e.g., fish, macroinvertebrates, macrophytes, and algae) are costly, both in terms of time and financially. However, in the last 10 years, there have been significant developments in biodiversity assessment with molecular tools. This revolution in molecular biodiversity assessment has been termed a “game changer” (Lawson Handley, 2015), where High-Throughput Sequencing (HTS) platforms are used to generate millions of DNA sequences from bulk tissue or environmental samples in a process known as metabarcoding.

Environmental DNA (eDNA) metabarcoding is now commonly used for biodiversity assessment in aquatic environments (e.g., Blackman et al., 2022; Brown et al., 2016; Deiner et al., 2016; Hänfling et al., 2016; Mächler et al., 2019; Pawlowski et al., 2018). DNA extracted from a water sample (Keck et al., 2022; Pawlowski et al., 2020; Taberlet et al., 2012) is amplified using PCR with relatively universal primers. This allows whole taxonomic groups to be targeted and sequenced simultaneously, without the need for physical collection of specimens. Large datasets of biodiversity information for a targeted taxonomic group from water samples are generated which can uncover biodiversity structure of whole lake or riverine systems (e.g., Altermatt et al., 2020; Hänfling et al., 2016). These data, in the form of sequence reads, are processed, quality checked, filtered, merged, error corrected, chimera cleaned and the resulting zero-radius Operational Taxonomic Units (zOTUs) or clustered Operational Taxonomic Units (OTUs) are then assigned to taxonomic names based on reference sequences (Antich et al., 2021; Brandt et al., 2021; Deiner et al., 2017; Mathon et al., 2021). Although there are several instances where taxonomy-free approaches may suffice (see Mächler et al., 2021), it is particularly important to assign a species, genus or family name for conservation purposes, such as the identification of invasive or endangered species. Ultimately, this linkage of sequence to taxonomic name allows researchers and practitioners to gain unprecedented information into biodiversity richness (Leray et al., 2022) and the ability to monitor ecosystems with a new perspective and minimal impact. However, as with any new method or technological advancement, the approaches that individual studies have used hitherto vary significantly. This is true for bioinformatic processing steps, with no clear consensus as to the best approach (e.g., Antich et al., 2021; Brandt et al., 2021; Mathon et al., 2021), although recent comparisons suggest some programs offer advantages over others in terms of processing time (see Mathon et al., 2021). Recently, several pipelines have been developed specifically to handle eDNA data and tackle the issue of processing raw reads to taxonomic assignment in a consistent way, which go some way in making data processing standardized (e.g., *Anacapa* (Curd et al., 2019); *SLIM* (Dufresne et al., 2019); *PEMA* (Zafeiropoulos et al., 2020); or *eDNAFlow* (Mousavi-Derazmahalleh et al., 2021)).

Although standardized data processing may be desirable, it also has its drawbacks. Processing must be adapted to the data structure and data losses during data processing must be explainable (see also Keck & Altermatt, 2022). Especially, in eDNA projects, removal or loss of data can also mean a loss of information. Furthermore, classic approaches to processing sequence reads, such as a 97% similarity cluster, may underestimate diversity, while the zOTU approach often overestimates diversity (Brandt et al., 2021). Whichever approach is used, they all ultimately rely on the accuracy and completeness of a reference database for subsequent taxonomic assignment (Dugal et al., 2022; Jackman et al., 2021). Reference databases therefore should be a major consideration of any metabarcoding study as they underpin the successful and accurate taxonomic assignments (de Santana et al., 2021; Rodríguez-Ezpeleta et al., 2021). Ideally, these databases consist of verified barcodes, including taxa known to occur in the study area (Taberlet et al., 2007). Unfortunately, the perfect reference database does not exist (yet), and databases are scarce for large organismal groups. Often, studies are forced to use incomplete reference databases, which sometimes have inaccurate or mis-labelled sequences (Somervuo et al., 2017) that can ultimately prohibit detection or assignment of species level identification to a cluster or zOTUs (Blackman et al., 2021; Li et al., 2022; Weigand et al., 2019).

In most studies, authors build or select databases they see as particularly suited to their eDNA data, and the choice of database often has a strong pragmatic component as well, especially with respect to the use of openly accessible databases. Commonly, databases are selected based on their size (i.e., number of taxa or sequences included), implying that a larger number of sequences means a higher chance of more or better taxonomic assignment. Consequently, metabarcoding studies source their reference sequences from publicly available sequence repositories, such as GenBank or BOLD. However, these publicly available reference databases (primary or secondary) are not always the best solution (Pentinsaari et al., 2020). Sequences submitted to public (open access) sequence databases, such as GenBank, do not require a validation step after submission (Kozlov et al., 2016), such that taxonomically incorrect or incomplete entries occur. Also, they are biased toward well-studied taxa (Meiklejohn et al., 2019; Porter & Hajibabaei, 2018; Schroeter et al., 2020), which means their suitability may not always be appropriate to detect rare or understudied taxa. It is also not advisable to merge sequences from several databases because duplications and deviating nomenclatures could counteract the increase in sequence diversity. Perhaps, the most widely used database of open access sequences is GenBank (Benson et al., 2013). Until recently, it was widely thought to have a high proportion of errors, yet analysis suggests that—although discrepancies are present—when examining taxonomic assignment above species level these errors are less than previously thought (Leray et al., 2019; Locatelli et al., 2020), yet geographic biases still exist (Li et al., 2022). These findings are encouraging; but similarly other studies have conversely highlighted issues when examining species level annotation, such as Conte-Grande and colleagues (2017) who found 16.3% of all snakehead (Channidae)

sequences (COI) on GenBank were in fact incorrect. Stringent curation steps are therefore vital when using sequences from large databases to ensure against incorrectly identified or poor-quality sequences which will lead to incorrect assignments, particularly if the aim is species level identification (Locatelli et al., 2020).

A solution to the above shortcoming is using a target or group specific reference databases where sequences have been verified. These reference databases, however, may also have varying degrees of coverage in terms of species' representation. Weigand et al. (2019) carried out a gap-analysis of the sequences available for European bioindicator groups (fish, macroinvertebrates, diatoms and plants). Of all groups, fish are generally best represented in BOLD/Genbank, and MitoFish, respectively, yet with both substantial geographic and taxonomic gaps remaining (Froese & Pauly, 2021; Iwasaki et al., 2013; Marques et al., 2021; Polanco et al., 2021). It is therefore not surprising that fish eDNA studies often include additional specimens collected within their study area to increase their reference database and fully exploit the benefits of metabarcoding (Cilleros et al., 2019; Schenekar et al., 2020). This was well shown by Schenekar et al. (2020), who's reanalysis of their metabarcoding data with a larger study specific reference database not only revealed further biodiversity but also changed some taxonomic assignment from one species to another (Schenekar et al., 2020). The change in taxonomic assignment is of particular importance as it also demonstrates that using limited sequence diversity and/or incomplete species-level taxonomic investigation, below a certain dissimilarity threshold, false positives can be generated due to "forcing" assignments by sequence similarities during taxonomic assignment processing. Similarly, relatively closely related species (i.e., in the first one to two million years after speciation) commonly found in diverse systems can usually not be distinguished by single stretches of sequence variation (e.g., Jackman et al., 2021), such as short barcodes. If not fully represented in the reference databases, groups of closely related species in the data will erroneously be assigned to a single species, potentially leading to wrong estimates of species diversity.

To explore the effect of reference databases on taxonomic assignment we used (i) publicly available (open) databases and (ii) a curated study specific (closed) reference database to compare the taxonomic assignment of freshwater fish from eDNA samples taken across Switzerland. The former are generally publicly accessible, and thus widely used, while the latter is a "close to perfect" version acting as a gold-standard, also to evaluate the formers' potential or limits. We use two publicly accessible reference databases as examples of easily accessible open references, namely MIDORI (Leray et al., 2018) and MitoFish (Iwasaki et al., 2013). We then supplemented the later with important consensus sequences, which we will refer to as MitoFish+. The closed or custom-built reference database used in this study is made up of sequences from tissue samples collected from two very broad quantitative fish biodiversity surveys (Alexander & Seehausen, 2021) across Switzerland. These surveys contain extensive collections and sequences of multiple specimens from nearly all fish species in Switzerland and included

multiple geographical populations of each. As such, it represents an example of what would be considered by many as the ideal reference database for taxonomic assignment.

2 | METHODS

2.1 | Sample collection

Environmental DNA samples were collected from 92 river sites in 2019 as part of a nationwide fish monitoring campaign in Switzerland (see Brantschen et al., 2021, Figure S1 and Table S1 for further details). For each of the sites, a total of 2 L of water was filtered using 4 Sterivex filters with a 0.22 µm pore size (Merck Millipore, Merck KgaA, Darmstadt, Germany). Filters were sealed with luer fitting and placed in a cool box for transporting to the laboratory where samples were stored at -20°C until further processing. Field negative controls, consisting of 2 L of ddH₂O, were filtered and stored in the same manner as the samples.

2.2 | eDNA extraction and library preparation

DNA extractions from filters were performed in a clean room environment at Eawag, Switzerland (Deiner et al., 2015). The DNA was extracted using the Qiagen PowerWater Sterivex Extraction Kit (Qiagen, Hilden, Germany). Filters from different sites were extracted in random batches including field and filter control that were treated equally to the samples. Extractions were performed as described by the manufacturer protocol. DNA was eluted into 100 µl of elution buffer and stored until further processing at -20°C. A two-step library preparation method was used targeting the hypervariable region of 12S rRNA gene which amplicon ranges from 163 to 185 bp using MiFish-U-F/R primer pair (forward primer sequence: 5'-GTCGGTAAACTCGTGCCAGC-3' and reverse primer sequence: 5'-CATAGTGGGGTATCTAATCCCAGTTTG-3') (Miya et al., 2015) hereafter known as the MiFish primer pair. These primers were modified to include the Nextera transposase sequences. Negative controls (field and PCRs), positive controls (PCP, containing a tissue DNA extract from Atlantic Cod, *Gadus morhua*, see Table S3), and samples were randomized over each 96-well PCR plates.

The first PCR reaction contained 0.5 µM each primer, 0.4 mg/ml BSA, 12.5 µl Q5 High Fidelity 2x Master Mix (New England Biolabs), and 2 µl template DNA. The PCR profiles were as follows: initial denaturation of 98°C for 5 min, followed by 35 cycles of 98°C for 10s, 65°C for 20s, and 72°C for 30s, and a final extension step of 72°C for 7 min. The first PCR was carried out in triplicate and samples were pooled and cleaned using SPRI beads (Applied Biological Materials Inc.) prior to the second PCR. Second PCRs were carried out using 15 µl of cleaned PCR product and the Nextera XT Index Kit v2 (Illumina), following the subsequent PCR profile: initial denaturation 95°C for 3 min followed by 10 cycles of 95°C for 30s, 58°C

for 30s, and 72°C for 30s, and a final extension step of 72°C for 5min. Negative controls and the addition of positive controls were processed in parallel with all samples. All PCR products were visualised using QiAxcel Advanced System by using a High-Resolution Cartridge (Qiagen, Hilden, Germany) and cleaned once more using SPRI beads. Samples were then quantified using the Spark 10M Multimode Microplate Reader (Tecan Group Ltd.) using the Qubit dsDNA BR assay (Thermo Fisher) and pooled equimolar. The libraries were loaded at 17.6pM concentration, with 10% PhiX control. A paired-end 600 cycle (2×300nt) sequencing was performed on an Illumina MiSeq (MiSeq Reagent Kit v3) following the manufacturer's run protocols (Illumina).

2.3 | Bioinformatics

To process the raw data, we used a standardized but parameter-optimized workflow (Figure 1). In short, the raw reads are first filtered to remove PhiX related (an internal standard), and low complexity reads. In a next step, the low-quality 3'-end are trimmed to improve read merging. We used an in-silico PCR approach to remove the primer site from the merged reads (amplicons). Subsequently, the amplicons were subjected to a quality and a size-range filter. The cleaned amplicons were de-replicated prior to clustering with Usearch::UNOISE (Edgar, 2016a). UNOISE3 includes error correction, zero-radius clustering and chimera removal (Figure S2). We used an abundance threshold of 10 to remove artificially created and therefore untrustworthy singletons and rare zOTUs.

2.4 | Taxonomic associations of zOTUs

We used two different methods to generate count tables with taxonomic associations. The first method is the classic approach of using a sequence reference to annotate the zOTUs. For this method, we used Usearch::SINTAX (Edgar, 2016b). SINTAX is a fast and reliable tool in combination with the right reference database and appropriate confidence cut-off. We used two publicly available 12S ribosomal RNA databases (i.e., MitoFish and MIDORI). MitoFish is a fish-specific mitochondrial genome reference database, while MIDORI has a broader species range including Eukaryota. Both references databases are well maintained sequence collections and are extremely helpful for taxonomic annotation. Nevertheless, both also have limits in terms of species diversity, accuracy and completeness, depending on the research question. We wanted to increase the annotation range of the MitoFish reference database by adding more nonfish-related sequences. The primary purpose of the reference extension was to better understand the high numbers of zOTUs with missing taxonomic assignments. For this reason, we blasted (blastn) all the badly annotated zOTUs against the NCBI nucleotide database (Altschul et al., 1990). The Blast hits were bit-score (>200) and identity (>80%) filtered and only fully annotated sequences were used.

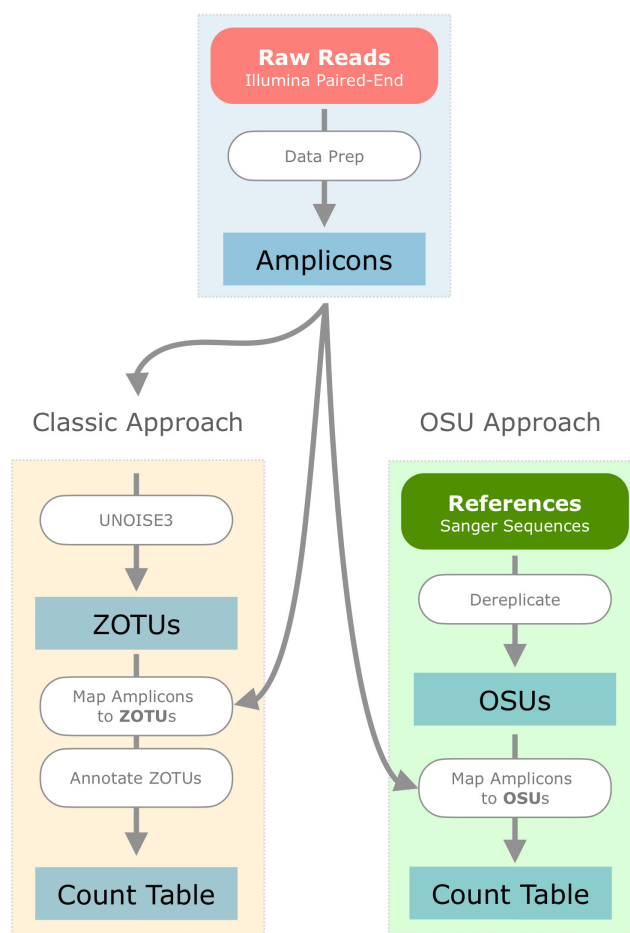


FIGURE 1 Data preparation workflow. The flow diagram shows the bioinformatic steps from the classic approach (raw read output from the MiSeq) and OSU approach (reference sequences) to the taxonomic assignment.

By expanding the diversity of the existing MitoFish reference, we could characterize nontarget (i.e., co-amplified) species.

2.5 | Reference databases summary

MitoFish (v362) (<http://mitofish.aori.u-tokyo.ac.jp>, Iwasaki et al., 2013) is a fish-specific mitochondrial genome database with complete and partial mtDNA sequences.

MitoFish+ is an extended version of the MitoFish database, containing additional 372 nonfish-related sequences from NCBI, including bacteria.

MIDORI 12S (GB240) (<http://reference-midori.info>, Machida et al., 2017) is a reference database of DNA sequences, which can be used for taxonomic assignments of all Eukaryota mitochondrial DNA sequences (Leray et al., 2018; Machida et al., 2017). The reference database contains 7590 fish species (Actinopteri = 7355, Cladistia = 44 and Chondrichthyes = 191) and 7797 nonfish species (Amphibia = 1981, Avea = 690, Dipnoi = 3, Hyperoartia = 14,

Insecta = 3, Lepidosauria = 2911, Mammalia = 2112, and Unknown Class = 83).

2.6 | Operational sequence unit (OSU) approach

As an alternative to the classic approach from above, we explored the use of a project-specific custom build 12S fish reference database. To do this, we barcoded tissue samples of fish species collected within the same geographical range of our surveys (See *Closed reference database* below for more details). The full sequenced fragment was 736 bp long and then trimmed to the MiFish fragment of 192 bp long. We de-replicate the reference sequences ($n = 912$) to get unique sequences ($n = 107$, 11.9%), which we call Operational Sequence Units (OSUs) (Figure S3). Sequences generated for the closed reference database have been deposited in BOLD and GenBank (See Data availability statement). We supplemented these sequences with 12S reference for the following sequences: a consensus sequence for *Gadus chalcogrammus* and *Gadus morhua* from the MitoFish reference, which is used in as our positive control and not found within the geographic range of our surveys. Then, we included, based on the NCBI Blast results of zOTUs, nonfish sequences to act as outgroups, namely: *Homo sapiens*, *Sus scrofa*, *Sturnus vulgaris*, *Rupicapra rupicapra*, *Ichthyosaura alpestris*, *Bos taurus* and *Turdus merula*. We therefore have a total of 116 OSUs to map the reads to (Table 1). For each OSU label, there are three elements to demonstrate the support for that taxonomic assignment derived from the reference database (see Table 1). First, the taxonomic assignment is derived from the consensus of the sequences which have clustered to form that OSU. If all sequences are from the same species, the species name is used for the taxonomic assignment, if the sequences differ, in that there are several different species from the same genus, the genus name is used for taxonomic assignment. Second, the remaining elements of the name are derived from the number of repeating clusters with that name within the dataset, for example: 1 of 2 and 2 of 2 would indicate that there are two OSUs with the same taxonomic assignment in the dataset. Thirdly, the number of sequences used to form the cluster for example: n8 would include 8 sequences. Once the OSUs were formed, reads were mapped to the OSUs using Usearch:: cluster_fast with two different parameters: sequence identity (ID) and query coverage (QC). To test mapping performance, different parameters were tested to demonstrate the variation in reads mapped caused by these parameters (Table 2).

2.7 | Closed reference database

Recently, two large-scale projects were carried out to quantitatively survey the fish biodiversity in the lakes of Switzerland and adjacent perialpine regions (Projet Lac, 2010–2018; Alexander & Seehausen, 2021), and in the rivers of Switzerland (Progetto Fiumi, 2013–2018). More than 150,000 fish individuals were sampled and identified. Population genetic studies were performed for many

complexes of closely related species to assess species boundaries (e.g., for the genera *Coregonus*, *Salvelinus*, *Cottus* and *Phoxinus*). From both projects, specimen and tissue reference collections were established for over 18,000 specimens at the Natural History Museum of Bern (NMBE). The closed reference database used in this study is a collection of 12S rRNA sequences of fish species from 912 of these specimens, representing all fish species, and multiple populations of most, collected in the Projet Lac and Progetto Fiumi. The collection represents 110 of the 124 species known to occur in Switzerland (Table S4). Total genomic DNA was extracted from muscle tissue or fin clips preserved in 100% ethanol and stored at -80°C using the DNeasy Blood and Tissue Kit on a QIAcube robotic workstation following manufacturer's instructions (Qiagen, Hilden, Germany). Partial 12S rRNA fragments were amplified using the primers MiFish_U_F (Miya et al., 2015) and Valentini_tele01_H1913 (Valentini et al., 2016). PCR protocol and PCR conditions follow Conte-Grand et al. (2017). PCR products were cleaned, and Sanger sequenced using both PCR primers by LGC Genomics, Berlin. Raw reads were edited and assembled into contigs using Geneious Prime v2022.0.2 (<https://www.geneious.com>) and consensus sequences were aligned using MAFFT v7.017 (Katoh & Standley, 2013), as implemented in Geneious Prime. The alignment was trimmed to correspond to the MiFish fragment and unique haplotypes were extracted. RAXML (Stamatakis, 2014) was used to reconstruct a Maximum Likelihood (ML) tree.

3 | RESULTS

The MiSeq run from the eDNA samples generated 14.27 million reads passing the default quality filter with Q30 of 79.13%. We used the MiSeq Control Software (MCS v3.1) to de-multiplex the data, which resulted in 10,966,567 paired-end raw reads. The median number of read per sample was 20,450 (range 8 to 93,441). After data processing, we had 8,889,951 (81%) high-quality amplicons for the downstream analysis (Figure 1). The percentage of high-quality amplicons that could be mapped back to the zOTUs was high, 98.5% (Figure S2). Overall, the abundance of reads assigned to each zOTUs and OSUs was relatively similar, with only minor variation (Figure S3). An overview of the quality control and information on data lost through filtering in the bioinformatic steps can be found in the DRYAD repository (See data availability statement).

3.1 | Comparison between the three reference databases

The number of zOTUs with annotation to species, genus or family level was similar for the three reference databases (Figure 2). The most common zOTUs were the same with each reference database, both in terms of identity and level of taxonomic resolution (i.e., *Cottus*, *Salmo*, *Squalis cephalus*, other zOTUs assigned to Cyprinidae). All zOTUs were annotated, though some of these annotations are

TABLE 1 Closed reference database fish OSUs

OSU name	Genus or species level	Number of haplotypes within OSU	No. of supporting sequences	Present in dataset	Description
Abramis_brama_1of1_n11	<i>Abramis brama</i>	1	11	Yes	Fish
Alburnoides_bipunctatus_1of2_n9	<i>Alburnoides bipunctatus</i>	2	9	Yes	Fish
Alburnoides_bipunctatus_2of2_n1	<i>Alburnoides bipunctatus</i>	2	1	Yes	Fish
Alburnus_alburnus_1of1_n11	<i>Alburnus alburnus</i>	1	11	Yes	Fish
Alburnus_arborella_1of2_n9	<i>Alburnus arborella</i>	2	9	Yes	Fish
Alburnus_arborella_2of2_n1	<i>Alburnus arborella</i>	2	1	No	Fish
Alosa_sp_1of1_n8	<i>Alosa</i> sp.	1	8	No	Fish
Ameiurus_melas_1of1_n5	<i>Ameiurus melas</i>	1	5	No	Fish
Anguilla_anguilla_1of1_n21	<i>Anguilla anguilla</i>	1	21	Yes	Fish
Barbatula_sp_lineage_I_1of5_n18	<i>Barbatula</i> sp. lineage I	5	18	Yes	Fish
Barbatula_sp_lineage_I_2of5_n1	<i>Barbatula</i> sp. lineage I	5	1	Yes	Fish
Barbatula_sp_lineage_I_3of5_n2	<i>Barbatula</i> sp. lineage I	5	2	Yes	Fish
Barbatula_sp_lineage_I_4of5_n1	<i>Barbatula</i> sp. lineage I	5	1	Yes	Fish
Barbatula_sp_lineage_I_5of5_n2	<i>Barbatula</i> sp. lineage I	5	2	Yes	Fish
Barbatula_sp_lineage_II_1of1_n18	<i>Barbatula</i> sp. lineage II	1	18	Yes	Fish
Barbatula_quignardi_1of1_n7	<i>Barbatula quignardi</i>	1	7	Yes	Fish
Barbus_barbus_1of3_n10	<i>Barbus barbus</i>	3	10	Yes	Fish
Barbus_barbus_2of3_n1	<i>Barbus barbus</i>	3	1	Yes	Fish
Barbus_barbus_3of3_n1	<i>Barbus barbus</i>	3	1	Yes	Fish
Barbus_caninus_1of1_n1	<i>Barbus caninus</i>	1	1	Yes	Fish
Barbus_plebejus_1of2_n6	<i>Barbus plebejus</i>	2	6	Yes	Fish
Barbus_plebejus_2of2_n2	<i>Barbus plebejus</i>	2	2	Yes	Fish
Blicca_bjoerkna_1of1_n8	<i>Blicca bjoerkna</i>	1	8	Yes	Fish
Bos_taurus	<i>Bos taurus</i>	NA	NA	Yes	Cow
Bos_taurus	<i>Bos taurus</i>	NA	NA	No	Cow
Carassius_gibelio_1of1_n7	<i>Carassius gibelio</i>	1	7	Yes	Fish
Chondrostoma_nasus_1of2_n2	<i>Chondrostoma nasus</i>	2	2	Yes	Fish
Chondrostoma_nasus_2of2_n4	<i>Chondrostoma nasus</i>	2	4	Yes	Fish
Chondrostoma_soetta_1of1_n2	<i>Chondrostoma soetta</i>	1	2	Yes	Fish
Cobitis_bilineata_1of1_n18	<i>Cobitis bilineata</i>	1	18	No	Fish
Coregonus_heglingus_1of1_n1	<i>Coregonus heglingus</i>	1	1	Yes	Fish
Coregonus_sp_1of3_n70	<i>Coregonus</i> sp.	3	70	Yes	Fish
Coregonus_sp_2of3_n1	<i>Coregonus</i> sp.	3	1	Yes	Fish
Coregonus_sp_3of3_n1	<i>Coregonus</i> sp.	3	1	Yes	Fish
Cottus_1of3_n37	<i>Cottus</i> sp.	3	37	Yes	Fish
Cottus_2of3_n22	<i>Cottus</i> sp.	3	22	Yes	Fish
Cottus_3of3_n4	<i>Cottus</i> sp.	3	4	Yes	Fish
Cyprinus_carpio_1of1_n11	<i>Cyprinus carpio</i>	1	11	Yes	Fish
Esox_cisalpinus_1of2_n7	<i>Esox cisalpinus</i>	2	7	Yes	Fish
Esox_cisalpinus_2of2_n1	<i>Esox cisalpinus</i>	2	1	Yes	Fish
Esox_lucius_1of2_n19	<i>Esox lucius</i>	2	19	Yes	Fish
Esox_lucius_2of2_n1	<i>Esox lucius</i>	2	1	Yes	Fish
Gadus_chalcogrammus_Consensus	<i>Gadus chalcogrammus</i>	NA	NA	Yes	Fish

- positive

TABLE 1 (Continued)

OSU name	Genus or species level	Number of haplotypes within OSU	No. of supporting sequences	Present in dataset	Description
Gadus_chalcogrammus_Consensus	<i>Gadus chalcogrammus</i>	NA	NA	No	Fish - positive
Gadus_morhua_Consensus	<i>Gadus morhua</i>	NA	NA	Yes	Fish - positive
Gadus_morhua_Consensus	<i>Gadus morhua</i>	NA	NA	No	Fish - positive
Gasterosteus_sp_1of1_n12	<i>Gasterosteus</i> sp.	1	12	Yes	Fish
Gobio_gobio_1of1_n14	<i>Gobio gobio</i>	1	14	Yes	Fish
Gobio_obtusirostris_1of2_n5	<i>Gobio obtusirostris</i>	2	5	Yes	Fish
Gobio_obtusirostris_2of2_n1	<i>Gobio obtusirostris</i>	2	1	Yes	Fish
Gymnocephalus_cernua_1of2_n11	<i>Gymnocephalus cernua</i>	2	11	Yes	Fish
Gymnocephalus_cernua_2of2_n10	<i>Gymnocephalus cernua</i>	2	10	Yes	Fish
Homo_sapiens_Consensus	<i>Homo sapiens</i>	NA	NA	Yes	Human
Homo_sapiens_Consensus	<i>Homo sapiens</i>	NA	NA	No	Human
Ichthyosaura_alpestris	<i>Ichthyosaura alpestris</i>	NA	NA	Yes	Newt
Ichthyosaura_alpestris	<i>Ichthyosaura alpestris</i>	NA	NA	No	Newt
Lampetra_planeri_1of1_n4	<i>Lampetra planeri</i>	1	4	No	Fish
Lepomis_gibbosus_1of1_n10	<i>Lepomis gibbosus</i>	1	10	Yes	Fish
Leucaspius_delineatus_1of1_n2	<i>Leucaspius delineatus</i>	1	2	Yes	Fish
Leuciscus_leuciscus_1of3_n16	<i>Leuciscus leuciscus</i>	3	16	Yes	Fish
Leuciscus_leuciscus_2of3_n3	<i>Leuciscus leuciscus</i>	3	3	Yes	Fish
Leuciscus_leuciscus_3of3_n1	<i>Leuciscus leuciscus</i>	3	1	Yes	Fish
Lota_lota_1of1_n16	<i>Lota lota</i>	1	16	Yes	Fish
Micropterus_salmoides_1of1_n5	<i>Micropterus salmoides</i>	1	5	No	Fish
Neogobius_kessleri_1of1_n2	<i>Neogobius kessleri</i>	1	2	No	Fish
Neogobius_melanostomus_1of1_n3	<i>Neogobius melanostomus</i>	1	3	Yes	Fish
Oncorhynchus_mykiss_1of2_n1	<i>Oncorhynchus mykiss</i>	2	1	Yes	Fish
Oncorhynchus_mykiss_2of2_n7	<i>Oncorhynchus mykiss</i>	2	7	Yes	Fish
Padogobius_bonelli_1of1_n9	<i>Padogobius bonelli</i>	1	9	No	Fish
Perca_fluviatilis_1of1_n30	<i>Perca fluviatilis</i>	1	30	Yes	Fish
Phoxinus_csikii_1of4_n2	<i>Phoxinus csikii</i>	4	2	Yes	Fish
Phoxinus_csikii_2of4_n1	<i>Phoxinus csikii</i>	4	1	Yes	Fish
Phoxinus_csikii_3of4_n26	<i>Phoxinus csikii</i>	4	26	Yes	Fish
Phoxinus_csikii_4of4_n1	<i>Phoxinus csikii</i>	4	1	Yes	Fish
Phoxinus_lumaireul_1of2_n6	<i>Phoxinus lumaireul</i>	2	6	Yes	Fish
Phoxinus_lumaireul_2of2_n2	<i>Phoxinus lumaireul</i>	2	2	Yes	Fish
Phoxinus_septimaniae_1of1_n18	<i>Phoxinus septimaniae</i>	1	18	Yes	Fish
Phoxinus_sp_1of1_n2	<i>Phoxinus</i> sp.	1	2	Yes	Fish
Pseudorasbora_parva_1of1_n7	<i>Pseudorasbora parva</i>	1	7	Yes	Fish
Rhodeus_amarus_1of3_n3	<i>Rhodeus amarus</i>	3	3	No	Fish
Rhodeus_amarus_2of3_n11	<i>Rhodeus amarus</i>	3	11	No	Fish
Rhodeus_amarus_3of3_n1	<i>Rhodeus amarus</i>	3	1	No	Fish
Rupicapra_rupicapra	<i>Rupicapra rupicapra</i>	NA	NA	Yes	Chamois
Rupicapra_rupicapra	<i>Rupicapra rupicapra</i>	NA	NA	No	Chamois
Rutilus_aula_1of1_n11	<i>Rutilus aula</i>	1	11	Yes	Fish

(Continues)

TABLE 1 (Continued)

OSU name	Genus or species level	Number of haplotypes within OSU	No. of supporting sequences	Present in dataset	Description
Rutilus_pigus_1of2_n1	<i>Rutilus pigus</i>	2	1	No	Fish
Rutilus_pigus_2of2_n9	<i>Rutilus pigus</i>	2	9	No	Fish
Rutilus_rutilus_1of2_n20	<i>Rutilus rutilus</i>	2	20	Yes	Fish
Rutilus_rutilus_2of2_n3	<i>Rutilus rutilus</i>	2	3	Yes	Fish
Sabanejewia_larvata_1of1_n1	<i>Sabanejewia larvata</i>	1	1	No	Fish
Salaria_fluviatilis_1of3_n6	<i>Salaria fluviatilis</i>	3	6	Yes	Fish
Salaria_fluviatilis_2of3_n1	<i>Salaria fluviatilis</i>	3	1	Yes	Fish
Salaria_fluviatilis_3of3_n14	<i>Salaria fluviatilis</i>	3	14	Yes	Fish
Salmo_carpio_1of1_n1	<i>Salmo carpio</i>	1	1	Yes	Fish
Salmo_salar_1of1_n3	<i>Salmo salar</i>	1	3	Yes	Fish
Salmo_sp_1of3_n47	<i>Salmo</i> sp.	3	47	Yes	Fish
Salmo_sp_2of3_n11	<i>Salmo</i> sp.	3	11	Yes	Fish
Salmo_sp_3of3_n2	<i>Salmo</i> sp.	3	2	Yes	Fish
Salmo_trutta_1of2_n1	<i>Salmo trutta</i>	2	1	Yes	Fish
Salmo_trutta_2of2_n1	<i>Salmo trutta</i>	2	1	Yes	Fish
Salvelinus_fontinalis_1of1_n6	<i>Salvelinus fontinalis</i>	1	6	Yes	Fish
Salvelinus_namaycush_1of1_n4	<i>Salvelinus namaycush</i>	1	4	Yes	Fish
Salvelinus_sp_1of1_n21	<i>Salvelinus</i> sp.	1	21	Yes	Fish
Sander_lucioperca_1of1_n9	<i>Sander lucioperca</i>	1	9	Yes	Fish
Scardinius_erythrophthalmus_1of3_n11	<i>Scardinius erythrophthalmus</i>	3	11	Yes	Fish
Scardinius_erythrophthalmus_2of3_n1	<i>Scardinius erythrophthalmus</i>	3	1	Yes	Fish
Scardinius_erythrophthalmus_3of3_n1	<i>Scardinius erythrophthalmus</i>	3	1	Yes	Fish
Scardinius_hesperidicus_1of2_n5	<i>Scardinius hesperidicus</i>	2	5	Yes	Fish
Scardinius_hesperidicus_2of2_n6	<i>Scardinius hesperidicus</i>	2	6	Yes	Fish
Silurus_glanis_1of1_n12	<i>Silurus glanis</i>	1	12	No	Fish
Squalius_cephalus_1of1_n13	<i>Squalius cephalus</i>	1	13	Yes	Fish
Squalius_squalus_1of2_n9	<i>Squalius squalus</i>	2	9	Yes	Fish
Squalius_squalus_2of2_n1	<i>Squalius squalus</i>	2	1	Yes	Fish
Sturnus_vulgaris	<i>Sturnus vulgaris</i>	NA	NA	Yes	Starling
Sturnus_vulgaris	<i>Sturnus vulgaris</i>	NA	NA	No	Starling
Sus_scrofa	<i>Sus scrofa</i>	NA	NA	Yes	Pig
Sus_scrofa	<i>Sus scrofa</i>	NA	NA	No	Pig
Telestes_muticellus_1of1_n12	<i>Telestes muticellus</i>	1	12	Yes	Fish
Telestes_souffia_1of1_n14	<i>Telestes souffia</i>	1	14	Yes	Fish
Thymallus_thymallus_1of3_n1	<i>Thymallus thymallus</i>	3	1	Yes	Fish
Thymallus_thymallus_2of3_n9	<i>Thymallus thymallus</i>	3	9	Yes	Fish
Thymallus_thymallus_3of3_n2	<i>Thymallus thymallus</i>	3	2	Yes	Fish
Tinca_tinca_1of1_n29	<i>Tinca tinca</i>	1	29	Yes	Fish
Turdus_philomelos	<i>Turdus philomelos</i>	NA	NA	Yes	Thrush
Turdus_philomelos	<i>Turdus philomelos</i>	NA	NA	No	Thrush

Note: The 116 Operational Sequence Units derived from the 912 Sanger sequences collected as part of this study. To provide certainty, the label of each OSU is broken into the elements used to make that taxonomic assignment: the consensus taxonomic label derived from the sequences in the cluster (green) + No. of haplotypes (blue) + No. of supporting sequences (orange). For example, *Abramis brama*_1of1_n11 is an OSU to species level with only 1 OSU in the reference database and 9 sequences supporting that OSU level taxonomic name. *Salmo*_sp_1of3_n47 is an OSU for genus level because the sequences which support this OSU are from different *Salmo* species, this OSU is 1 of 3 *Salmo* sp. OSU and has 47 sequences supporting this OSU level taxonomic name.

not beyond Class level: 3.56%, 1.22% and 1.28% of the total abundance MitoFish, MitoFish+ and MIDORI, respectively (Figure 2). However, these taxonomic annotations can be misleading. As missing clades or outgroups are the Achilles' heel of classifiers, the actual percentage of missing annotation may be much higher. All records in the MitoFish reference have the same kingdom (Eukaryota), phylum

TABLE 2 Test for mapping parameter and efficiency

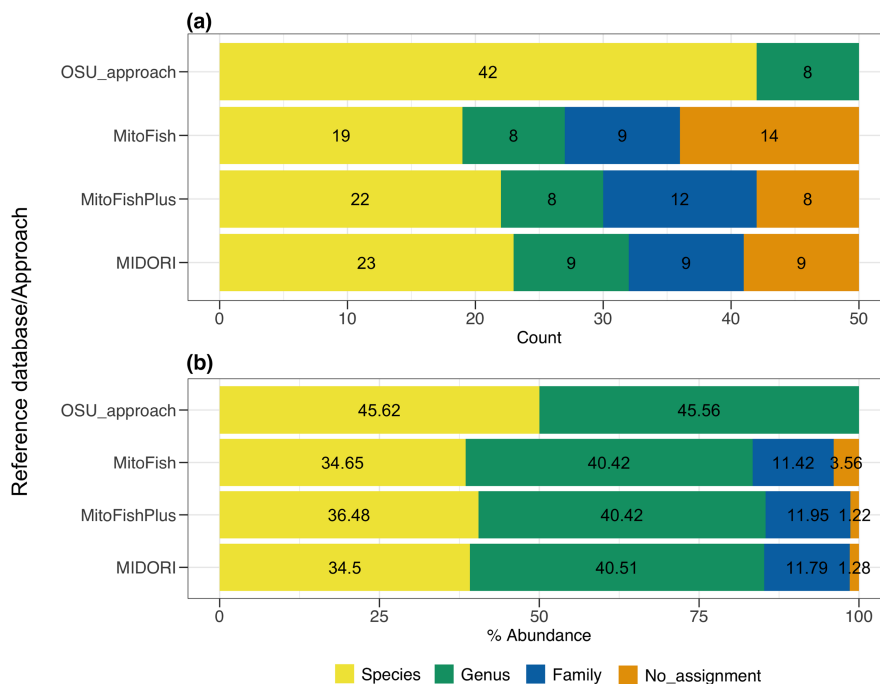
ID	QC	Reads mapped	% Mapped	No. of OSUs
1	1	4,683,758	52.7	95
1	0.99	4,684,009	52.7	95
1	0.98	4,684,012	52.7	95
1	0.97	4,684,012	52.7	95
0.99	1	6,991,312	78.6	99
0.99	0.99	6,991,768	78.6	99
0.99	0.98	6,991,773	78.6	99
0.99	0.97	6,991,775	78.6	99
0.98	1	7,753,236	87.2	101
0.98	0.99	7,753,792	87.2	101
0.98	0.98	7,753,802	87.2	101
0.98	0.97	7,753,806	87.2	101
0.97	1	7,877,395	88.6	101
0.97	0.99	7,878,060	88.6	101
0.97	0.98	7,878,082	88.6	101
0.97	0.97	7,878,086	88.6	101

Note: Number and percentage of sequences mapped to the total of 116 OSUs using Usearch::cluster_fast with different sequence identity (ID) and query coverage (QC) parameters. The parameters highlighted in bold were used to generate the data in this study.

(Chordata) and class (Actinopteri), therefore all zOTUs are annotated at least to class level "Actinopteri." The MIDORI reference is more diverse, but all records are Eukaryotes and so all unidentified non-eukaryotic zOTUs will be mis-labelled as "Eukaryotes" as the highest taxonomic grouping. Considering the diversity limits of the reference database, the annotation failed for 98.34% of the zOTUs using MitoFish and 96.96% using MIDORI as annotation reference beyond family level. This result shows that not all zOTUs generated in this study are fish related. To improve the annotation, we added nonfish related sequences to the MitoFish reference (MitoFish+) based on Blast hits using zOTUs with missing or bad (low taxonomic rank or unclear labels like "environmental samples") annotations as queries. This expanded MitoFish+ reference database included outgroups like bacteria. Surprisingly, 71% of the zOTUs are assigned to bacteria and are thus not of eukaryotic origin. This seems a lot, but these bacteria zOTUs represented only 6.9% of the total reads (abundance), therefore bacteria zOTUs are abundant (in terms of species number) but rare in our dataset in terms of abundances (Table S2). It also shows that the annotations should be evaluated with caution and references might be adjusted to get more meaningful results.

Overall, we find that the predicted annotations for the most abundant zOTUs (those zOTUs with the highest read number) are similar to Mitofish, Mitofish+ or MIDORI as the annotation reference, with a few interesting exceptions. Although the MiFish primers are fish-specific, nontarget amplification is difficult to avoid. For this reason, the reference databases should cover a broader diversity (e.g., MIDORI) or be extended (MitoFish+). For example, zOTU13 and zOTU24 had no correct assignment with MitoFish. Using the MIDORI as a reference, zOTU13 was associated with Bovidae (Cow), possibly because of the addition of BSA in the PCR reaction and zOTU24 was identified as possible human contamination. There are also minor, but perhaps intriguing, differences in

FIGURE 2 A comparison (e.g., number and abundance) of the taxonomic annotation depth of the 50 most common zOTUs and OSUs. Either MitoFish, MitoFishPlus or MIDORI was used to annotate the zOTUs. (a) Number; (b) Abundance. The colors represent the taxonomic assignment level (species—yellow, genus—green and family—blue), with orange indicating the proportion of zOTUs without taxonomic assignment in the corresponding reference database.



annotation between the open reference databases. One prime example to demonstrate this is the annotation of zOTU27, which is associated with *Phoxinus phoxinus* using MitoFish, while using MIDORI the annotation concurs but is only to family level (Leuciscidae). As the species *P. phoxinus* is not found in the study area (Alexander & Seehausen, 2021) it showcases the differing outcomes and reliability of using closed reference databases.

3.2 | Closed reference database and OSU mapping approach

The number of unique OSUs per species varied from 1 (e.g., *Tinca tinca*) to maximal 5 (e.g., *Barbatula* sp. lineage I). The number of sequences support unique haplotypes varies from 1 (e.g., *Salaria fluviatilis*) to 47 (e.g., *Salmo* sp.) with a mean of 7.75. The sequence resolution for the 12S amplicon is good and most unique haplotype groups contain only one species exception occurred (e.g., *Phoxinus* or *Salmo*, see Table 1 for further OSU details). The ML tree of the

unique haplotypes is shown in Figure S4. By using a closed reference database with the target taxa and supplemented with possible outgroups, here we assigned amplicons reads to a single OSU with the appropriate species or genus level taxonomic information (Figures 2 and 3). For the OSU read mapping, we use two parameters: Sequence identity (ID) and query coverage (QC). Stringent mapping criteria reduces mapping efficiency, while relaxed ones will cause false assignment. In Table 2, the percentage of reads mapped ranges from 52.7% to 88.6%, with the ID threshold having a bigger influence on the mapping rate than QC and therefore a higher number of OSUs being found when the ID threshold was lowered. The best results were obtained with the following parameters and used for subsequent mapping: ID: 97 and QC: 100. This resulted in 7,877,395 (88.6%) reads mapped to 101 OSUs (87%, shown in Figure S5). Compared to the open reference database taxonomic assignments, was entirely to genus/species level (8 taxonomic assignments were to genus and could not be resolved to species because of insufficient amplicon resolution, see Figure 3). Most of the reads which could not be mapped to any of the OSUs were bacteria.

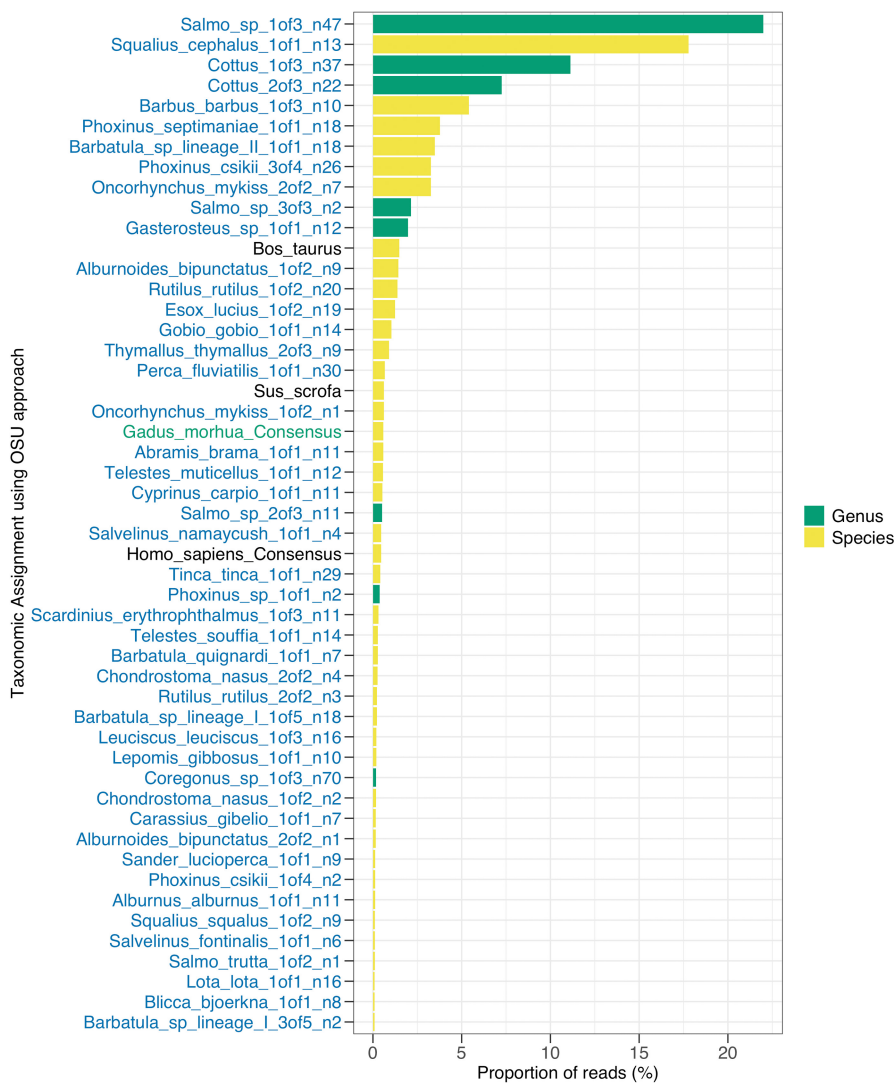


FIGURE 3 The top 50 OSU. Taxonomic assignment using this approach was to either genus or species level. The taxonomic assignment includes the species/genus name, the number of OSUs with the same annotation in the dataset (i.e., 1 of 2, 2 of 2) and number of sequences supporting this taxonomic assignment (genus—green, species—yellow). See also Table 1. Text color indicates taxa type: blue—fish species or genus, black—mammals and green positive control.

3.3 | Positive controls

We used four positive control samples (PCP) in this library, containing a tissue DNA extract from Atlantic Cod, *Gadus morhua* (See Table S3 for sequence). Using positive control samples in this way enables us to identify possible contamination levels between samples (cross-contamination or tag-switching) and is a more effective method compared to using negative controls. The classic zOTU (MIDORI) and the OSU approach show a similar relative composition (Figure 4). In the classic approach, 72.9%–85.3% of reads are assigned to Family level Gadidae and in the OSU approach 71.7%–86.3% of the reads were assigned to *Gadus morhua*. The difference between the approaches lies in the taxonomic depth of the annotation.

In both approaches, minor contamination (in the form of other taxa) was found, but as shown in Figure 4, the relative read abundance of these other taxa was low compared to the Gadidae or *Gadus morhua* assigned reads. Interestingly, when comparing the different taxa found in the positive samples with each of the two approaches, there is only a minor deviation between the two approaches in terms of taxonomic assignments. Using a minimum abundance threshold (200 reads), a total of 11 and 10 other taxa were found using the zOTU-MIDORI and OSU approach, respectively. Most of the annotations were identical (i.e., *Salmo*, *Cottus*, *Squalis cephalus*, *Bos tarus*, *Esox lucius*, *Oncorhynchus mykiss*, *Sus scrofa*, *Homo sapiens*) with a few

exceptions: *Phoxinus phoxinus* (zOTU-MIDORI) vs. *P. csikii* (OSU) (See Figure 4) and *Barbonymus schwanefeldii* with zOTU-MIDORI, but not found with the OSU approach. As the detection of *B. schwanefeldii* was only found in the positive controls (206 reads on average, ranging from 0 to 523), we can assume this represents a true contamination associated with our positive control spike, and therefore would not be assigned using the OSU approach, as it was not in the closed reference database. In terms of the abundances of taxa found in all positive control samples, there is a striking similarity between the open and closed reference databases for those taxa with high abundance as is also the case for those taxa found in the actual samples (Figures S6 and S7). The variation in read % per taxon between the two approaches for those common taxa can be explained by the difference in species contained in each of the reference database approaches. Similarly, we find the detection of some taxa which could be considered rare (in low abundance) is more variable and often only found in one of the two approaches, as the species is likely absent from one of the reference databases.

3.4 | Negative controls

As with the positive controls, minor contamination was also found in the negative controls. This is likely to have occurred during library

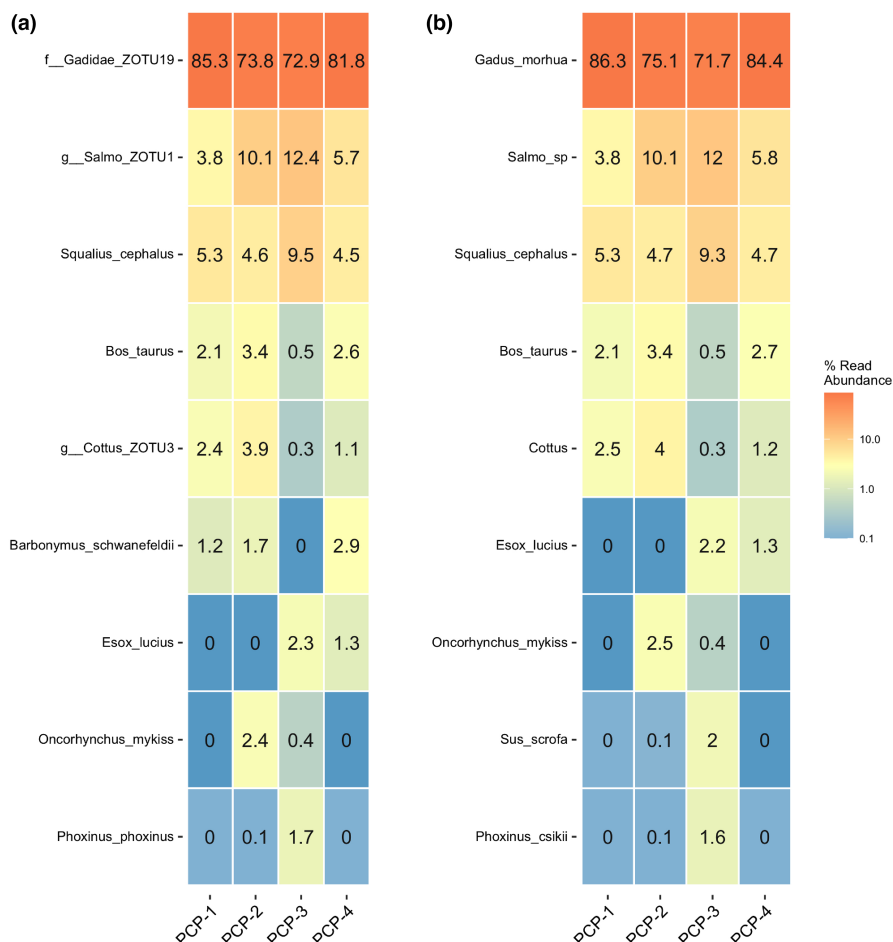


FIGURE 4 Relative abundance heat map for the four positive control PCR samples (PCP). a. using the zOTU approach and the MIDORI reference database, b. using OSU approach. As a positive control we used *Gadus morhua* (Atlantic cod) tissue extract.

preparation and is typical of such eDNA studies where PCR amplification is highly efficient in those samples with minimal amplification competition, such as that from minor contamination. Only two of the eight negative control samples contained a total of >350 reads. The distribution of reads for these samples are shown in Figure S8 and are strikingly similar across each of the taxonomic assignment methods. The majority of the reads found belong to some of the most common taxa found in the samples, including the most and second most occurring species found in the dataset (*Squalius cephalus* and *Salmo* sp.). The six other samples contain on average 74 (range 12–374), 74 (range 12–374), and 14 (range 10–19) using the MIDORI, MitoFish+ and OSU reference databases, respectively.

3.5 | Phylogeny

To compare the sequences derived from each approach, classic and OSU, we carried out a phylogenetic comparison. Figure 5 shows the top 20 most abundant zOTU (MIDORI) alongside the 20 most abundant OSU hits. The tree highlights both the disparity (e.g., Leuciscidae) and similarity (e.g., *Squalius cephalus*) between taxonomic annotation levels using the classic zOTU (MIDORI) and OSU approaches. However, with *Salmo* sp. and *Cottus* sp. both approaches are the same to genus level only. This is because of the 12S amplicon length

(in this case 192 bp). Species detection is the aim of most eDNA studies, however sequencing technologies often only target a small fragment of DNA, which is insufficient for species determination for some taxonomy.

4 | DISCUSSION

Species detection derived from eDNA samples relies on reference sequences and involves linking genetic sequences (barcode regions) with taxonomic names. Here, we examined the influence of different types of reference databases (open/publicly available vs. closed/curated) and annotation approaches. Our study takes advantage of a curated fish reference database generated from multiple specimens of all fish species occurring across the study area (i.e., a nationwide collection in Switzerland) to explore fish communities derived from eDNA water samples collected in rivers and acts as a proof of principle study for the wider use of curated and study specific reference databases. Using a curated reference database, combined with a proposed OSU approach, we gained a higher proportion of species and genus level taxonomic assignment compared to the open reference databases used in this study (MitoFish, MitoFish+ and MIDORI). Although the annotation may be to different levels, encouragingly, the top 5 taxa generated from the classic zOTU and OSU approach

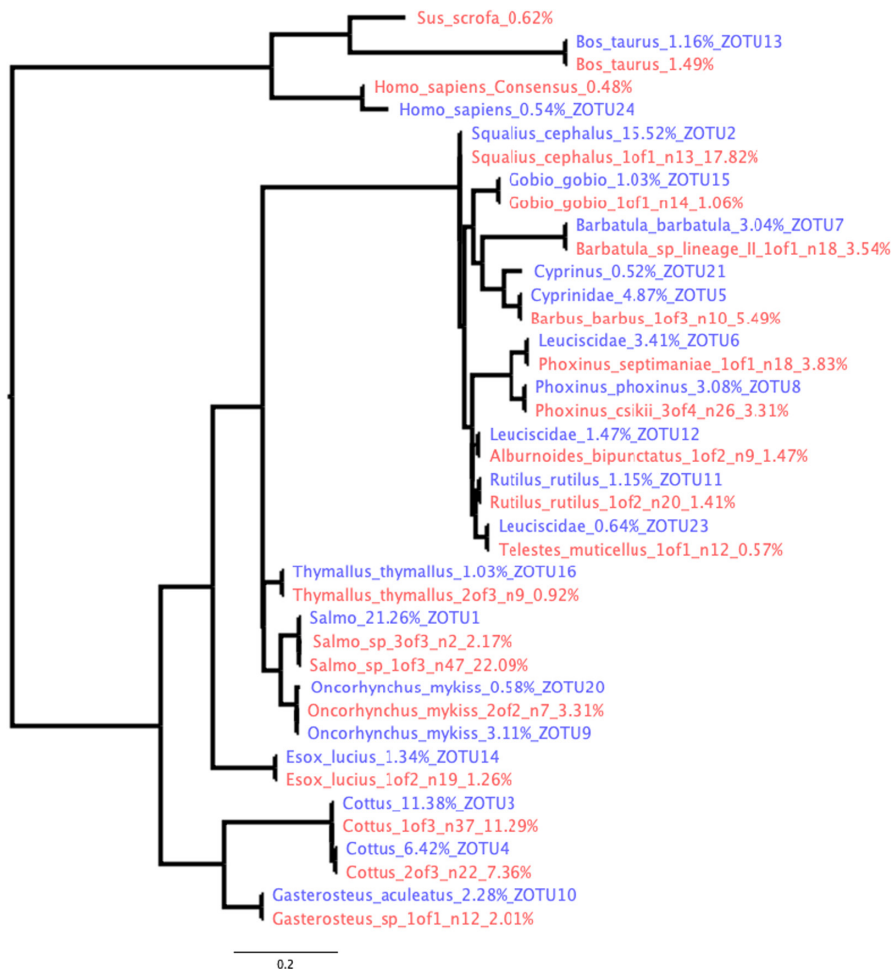


FIGURE 5 A phylogenetic comparison of the top 20 zOTUs and OSUs. zOTUs are presented in blue and OSUs are in red, relative abundance is added to the label for both approaches.

TEXT BOX 1 The perfect reference database is difficult to create, but here are some important factors to consider when constructing your own study specific reference database:

General considerations

- Species should be represented by multiple sequences.
- Species must be correctly identified.
- Sequences should be derived from reliable tissue samples.
- All species found in the geographic region of the study (from the target groups) should have representative sequences in the reference database.
- Sequences should be phylogenetically placed to ensure against incorrect labelling.
- All reference sequences used should cover the full length of the amplicon.

Specific considerations

- Reference databases must include outgroups to avoid problematic assignments due to co-amplification of non-target groups.
- Consider an OSU approach for a more specific survey if sequences for target species are available.
- After assigning taxonomic names to your data, ensure its biological/ecological plausibility.

are identical and equate to 53.03% and 63.56% of the taxonomically assigned reads generated from the two approaches, respectively. The quality and depth of the annotation in all reference databases depends on several factors, including species coverage, sequence representation, and reliability of the sequence. Consequently, we have synthesized the main points that should be considered when curating a reference database to ensure appropriate taxonomic assignment with full, verified taxonomic annotation (See Text Box 1). Here, we discuss how this will lead to more accurate and informative biodiversity data.

4.1 | Reference database construction: General considerations

Basic considerations when constructing a reference database include species coverage, sequence representation, and reliability of the sequence. Firstly, a list of the target species within the study area should be made and the correct sequences available (in terms of length and marker) found for those species should be assembled (Blackman et al., 2021). This allows any gaps to be identified and therefore an immediate understanding of the limitation of your reference database and data interpretation (Li et al., 2022; Weigand et al., 2019). Secondly, multiple sequences should be used for each

species to ensure haplotypic diversity (Leite et al., 2020). Notably, of the top 5 most abundant taxa identified in our study, *Barbus barbus* is identified at species level using the OSU (closed reference), but only at family level using the zOTU approach (open reference). The lack of species determination for *B. barbus* could be due to the lack of local species sequence in the open reference databases and the high intraspecific variation not represented in these databases (Leite et al., 2020; Weigand et al., 2019). The representation of the genetic diversity within a species is an important aspect to consider when curating a reference library. Thus, not only should multiple sequences for each species be included in the database, but also the origins of those sequences must be considered, in order to cover geographic and taxonomic breadth. By including geographic information when submitting sequences, studies can ensure species are represented by appropriate sequences (see also Li et al., 2022), which cover local genetic diversity and are relevant to the study area and the resulting data (Bergsten et al., 2012).

Once the sequences have been sourced for a reference database, essential taxonomic annotation checks should be applied to ensure their identity is correct and the annotations is complete. Leray et al. (2019) examined the potential reasons for these discrepancies in GenBank. Their study documented several reasons why the taxonomic annotation could be incorrect; firstly, the amplified sequence may not be from the target organism, either due to contamination or intimate association (i.e., bacteria). Secondly, pseudogene amplification and thirdly, incorrect identification of the target taxa initially. To address these issues when curating a reference database, sequences should be phylogenetically placed to ensure their annotation is correct. Mislabeled sequences can then be identified and removed from the database (for example, by using SATIVA; Kozlov et al., 2016). Thirdly, by systematically screening sequences based on their phylogeny, any sequences with inadequate annotation may be placed phylogenetically within the contexts of the other sequences within the database.

4.2 | Reference database construction: Specific considerations

As with several metabarcoding primers, the MiFish primers are known to amplify other nontarget organismal groups, including mammal and bird species (Ritter et al., 2022). In our study, we further examined those zOTUs that received a poor taxonomic annotation. The Blast hit results from these sequences showed that a proportion of our data was in fact assigned to Bacteria. We therefore recommend that reference databases routinely encompass outgroups, that is, a wider diversity than the study target group, as such sequences would otherwise be incorrectly (and phylogenetically too narrowly) assigned. Here, our closed reference database MitoFish+ and MIDORI included further diversity, as opposed to MitoFish, which is fish only. This approach will allow reads from nontarget groups to be assigned correctly rather than “forcing” taxonomic assignment to the represented taxa incorrectly. With this approach we revealed taxa from different classes and orders (e.g., see nontarget

assignments in Figures S6 and S7). Other studies of fish communities derived from eDNA have previously demonstrated a shift in taxonomic assignment when reference databases are broadened and further diversity is included (Schenekar et al., 2020); therefore, the inclusion of such outgroups are not only advised, but essential in preventing false positives results. Confidence thresholds for the annotation are also crucial in this respect. We recommend using higher cut-offs (>85%) and a closer examination of the unannotated zOTU instead of lowering the confidence threshold.

By employing an OSU approach and assigning (mapping) reads to individual sequence representatives, we ensured the full taxonomic annotation and certainty of each assignment. In this study, we had access to a large custom-built reference database, not only encompassing the geographic range of our study sites in which eDNA samples were taken, but also with complete taxonomic annotation carried out by expert ichthyologists. The OSU approach worked well in mapping sequences to the OSUs in the database, but back-mapping efficiency varied (52.7%–88.6%) depending on the parameters used. The results show that identity (ID) had a bigger influence on efficiency than query coverage (QC). However, although more reads map to the OSUs with a lower ID, the fish-species composition does not change. There are also limitations to the OSU approach to consider. There are different ways of handling sequences which cannot be mapped clearly to one OSU, but to several. For example, Usearch::Sintax assigns these sequences to the first OSU, so the order of similar OSUs could play a role. In addition, different ways of calculating similarity might be considered when mapping the sequence to an OSUs. The position of a mismatch could be decisive in assigning a sequence to the correct OSU. Furthermore, the codon structure of amplicons of coding gene regions (e.g., COI) could improve the mapping. That said, the reference database we used here represents a near perfect set of sequences across our study area and target group. This form of reference database is close to the ideal, but required extensive investment, both in terms of time and financially. Attempts to replicate this form of reference database should not be disregarded and should (long term) be an aim of those researchers and end-users who can collect specimens and contribute to these databases.

In terms of reference database curation, a last but vital consideration when reviewing data should be its biological or ecological setting. In all metabarcoding and biodiversity studies, particularly those from eDNA samples, scientists must always contextualize their data to determine the success or failure of taxonomic assignment, particularly if a species is found outside its known habitat or range. As advice for data analysis, derived either from an open or closed reference database, the possible reasons other than natural occurrences must be considered, such as zoo or restaurant outflows into the water being sampled.

4.3 | Further considerations when using eDNA

Although the OSUs represented 116 species/genera of fish that are known to occur in the sites we sampled, 15 autochthonous species

known from the study area were not detected in this sampling campaign. This level of discrepancy is comparable to many other eDNA studies (e.g., see Keck et al., 2022). These absences could be either true negatives of the species at the sampling site or false negatives. The latter can derive from several biological and methodological constraints of collecting eDNA in flowing water, such as insufficient amounts of DNA shed by the target organism into the water body or similarly the heterogenous nature of eDNA within the water column meant the sample did not contain any of that species' DNA (as discussed in Bruce et al., 2021). Environmental DNA sampling, although not a perfect method, has vastly improved in the last 10 years, in terms of both usability for biomonitoring and our interpretation and understanding of the results we glean from collecting these samples. Nevertheless, one noticeable and further constraint of eDNA is the reliance on the amplification of small fragments of DNA. Typically, fragments <400 bp are targeted due to sequencing platform requirements (the 12S fragment here ~170 bp). By using such small fragments, confidence in assignment and taxonomic resolution is limited (Deiner et al., 2016; Port et al., 2016), as demonstrated in the genus level-only determination of certain taxa using both the classic and OSU approach (e.g., *Salmo*, *Cottus*, *Coregonus* and *Phoxinus*). With the 12S amplicon we can distinguish between genera, but not necessarily within. For example, relatively closely related species (i.e., in the first one to two million years after speciation) can usually not be distinguished by single stretches of sequence variation, such as short barcodes. Therefore, we cannot distinguish certain species, such as several *Salmo* species which are known from Switzerland (i.e., *Salmo labrax* and *Salmo marmoratus*), or virtually all the phylogenetically very young *Coregonus* species (which radiated within the last few ten thousand years, Jardim de Queiroz et al., 2022). With all reference approaches, the reads of the former all received the *Salmo* sp. assignments. This un-resolved assignment was therefore due to the methodological constraints of using a short DNA fragment unable to distinguish young species within this genus. Focusing on larger fragments or different regions may become more common practice in the future, but currently, when using such small fragments of DNA for fish biomonitoring, it is an important aspect to consider, particularly if the aim is species determination.

5 | CONCLUSION

Incomplete or inaccurate reference databases have long been highlighted as a limiting factor to correctly assign metabarcoding data. Here, we demonstrated how locally derived curated reference databases improve species level assignments, which is a key goal of eDNA metabarcoding studies. However, we have also identified potential ways to improve and utilize current open reference databases with simple curation steps (i.e., phylogenetic placement, the addition of geographically relevant sequences, including outgroups, and removing incomplete or problematic annotated references), which will likely improve family or genus level assignment and help to explain data losses. We encourage researchers to consider the implications

of their reference databases and curate them following our simple suggestions (See Text Box 1). We further support the improvement of reference sequence generation, including efforts to collect, morphologically identify, and sequence more specimens to fill gaps in current reference databases. These steps will lead to better reference databases and taxonomic assignment for biodiversity.

AUTHOR CONTRIBUTIONS

Rosetta C. Blackman, Jean-Claude Walser, and Florian Altermatt conceived the study. Jakob Brodersen performed the genomic lab analyses and Jean-Claude Walser performed the bioinformatics. Lukas Rüber, Ole Seehausen, Soraya Villalba, and Jakob Brodersen collected the fish, sequenced them, and constructed the closed reference database. Jean-Claude Walser performed the analysis. Rosetta C. Blackman and Jean-Claude Walser produced all figures and wrote the first draft. All authors contributed to the interpretation data and commented on the paper.

ACKNOWLEDGMENTS

We thank the Swiss Federal Office for the Environment (BAFU/FOEN) for access to the eDNA samples. We thank two anonymous reviewers for their comments on the manuscript. Funding is from the Swiss National Science Foundation (grant nr. 31003A_173074) and the University of Zurich Research Priority Programme in Global Change and Biodiversity (URPP GCB) to FA. The Federal Office for the Environment/Bundesamt für Umwelt (FOEN/BAFU) financed the establishment of the 12S DNA reference library of Swiss fishes under the contract 00.5058.PZ/6B1725F08 to LR and OS. Genetic analyses were done in collaboration with the Genetic Diversity Centre (GDC) at ETH Zurich.

CONFLICT OF INTEREST

The authors of this study have no conflict of interest to declare. The manuscript is original work. All contributing authors have seen and agreed with the content before submission. Furthermore, we guarantee that this study has not been submitted to any other journal.

DATA AVAILABILITY STATEMENT

Sequencing data generated during this study, the data analysis scripts and an overview of the quality control and information on data lost through filtering in the bioinformatic steps can be found on the DRYAD repository <https://doi.org/10.5061/dryad.1g1jwsv15>. Sequences used for the closed reference database are available on GenBank and Bold under accession numbers OP930966-OP931877.

ORCID

Rosetta C. Blackman  <https://orcid.org/0000-0002-6182-8691>

Jean-Claude Walser  <https://orcid.org/0000-0003-1513-0783>

Lukas Rüber  <https://orcid.org/0000-0003-0125-008X>

Jeanine Brantschen  <https://orcid.org/0000-0002-2945-3607>

Jakob Brodersen  <https://orcid.org/0000-0003-2060-6379>

Ole Seehausen  <https://orcid.org/0000-0001-6598-1434>

Florian Altermatt  <https://orcid.org/0000-0002-4831-6958>

REFERENCES

- Alexander, T., & Seehausen, O. (2021). *Diversity, distribution and community composition of fish in perialpine lakes—“Projet Lac” synthesis report* (p. 282). Swiss Federal Institute of Aquatic Science and Technology.
- Altermatt, F., Little, C. J., Mächler, E., Wang, S., Zhang, X., & Blackman, R. C. (2020). Uncovering the complete biodiversity structure in spatial networks—The example of riverine systems. *Oikos*, 129, 607–618.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215, 403–410.
- Antich, A., Palacin, C., Wangenstein, O. S., & Turon, X. (2021). To denoise or to cluster, that is not the question: Optimizing pipelines for COI metabarcoding and metaphylogeography. *BMC Bioinformatics*, 22(1), 177.
- Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, 41, D36–D42.
- Bergsten, J., Bilton, D. T., Fujisawa, T., Elliott, M., Monaghan, M. T., Balke, M., Hendrich, L., Geijer, J., Herrmann, J., Foster, G. N., Ribera, I., Nilsson, A. N., Barraclough, T. G., & Vogler, A. P. (2012). The effect of geographical scale of sampling on DNA barcoding. *Systematic Biology*, 61(5), 851–869.
- Blackman, R. C., Ho, H. C., Walser, J. C., & Altermatt, F. (2022). Spatio-temporal patterns of multi-trophic biodiversity and food-web characteristics uncovered across a river catchment using environmental DNA. *Communications Biology*, 5, 259.
- Blackman, R. C., Osathanunkula, M., Brantschen, J., Di Muri, C., Harper, L. R., Mächler, E., Hänfling, B., & Altermatt, F. (2021). Mapping biodiversity hotspots of fish communities in subtropical streams through environmental DNA. *Scientific Reports*, 11, 10375.
- Brandt, M. I., Trouche, B., Quintric, L., Günther, B., Wincker, P., Poulain, J., & Arnaud-Haond, S. (2021). Bioinformatic pipelines combining denoising and clustering tools allow for more comprehensive prokaryotic and eukaryotic metabarcoding. *Molecular Ecology Resources*, 21(6), 1904–1921.
- Brantschen, J., Blackman, R. C., Walser, J.-C., & Altermatt, F. (2021). Environmental DNA gives comparable results to morphology-based indices of macroinvertebrates in a large-scale ecological assessment. *PLOS ONE*, 16, e0257510.
- Brown, E. A., Chain, F. J. J., Zhan, A., MacIsaac, H. J., & Cristescu, M. E. (2016). Early detection of aquatic invaders using metabarcoding reveals a high number of non-indigenous species in Canadian ports. *Diversity & Distributions*, 22(10), 1045–1059.
- Bruce, K., Blackman, R. C., Bourlat, S. J., Hellström, M., Bakker, J., Bista, I., Bohmann, K., Bouchez, A., Brys, R., Clark, K., Elbrecht, V., Fazi, S., Fonseca, V. G., Hänfling, B., Leese, F., Mächler, E., Mahon, A. R., Meissner, K., Panksep, K., ... Deiner, K. (2021). *A practical guide to DNA based methods for biodiversity assessment*. Pensoft.
- Cilleros, K., Valentini, A., Allard, L., Dejean, T., Etienne, R., Grenouillet, G., Iribar, A., Taberlet, P., Vigouroux, R., & Brosse, S. (2019). Unlocking biodiversity and conservation studies in high-diversity environments using environmental DNA (eDNA): A test with Guianese freshwater fishes. *Molecular Ecology Resources*, 19(1), 27–46.
- Conte-Grand, C., Britz, R., Dahanukar, N., Raghavan, R., Pethiyagoda, R., Tan, H. H., Hadiaty, R. K., Yaakob, N. S., & Rüber, L. (2017). Barcoding snakeheads (Teleostei, Channidae) revisited: Discovering greater species diversity and resolving perpetuated taxonomic confusions. *PLoS ONE*, 12(9), e0184017.
- Curd, E. E., Gold, Z., Kandlikar, G. S., Gomer, J., Ogden, M., O'Connell, T., Pipes, L., Schweizer, T. M., Rabichow, L., Lin, M., Shi, B., Barber, P. H., Kraft, N., Wayne, R., & Meyer, R. S. (2019). Anacapa Toolkit: An environmental DNA toolkit for processing multilocus metabarcode datasets. *Methods in Ecology and Evolution/British Ecological Society*, 10(9), 1469–1475.

- de Santana, C. D., Parenti, L. R., Dillman, C. B., Coddington, J. A., Bastos, D. A., Baldwin, C. C., Zuanon, J., Torrente-Vilara, G., Covain, R., Menezes, N. A., Datovo, A., Sado, T., & Miya, M. (2021). The critical role of natural history museums in advancing eDNA for biodiversity studies: A case study with Amazonian fishes. *Scientific Reports*, *11*(1), 18159.
- Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895.
- Deiner, K., Fronhofer, E. A., Mächler, E., Walser, J.-C., & Altermatt, F. (2016). Environmental DNA reveals that rivers are conveyor belts of biodiversity information. *Nature Communications*, *7*, 12544.
- Deiner, K., Walser, J.-C., Mächler, E., & Altermatt, F. (2015). Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, *183*, 53–63.
- Dufresne, Y., Lejzerowicz, F., Perret-Gentil, L. A., Pawlowski, J., & Cordier, T. (2019). SLIM: A flexible web application for the reproducible processing of environmental DNA metabarcoding data. *BMC Bioinformatics*, *20*(1), 88.
- Dugal, L., Thomas, L., Wilkinson, S. P., Richards, Z. T., Alexander, J. B., Adam, A. A. S., Kennington, W. J., Jarman, S., Ryan, N. M., Bunce, M., & Gilmour, J. P. (2022). Coral monitoring in northwest Australia with environmental DNA metabarcoding using a curated reference database for optimized detection. *Environmental DNA*, *4*(1), 63–76.
- Edgar, R. C. (2016a). UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon reads. *bioRxiv*. <https://doi.org/10.1101/081257>
- Edgar, R. C. (2016b). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences. *bioRxiv*. <https://doi.org/10.1101/074161>
- Froese, R., & Pauly, D. (2021). *FishBase*. World Wide Web Electronic Publication. www.fishbase.org
- Hänfling, B., Lawson Handley, L., Read, D. S., Hahn, C., Li, J., Nichols, P., Blackman, R. C., Oliver, A., & Winfield, I. J. (2016). Environmental DNA metabarcoding of lake fish communities reflects long-term data from established survey methods. *Molecular Ecology*, *25*(13), 3101–3119.
- Iwasaki, W., Fukunaga, T., Isagozawa, R., Yamada, K., Maeda, Y., Satoh, T. P., Sado, T., Mabuchi, K., Takeshima, H., Miya, M., & Nishida, M. (2013). MitoFish and MitoAnnotator: A mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Molecular Biology and Evolution*, *30*(11), 2531–2540.
- Jackman, J. M., Benvenuto, C., Coscia, I., Oliveira Carvalho, C., Ready, J. S., Boubli, J. P., Magnusson, W. E., McDevitt, A. D., & Guimarães Sales, N. (2021). eDNA in a bottleneck: Obstacles to fish metabarcoding studies in megadiverse freshwater systems. *Environmental DNA*, *3*, 837–849.
- Jardim de Queiroz, L., Dösz, C., Altermatt, F., Alther, R., Borko, Š., Brodersen, J., Gossner, M., Graham, C., Matthews, B., McFadden, I. R., Pellissier, L., Schmitt, T., Selz, O. M., Villalba, S., Rüber, L., Zimmermann, N., & Seehausen, O. (2022). Climate, immigration and speciation shape terrestrial and aquatic biodiversity in the European Alps. *Proceedings of the Royal Society B: Biological Sciences*, *289*, 20221020.
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, *30*(4), 772–780.
- Keck, F., & Altermatt, F. (2022). Management of DNA reference libraries for barcoding and metabarcoding studies with the R package redb. *Molecular Ecology Resources*. in press. <https://doi.org/10.1111/1755-0998.13723>
- Keck, F., Blackman, R. C., Bossart, R., Brantschen, J., Couton, M., Hürlemann, S., Kirschner, D., Locher, N., Zhang, H., & Altermatt, F. (2022). Meta-analysis shows both congruence and complementarity of DNA and eDNA metabarcoding to traditional methods for biological community assessment. *Molecular Ecology*, *31*, 1820–1835.
- Kozlov, A. M., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, *44*(11), 5022–5033.
- Lawson Handley, L. (2015). How will the “molecular revolution” contribute to biological recording? *Biological Journal of the Linnean Society*, *115*(3), 750–766.
- Leite, B. R., Vieira, P. E., Teixeira, M. A. L., Lobo-Arteaga, J., Hollatz, C., Borges, L. M. S., Duarte, S., Troncoso, J. S., & Costa, F. O. (2020). Gap-analysis and annotated reference library for supporting macroinvertebrate metabarcoding in Atlantic Iberia. *Regional Studies in Marine Science*, *36*, 101307.
- Leray, M., Ho, S.-L., Lin, I.-J., & Machida, R. J. (2018). MIDORI server: A webserver for taxonomic assignment of unknown metazoan mitochondrial-encoded sequences using a curated database. *Bioinformatics*, *34*(21), 3753–3754.
- Leray, M., Knowlton, N., Ho, S.-L., Nguyen, B. N., & Machida, R. J. (2019). GenBank is a reliable resource for 21st century biodiversity research. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(45), 22651–22656.
- Leray, M., Knowlton, N., & Machida, R. J. (2022). MIDORI2: A collection of quality controlled, preformatted, and regularly updates references databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environmental DNA*, early view. <https://doi.org/10.1002/edn3.303>
- Li, F., Zhang, Y., Altermatt, F., Zhang, X., Cai, Y., & Yang, Z. (2022). Gap analysis for DNA-based biomonitoring of aquatic ecosystems in China. *Ecological Indicators*, *137*, 108732.
- Locatelli, N. S., McIntyre, P. B., Therikildsen, N. O., & Baetscher, D. S. (2020). GenBank's reliability is uncertain for biodiversity researchers seeking species-level assignment for eDNA. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(51), 32211–32212.
- Machida, R. J., Leray, M., Ho, S.-L., & Knowlton, N. (2017). Metazoan mitochondrial gene sequence reference datasets for taxonomic assignment of environmental samples. *Scientific Data*, *4*, 170027.
- Mächler, E., Little, C. J., Wüthrich, R., Alther, R., Fronhofer, E. A., Gounand, I., Harvey, E., Hürlemann, S., Walser, J., & Altermatt, F. (2019). Assessing different components of diversity across a river network using eDNA. *Environmental DNA*, *1*(3), 290–301.
- Mächler, E., Walser, J. C., & Altermatt, F. (2021). Decision making and best practices for taxonomy-free eDNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology*, *30*, 3326–3339.
- Marques, V., Milhau, T., Albouy, C., Dejean, T., Manel, S., Mouillot, D., & Juhel, J.-B. (2021). GAPeDNA: Assessing and mapping global species gaps in genetic databases for eDNA metabarcoding. *Diversity & Distributions*, *27*(10), 1880–1892.
- Mathon, L., Valentini, A., Guérin, P.-E., Normandeau, E., Noel, C., Lionnet, C., Boulanger, E., Thuiller, W., Bernatchez, L., Mouillot, D., Dejean, T., & Manel, S. (2021). Benchmarking bioinformatic tools for fast and accurate eDNA metabarcoding species identification. *Molecular Ecology Resources*, *21*(7), 2565–2579.
- Meiklejohn, K. A., Damaso, N., & Robertson, J. M. (2019). Assessment of BOLD and GenBank—Their accuracy and reliability for the identification of biological materials. *PLoS ONE*, *14*(6), e0217084.
- Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7), 150088.
- Mousavi-Derazmahalleh, M., Stott, A., Lines, R., Peverley, G., Nester, G., Simpson, T., Zawiarta, M., De La Pierre, M., Bunce, M., & Christophersen, C. T. (2021). eDNAFlow, an automated, reproducible and scalable workflow for analysis of environmental DNA

- sequences exploiting nextflow and Singularity. *Molecular Ecology Resources*, 21(5), 1697–1704.
- Pawlowski, J., Apothéoz-Perret-Gentil, L., & Altermatt, F. (2020). Environmental (e)DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, 29, 4258–4264.
- Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéoz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., Feio, M. J., Filipe, A. F., Fornaroli, R., Graf, W., Herder, J., van der Hoorn, B., Jones, I. J., Sagova-Mareckova, M., Moritz, C., ... Kahlert, M. (2018). The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. *Science of The Total Environment*, 637, 1295–1310.
- Pentinsaari, M., Ratnasingham, S., Miller, S. E., & Hebert, P. D. N. (2020). BOLD and GenBank revisited—Do identification errors arise in the lab or in the sequence libraries? *PLoS ONE*, 15(4), e0231814.
- Polanco, A., Richards, E., Valentini, A., Flueck, B., Altermatt, F., Brosse, S., Walser, J. C., Eme, D., Marques, V., Manel, S., Albouy, C., Dejean, T., & Pellissier, L. (2021). Comparing the performance of 12S mitochondrial primers for fish environmental DNA across ecosystems. *Environmental DNA*, 3, 1113–1127.
- Port, J. A., O'Donnell, J. L., Romero-Maraccini, O. C., Leary, P. R., Litvin, S. Y., Nickols, K. J., Yamahara, K. M., & Kelly, R. P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental DNA. *Molecular Ecology*, 25(2), 527–541.
- Porter, T. M., & Hajibabaei, M. (2018). Scaling up: A guide to high-throughput genomic approaches for biodiversity analysis. *Molecular Ecology*, 27(2), 313–338.
- Ritter, C. D., Dal Pont, G., Stica, P. V., Horodesky, A., Cozer, N., Netto, O. S. M., Henn, C., Ostrensky, A., & Pie, M. R. (2022). Wanted not, wasted not: Searching for non-target taxa in environmental DNA metabarcoding by-catch. *Environmental Advances*, 7, 100169.
- Rodríguez-Ezpeleta, N., Zinger, L., Kinziger, A., Bik, H. M., Bonin, A., Coissac, E., Emerson, B. C., Lopes, C. M., Pelletier, T. A., Taberlet, P., & Narum, S. (2021). Biodiversity monitoring using environmental DNA. *Molecular Ecology Resources*, 21(5), 1405–1409.
- Schenecker, T., Schletterer, M., Lecaudey, L. A., & Weiss, S. J. (2020). Reference databases, primer choice, and assay sensitivity for environmental metabarcoding: Lessons learnt from a re-evaluation of an eDNA fish assessment in the Volga headwaters. *River Research and Applications*, 36(7), 1004–1013.
- Schroeter, J. C., Maloy, A. P., Rees, C. B., & Bartron, M. L. (2020). Fish mitochondrial genome sequencing: Expanding genetic resources to support species detection and biodiversity monitoring using environmental DNA. *Conservation Genetics Resources*, 12(3), 433–446.
- Somervuo, P., Yu, D. W., Xu, C. C. Y., Ji, Y., Hultman, J., Wirta, H., & Ovaskainen, O. (2017). Quantifying uncertainty of taxonomic placement in DNA barcoding and metabarcoding. *Methods in Ecology and Evolution/British Ecological Society*, 8(4), 398–407.
- Stamatakis, A. (2014). RAXML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9), 1312–1313.
- Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, 21(8), 2045–2050.
- Taberlet, P., Coissac, E., Pompanon, F., Gielly, L., Miquel, C., Valentini, A., Vermet, T., Corthier, G., Brochmann, C., & Willerslev, E. (2007). Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. *Nucleic Acids Research*, 35(3), e14.
- Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P. F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F., Gaboriaud, C., Jean, P., Poulet, N., Roset, N., Copp, G. H., Geniez, P., Pont, D., Argillier, C., Baudoin, J. M., ... Dejean, T. (2016). Next generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, 25(4), 929–942.
- Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M. F., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wylter, S. A., Bouchez, A., Borja, A., Čiamporová-Zatovičová, Z., Ferreira, S., ... Ekrem, T. (2019). DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. *The Science of the Total Environment*, 678, 499–524.
- Zafeiropoulos, H., Viet, H. Q., Vasileiadou, K., Potirakis, A., Arvanitidis, C., Topalis, P., Pavlodi, C., & Pafilis, E. (2020). PEMA: A flexible pipeline for environmental DNA metabarcoding analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience*, 9(3), gaa022.

SUPPORTING INFORMATION

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

How to cite this article: Blackman, R. C., Walser, J.-C., Rüber, L., Brantschen, J., Villalba, S., Brodersen, J., Seehausen, O., & Altermatt, F. (2022). General principles for assignments of communities from eDNA: Open versus closed taxonomic databases. *Environmental DNA*, 00, 1–17. <https://doi.org/10.1002/edn3.382>