Received: 22 July 2022 | Accepted: 5 December 2022

DOI: 10.1111/1755-0998.13746

MOLECULAR ECOLOGY RESOURCES WILEY

RESOURCE ARTICLE

Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses

François Keck[1] | Marjorie Couton[1] | Florian Altermatt[1,2]

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

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

Correspondence
François Keck and Florian Altermatt, Department of Aquatic Ecology, Eawag: Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland.
Email: francois.keck@gmail.com and florian.altermatt@eawag.ch

Funding information
Schweizerischer Nationalfonds zur Förderung der Wissenschaftlichen Forschung, Grant/Award Number: 31003A_173074; Swiss Federal Office for the Environment (FOEN/BAFU); University of Zurich Research Priority Programme in Global Change and Biodiversity

Handling Editor: Carla Martins Lopes

Abstract

Assessment of biodiversity using metabarcoding data, such as from bulk or environmental DNA sampling, is becoming increasingly relevant in ecology, biodiversity sciences and monitoring. Thereby, the taxonomic identification of species from their DNA sequences relies strongly on reference databases that link genetic sequences to taxonomic names. These databases vary in completeness and availability, depending on the taxonomic group studied and the genetic region targeted. The incompleteness of reference databases is an important argument to explain the nondetection by metabarcoding of species supposedly present. However, there exist further and generally overlooked problems with reference databases that can lead to false or inaccurate inferences of taxonomic assignment. Here, we synthesize all possible problems inherent to reference databases. In particular, we identify a complete, mutually nonexclusive list of seven classes of challenges when it comes to selecting, developing and using a reference database for taxonomic assignment. These are: (i) mislabelling, (ii) sequencing errors, (iii) sequence conflict, (iv) taxonomic conflict, (v) low taxonomic resolution, (vi) missing taxa and (vii) missing intraspecific variants. For each problem identified, we provide a description of possible consequences on the taxonomic assignment process. We illustrate the respective problem with examples taken from the literature or obtained by quantitative analyses of public databases, such as GenBank or BOLD. Finally, we discuss possible solutions to the identified problems and how to navigate them. Only by raising users' awareness of the limitations of metabarcoding data and DNA reference databases will adequate interpretations of these data be achieved.

KEYWORDS
barcoding, DNA reference library, metabarcoding, reference database, taxonomic affiliation

1 | INTRODUCTION

DNA analysis is a powerful approach to detect and identify biological organisms and a potential alternative to traditional methods based on morphological characters (Deiner et al., 2017; Gibson et al., 2015). Pioneering works conducted in the 1990s have applied DNA-based identification systems to a variety of organisms (e.g., Baker & Palumbi, 1994; Brown et al., 1999; Bucklin et al., 1999; Sperling et al., 1994). In 2003, Hebert introduced DNA barcoding as a universal system for animal species identification by using a standard marker (i.e., COI). Building on this idea, DNA metabarcoding allows identification of multiple species from the DNA present in a sample and is becoming a standard across many fields in ecology and biodiversity sciences (Blaxter, 2004; Creer et al., 2016; Galan et al., 2018; Pedersen et al., 2015). By enabling the analysis of large numbers of samples at very high throughput and fine spatial, temporal and taxonomic resolution, these approaches are currently revolutionizing the way we study and monitor biodiversity (Deiner

et al., 2017; Euclide et al., 2021; Ji et al., 2013; Keck et al., 2017), a change which is necessary to address the numerous threats to the biosphere in the Anthropocene.

In both barcoding and metabarcoding approaches, DNA is successively extracted, amplified and sequenced. As opposed to barcoding, for which DNA is directly extracted from isolated specimens, metabarcoding extractions are performed either from bulk samples (i.e., sorted individuals, potentially belonging to multiple species) or from environmental samples, that is, environmental DNA (eDNA) from water, soil, sediment or air (Pawlowski et al., 2020). In metabarcoding, extracted DNA is then amplified using a set of primers delineating a barcode region targeting a specific taxonomic group of interest (Pawlowski et al., 2020; Taberlet et al., 2012) and the product of amplification is sequenced on a high-throughput sequencing (HTS) platform. The choice of the marker as well as the primers is decisive in the conduct of a metabarcoding project and results from a compromise between the specificity of the barcode for the targeted group, its capacity to discriminate taxonomic units at the desired level and its representativeness in available reference databases (Casey et al., 2021; Clarke et al., 2017; Ficetola et al., 2021).

After sequencing, data are processed using a set of bioinformatics tools and methods to prepare and conform them for subsequent analysis. A critical step in bioinformatics pipelines is taxonomic assignment, which is the inference of a taxonomic classification for the analysed sequences. Although the direct analysis of raw genetic sequences is possible (see, e.g., Cordier et al., 2020; Mächler et al., 2021; Marques et al., 2020; Tapolczai et al., 2019), it is often important to identify the species by their taxonomic names. This allows the use of attributes specific to these taxa (e.g., their biological traits and ecological preferences) and, to a certain extent, to link the results of (meta)barcoding analyses with results obtained by traditional approaches. Furthermore, from a conservation biology perspective, including aspects of environmental law, a taxonomic assignment is often essential not only for communication but also for implementing environmental regulations. In fact, "species" is most often the only valued unit when implementing management strategies of endangered species (e.g., the U.S. Endangered Species Act, 1973), IUCN lists of threatened European species (e.g., Bilz et al., 2011), invasive species or pest species. Given that the results of the taxonomic classification will serve as the starting point for downstream analyses, it is important to ensure that data sets based on the assignment of species (or other taxonomic levels) to genetic sequences obtained from heterogeneous and broad metabarcoding are as complete and accurate as possible.

Taxonomic assignment relies on two main elements: first, a reference database, which links DNA sequences to a known taxonomic classification; and second, an algorithm, which uses the reference database to classify new sequences. Although the choice of the algorithm may have an effect on the speed and the outcome of the process (Murali et al., 2018; Allard et al., 2015; Edgar, 2016; Barbera et al., 2019), the performance of these algorithms relies entirely on reference data linking DNA sequence to taxonomy. Therefore, it is obvious that the quality of the taxonomic assignment depends largely on the quality of the reference base being used, which strictly constrains the prediction domain of the algorithms.

Given the importance of reference databases, many projects have been developed over the years to collect, store and distribute reference sequences. GenBank (Benson et al., 2008) is one of the oldest and best known of these projects, and indexes several billions of annotated sequences across the entire tree of life. Other large-scale projects target specific groups of organisms, such as BOLD (Ratnasingham & Hebert, 2007), which targets animals through the COI marker, Silva (Pruesse et al., 2007), which is a generalist database focused on ribosomal RNA sequences (16S, 18S, 23S and 28S markers) or Unite (Nilsson et al., 2019), which is mainly focused on fungi through the ITS marker.

Despite the variety of available choices, reference database quality is often questioned by both expert taxonomists as well as practitioners, and, in particular, the completeness of reference databases is almost systematically mentioned to explain the nondetection of a species by DNA that is otherwise known to be present (false negative). This issue has been reported in many publications investigating diversity in a wide range of biological groups such as vertebrates (e.g., Cilleros et al., 2019; Gold et al., 2021; Lynggaard et al., 2019), invertebrates (e.g., Dowle et al., 2016; Elbrecht et al., 2017; Watts et al., 2019), microbes (e.g., Minerovic et al., 2020; Vasselon et al., 2017) and plants (e.g., Arstingstall et al., 2021; Bell et al., 2017; Gous et al., 2019). Moreover, the completeness and quality of reference databases are known to be geographically and taxonomically biased. For example, Marques et al. (2021) showed that gaps in reference databases of fish species increase towards the tropics; Weigand et al. (2019) found that some taxonomic groups used for the ecological monitoring of aquatic environments in Europe were much better represented in reference databases (e.g., fish, true bugs and freshwater vascular plants) than others (e.g., freshwater diatoms and marine molluscs); and Li et al. (2022) showed that barcode reference libraries covering aquatic taxa in China are often geographically biased to where the specimens/sequences come from. While several factors influence the ability of DNA methods to detect particular species, including primer biases or DNA degradation (Barnes & Turner, 2016; Deiner et al., 2017; Schenekar et al., 2020), it is certain that the limitations of reference databases account for a large proportion of unclassified reads frequently reported (e.g., Haenel et al., 2017; Nunes et al., 2019; Rivera et al., 2018) and the poor congruence observed between DNA-based methods and traditional methods for some taxonomic groups (Keck et al., 2022). Resolving possible limitations or challenges associated with databases is thus crucial to advance the use of metabarcoding in general, and is in particular becoming more relevant in the increasing use of eDNA in ecology and conservation biology. Importantly, there are several factors specific to reference databases that affect taxonomic inference. These are sometimes (or even often) overlooked and ignored, or their effects are simply not well understood by users.

Here, we synthesize the problems associated with reference databases that can compromise the success of the taxonomic identification process. We identify a complete, mutually nonexclusive list of

seven classes of challenges when it comes to selecting, developing and using a reference database for taxonomic assignment, including (i) mislabelling, (ii) sequencing errors, (iii) sequence conflict, (iv) taxonomic conflict, (v) low taxonomic resolution, (vi) missing taxa and (vii) missing intraspecific variants. We first describe each of these seven challenges in detail, both linking to existing literature but also illustrating them with examples from the largest available and broadly used reference databases (GenBank and BOLD). For coherence, the examples given mainly target metazoans with a focus on a 313-bp COI fragment amplified by the primers designed by Leray et al. (2013), as these are the most commonly used COI primers for invertebrates and other groups of animals (Arribas et al., 2022; Duarte et al., 2021). Nevertheless, the ideas discussed are relevant for all groups of organisms and all markers, covering microbes, plants and animals. Only the respective relevancy of these challenges may vary, for example due to different completeness of the respective databases. We explain for each challenge why it exists and why it may cause inaccurate, false or impossible inferences. Finally, we summarize solutions to overcome these problems or limit their impact for each of the seven challenges.

Our goal is to increase awareness of these challenges to those generating and using such metabarcoding data and to give a better understanding of their potential limitations. This applies to scientists who generate reference sequences, to those who develop and maintain reference databases, to those who use these databases to infer taxonomy from their sequencing data, and, ultimately, to anyone who needs to interpret results obtained by (meta)barcoding. Such knowledge is needed in order to create accurate, trustworthy and replicable results from any metabarcoding analysis in the field of ecology and biodiversity sciences, and will be a benchmark for the credibility of the field as a whole.

2 | THE SEVEN CHALLENGES AND POSSIBLE SOLUTIONS

2.1 | Perfect world

In a perfect world, a scientist performing taxonomic classification of unknown DNA sequences generated by barcoding or metabarcoding would have access to a comprehensive and error-free reference database, as depicted in Figure 1 (upper box). This database would contain all the taxa potentially included in the samples analysed (and all sequences generated). These taxa would be identified to the taxonomic level targeted by the scientist and would be classified according to a hierarchical taxonomic system that would closely reflect the phylogenetic signal contained in the DNA barcode sequences. The barcode sequences would be long and variable enough to ensure a unique genetic signature for each taxon. The database would contain no errors in linking DNA sequences to taxonomic labels.

This perfect world, however, does not exist, nor do reference databases that combine all these qualities. As outlined above, we identify a complete, mutually nonexclusive list of seven classes of challenges when selecting, developing and using a reference database for taxonomic assignment. The lower box of Figure 1 gives an overview of these seven challenges, their possible negative impacts on taxonomic inference and the potential solutions that can be implemented to mitigate them. In the following sections, each of them is individually discussed in detail and possible solutions are presented.

In all of these seven cases, a good understanding of the potential but also limitations of the taxonomic assignment conducted is fundamental for an adequate and correct scientific interpretation of the data, and subsequent decisions with respect to management or policy. In the following we describe these challenges and how they can be addressed.

2.2 | Taxonomic mislabelling

2.2.1 | Description

Taxonomic mislabelling corresponds to an error in the taxonomic identification of the biological material deposited in the reference database. The sequence in the database corresponds to the target organism, but its label (and therefore its taxonomy) is incorrect. This can happen at all taxonomic levels (and will then downscale from the level it occurs to all finer levels, respectively), yet is more likely to occur at finer taxonomic levels (Leray et al., 2019). For example, one taxon may be confused with another at the species level because the latter is closely related to it, making the two taxa difficult to distinguish on the basis of morphological characters (e.g., Viard et al., 2019). This problem is also more likely to occur when the person doing the identification is not a skilled taxonomist or when the organism being studied belongs to a particularly complex group, and it is thus more prevalent for taxonomically understudied groups. Finally, taxonomic mislabelling may arise when taxonomy is revised and when taxa are separated or merged into new taxa, yet this is not updated in the reference databases (e.g., Gissi et al., 2017).

Taxonomic mislabelling is a typical example where a mistake done at one step in the process chain has cascading effects and will cause incorrect inferences. We argue that it is one of the most fundamental problems to avoid: it may be better to have gaps (that can be filled later) than taxonomically incorrectly labelled data in a database. The global decline in taxonomists, and the decline in universities and other institutions training taxonomists, will aggravate this problem, yet solving it can also offer new significance and value to taxonomic knowledge (Sheth & Thaker, 2017).

2.2.2 | Possible outcome

Since the sequence of concern is linked to an incorrect taxonomic label, there is a risk of misclassification. In this case, the query sequence correctly matches a sequence in the reference database, but the taxonomic assignment linking them is incorrect because the

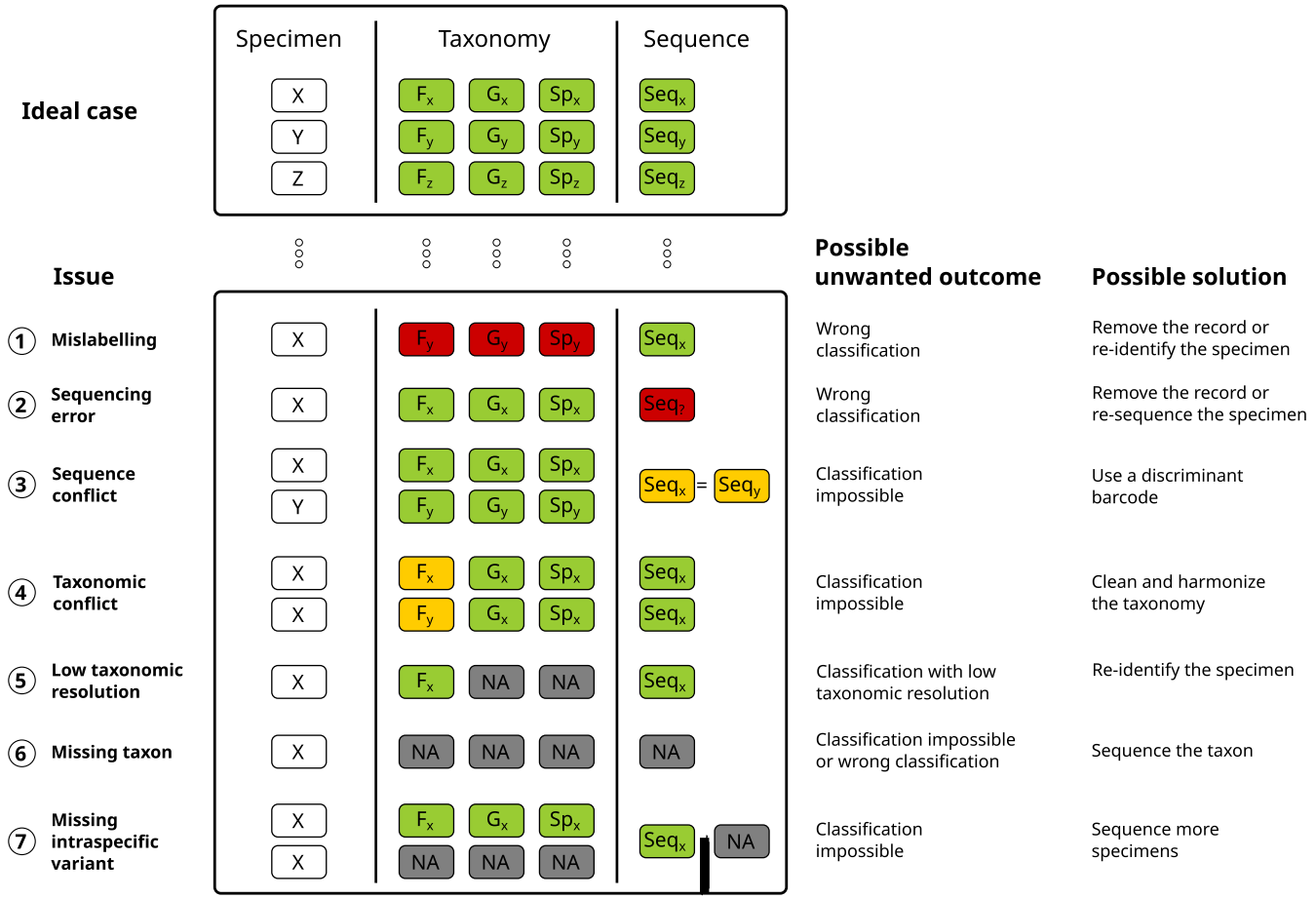

FIGURE 1 Illustration of the ideal case ("perfect world") for which all sequencing information from metabarcoding matches completely to mutually exclusive taxonomic units (upper box). The lower box gives an overview and illustration of all seven possible, mutually nonexclusive, potential challenges with reference databases for taxonomic classification. The frame represents a reference database linking taxonomy to sequences for different specimens. Green boxes show correct data and red boxes incorrect data. Yellow boxes represent correct data that nonetheless result in classification issues. As an example, real taxonomic units, based on specimens, are indexed and labelled as X, Y and Z (each representing a unique species), and their respective taxonomic classification at the level of families (F), genera (G) and species (Sp), as well as the corresponding sequence (seq), is illustrated. While the example here covers for simplicity only a few units (species), the indexing of X, Y and Z in metabarcoding data and databases usually covers thousands to tens of thousands of units, as do the possible challenges described here.

reference database entry is taxonomically not correct. However, given that mislabelling problems usually occur at fine taxonomic levels, the error does not necessarily compromise higher levels of classification (Leray et al., 2019).

2.2.3 | Possible solution

The only solution to this challenge is to have error-free databases, in which each sequence is correctly assigned to an organism with a valid taxonomic name and classification, and in which the database is constantly updated with respect to taxonomic changes (i.e., splitting of species). While the current discussion has been largely about how complete databases are (Li et al., 2022; Weigand et al., 2019), the discussion and focus should actually be about how complete and correct these databases are, and specifically the latter, as detecting and correcting taxonomic mislabelling in existing databases is arduous.

One solution to identify rogue taxa is to perform a cross-validation of the database, where the taxonomy of each sequence is re-inferred using the taxonomy of other sequences and compared to the original. However, this approach can be time- and resource-consuming and requires good coverage of the finest taxonomic levels. Alternatively, if a reference phylogeny is available, one can replace the sequences of the database in the phylogenetic tree and assess if their phylogenetic position is consistent with their taxonomic label (Kozlov et al., 2016). A mislabelled sequence should be deleted from the database unless the original biological material has been preserved and allows for re-identification and revision of the entry concerned. Additionally, user-friendly algorithms to detect taxonomic mislabelling could be implemented upstream, at the deposition step, to help database contributors to rigorously check the sequences they are about to submit against existing databases and verify suspicious results. Performing multilocus sequencing from the same biological material can also improve the confidence in the taxonomic affiliation, provided that at

least one of the sequenced loci is already represented and correctly identified in the reference database.

2.2.4 | Examples

Documenting taxonomic mislabelling requires significant effort and careful investigation, yet is critical for the long-term validity of metabarcoding studies. A general consensus and support (both taxonomically and financially) to reach complete and error-free databases should be the prime goal of the field of metabarcoding research. An interesting example is reported by Viard et al. (2019) who investigated the GenBank labels of three species within the genus *Botrylloides* by comparing them to new sequences obtained from more than 750 colonies. Using phylogenetic clustering analyses, they found that one species, in particular, *B. diegensis*, was systematically mislabelled as *B. leachii* with possible implications for the management of these two non-native species in the Mediterranean Sea. Similarly, Seah et al. (2017) performed a phylogenetic analysis of publicly available sequences for 24 species of the fish family Leiognathidae. From 232 sequences downloaded from BOLD and GenBank, they reported up to 88 sequences with potential misidentification problems.

2.3 | Sequencing error

2.3.1 | Description

Sequencing error is a situation where the biological material is correctly identified but the DNA sequence attached to it is erroneous. This problem can have several origins such as PCR errors and PCR-induced chimeras (Potapov & Ong, 2017) or the amplification of pseudogenes (Buhay, 2009). These are particularly problematic when targeting mitochondrial markers because of the presence of nonfunctional copies of mitochondrial genes in the nuclear genome (NUMTs; Bensasson et al., 2001). These copies, although in theory less abundant than the targeted marker, can accumulate mutations and be as divergent as 36% from their parent sequence (Schultz & Hebert, 2022). The dominance of mitochondrial DNA (mtDNA) in individual-based sequencing should, however, limit the publication of such sequences as references in public databases. Thus the most probable cause of error is the amplification and sequencing of a contaminant. Typical laboratory contaminations are of human and bacterial origin (Siddall et al., 2009), and of organisms that are also used/studied at the respective facilities.

2.3.2 | Possible outcomes

Similar to taxonomic mislabelling, when the link between the sequence and its taxonomy is incorrect, there is a risk of misclassification. In the case of external contamination, the error is likely to occur at a high taxonomic level, which can theoretically allow this type of error to be separated from taxonomic mislabelling. PCR errors, including polymerase misincorporation, structure-induced template-switching, PCR-mediated recombination and DNA damage (Potapov & Ong, 2017), can affect a single nucleotide and be virtually undetectable, or be at the origin of major recombinations, potentially easy to detect but with unpredictable consequences for the taxonomic classification process.

2.3.3 | Possible solutions

Similar to taxonomic mislabelling, it is possible to search for incorrect sequences by reassessing the taxonomy of each entry in the database using all the other entries. It can be hard to disentangle sequencing errors from taxonomic mislabelling, but the two kinds of errors would probably occur at different taxonomic levels. For example, a bacterial sequence identified as an arthropod species would probably be the result of contamination as it is very unlikely that these two organisms can be mixed up. Especially when creating sequences, bacterial contaminations (incorrectly) assigned to the taxon are not uncommon. When detected, it is strongly recommended to remove such entries from the database unless there is a possibility to resequence the original material to fix the error. Here again, making error detection tools available to researchers during sequence submission could help prevent the multiplication of this type of problem in public databases.

2.3.4 | Examples

Leray et al. (2019) assessed the NCBI GenBank database by clustering metazoan mitochondrial encoded sequences at a threshold of 97% similarity. Assuming this threshold should theoretically generate clusters of congeneric species only, they flagged as potential errors all the clusters that regrouped more than one taxon at different taxonomic levels. Although they concluded that most GenBank sequences are correctly labelled (at genus level and above), they nonetheless detected many clusters regrouping sequences belonging to different phyla (375), classes (537) and orders (1610). Heterogeneous clusters at these taxonomic levels could be the result of sequencing errors, including PCR errors, pseudogenes and contaminations. Screening for human or bacterial DNA sequences in NCBI GenBank is sufficient to find several suspicious records. To demonstrate this, we searched for both human and bacterial sequences (two typical types of external contaminations) in NCBI. We blasted one sequence of human COI (MT242596.1) corresponding to the 313-bp fragment mlCOIintF/HCO2198 (Folmer et al., 1994; Leray et al., 2013) against the GenBank database (see Supporting Information S1.1 for detailed methods). The BLAST algorithm detected 15 sequences with >99% identity with MT242596.1 and which are labelled as insect (12), polychaetes (two) or protura (one) (see detailed results in the Supporting Information). These sequences are probably of human origin despite their taxonomic labels which point to very distant

groups. Similarly, we searched for bacterial sequences in three taxonomic groups: Crustacea, Gastropoda and Bivalvia using the DARN tool (Zafeiropoulos et al., 2021) applied on the 313-bp fragment mlCOIintF/HCO2198 of COI. DARN assesses the taxonomic assignment of the sequences against a reference phylogenetic tree of 1593 consensus sequences covering all the three domains of life and thus allows us to classify sequences as being Eukaryotes, Bacteria or Archaea. In total, we found 41, three and eight sequences phylogenetically identified as bacteria but labelled as Crustacea, Gastropoda and Bivalvia, respectively (Supporting Information S1.2).

2.4 | Sequence conflict

2.4.1 | Description

A sequence conflict is a situation where several different taxa are assigned to the exact same genetic sequence. This occurs when the barcode region is not of sufficient resolution to discriminate between two or more species or after an introgression, when part of the genome of one species is integrated in the genome of another.

This challenge can arise because the barcode region chosen is not suitable for the targeted taxonomic group. This can be especially an issue for barcode regions and generic primers used to cover a very (too) broad range of organisms, such that for individual subgroups the resolution is no longer optimal. The challenge is especially prominent for phylogenetically young species that have diverged only recently (as for example in recent radiations), and for which few to no diagnostic barcode regions may exist at all. For example, recent radiations of whitefish (*Coregonus* spp.) are so young that single barcode regions are not able to diagnose distinct species (Feulner & Seehausen, 2019).

2.4.2 | Possible outcomes

Classification may be impossible, especially if there are no other genetic variants in the database for the taxa concerned. For recent radiations, and taxon groups that are nondistinct at the barcode region looked at, a classification is fundamentally impossible. In the case of an incomplete database, sequence conflicts can lead to incorrect taxonomic assignment. To prevent classification error due to lack of taxonomic resolution, West et al. (2020) treated fish species detection as putative unless all other species from the same genus were sequenced.

2.4.3 | Possible solutions

The detection of this problem is simple: it consists in searching for sequences (or clusters of sequences) to which more than one taxon is associated. Since this problem is not the result of an error per se, deleting these entries from the database is not recommended. The solution to this problem is to use a different and better resolving barcode (Leese et al., 2021) or to combine several markers (Corse et al., 2019; Hajibabaei et al., 2019). If conflictual species are known to occur in different geographical locations, a geographical filtering can also be applied to optimize a reference database by excluding species whose range is known to be outside the study area. Finally, an apparent sequence conflict may also be the consequence of a taxonomic misidentification (see 2.2). In that case, the solution is to correct the erroneous taxonomic label or to remove the misidentified sequence from the database.

2.4.4 | Examples

To demonstrate the prevalence of sequence conflicts on short fragments traditionally used in metabarcoding studies, we searched for such conflicts in the sequences labelled as molluscs in the NCBI GenBank database. More specifically, across a data set of 35,008 sequences corresponding to the 313-bp COI fragment mlCOIintF/HCO2198 (Folmer et al., 1994; Leray et al., 2013) commonly used in barcoding, we looked for the identical sequence that was linked to different taxonomic labels (for details on the methods of this example, see Supporting Information S2). We found a total of 434 sequence conflicts. Most of the conflicts happened at the species (294) and genus (110) levels, but some also appeared at higher taxonomic levels, at the family (30), and even order (two) levels. Note that, as mentioned above, it is possible that some of these conflicts are the result of taxonomic mislabelling.

2.5 | Taxonomic conflict

2.5.1 | Description

The same organism is registered several times in the reference database with different upstream taxonomy, hence resulting in a taxonomic conflict. This can be caused by the presence of synonyms in taxonomic names, by different taxonomic systems coexisting in the database, or by different versions of the same taxonomic system (following revisions) coexisting in the database.

2.5.2 | Possible outcomes

Classification may become impossible and classification confidence values may be negatively impacted.

2.5.3 | Possible solutions

The solution to this problem is to harmonize the taxonomy across the whole reference database. This must be done by choosing an appropriate taxonomic system to be applied systematically and by resolving synonymies in a consistent way. Taxonomic databases and catalogues of taxonomic names can help in this process, which can

be long and tedious, depending on the size and diversity of the taxonomic groups included. However, several tools exist that have been specifically developed to assist scientists in this process (Balvočiūtė & Huson, 2017; Grenié et al., 2022).

2.5.4 | Examples

Taxonomic conflicts can happen within the same database. To show this we screened 172,003 sequences of molluscs from the GenBank reference database for potential conflicts (see Supporting Information S3). We found one taxonomic conflict between the two congeneric mollusc species *Lepeta concentrica* (e.g., MZ580712.1) and *Lepeta caeca* (e.g., AB543977.1), which (as of May 2022) have different upstream taxonomies diverging from the subclass level. The fact that only one taxonomic conflict was detected among 172,003 sequences demonstrates the robustness of the GenBank taxonomy, yet also may vary depending on the taxonomic groups looked at. Especially for taxa that are understudied (such as many single-celled eukaryotes covering a wide range of taxonomic groups), this may be more prevalent (see Adl et al., 2012).

However, taxonomic conflicts also and mainly exist between databases, which can generate problems in the case of merging records from different sources. For example, the species labelled as *Alainites muticus* in BOLD are labelled with the genus name *Takobia* in GenBank, with *Takobia* being a synonym of *Alainites*. Sometimes taxonomic conflicts are the result of a fundamental mistake about the type of organism. For example, the species *Psephurus gladius* is identified as a mollusc in BOLD (GBMNA14336-19) but its corresponding record in GenBank (AY571339) is labelled as a fish (Actinopterygii). To illustrate the extent of the problem, we converted the taxonomic classification of 198,445 animals from BOLD to the taxonomic system (i.e., taxonomic names and lineages) of NCBI (Supplementary Information S3). Although most of the taxonomic names remain the same, a significant number (17.1% and 2%, respectively) of species and genus names used in BOLD were not found in the NCBI database and a significant proportion of taxonomic names (3.4% for species names and up to 10% for class names) are different between the two databases (Figure 2).

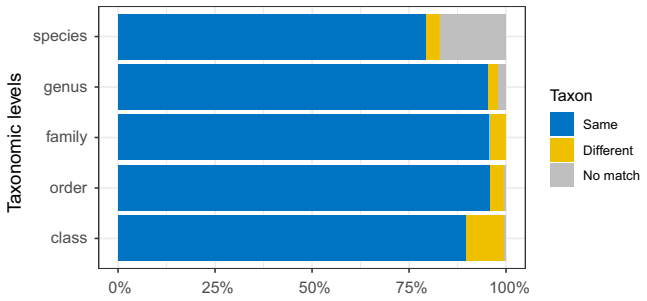

FIGURE 2 Conversion of taxonomic names of animals from BOLD to NCBI taxonomy. Stacked bars represent the proportions, for each taxonomic level, of taxonomic names of BOLD that are the same (blue), different (yellow) or not found (grey) in the NCBI taxonomic database.

2.6 | Low taxonomic resolution

2.6.1 | Description

Database entries have a low taxonomic resolution because organisms have been identified at higher ranks only. This problem is more likely to occur in groups where taxonomic identification is difficult and requires advanced taxonomic skills. Additionally, sequences are sometimes deposited before a formal description, while taxon names are not yet available, thus creating imprecise entries (Garg et al., 2019).

2.6.2 | Possible outcomes

Taxonomic inference at fine resolution is impossible with sequences labelled at higher taxonomic ranks only.

2.6.3 | Possible solutions

The solution to this problem is to re-identify the original material. This can be difficult, either because the original biological material no longer exists or is not accessible, or because the condition of the specimen does not allow a fine taxonomic identification. Importantly, it shows the necessity of long-term collections in which specimens of respective sequence entries are stored and can be re-assessed (Puillandre et al., 2012). If the objective of the study is to reach a fine taxonomic resolution and the clades concerned by these sequences are well represented in the database, removing them can be considered.

2.6.4 | Examples

The number of sequences identified at higher ranks than, for example, the species level depends on the database and the taxonomic group considered. For example, we found that 54% of the records labelled as animals are not identified at the species level in BOLD (see Figure 3a and Supporting Information S4). Similarly, in an extract of 3.06 million COI sequences of arthropods from GenBank, we found that 45.8% of the sequences were not identified at the species level (see Supporting Information S4 for detailed methods). This proportion was 23.6% at the genus level and 2.8% at the family level (Figure 3b).

2.7 | Missing taxa

2.7.1 | Description

Missing taxa refers to a challenge in which all existing taxa which are not present in the reference database constitute a significant limitation for taxonomic inference. There are important biases in

the taxonomic coverage of reference databases (Li et al., 2022; Weigand et al., 2019). The incompleteness of databases and missing taxa is a major problem, yet databases are currently being completed at an unprecedented rate, and thus this challenge may diminish over time.

2.7.2 | Possible outcomes

Taxa that are missing in the reference database cannot be detected in an unknown sample. It is possible to identify missing taxa by comparing the reference database with taxonomic databases and species catalogues. It is also possible that a sequence can be wrongly assigned to a closely related species if the reference for the true species is missing (Couton et al., 2022; Schenekar et al., 2020).

2.7.3 | Solution

The unique solution to this problem is to sequence the missing taxa and add them to the reference database. As new sequences for new organisms are continually added to public databases, it can be expected that the problem will gradually become less important provided that efforts are maintained, especially for the most underrepresented taxonomic groups and geographical regions. Large programmes aimed at genetically sequencing biodiversity are currently underway and mobilize significant financial and human resources such as BIOSCAN (Hobern & Hebert, 2019), Earth BioGenome (Lewin et al., 2022) and Darwin Tree of Life (The Darwin Tree of Life Project Consortium, 2022), which represent a positive sign for the completion of DNA reference databases.

2.7.4 | Examples

Since the completion of reference databases is an important topic, several studies have focused on evaluating their representativeness by comparing them to lists of described species. Schoch et al. (2020) compared the number of formal species in the NCBI Taxonomy database with the number of species in different catalogues. They found that, in 2019, 83% of the described invertebrate taxa were missing from the NCBI Taxonomy. This proportion was 83% for fungi, 62% for green plants, 32% for vertebrates and only 1% for bacteria. Note that these numbers are based only on the taxa already described and are far from representative of the actual availability of sequences for a given barcode and a given taxonomic group. Weigand et al. (2019) analysed gaps in the Barcode of Life Data Systems (BOLD) and NCBI GenBank databases, with a focus on the taxa most frequently used in aquatic biomonitoring. They found that database completeness varies strongly among taxonomic groups. For example, barcode sequences were lacking for many taxa for marine molluscs, ascidians and freshwater diatoms, while other groups (fish, true bugs, caddisflies and vascular plants) are better represented.

2.8 | Missing intraspecific variants

2.8.1 | Description

For a given barcode, it is common to observe different genetic variants (haplotypes) within a single species. For optimal taxonomic classification, it is important that this intraspecific diversity is sufficiently represented in the database. Missing intraspecific variants can be particularly problematic for cryptic species and/or invasive species whose geographical distribution and genetic diversity can be very wide (Rocha et al., 2021).

2.8.2 | Possible outcomes

If only one sequence is available for a species with high genetic variability, it is possible that some haplotypes cannot be correctly identified. Furthermore, some haplotypes of two closely related species may be indistinguishable due to a lack of barcode resolution (see section 2.4, sequence conflicts). In this case, it is important that these haplotypes are present in the database to limit the risk of incorrect classification.

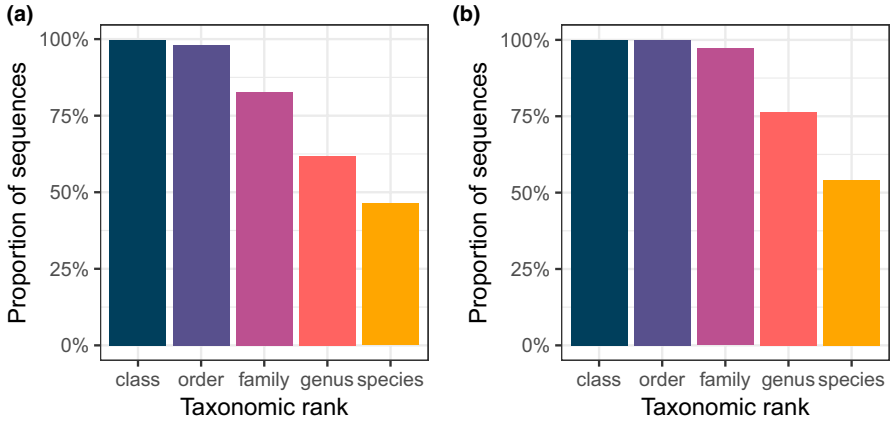

FIGURE 3 The proportion of labelled sequences per taxonomic rank in (a) BOLD specimen records labelled as animals and (b) GenBank sequences labelled as arthropods.

2.8.3 | Solution

The unique solution is to sequence more individuals of the same species, especially individuals from different locations and populations. The multiple studies trying to resolve broad species complexes across the world (e.g., Brunetti et al., 2020; Rocha et al., 2021) are contributing greatly to solving this issue, and continuing this type of work is essential. Using barcodes of individuals from the same region as the study sample can also limit the risks mentioned above.

2.8.4 | Examples

On a GenBank extract of 261,697 unique sequences of the 313-bp COI fragment mlCOIintF/HCO2198 (Folmer et al., 1994; Leray et al., 2013) attributed to 92,525 species of arthropods, we found that 47,762 species (51.6%) were represented by only one single sequence, while only 4% of the species were represented by 10 sequences or more.

In molluscs (30,543 sequences attributed to 4974 species), we found a median number of two sequences available per species (mean = 6.1, see also Figure 4). However, some species are much better represented than others. For example, the invasive snail *Pomacea canaliculata* is represented by 573 sequences, including 26 different variants with 43 variable nucleotide positions across the mlCOIintF/HCO2198 fragment.

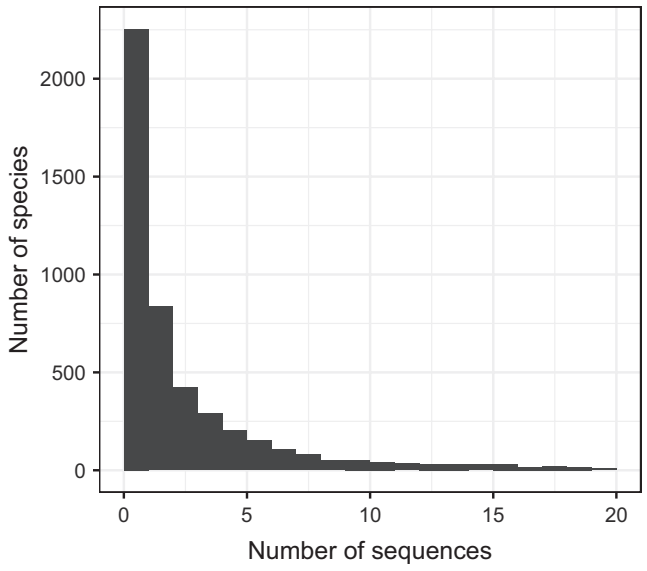

FIGURE 4 Distribution of the number of sequences available per species labelled as molluscs in NCBI GenBank (only species with fewer than 20 sequences are shown).

3 | CONCLUSIONS AND RECOMMENDATIONS

Compiling and using reference databases is a time-consuming and delicate operation and many problems can affect the results of taxonomic classification when using metabarcoding data. With this study, we outline these problems and their possible consequences on the taxa detected by DNA (meta)barcoding. Being familiar with these problems helps to develop the critical thinking skills needed to understand the data and results generated by metabarcoding methods. With the generalization of molecular biology methods for the study of biodiversity, more and more genetic data are being circulated. It is important to remember that these genetic data must be the subject of particular attention. Researchers using and interpreting such molecular data must understand what reference databases were used to produce the taxonomic classification and what their limitations are.

Some challenges may be overcome, such as through the removal of mislabelled sequences in databases. Others, however, such as an incomplete taxonomic resolution of some barcode regions and groups of taxa, are harder to overcome. The latter could be solved by changing the barcode regions looked at, or could be—as in the case of phylogenetically very young species—fundamentally impossible to resolve, as no single barcode region will be able to identify and resolve the taxa of interest. Moreover, the issues discussed here may combine and the intersection of several of them can make the situation even more complex. For example, for species with intraspecific diversity for the barcode studied (missing intraspecific variants), some variants may also be shared with other closely related species (sequence conflict). Some taxonomic groups, especially microbial groups, are particularly challenging. This is due to the great diversity of species they contain, to our limited taxonomic knowledge of them, and to the very concept of species which is not always well defined or compatible with molecular data. However, the requirements for taxonomic resolution vary greatly from study to study, depending on the goals, organisms and ecosystems being studied. Thus, although identification to the species level may be difficult or impossible for some taxonomic groups, the ability to characterize biodiversity at the genus or family level remains extremely useful. In general, the improvement of reference databases requires the development of our knowledge of the taxonomic groups studied, of their taxonomic and phylogenetic classification but also of their ecology and distribution. This also highlights the need for well-trained people with sufficient knowledge of the organismal groups being studied.

The results of the taxonomic classification are a source of valuable information. A very large proportion of unassigned sequences or sequences assigned to very high taxonomic levels may be caused by a reference base of poor quality. Field knowledge and ecological expertise are also important. For example, it may be useful to compare the results of the taxonomic classification with checklists of locally known species, in order to identify doubtful taxa. Knowledge of the ecology of the species studied can also help to identify aberrant results. Regardless, incorrect assignments can have dramatic effects, especially if it relates to species of particular interest such as threatened or invasive species that are subject to management policies (Darling et al., 2020), and all new records of species in a given

area must be checked manually and further proven by a different approach. Finally, it is important that researchers understand the importance and meaning of statistical confidence measures associated with inferred taxonomic groups and use them appropriately (Cristescu & Hebert, 2018).

More generally, addressing the challenges specific to reference databases is a process that will take time and must be done vertically at several levels. First, the completion of reference databases is everyone's concern and it is everyone's responsibility to share publicly the genetic data they produce. More specifically, projects and scientists who add new high-quality entries to public databases must be supported. This includes the entire data production chain, from good laboratory practices to taxonomic expertise and the publication of rich data and metadata adhering to FAIR principles (Rimet et al., 2021). Second, it is important to support consortia and working groups that develop and maintain high-quality reference databases at different scales and for different applications, such as PhytoREF (Decelle et al., 2015) for photosynthetic eukaryotes, PR2 (Guillou et al., 2013) for unicellular eukaryotes, PFR2 (Morard et al., 2015) for foraminifera, dinoref (Mordret et al., 2018) for dinoflagellates and diat.barcode (Rimet et al., 2019) for diatoms. Finally, at the most local level, it is necessary to increase awareness and train bioinformaticians and end-users on the problems specific to reference databases. For example, tools can be implemented to improve the quality control and curation workflows of reference databases such as TAXCI (Rulik et al., 2017), MetaCurator (Richardson et al., 2020), Anacapa (Curd et al., 2019), BCdatabaser (Keller et al., 2020), RESCRIPt (Robeson et al., 2021), DB4Q2 (Dubois et al., 2022), NEA_fish_DB (Claver et al., 2022), CRABS (Jeunen et al., 2022) and refdb (Keck & Altermatt, 2022). The technical solutions discussed in this paper should be used by scientists willing to compile their own database. They could also be implemented and provided by large database repositories to ensure minimum quality control at the deposition step and to prevent the accidental submission of new erroneous sequences. Some databases such as PR2 (Guillou et al., 2013) and MIDORI2 (Leray et al., 2022) already implement such automated tests and filtering.

Taking all these aspects into account should allow us to continue to improve genetic reference databases and to improve the results of taxonomic classification of DNA sequences obtained through (meta)barcoding.

AUTHOR CONTRIBUTIONS

FA, FK and MC designed the research, FK and MC performed the analyses, and FK wrote the manuscript with significant inputs from FA and MC.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for comments on the manuscript. We thank the Swiss National Science Foundation (grant no. 31003A_173074), the Swiss Federal Office for the Environment (FOEN/BAFU), and the University of Zurich Research Priority Programme in Global Change and Biodiversity (URPP GCB) for funding.

CONFLICT OF INTEREST

The authors declared no conflict of interest for this article.

DATA AVAILABILITY STATEMENT

All the data used in this article were retrieved from the NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and BOLD (https://www.boldsystems.org/) public repositories.

ORCID

François Keck https://orcid.org/0000-0002-3323-4167
Marjorie Couton https://orcid.org/0000-0001-9880-8646
Florian Altermatt https://orcid.org/0000-0002-4831-6958

REFERENCES

Adl, S. M., Simpson, A. G. B., Lane, C. E., Lukeš, J., Bass, D., Bowser, S. S., Brown, M. W., Burki, F., Dunthorn, M., Hampl, V., Heiss, A., Hoppenrath, M., Lara, E., Le Gall, L., Lynn, D. H., McManus, H., Mitchell, E. A. D., Mozley-Stanridge, S. E., Parfrey, L. W., … Spiegel, F. W. (2012). The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology*, *59*(5), 429–514. https://doi.org/10.1111/j.1550-7408.2012.00644.x

Allard, G., Ryan, F. J., Jeffery, I. B., & Claesson, M. J. (2015). SPINGO: A rapid species-classifier for microbial amplicon sequences. *BMC Bioinformatics*, *16*(1), 324. https://doi.org/10.1186/s12859-015-0747-1

Arribas, P., Andújar, C., Bohmann, K., deWaard, J. R., Economo, E. P., Elbrecht, V., Geisen, S., Goberna, M., Krehenwinkel, H., Novotny, V., Zinger, L., Creedy, T. J., Meramveliotakis, E., Noguerales, V., Overcast, I., Morlon, H., Papadopoulou, A., & Emerson, B. C. (2022). Toward global integration of biodiversity big data: A harmonized metabarcode data generation module for terrestrial arthropods. *GigaScience*, *11*, giac065. https://doi.org/10.1093/gigascience/giac065

Arstingstall, K. A., DeBano, S. J., Li, X., Wooster, D. E., Rowland, M. M., Burrows, S., & Frost, K. (2021). Capabilities and limitations of using DNA metabarcoding to study plant–pollinator interactions. *Molecular Ecology*, *30*(20), 5266–5297. https://doi.org/10.1111/mec.16112

Baker, C. S., & Palumbi, S. R. (1994). Which whales are hunted? A molecular genetic approach to monitoring whaling. *Science*, *265*(5178), 1538–1539. https://doi.org/10.1126/science.265.5178.1538

Balvočiūtė, M., & Huson, D. H. (2017). SILVA, RDP, Greengenes, NCBI and OTT—How do these taxonomies compare? *BMC Genomics*, *18*(2), 114. https://doi.org/10.1186/s12864-017-3501-4

Barbera, P., Kozlov, A. M., Czech, L., Morel, B., Darriba, D., Flouri, T., & Stamatakis, A. (2019). EPA-ng: Massively parallel evolutionary placement of genetic sequences. *Systematic Biology*, *68*(2), 365–369. https://doi.org/10.1093/sysbio/syy054

Barnes, M. A., & Turner, C. R. (2016). The ecology of environmental DNA and implications for conservation genetics. *Conservation Genetics*, *17*(1), 1–17. https://doi.org/10.1007/s10592-015-0775-4

Bell, K. L., Fowler, J., Burgess, K. S., Dobbs, E. K., Gruenewald, D., Lawley, B., Morozumi, C., & Brosi, B. J. (2017). Applying pollen DNA metabarcoding to the study of plant–pollinator interactions. *Applications in Plant Sciences*, *5*(6), 1600124. https://doi.org/10.3732/apps.1600124

Bensasson, D., Zhang, D.-X., Hartl, D. L., & Hewitt, G. M. (2001). Mitochondrial pseudogenes: Evolution's misplaced witnesses. *Trends in Ecology & Evolution*, *16*(6), 314–321. https://doi.org/10.1016/S0169-5347(01)02151-6

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2008). GenBank. *Nucleic Acids Research*, *36*(Database Issue), D25–D30. https://doi.org/10.1093/nar/gkm929

Bilz, M., Kell, S. P., Maxted, N., & Lansdown, R. V. (2011). European Red List of vascular plants. https://doi.org/10.2779/8515

Blaxter, M. L. (2004). The promise of a DNA taxonomy. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *359*(1444), 669–679. https://doi.org/10.1098/rstb.2003.1447

Brown, B., Emberson, R. M., & Paterson, A. M. (1999). Mitochondrial COI and II provide useful markers for Wiseana (lepidoptera: Hepialidae) species identification. *Bulletin of Entomological Research*, *89*(4), 287–293. https://doi.org/10.1017/S0007485399000437

Brunetti, R., Griggio, F., Mastrototaro, F., Gasparini, F., & Gissi, C. (2020). Toward a resolution of the cosmopolitan Botryllus schlosseri species complex (Ascidiacea, Styelidae): Mitogenomics and morphology of clade E (Botryllus gaiae). *Zoological Journal of the Linnean Society*, *190*(4), 1175–1192. https://doi.org/10.1093/zoolinnean/zlaa023

Bucklin, A., Guarnieri, M., Hill, R. S., Bentley, A. M., & Kaartvedt, S. (1999). Taxonomic and systematic assessment of planktonic copepods using mitochondrial COI sequence variation and competitive, species-specific PCR. In J. P. Zehr & M. A. Voytek (Eds.), *Molecular ecology of aquatic communities* (pp. 239–254). Springer Netherlands. https://doi.org/10.1007/978-94-011-4201-4_18

Buhay, J. E. (2009). "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. *Journal of Crustacean Biology*, *29*(1), 96–110. https://doi.org/10.1651/08-3020.1

Casey, J. M., Ransome, E., Collins, A. G., Mahardini, A., Kurniasih, E. M., Sembiring, A., Schiettekatte, N. M. D., Cahyani, N. K. D., Anggoro, A. W., Moore, M., Uehling, A., Belcaid, M., Barber, P. H., Geller, J. B., & Meyer, C. P. (2021). DNA metabarcoding marker choice skews perception of marine eukaryotic biodiversity. *Environmental DNA*, *3*(6), 1229–1246. https://doi.org/10.1002/edn3.245

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. https://doi.org/10.1111/1755-0998.12900

Clarke, L. J., Beard, J. M., Swadling, K. M., & Deagle, B. E. (2017). Effect of marker choice and thermal cycling protocol on zooplankton DNA metabarcoding studies. *Ecology and Evolution*, *7*(3), 873–883. https://doi.org/10.1002/ece3.2667

Claver, C., Canals, O., de Amézaga, L. G., Mendibil, I., & Rodriguez-Ezpeleta, N. (2022). An automated workflow to assess completeness and curate GenBank for eDNA metabarcoding: The marine fish assemblage as case study. *bioRxiv*. https://doi.org/10.1101/2022.10.26.513819

Cordier, T., Alonso-Sáez, L., Apothéloz-Perret-Gentil, L., Aylagas, E., Bohan, D. A., Bouchez, A., Chariton, A., Creer, S., Frühe, L., Keck, F., Keeley, N., Laroche, O., Leese, F., Pochon, X., Stoeck, T., Pawlowski, J., & Lanzén, A. (2020). Ecosystems monitoring powered by environmental genomics: A review of current strategies with an implementation roadmap. *Molecular Ecology*, *30*, 2937–2958. https://doi.org/10.1111/mec.15472

Corse, E., Tougard, C., Archambaud-Suard, G., Agnèse, J.-F., Messu Mandeng, F. D., Bilong Bilong, C. F., Duneau, D., Zinger, L., Chappaz, R., Xu, C. C. Y., Meglécz, E., & Dubut, V. (2019). One-locus-several-primers: A strategy to improve the taxonomic and haplotypic coverage in diet metabarcoding studies. *Ecology and Evolution*, *9*(8), 4603–4620. https://doi.org/10.1002/ece3.5063

Couton, M., Lévêque, L., Daguin-Thiébaut, C., Comtet, T., & Viard, F. (2022). Water eDNA metabarcoding is effective in detecting non-native species in marinas, but detection errors still hinder its use for passive monitoring. *Biofouling*, *38*(4), 367–383. https://doi.org/10.1080/08927014.2022.2075739

Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008–1018. https://doi.org/10.1111/2041-210X.12574

Cristescu, M. E., & Hebert, P. D. N. (2018). Uses and misuses of environmental DNA in biodiversity science and conservation. *Annual Review of Ecology, Evolution, and Systematics*, *49*(1), 209–230. https://doi.org/10.1146/annurev-ecolsys-110617-062306

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*, *10*(9), 1469–1475. https://doi.org/10.1111/2041-210X.13214

Darling, J. A., Pochon, X., Abbott, C. L., Inglis, G. J., & Zaiko, A. (2020). The risks of using molecular biodiversity data for incidental detection of species of concern. *Diversity and Distributions*, *26*(9), 1116–1121. https://doi.org/10.1111/ddi.13108

Decelle, J., Romac, S., Stern, R. F., Bendif, E. M., Zingone, A., Audic, S., Guiry, M. D., Guillou, L., Tessier, D., Le Gall, F., Gourvil, P., Dos Santos, A. L., Probert, I., Vaulot, D., de Vargas, C., & Christen, R. (2015). PhytoREF: A reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources*, *15*(6), 1435–1445. https://doi.org/10.1111/1755-0998.12401

Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., de Vere, N., Pfrender, M. E., & Bernatchez, L. (2017). Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. *Molecular Ecology*, *26*(21), 5872–5895. https://doi.org/10.1111/mec.14350

Dowle, E. J., Pochon, X. C., Banks, J., Shearer, K., & Wood, S. A. (2016). Targeted gene enrichment and high-throughput sequencing for environmental biomonitoring: A case study using freshwater macroinvertebrates. *Molecular Ecology Resources*, *16*(5), 1240–1254. https://doi.org/10.1111/1755-0998.12488

Duarte, S., Leite, B. R., Feio, M. J., Costa, F. O., & Filipe, A. F. (2021). Integration of DNA-based approaches in aquatic ecological assessment using benthic macroinvertebrates. *Water*, *13*(3), 331. https://doi.org/10.3390/w13030331

Dubois, B., Debode, F., Hautier, L., Hulin, J., Martin, G. S., Delvaux, A., Janssen, E., & Mingeot, D. (2022). A detailed workflow to develop QIIME2-formatted reference databases for taxonomic analysis of DNA metabarcoding data. *BMC Genomic Data*, *23*(1), 53. https://doi.org/10.1186/s12863-022-01067-5

Edgar, R. C. (2016). SINTAX: A simple non-Bayesian taxonomy classifier for 16S and ITS sequences (p. 074161). *bioRxiv*. https://doi.org/10.1101/074161v1

Elbrecht, V., Vamos, E. E., Meissner, K., Aroviita, J., & Leese, F. (2017). Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. *Methods in Ecology and Evolution*, *8*(10), 1265–1275. https://doi.org/10.1111/2041-210X.12789

ESA. (1973). US Endangered Species Act of 1973.

Euclide, P. T., Lor, Y., Spear, M. J., Tajjioui, T., Vander Zanden, J., Larson, W. A., & Amberg, J. J. (2021). Environmental DNA metabarcoding as a tool for biodiversity assessment and monitoring: Reconstructing established fish communities of north-temperate lakes and rivers. *Diversity and Distributions*, *27*(10), 1966–1980. https://doi.org/10.1111/ddi.13253

Feulner, P. G. D., & Seehausen, O. (2019). Genomic insights into the vulnerability of sympatric whitefish species flocks. *Molecular Ecology*, *28*(3), 615–629. https://doi.org/10.1111/mec.14977

Ficetola, G. F., Boyer, F., Valentini, A., Bonin, A., Meyer, A., Dejean, T., Gaboriaud, C., Usseglio-Polatera, P., & Taberlet, P. (2021).

Comparison of markers for the monitoring of freshwater benthic biodiversity through DNA metabarcoding. *Molecular Ecology*, *30*(13), 3189–3202. https://doi.org/10.1111/mec.15632

Folmer, O., Black, M., Hoeh, W., Lutz, R., & Vrijenhoek, R. (1994). DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Molecular Marine Biology and Biotechnology*, *3*(5), 294–299.

Galan, M., Pons, J.-B., Tournayre, O., Pierre, É., Leuchtmann, M., Pontier, D., & Charbonnel, N. (2018). Metabarcoding for the parallel identification of several hundred predators and their prey: Application to bat species diet analysis. *Molecular Ecology Resources*, *18*(3), 474–489. https://doi.org/10.1111/1755-0998.12749

Garg, A., Leipe, D., & Uetz, P. (2019). The disconnect between DNA and species names: Lessons from reptile species in the NCBI taxonomy database. *Zootaxa*, *4706*(3), 401–407. https://doi.org/10.11646/zootaxa.4706.3.1

Gibson, J. F., Shokralla, S., Curry, C., Baird, D. J., Monk, W. A., King, I., & Hajibabaei, M. (2015). Large-scale biomonitoring of remote and threatened ecosystems via high-throughput sequencing. *PLoS One*, *10*(10), e0138432. https://doi.org/10.1371/journal.pone.0138432

Gissi, C., Hastings, K. E. M., Gasparini, F., Stach, T., Pennati, R., & Manni, L. (2017). An unprecedented taxonomic revision of a model organism: The paradigmatic case of Ciona robusta and Ciona intestinalis. *Zoologica Scripta*, *46*(5), 521–522. https://doi.org/10.1111/zsc.12233

Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., Walker, H. J., Jr., Burton, R. S., Kacev, D., Martz, L. D., & Barber, P. H. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resources*, *21*(7), 2546–2564. https://doi.org/10.1111/1755-0998.13450

Gous, A., Swanevelder, D. Z. H., Eardley, C. D., & Willows-Munro, S. (2019). Plant–pollinator interactions over time: Pollen metabarcoding from bees in a historic collection. *Evolutionary Applications*, *12*(2), 187–197. https://doi.org/10.1111/eva.12707

Grenié, M., Berti, E., Carvajal-Quintero, J., Dädlow, G. M. L., Sagouis, A., & Winter, M. (2022). Harmonizing taxon names in biodiversity data: A review of tools, databases and best practices. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13802

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., Del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., … Christen, R. (2013). The protist ribosomal reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/nar/gks1160

Haenel, Q., Holovachov, O., Jondelius, U., Sundberg, P., & Bourlat, S. (2017). NGS-based biodiversity and community structure analysis of meiofaunal eukaryotes in shell sand from Hållö Island, Smögen, and soft mud from Gullmarn Fjord, Sweden. *Biodiversity Data Journal*, *5*, e12731. https://doi.org/10.3897/BDJ.5.e12731

Hajibabaei, M., Porter, T. M., Wright, M., & Rudar, J. (2019). COI metabarcoding primer choice affects richness and recovery of indicator taxa in freshwater systems. *PLoS One*, *14*(9), e0220953. https://doi.org/10.1371/journal.pone.0220953

Hobern, D., & Hebert, P. (2019). BIOSCAN-revealing eukaryote diversity, dynamics, and interactions. *Biodiversity Information Science and Standards*, *3*, e37333.

Jeunen, G.-J., Dowle, E., Edgecombe, J., von Ammon, U., Gemmell, N. J., & Cross, H. (2022). CRABS—A software program to generate curated reference databases for metabarcoding sequencing data. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13741

Ji, Y., Ashton, L., Pedley, S. M., Edwards, D. P., Tang, Y., Nakamura, A., … Yu, D. W. (2013). Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, *16*(10), 1245–1257. https://doi.org/10.1111/ele.12162

Keck, F., & Altermatt, F. (2022). Management of DNA reference libraries for barcoding and metabarcoding studies with the R package refdb. *Molecular Ecology Resources*. 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*(6), 1820–1835. https://doi.org/10.1111/mec.16364

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F., & Bouchez, A. (2017). Freshwater biomonitoring in the information age. *Frontiers in Ecology and the Environment*, *15*(5), 266–274. https://doi.org/10.1002/fee.1490

Keller, A., Hohlfeld, S., Kolter, A., Schultz, J., Gemeinholzer, B., & Ankenbrand, M. J. (2020). BCdataBaser: On-the-fly reference database creation for (meta-)barcoding. *Bioinformatics*, *36*(8), 2630–2631. https://doi.org/10.1093/bioinformatics/btz960

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. https://doi.org/10.1093/nar/gkw396

Leese, F., Sander, M., Buchner, D., Elbrecht, V., Haase, P., & Zizka, V. M. A. (2021). Improved freshwater macroinvertebrate detection from environmental DNA through minimized nontarget amplification. *Environmental DNA*, *3*(1), 261–276. https://doi.org/10.1002/edn3.177

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*, *116*(45), 22651–22656. https://doi.org/10.1073/pnas.1911714116

Leray, M., Knowlton, N., & Machida, R. J. (2022). MIDORI2: A collection of quality controlled, preformatted, and regularly updated reference databases for taxonomic assignment of eukaryotic mitochondrial sequences. *Environmental DNA*, *4*(4), 894–907. https://doi.org/10.1002/edn3.303

Leray, M., Yang, J. Y., Meyer, C. P., Mills, S. C., Agudelo, N., Ranwez, V., Boehm, J. T., & Machida, R. J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: Application for characterizing coral reef fish gut contents. *Frontiers in Zoology*, *10*(1), 34. https://doi.org/10.1186/1742-9994-10-34

Lewin, H. A., Richards, S., Lieberman Aiden, E., Allende, M. L., Archibald, J. M., Bálint, M., Barker, K. B., Baumgartner, B., Belov, K., Bertorelle, G., Blaxter, M. L., Cai, J., Caperello, N. D., Carlson, K., Castilla-Rubio, J. C., Chaw, S.-M., Chen, L., Childers, A. K., Coddington, J. A., … Zhang, G. (2022). The earth BioGenome project 2020: Starting the clock. *Proceedings of the National Academy of Sciences*, *119*(4), e2115635118. https://doi.org/10.1073/pnas.2115635118

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. https://doi.org/10.1016/j.ecolind.2022.108732

Lynggaard, C., Nielsen, M., Santos-Bay, L., Gastauer, M., Oliveira, G., & Bohmann, K. (2019). Vertebrate diversity revealed by metabarcoding of bulk arthropod samples from tropical forests. *Environmental DNA*, *1*(4), 329–341. https://doi.org/10.1002/edn3.34

Mächler, E., Walser, J.-C., & Altermatt, F. (2021). Decision-making and best practices for taxonomy-free environmental DNA metabarcoding in biomonitoring using Hill numbers. *Molecular Ecology*, *30*(13), 3326–3339. https://doi.org/10.1111/mec.15725

Marques, V., Guérin, P.-É., Rocle, M., Valentini, A., Manel, S., Mouillot, D., & Dejean, T. (2020). Blind assessment of vertebrate taxonomic diversity across spatial scales by clustering environmental DNA

metabarcoding sequences. *Ecography*, *43*(12), 1779–1790. https://doi.org/10.1111/ecog.05049

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 and Distributions*, *27*(10), 1880–1892. https://doi.org/10.1111/ddi.13142

Minerovic, A. D., Potapova, M. G., Sales, C. M., Price, J. R., & Enache, M. D. (2020). 18S-V9 DNA metabarcoding detects the effect of water-quality impairment on stream biofilm eukaryotic assemblages. *Ecological Indicators*, *113*, 106225. https://doi.org/10.1016/j.ecolind.2020.106225

Morard, R., Darling, K. F., Mahé, F., Audic, S., Ujiié, Y., Weiner, A. K. M., André, A., Seears, H., Wade, C. M., Quillévéré, F., Douady, C. J., Escarguel, G., de Garidel-Thoron, T., Siccha, M., Kucera, M., & de Vargas, C. (2015). PFR2: A curated database of planktonic foraminifera 18S ribosomal DNA as a resource for studies of plankton ecology, biogeography and evolution. *Molecular Ecology Resources*, *15*(6), 1472–1485. https://doi.org/10.1111/1755-0998.12410

Mordret, S., Piredda, R., Vaulot, D., Montresor, M., Kooistra, W. H. C. F., & Sarno, D. (2018). Dinoref: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. *Molecular Ecology Resources*, *18*(5), 974–987. https://doi.org/10.1111/1755-0998.12781

Murali, A., Bhargava, A., & Wright, E. S. (2018). IDTAXA: A novel approach for accurate taxonomic classification of microbiome sequences. *Microbiome*, *6*(1), 140. https://doi.org/10.1186/s40168-018-0521-5

Nilsson, R. H., Larsson, K.-H., Taylor, A. F. S., Bengtsson-Palme, J., Jeppesen, T. S., Schigel, D., Kennedy, P., Picard, K., Glöckner, F. O., Tedersoo, L., Saar, I., Kõljalg, U., & Abarenkov, K. (2019). The UNITE database for molecular identification of fungi: Handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Research*, *47*(D1), D259–D264. https://doi.org/10.1093/nar/gky1022

Nunes, M., Adams, J., Van Aswegen, S., & Matcher, G. (2019). A comparison between the morphological and molecular approach to identify the benthic diatom community in the St Lucia estuary, South Africa. *African Journal of Marine Science*, *41*(4), 429–442. https://doi.org/10.2989/1814232X.2019.1689169

Pawlowski, J., Apothéloz-Perret-Gentil, L., & Altermatt, F. (2020). Environmental DNA: What's behind the term? Clarifying the terminology and recommendations for its future use in biomonitoring. *Molecular Ecology*, *29*(22), 4258–4264. https://doi.org/10.1111/mec.15643

Pedersen, M. W., Overballe-Petersen, S., Ermini, L., Sarkissian, C. D., Haile, J., Hellstrom, M., Spens, J., Thomsen, P. F., Bohmann, K., Cappellini, E., Schnell, I. B., & Cappellini, E. (2015). Ancient and modern environmental DNA. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1660), 20130383.

Potapov, V., & Ong, J. L. (2017). Examining sources of error in PCR by single-molecule sequencing. *PLoS One*, *12*(1), e0169774. https://doi.org/10.1371/journal.pone.0169774

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., & Glöckner, F. O. (2007). SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, *35*(21), 7188–7196. https://doi.org/10.1093/nar/gkm864

Puillandre, N., Bouchet, P., Boisselier-Dubayle, M.-C., Brisset, J., Buge, B., Cas^{TEL}in, M., Chagnoux, S., Christophe, T., CORBARI, L., Lambourdière, J., Lozouet, P., Marani, G., Rivasseau, A., Silva, N., Terryn, Y., Tillier, S., Utge, J., & Samadi, S. (2012). New taxonomy and old collections: Integrating DNA barcoding into the collection curation process. *Molecular Ecology Resources*, *12*(3), 396–402. https://doi.org/10.1111/j.1755-0998.2011.03105.x

Ratnasingham, S., & Hebert, P. D. N. (2007). Bold: The barcode of life data system (http://www.barcodinglife.org). *Molecular Ecology Notes*, *7*(3), 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Richardson, R. T., Sponsler, D. B., McMinn-Sauder, H., & Johnson, R. M. (2020). MetaCurator: A hidden Markov model-based toolkit for extracting and curating sequences from taxonomically-informative genetic markers. *Methods in Ecology and Evolution*, *11*(1), 181–186. https://doi.org/10.1111/2041-210X.13314

Rimet, F., Aylagas, E., Borja, A., Bouchez, A., Canino, A., Chauvin, C., Teofana Chonova, F. C., Jr., Costa, F. O., Ferrari, B. J. D., Gastineau, R., Goulon, C., Gugger, M., Holzmann, M., Jahn, R., Kahlert, M., Kusber, W.-H., Laplace-Treyture, C., Leese, F., Leliaert, F., … Ekrem, T. (2021). Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. *Metabarcoding and Metagenomics*, *5*, e58056. https://doi.org/10.3897/mbmg.5.58056

Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., Mann, D. G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., & Bouchez, A. (2019). Diat.Barcode, an open-access curated barcode library for diatoms. *Scientific Reports*, *9*(1), 15116. https://doi.org/10.1038/s41598-019-51500-6

Rivera, S. F., Vasselon, V., Jacquet, S., Bouchez, A., Ariztegui, D., & Rimet, F. (2018). Metabarcoding of lake benthic diatoms: From structure assemblages to ecological assessment. *Hydrobiologia*, *807*(1), 37–51. https://doi.org/10.1007/s10750-017-3381-2

Robeson, M. S., O'Rourke, D. R., Kaehler, B. D., Ziemski, M., Dillon, M. R., Foster, J. T., & Bokulich, N. A. (2021). RESCRIPt: Reproducible sequence taxonomy reference database management. *PLoS Computational Biology*, *17*(11), e1009581. https://doi.org/10.1371/journal.pcbi.1009581

Rocha, R. M., Teixeira, J. A., & de Barros, R. C. (2021). Genetic diversity in the Diplosoma listerianum complex (Ascidiacea: Didemnidae) from the Western Atlantic. *Systematics and Biodiversity*, *19*(8), 1149–1163. https://doi.org/10.1080/14772000.2021.1988003

Rulik, B., Eberle, J., von der Mark, L., Thormann, J., Jung, M., Köhler, F., Apfel, W., Weigel, A., Kopetz, A., Köhler, J., Fritzlar, F., Hartmann, M., Hadulla, K., Schmidt, J., Hörren, T., Krebs, D., Theves, F., Eulitz, U., Skale, A., … Ahrens, D. (2017). Using taxonomic consistency with semi-automated data pre-processing for high quality DNA barcodes. *Methods in Ecology and Evolution*, *8*(12), 1878–1887. https://doi.org/10.1111/2041-210X.12824

Schenekar, 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. https://doi.org/10.1002/rra.3610

Schoch, C. L., Ciufo, S., Domrachev, M., Hotton, C. L., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J. P., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI taxonomy: A comprehensive update on curation, resources and tools. Database, 2020, baaa062. https://doi.org/10.1093/database/baaa062

Schultz, J. A., & Hebert, P. D. N. (2022). Do pseudogenes pose a problem for metabarcoding marine animal communities? *Molecular Ecology Resources*, *22*(8), 2897–2914. https://doi.org/10.1111/1755-0998.13667

Seah, Y. G., Ariffin, A. F., & Jaafar, T. N. A. M. (2017). Levels of COI divergence in family Leiognathidae using sequences available in GenBank and BOLD systems: A review on the accuracy of public databases. *Aquaculture, Aquarium, Conservation & Legislation*, *10*(2), 391–401.

Sheth, B. P., & Thaker, V. S. (2017). DNA barcoding and traditional taxonomy: An integrated approach for biodiversity conservation. *Genome*, *60*(7), 618–628. https://doi.org/10.1139/gen-2015-0167

Siddall, M. E., Fontanella, F. M., Watson, S. C., Kvist, S., & Erséus, C. (2009). Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. *Systematic Biology*, *58*(4), 445–451. https://doi.org/10.1093/sysbio/syp033

Sperling, F. A., Anderson, G. S., & Hickey, D. A. (1994). A DNA-based approach to the identification of insect species used for postmortem interval estimation. *Journal of Forensic Sciences*, *39*(2), 418–427.

Taberlet, P., Coissac, E., Hajibabaei, M., & Rieseberg, L. H. (2012). Environmental DNA. *Molecular Ecology*, *21*(8), 1789–1793. https://doi.org/10.1111/j.1365-294X.2012.05542.x

Tapolczai, K., Keck, F., Bouchez, A., Rimet, F., Kahlert, M., & Vasselon, V. (2019). Diatom DNA metabarcoding for biomonitoring: Strategies to avoid major taxonomical and bioinformatical biases limiting molecular indices capacities. *Frontiers in Ecology and Evolution*, *7*, 1–15. https://doi.org/10.3389/fevo.2019.00409

The Darwin Tree of Life Project Consortium. (2022). Sequence locally, think globally: The Darwin tree of life project. *Proceedings of the National Academy of Sciences*, *119*(4), e2115642118. https://doi.org/10.1073/pnas.2115642118

Vasselon, V., Rimet, F., Tapolczai, K., & Bouchez, A. (2017). Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte Island, France). *Ecological Indicators*, *82*, 1–12. https://doi.org/10.1016/j.ecolind.2017.06.024

Viard, F., Roby, C., Turon, X., Bouchemousse, S., & Bishop, J. (2019). Cryptic diversity and database errors challenge non-indigenous species surveys: An illustration with Botrylloides spp. in the English Channel and Mediterranean Sea. *Frontiers in Marine Science*, *6*, 1–13. https://doi.org/10.3389/fmars.2019.00615

Watts, C., Dopheide, A., Holdaway, R., Davis, C., Wood, J., Thornburrow, D., & Dickie, I. A. (2019). DNA metabarcoding as a tool for invertebrate community monitoring: A case study comparison with conventional techniques. *Austral Entomology*, *58*(3), 675–686. https://doi.org/10.1111/aen.12384

Weigand, H., Beermann, A. J., Čiampor, F., Costa, F. O., Csabai, Z., Duarte, S., Geiger, M., Grabowski, M., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A., Willassen, E., Wyler, S., Bouchez, A., Borja, Á., Čiamporová-Zaťovičová, 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. *Science of the Total Environment*, *678*, 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

West, K. M., Stat, M., Harvey, E. S., Skepper, C. L., DiBattista, J. D., Richards, Z. T., Travers, M. J., Newman, S. J., & Bunce, M. (2020). EDNA metabarcoding survey reveals fine-scale coral reef community variation across a remote, tropical Island ecosystem. *Molecular Ecology*, *29*(6), 1069–1086. https://doi.org/10.1111/mec.15382

Zafeiropoulos, H., Gargan, L., Hintikka, S., Pavloudi, C., & Carlsson, J. (2021). The dark mAtteR iNvestigator (DARN) tool: Getting to know the known unknowns in COI amplicon data. *Metabarcoding and Metagenomics*, *5*, e69657. https://doi.org/10.3897/mbmg.5.69657

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: Keck, F., Couton, M., & Altermatt, F. (2022). Navigating the seven challenges of taxonomic reference databases in metabarcoding analyses. *Molecular Ecology Resources*, *00*, 1–14. https://doi.org/10.1111/1755-0998.13746