Collins Rupert A (Orcid ID: 0000-0002-9135-1169)

Meta-Fish-Lib: A generalised, dynamic DNA reference library pipeline for metabarcoding of fishes

Rupert A. Collins[1], Giulia Trauzzi[1,2], Katherine M. Maltby[3], Thomas I. Gibson[4], Frances C. Ratcliffe[5], Jane Hallam[6], Sophie Rainbird[7], James Maclaine[8], Peter A. Henderson[9], David W. Sims[7,10], Stefano Mariani[11,12], and Martin J. Genner[1]

[1] School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK

[2] School of Biological Sciences, Victoria University of Wellington, Wellington, NZ

[3] Centre for Environment, Fisheries and Aquaculture Science (Cefas), Pakefield Road, Lowestoft, NR33 0HT, UK

[4] Molecular Ecology and Fisheries Genetics Laboratory, Bangor University School of Natural Sciences, Environment Centre Wales, Deiniol Road, Bangor, Gwynedd, LL57 2UW, UK

[5] Centre for Sustainable Aquatic Research (CSAR), Swansea University, Swansea, UK

[6] School of Biological and Chemical Sciences, Queen Mary University of London, London, E1 4NS, UK

[7] Marine Biological Association of the United Kingdom, The Laboratory, Citadel Hill, Plymouth PL1 2PB, UK

[8] Department of Life Sciences, The Natural History Museum, Cromwell Road, South Kensington, London, SW7 5BD, UK

[9] Pisces Conservation Ltd, IRC House, The Square, Pennington, Lymington, Hampshire, SO41 8GN, UK

[10] Ocean and Earth Science, University of Southampton, National Oceanography Centre Southampton, European Way, Southampton SO14 3ZH, UK

[11] Ecosystems & Environment Research Centre, School of Environment & Life Sciences, University of Salford, Salford M5 4WT, UK

[12] School of Biological & Environmental Sciences, Liverpool John Moores University, Liverpool, L3 3AF, UK

Corresponding author: Rupert A. Collins, School of Biological Sciences, University of Bristol, Life Sciences Building, Tyndall Avenue, Bristol BS8 1TQ, UK. Email: rupertcollins@gmail.com

Funding information: This work was funded by the Natural Environment Research Council grant NE/N005937/1 (project SeaDNA).

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/jfb.14852

Abstract

The accuracy and reliability of DNA metabarcoding analyses depend on the breadth and quality of the reference libraries that underpin them. However, there are limited options available to obtain and curate the huge volumes of sequence data that are available on public repositories such as NCBI and BOLD. Here, we provide a pipeline to download, clean, and annotate mitochondrial DNA sequence data for a given list of fish species. Features of this pipeline includes: (i) support for multiple metabarcode markers; (ii) searches on species synonyms and taxonomic name validation; (iii) phylogeny assisted quality control for identification and removal of misannotated sequences; (iv) automatically generated coverage reports for each new GenBank release update; and (v) citable, versioned DOIs. As an example we provide a ready-to-use curated reference library for the marine and freshwater fishes of the United Kingdom. To augment this reference library for environmental DNA metabarcoding specifically, we generated 241 new MiFish-12S sequences for 88 UK marine species, and make available new primer sets useful for sequencing these. This brings the coverage of common UK species for the MiFish-12S fragment to 93%, opening new avenues for scaling up fish metabarcoding across wide spatial gradients. The Meta-Fish-Lib reference library and pipeline is hosted at https://github.com/genner-lab/meta-fish-lib.

1 Introduction

DNA barcoding and DNA metabarcoding are increasingly important genetic techniques now employed widely in ecological, biomonitoring, biosecurity, and fisheries research (Gilbey et al., 2021). Both methods allow unique insights into species compositions of a wide range of biological material from aquatic environments. For example, DNA barcoding can confirm the identity of monospecies samples such as seafoods (Wong and Hanner, 2008) or exotic pets (Collins et al., 2012), while DNA metabarcoding can elucidate the composition of complex multispecies substrates such as gut contents or environmental water samples (Taberlet et al., 2012). As techniques are refined and working protocols standardised, environmental DNA (eDNA) analyses are increasingly considered as biomonitoring methodologies appropriate under legal frameworks such as the EU Water Framework Directive and Marine Strategy Framework Directive (Gilbey et al., 2021; Hering et al., 2018). A critically important but neglected aspect of protocol standardisation, however, is that of the sequence reference library (Arranz et al., 2020; Cristescu and Hebert, 2018; Weigand et al., 2019).

Ascertaining the species or higher taxonomic identity of unknown DNA sequences requires a training or reference set of sequences that have a known *a priori* taxonomic structure; these are the

"reference sequences" or the "reference library" (Collins and Cruickshank, 2014). This dataset can be generated directly from tissue samples, but most studies reuse sequence data obtained from public nucleotide sequence databases (Leray et al., 2019). The most commonly used repositories are NCBI GenBank (https://www.ncbi.nlm.nih.gov/genbank/) and Barcode of Life Data System (BOLD; http://www.boldsystems.org/). Taxonomic assignments can be made by querying these databases directly using online tools such as *Blast* (https://blast.ncbi.nlm.nih.gov/Blast.cgi). Here, the user can search the most up-to-date database version, but there are implications for repeatability because the set of reference sequences used to generate matches are not known, and change with each update (Federhen, 2011). There is also no quality control of reference sequences, and unexpected results must be rationalised *post hoc* (Axtner et al., 2019). A more defensible approach is to generate a bespoke reference library for each study, from sequences downloaded from the public databases (Valentini et al., 2016). These studies are repeatable because a copy of the resulting reference library can be deposited as supporting information unique to that publication, and sequences can also be evaluated by the user to ensure quality. However, the methods used to obtain, filter and archive sequences obtained from the public databases are often poorly documented, while the scope for updating and reusing these data are limited. Improvement in these aspects will increase the reliability, flexibility, and transparency of metabarcoding protocols.

To address some of the problems associated with reference library repeatability, a number of excellent tools have been developed in order to create a set of sequences from version-controlled code bases. These include: *Midori* (Machida et al., 2017), *CRUX* (Curd et al., 2019), *BAGS* (Fontes et al., 2021), *CO-ARBitrator* (Heller et al., 2018), *MetaCurator* (Richardson et al., 2020), *Metaxa2* (Bengtsson-Palme et al., 2018), *MARES* (Arranz et al., 2020), and *MitoFish* (Sato et al., 2018). Of these, some solutions are restricted to particular markers such as the standard COI barcode (*BAGS*, *CO-ARBitrator*, *MARES*). Others, such as *Midori* and *MitoFish* contain all mitochondrial loci, but as such it can then be challenging to subsequently subset the sequences that are representative of the metabarcode region of interest. *MetaCurator* and *CRUX* provide targeted reference libraries for user-specified markers, but without an *a priori* set of sequence data, users must download entire copies of NCBI databases locally and run *in silico* PCR, which may become computationally prohibitive to store and process on some machines.

In contrast, obtaining and curating sequence data directly from a restricted list of regional study species is desirable because: (i) taxonomic misassignment increases with geographic scale (Bergsten et al., 2012); (ii) accounting for the species not present in the reference library can increase the reliability of taxonomic assignments (Collins and Cruickshank, 2014; Somervuo et al., 2017); and (iii) searching for species and then subsequently extracting metabarcodes should be computationally

tractable. Additionally, many of the current reference library pipelines produce outputs in formats specific to particular taxonomic assignment software, and also provide no sample metadata together with the sequences, thus limiting options for the further quality control of reference data.

Gap analyses of DNA reference libraries have shown that fishes are among the best represented taxonomic groups for the COI barcode marker, at 82-88% coverage across Europe (Weigand et al., 2019), and 91% in the United Kingdom (Collins et al., 2019). Unfortunately, however, while COI is an excellent marker for many applications, for eDNA metabarcoding of water where target DNA is in low abundance and off-target DNA is high relative abundance, current COI assays offer poor specificity and substantial off-target amplification (Collins et al., 2019). Ribosomal 12S markers—and particularly the MiFish (Miya et al., 2015) and Tele02 (Taberlet et al., 2018) primer sets—perform better, with less off-target amplification and a desirable combination of amplification universality, amplification specificity, and taxonomic discrimination (Collins et al., 2019; Miya et al., 2020). But unlike COI, reference library coverage is poor for MiFish-12S (Collins et al., 2019), with 62% of common UK species represented (versus 97% for COI), and fewer individuals per species available (median of three versus 38). Across Europe only around a third of freshwater fish species have 12S reference sequences (Weigand et al., 2019). In the UK there is a demand for high quality public reference databases for all taxonomic groups, but only around 4% of sequences come from UK specimens held in UK repositories, meaning “the UK lags behind several countries in Europe and North America in that we lack trusted, reliable and openly accessible reference sequences for key UK taxa” (Price et al., 2020). Therefore, until broader initiatives are put in place, there is a need to increase the species coverage of the MiFish-12S marker and facilitate wide-scale eDNA metabarcoding of aquatic environments.

Here, we help to address some of these issues by: (i) developing a reference library pipeline that is generalised (i.e. not specific to any particular metabarcode marker or taxonomy assignment software), dynamic (i.e. easy to update, archive, and cite), annotated, and quality controlled; (ii) providing a curated reference library for the fishes of the United Kingdom as a demonstration of the software; (iii) developing new primer sets to amplify MiFish-12S reference sequences of fishes; and (iv) filling gaps in the UK’s MiFish-12S reference library with new sequence data.

2 Methods

2.1 Data availability

All data, scripts, and instructions to reproduce this work are available from a public repository hosted at https://github.com/genner-lab/meta-fish-lib. The generic DOI “10.5281/zenodo.4443447” resolves

to the most recent version of the repository, while version specific DOIs are found at https://github.com/genner-lab/meta-fish-lib/releases. Sequence data generated as part of this study are available on the NCBI nucleotide database at https://ncbi.nlm.nih.gov/nucleotide (accessions MW818192:MW818432).

2.2 Reference library pipeline

2.2.1 System requirements

Accessing the ready-to-use UK fish reference library in FASTA and CSV format requires only a working R installation (R Core Team, 2020) on any operating system, two loaded packages, and just ten lines of code. The pipeline to assemble a new reference library from scratch runs on Mac and Linux as five executable R scripts in the bash terminal, and is supplied with a tutorial and FAQ. An overview of the pipeline is provided in Figure 1. Scripts are multithreaded and the user is given the option of the number of parallel processing cores to run. In addition to R, the following software is required to be available on the system: *HMMER* (Eddy, 1998), *RAxML* (Stamatakis et al., 2008), and *MAFFT* (Katoh and Standley, 2013). The R package requirements are managed by *renv* (Ushey, 2021), meaning that it is possible to recreate an exact replica of the pipeline independently of any other R package versions installed on the system. An API key from NCBI is also required in order to access their database at a faster rate than a regular user.

2.2.2 Reference library assembly

Assembly of a reference library from scratch broadly follows Collins et al. (2019). First, a species list is required to search against. This list can be provided manually by the user, or via a provided tutorial that automatically creates and formats a species list from FishBase (https://www.fishbase.de/home.htm) for a given country using *rfishbase* (Boettiger et al., 2012), and including all species synonyms. The NCBI GenBank and BOLD databases are then searched using *rentrez* (Winter, 2017) and *bold* (Chamberlain, 2020). The NCBI search uses liberal terms to target mitochondrial loci of interest, e.g. “COI, CO1, cox1,12S, 16S, rRNA, ribosomal, cytochrome, subunit, cytb, COB, CYB, mitochondrial, mitochondrion”. The fragments of DNA homologous to the metabarcode primer sets are then extracted from the dump of sequence data using hidden Markov models in *HMMER* (Eddy, 1998). The marker fragments are then compiled into a single table and annotated with NCBI metadata using *traits* (Chamberlain et al., 2020). Next, the sequences are collapsed to haplotypes by species (taxonomically aware dereplication) and phylogenetic trees are generated for each marker fragment using *RAxML* (Figure 2). The branch tips are then annotated with

the number of haplotypes and coloured according to species monophyly and haplotype sharing, and the trees exported as PDF. These trees must be reviewed manually by the user in order to identify misannotated sequences, i.e. those with incorrect species names. Accessions that are deemed dubious can then be added to a "blacklist" file. This blacklist is then called each time the reference library is loaded, and the misannotated accessions are automatically removed along with those that contain terms in the GenBank description such as "unverified", "similar to", and "-like".

2.3 UK fish reference library

2.3.1 Species list

We compiled a list of marine and freshwater species from the United Kingdom from three sources: (i) the Global Biodiversity Information Facility (https://www.gbif.org/); (ii) FishBase; and (iii) the European Water Framework Directive United Kingdom Technical Advisory Group list of transitional fish species (https://www.wfduk.org/resources/transitional-waters-fish; Annex 1). This list was then validated in the pipeline following FishBase taxonomy and synonyms, and searched and quality controlled as outlined above. To provide a digestible summary, species were arbitrarily labelled as "common" if they are frequently encountered inshore marine species or widespread freshwater species, and otherwise as "rare" (generally deep sea, oceanic or range restricted species).

2.3.2 Tissue sampling, PCR and sequencing

Fin and muscle tissue samples of UK marine fishes were obtained from: (i) Marine Biological Association standard trawl surveys (for methodology see Genner et al., 2010); (ii) fish impingement surveys of power stations by Pisces Conservation Ltd. (for methodology see Collins et al., 2019); and (iii) CEFAS (Centre for Environment, Fisheries, and Aquaculture Science) 2017 Eastern English Channel Beam Trawl Survey (ICES, 2019). Tissues were either preserved in absolute ethanol, or frozen directly at −20°C. Voucher material for these tissues was fixed in either absolute ethanol or 5% formalin solution, and subsequently stored in 70% ethanol. Taxonomic identification of the voucher material followed Henderson (2014).

Isolation of genomic DNA followed a simple lysis-precipitation protocol (https://github.com/genner-lab/Molecular-Lab-Protocols). PCR reactions were then conducted in 20 μL reactions using 10 μL GoTaq G2 Green Master Mix (Promega M7822), 2 μL forward and reserve primer (2 μM), and 50 ng template DNA. Primer pairs to amplify a partial 12S fragment containing the MiFish-12S fragment are presented in Table 1. We first used the Aa22-PheF/Aa633-12sR primer pair, followed by the MiFish-U primer pair if those amplifications failed. Thermocycling parameters

followed polymerase manufacturer's instructions with annealing temperatures from Table 1, and were carried out on an Eppendorf Nexus machine. Amplicons were then purified using spin columns (Zymo C1004-250), and Sanger sequenced using the Aa633-12sR primer and the Eurofins Genomics PlateSeq service, according to manufacturer's instructions. Amplicons sequenced with the MiFish-U primer pair were sequenced in both directions. Cromatograms were assembled into contigs with *Geneious* v8.8.1 (Kearse et al., 2012), and checked for contamination or mislabelling using phylogenetic trees and NCBI *Blast*.

2.3.3 Ethical statement

The collection of animals for study was part of standard fish surveying procedures and complied with the guidelines and policies as approved by the Marine Biological Association, Pisces Conservation Ltd., and the Centre for Environment, Fisheries, and Aquaculture Science (CEFAS).

3 Results

3.1 UK fish reference library

3.1.1 Database search and library coverage

Our compilation of UK marine and freshwater species identified a total of 530 accepted scientific names, and a further 3,733 synonyms. The NCBI GenBank and BOLD databases were searched on 13 January 2021 (GenBank release 241), and retrieved 51,748 accessions. After quality control 49,233 accessions from 492 unique species corresponding to the eight primer sets listed in Table 2 remained (2,515 removed). Search, assembly and annotation of the reference library took around two hours on an average specification Ubuntu Linux desktop machine (i7-3820; 8×3.60 GHz; 16 GB RAM). The phylogenetic quality control step took around eight hours to complete on the same machine using eight processing cores, with the COI datasets taking the longest amount of time.

The Ward et al. (2005) fish barcoding primers for COI yielded the greatest number of sequences at 28,297, resulting in 98% of common species and 92% of all species covered, with a median of 10 haplotypes per species and 4% represented by one sequence (Table 2). The MiFish-12S primer set covered 76% of common species, 70% of all species, median of one haplotype per species, and 25% of species represented by one sequence.

3.1.2 MiFish-12S sequencing

A total of 241 MiFish-12S sequences were obtained from the identified voucher specimens collected. These sequences represent 88 species, 30 of which were not available on GenBank. This raises the coverage of MiFish-12S references for UK species from 370 to 400, and represents an increase from 70% to 75% of the 530 total species. Common species increased from 134 to 164 (76% to 93% of the 176 common species). The sequences were uploaded to GenBank under the accession numbers MW818192:MW818432 (Table S1).

The Aa22-PheF/Aa633-12sR primer set developed here (Table 1) amplifies the MiFish-12S region and its priming sites in one sequencing reaction. This set successfully amplified all marine species sampled here, with the exception of callionymids and syngnathids, which needed to be amplified with the MiFish-U primers in two sequencing reactions in order to capture the full metabarcode (minus the priming sites). The Aa22-PheF/Aa633-12sR primer set can amplify the 12S-V5 region (Riaz et al., 2011) if additionally sequenced in the forward direction. The 12S primer sets of Hänfling et al. (2016) and Stoeckle et al. (2018) amplify the MiFish-12S fragment, the 12S-V5 fragment, and the Valentini et al. (2016) L1848/H1913 teleo region if sequenced in both directions. These primer sets allow several options for *de novo* generation of reference sequence data from new and archived tissue collections.

4 Discussion

Here, we provide the Meta-Fish-Lib pipeline to obtain, curate and archive reference sequence data from GenBank and BOLD for a given list of fish species. A species-focused pipeline is most useful for situations where a reasonably large proportion of the species expected in a study have data present on public repositories. In situations where the target community is poorly represented or not known, such as in the case of some hyperdiverse tropical systems, the user can search for genera rather than species in order to build a reference library better suited to higher taxonomic assignment. When a species list is not available to the user, the pipeline can use FishBase to generate taxonomically validated lists of species names for a given country.

While the NCBI taxonomy database can resolve synonyms, the step of including these in the search stage and subsequently also validating names makes the pipeline robust to changes in taxonomy. The user is informed where changes have taken place, and can also make their own custom changes. The pipeline uses generous search terms to obtain mitochondrial sequence data, thereby making the search process additionally robust to idiosyncrasies in sequence annotation, such as “COI” versus “CO1” or “COX1”. Since several different primer sets can be used to amplify loci such as 12S

(Table 1), and because there are several different metabarcode regions within it (Table 2), it is difficult to know if a sequence annotated as "12S" contains the marker fragment of interest. Here, hidden Markov models isolate specific metabarcode fragments from the dump of sequence data, thus eliminating superfluous nucleotides that can increase classification errors (Richardson et al., 2020). The pipeline is not limited to any particular DNA barcode or DNA metabarcode primer set. Currently there are eight popular primer sets for fishes implemented (Table 2), but additional sets can be added as necessary, or removed to reduce computational load.

A phylogenetic quality control step is also included, and while this is the most time consuming step to perform, it is arguably one of the most important given the sensitivity of species level assignments to misannotated reference data (Locatelli et al., 2020). Here, phylogenetic trees are generated for each primer set, and the trees are annotated by interspecific haplotype sharing and non-monophyly (Figure 2). It is then the task of the user to employ these resources to assist in identifying accessions that have potentially been misannotated. In the examples illustrated in Figure 2, the shared COI haplotypes of *Conger conger*, *Scyliorhinus stellaris*, *Squalus acanthias* and *Galeus melastomus* are likely to have been misannotated based on the evidence that their putative conspecifics are represented elsewhere in the tree by sequences from multiple studies; for the shared 12S *Alosa* haplotypes, in the absence of information from other sequence data, these are most likely explained by them being closely related congeneric species (Bloom et al., 2018). Advantages of this manual approach are that the user visualises the data, can focus on the taxa that are of particular importance to them, and can flexibly apply their own criteria to exclude sequences. However, as indicated in the examples in Figure 2, there are disadvantages: (i) determining misidentifications among closely related congeneric species can be challenging because some may naturally share haplotypes—especially so for the rRNA loci that are less variable than COI—and may require specialist taxonomic expertise to clarify (Leray et al., 2020); (ii) there may be insufficient numbers of sequences available for some species to reach a reasonable conclusion; (iii) while blacklisted sequences are stored and archived as part of the reference library, the determinations are subjective according to the user's criteria; and (iv) generating phylogenetic trees is not scalable for very large numbers of sequences. Regarding this final point, for datasets with hundreds of thousands of sequences, users are recommended to remove metabarcodes that they are not interested in, and to split their input list of species into more manageable partitions and merge the tables once completed. In terms of automating quality control to improve repeatability, a barcode audit and grading system (Fontes et al., 2021; Oliveira et al., 2016) or software to detect misannotations (Kozlov et al., 2016) would be possible to implement in future versions.

Unlike other software, reference libraries produced by this pipeline are additionally annotated with metadata from NCBI, BOLD, and FishBase, including higher taxonomic ranks, voucher information (institution, catalogue number), collection information (country, longitude, latitude), publication information (journal, title, lead author), and accession information (date uploaded, GenBank release version). As the dataset grows, users can be increasingly selective over which accessions are used as references, preferring those for example that have voucher material, are published, or were collected in the study region (Price et al., 2020). Once the reference library is assessed and quality controlled, a summary document is compiled containing important statistics. This includes a primer set coverage table, a table of species and the number of sequences for each primer set, and a table of new sequences that were not present in the previous version of the library. At this point, the library can be archived to a GitHub repository and a DOI obtained to enable that exact version of the library to be cited in a publication or report.

For the fishes of the United Kingdom we assembled an extensive reference library for eight metabarcode markers from COI, 12S, 16S and cytochrome *b*, with a total of 49,233 accessions and 492 species after quality control. This UK fish reference library is quality controlled and ready-to-use, and is archived with DOIs for recent and previous GenBank releases. With 98% of common species covered and a median of 10 haplotypes per species, the COI references represent an unmatched resource for DNA barcoding and metabarcoding of the regional ichthyofauna. Coverage for the MiFish-12S primer set, however, which is more effective for eDNA metabarcoding than COI (Collins et al., 2019; Miya et al., 2020), was considerably lower, at 76% common species coverage, with 25% of those species represented by only one sequence. In order to help address this deficit we obtained 241 MiFish-12S sequences from 88 marine species, 30 of which were previously unrepresented in the most recent GenBank release (241), and includes common species such as *Platichthys flesus*, *Scophthalmus rhombus*, *Lipophrys pholis*, *Callionymus lyra* and *Scyliorhinus stellaris*. These new sequences, plus the recent contributions to GenBank, translate to an increase in common species coverage from 62% (Collins et al., 2019), to 93% here. We also present new and previously published primer sets to facilitate amplification of both the MiFish-12S reference sequences and the priming sites which are useful for quantifying primer bias (Collins et al., 2019).

Implementation of eDNA metabarcoding as a standardised aquatic survey tool is impeded by the availability of suitable sequence reference libraries, and particularly so for its deployment as part of a legal monitoring framework (Weigand et al., 2019). Quality controlled and densely sampled region-specific reference libraries detect more taxa, more reliably (Stoeckle et al., 2020), and large scale metabarcoding projects across temporal or spatial gradients require complete species coverage to allow for the reliable detection of taxa characteristic of different environments. As well as providing a

general solution for assembling and curating reference libraries for fishes, this study builds on previous reference libraries for UK and European fishes (Knebelsberger et al., 2014; Oliveira et al., 2016) by both expanding the choice of metabarcode marker beyond the standard COI barcode, and by additionally providing new MiFish-12S reference sequences for 88 European marine species. We therefore expect this resource to significantly expand the reach and accuracy of DNA metabarcoding studies in the North-East Atlantic, and pave the way for a more robust approach to DNA-based biomonitoring across the globe.

Acknowledgements

This work was funded by the Natural Environment Research Council grant NE/N005937/1 (project SeaDNA). Thomas Gibson would like to acknowledge the work of Jim Drewery and the other scientists and crew of the *MRV Scotia* (Marine Scotland Science) in collecting his tissue samples, although they were not sequenced in time to be included in this publication.

Author contributions

M.J.G., S.M., D.W.S., and P.A.H. conceived the study and obtained funding; R.A.C., S.R., and K.M.M. conducted fieldwork. R.A.C., G.T., F.C.R., T.I.G., and J.H. conducted laboratory work. R.A.C. conducted the analyses and wrote the software. R.A.C. wrote the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Supporting information

Table S1. Darwin Core formatted CSV table of fish specimens sequenced as part of this work, including NCBI GenBank accessions and metadata.

References

Arranz, V., Pearman, W. S., Aguirre, J. D., & Liggins, L. (2020). MARES, a replicable pipeline and curated reference database for marine eukaryote metabarcoding. *Scientific Data*, *7*, 1–8. https://doi.org/10.1038/s41597-020-0549-9.

Axtner, J., Crampton-Platt, A., Hörig, L. A., Mohamed, A., Xu, C. C. Y., Yu, D. W., & Wilting, A. (2019). An efficient and robust laboratory workflow and tetrapod database for larger scale environmental DNA studies. *GigaScience*, *8*, 1–17. https://doi.org/10.1093/gigascience/giz029.

Bengtsson-Palme, J., Richardson, R. T., Meola, M., Wurzbacher, C., Tremblay, É. D., Thorell, K., Kanger, K., Eriksson, K. M., Bilodeau, G. J., Johnson, R. M., Hartmann, M., & Nilsson, R. H. (2018). Metaxa2 database builder: enabling taxonomic identification from metagenomic or metabarcoding data using any genetic marker. *Bioinformatics*, *34*, 4027–4033. https://doi.org/10.1093/bioinformatics/bty482.

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*, 851–869. https://doi.org/10.1093/sysbio/sys037.

Berry, T. E., Osterrieder, S. K., Murray, D. C., Coghlan, M. L., Richardson, A. J., Grealy, A. K., Stat, M., Bejder, L., & Bunce, M. (2017). Metabarcoding for diet analysis and biodiversity: A case study using the endangered Australian sea lion (*Neophoca cinerea*). *Ecology and Evolution*, *7*, 5435–5453. https://doi.org/10.1002/ece3.3123.

Bloom, D. D., Burns, M. D., & Schriever, T. A. (2018). Evolution of body size and trophic position in migratory fishes: a phylogenetic comparative analysis of Clupeiformes (anchovies, herring, shad and allies). *Biological Journal of the Linnean Society*, *125*, 302–314. https://doi.org/10.1093/biolinnean/bly106.

Boettiger, C., Lang, D. T., & Wainwright, P. C. (2012). rfishbase: exploring , manipulating and visualizing FishBase data from R. *Journal of Fish Biology*, *81*, 2030–2039. https://doi.org/10.1111/j.1095-8649.2012.03464.x.

Chamberlain, S. (2020). bold: Interface to Bold Systems API. https://cran.r-project.org/package=bold. R package version 1.1.0.

Chamberlain, S., Foster, Z., Bartomeus, I., LeBauer, D., Black, C., & Harris, D. (2020). traits: species trait data from around the web. https://github.com/ropensci/traits. R package version 0.5.0.

Collins, R. A., Armstrong, K. F., Meier, R., Yi, Y., Brown, S. D. J., Cruickshank, R. H., Keeling, S., & Johnston, C. (2012). Barcoding and border biosecurity: identifying cyprinid fishes in the aquarium trade. *PLoS ONE*, *7*, e28381. https://doi.org/10.1371/journal.pone.0028381.

Collins, R. A., Bakker, J., Wangensteen, O. S., Soto, A. Z., Corrigan, L., Sims, D. W., Genner, M. J., & Mariani, S. (2019). Non-specific amplification compromises environmental DNA metabarcoding with COI. *Methods in Ecology and Evolution*, *10*, 1985–2001. https://doi.org/10.1111/2041-210X.13276.

Collins, R. A. & Cruickshank, R. H. (2014). Known knowns, known unknowns, unknown unknowns and unknown knowns in DNA barcoding: A comment on Dowton et al. *Systematic Biology*, *63*, 1005–1009. https://doi.org/10.1093/sysbio/syu060.

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

Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics*, *14*, 755–763. https://doi.org/10.1093/bioinformatics/14.9.755.

Federhen, S. (2011). Comment on 'Birdstrikes and barcoding: can DNA methods help make the airways safer?'. *Molecular Ecology Resources*, *11*, 937–938. https://doi.org/10.1111/j.1755-0998.2011.03054.x.

Fontes, J. T., Vieira, P. E., Ekrem, T., Soares, P., & Costa, F. O. (2021). BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. *Molecular Ecology Resources*, *21*, 573–583. https://doi.org/10.1111/1755-0998.13262.

Genner, M. J., Sims, D. W., Southward, A. J., Budd, G. C., Masterson, P., McHugh, M., Rendle, P., Southall, E. J., Wearmouth, V. J., & Hawkins, S. J. (2010). Body size-dependent responses of a marine fish assemblage to climate change and fishing over a century-long scale. *Global Change Biology*, *16*, 517–527. https://doi.org/10.1111/j.1365-2486.2009.02027.x.

Gilbey, J., Carvalho, G., Castilho, R., Coscia, I., Coulson, M. W., Dahle, G., Derycke, S., Francisco, S. M., Helyar, S. J., Johansen, T., Junge, C., Layton, K. K. S., Martinsohn, J., Matejusova, I., Robalo, J. I., Rodríguez-Ezpeleta, N., Silva, G., Strammer, I., Vasemägi, A., & Volckaert, F. A. M. (2021). Life in a drop: Sampling

environmental DNA for marine fishery management and ecosystem monitoring. *Marine Policy*, *124*, 104331. https://doi.org/10.1016/j.marpol.2020.104331.

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*, 3101–3119. https://doi.org/10.1111/mec.13660.

Heller, P., Casaletto, J., Ruiz, G., & Geller, J. (2018). A database of metazoan cytochrome *c* oxidase subunit I gene sequences derived from GenBank with CO-ARBitrator. *Scientific Data*, *5*, 1–7. https://doi.org/10.1038/sdata.2018.156.

Henderson, P. A. (2014). *Identification Guide to the Inshore Fish of the British Isles*. Pisces Conservation Ltd., Pennington, UK.

Hering, D., Borja, A., Jones, J. I., Pont, D., Boets, P., Bouchez, A., Bruce, K., Drakare, S., Hänfling, B., Kahlert, M., Leese, F., Meissner, K., Mergen, P., Reyjol, Y., Segurado, P., Vogler, A., & Kelly, M. (2018). Implementation options for DNA-based identification into ecological status assessment under the European Water Framework Directive. *Water Research*, *138*, 192–205. https://doi.org/10.1016/j.watres.2018.03.003.

ICES (2019). *Manual for the Offshore Beam Trawl Surveys, Version 3.4, April 2019, Working Group on Beam Trawl Surveys*. DOI: http://doi.org/10.17895/ices.pub.5353.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, *30*, 772–780. https://doi.org/10.1093/molbev/mst010.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., Duran, C., Thierer, T., Ashton, B., Meintjes, P., & Drummond, A. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*, 1647–1649. https://doi.org/10.1093/bioinformatics/bts199.

Knebelsberger, T., Landi, M., Neumann, H., Kloppmann, M., Sell, A. F., Campbell, P. D., Laakmann, S., Raupach, M. J., Carvalho, G. R., & Costa, F. O. (2014). A reliable DNA barcode reference library for the identification of the North European shelf fish fauna. *Molecular Ecology Resources*, *14*, 1060–1071. https://doi.org/10.1111/1755-0998.12238.

Kozlov, A., Zhang, J., Yilmaz, P., Glöckner, F. O., & Stamatakis, A. (2016). Phylogeny-aware identification and correction of taxonomically mislabeled sequences. *Nucleic Acids Research*, *2*, 5022–5033. https://doi.org/10.1101/042200.

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*, 22651–22656. https://doi.org/10.1073/pnas.1911714116.

Leray, M., Knowlton, N., Ho, S. L., Nguyen, B. N., & Machida, R. J. (2020). Evaluating species-level accuracy of GenBank metazoan sequences will require experts' effort in each group. *Proceedings of the National Academy of Sciences of the United States of America*, *117*, 32213–32214. https://doi.org/10.1073/pnas.2019903117.

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*, 34. https://doi.org/10.1186/1742-9994-10-34.

Locatelli, N. S., McIntyre, P. B., Therkildsen, 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*, 32211–32212. https://doi.org/10.1073/pnas.2007421117.

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*, 1–7. https://doi.org/10.1038/sdata.2017.27.

Minamoto, T., Yamanaka, H., Takahara, T., Honjo, M. N., & Kawabata, Z. (2012). Surveillance of fish species composition using environmental DNA. *Limnology*, *13*, 193–197. https://doi.org/10.1007/s10201-011-0362-4

Miya, M., Gotoh, R. O., & Sado, T. (2020). MiFish metabarcoding: a high-throughput approach for simultaneous detection of multiple fish species from environmental DNA and other samples. *Fisheries Science*, *86*, 939–970. https://doi.org/10.1007/s12562-020-01461-x.

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*, 150088. https://doi.org/10.1098/rsos.150088.

Oliveira, L. M., Knebelsberger, T., Landi, M., Soares, P., Raupach, M. J., & Costa, F. O. (2016). Assembling and auditing a comprehensive DNA barcode reference library for European marine fishes. *Journal of Fish Biology*, *89*, 2741–2754. https://doi.org/10.1111/jfb.13169.

Price, B. W., Briscoe, A. G., Misra, R., & Broad, G. (2020). DEFRA Centre of Excellence for DNA Methods: Evaluation of DNA barcode libraries used in the UK and developing an action plan to fill priority gaps. Technical report, ISBN: 9781783546718, http://publications.naturalengland.org.uk/.

R Core Team (2020). R: A Language and Environment for Statistical Computing. https://www.r-project.org/. R version 4.0.3.

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

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*, 181–186. https://doi.org/10.1111/2041-210X.13314.

Sato, Y., Miya, M., Fukunaga, T., Sado, T., & Iwasaki, W. (2018). MitoFish and mifish pipeline: A mitochondrial genome database of fish with an analysis pipeline for environmental DNA metabarcoding. *Molecular Biology and Evolution*, *35*, 1553–1555. https://doi.org/10.1093/molbev/msy074.

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*, *8*, 398–407. https://doi.org/10.1111/2041-210X.12721.

Stamatakis, A., Hoover, P., & Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML Web servers. *Systematic Biology*, *57*, 758–771. https://doi.org/10.1080/10635150802429642.

Stoeckle, M. Y., Das Mishu, M., & Charlop-Powers, Z. (2020). Improved environmental DNA reference library detects overlooked marine fishes in New Jersey, United States. *Frontiers in Marine Science*, *7*, 226. https://doi.org/10.3389/fmars.2020.00226.

Stoeckle, M. Y., Mishu, M. D., & Charlop-Powers, Z. (2018). GoFish: a streamlined environmental DNA presence/absence assay for marine vertebrates. *PLoS ONE*, *13*, e0198717. https://doi.org/10.1101/331322.

Taberlet, P., Bonin, A., Zinger, L., & Coissac, E. (2018). *Environmental DNA: For Biodiversity Research and Monitoring*. Oxford University Press, Oxford. https://doi.org/10.1093/oso/9780198767220.001.0001.

Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C., & Willerslev, E. (2012). Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, *21*, 2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x.

Ushey, K. (2021). renv: Project Environments. https://cran.r-project.org/package=renv. R package version 0.12.5.

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., Peroux, T., Crivelli, A. J., Olivier, A., Acqueberge, M., Le Brun, M., Møller, P. R., Willerslev, E., & Dejean, T. (2016). Next-generation monitoring of aquatic biodiversity using environmental DNA metabarcoding. *Molecular Ecology*, *25*, 929–942. https://doi.org/10.1111/mec.13428.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. N. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 1847–1857. https://doi.org/10.1098/rstb.2005.1716.

Weigand, H., Beermann, A. J., Fedor, Č., Costa, F. O., Csabai, Z., Geiger, M. F., Rimet, F., Rulik, B., Strand, M., Szucsich, N., Weigand, A. M., Willassen, E., Wyler, A., Bouchez, A., Borja, A., Zuzana, Č., Ferreira, S., Dijkstra, K. B., Eisendle, U., Freyhof, J., Gadawski, P., Graf, W., Haegerbaeumer, A., Hoorn, B. B. V. D., Japoshvili, B., Keresztes, L., Keskin, E., Leese, F., Macher, J. N., Mamos, T., Paz, G., Pe, V., Maric, D., Andreas, M., Price, B. W., Rinkevich, B., Teixeira, M. A. L., Várbíró, G., & 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.

Winter, D. J. (2017). rentrez: an R package for the NCBI eUtils API. *The R Journal*, 9, 520–526.

Wong, E. H. K. & Hanner, R. H. (2008). DNA barcoding detects market substitution in North American seafood. *Food Research International*, *41*, 828–837. https://doi.org/10.1016/j.foodres.2008.07.005.

TABLE 1 Primers for amplifying the MiFish-12S metabarcode marker reference library. Positions and sizes are relative to the mitogenome of *Anguilla anguilla* (AP007233.1). Amplicon size includes primers.

Primer	Direction	Amplicon size (bp)	Position	Oligonucleotide 5′–3′	Annealing temp. (°C)	Reference
Aa22-PheF	Forward	612	22	AGCATAACACTGAAGATRYTARGA	53	This study
Aa633-12sR	Reverse	612	633	TTCTAGAACAGGCTCCTCTAG	53	This study
12S_30F	Forward	1,296	29	CACTGAAGMTGYTAAGAYG	50	Hänfling et al. (2016)
12S_1380R	Reverse	1,296	1,324	CTKGCTAAATCATGATGC	50	Hänfling et al. (2016)
MiFish-U-F	Forward	219	294	GTCGGTAAAACTCGTGCCAGC	60	Miya et al. (2015)
MiFish-U-R	Reverse	219	512	CATAGTGGGGTATCTAATCCCAGTTTG	60	Miya et al. (2015)
Li-F	Forward	721	294	GYCGGTAAAAYTCGTGCCAG	57	Stoeckle et al. (2018)
Li-R	Reverse	721	1,014	YCCAAGYGCACCTTCCGGTA	57	Stoeckle et al. (2018)

TABLE 2 Reference library coverage for eight commonly used metabarcode primer sets from 530 UK fish species accessed from GenBank (release 241) using the Meta-Fish-Lib pipeline. Cov. = coverage proportion of all, common (com.), and rare species; *n*=1 represents singleton species, i.e. proportion of species (*n*>0) represented with one sequence; Haps. = haplotypes per species (mean and median).

Locus	Primers	Reference	Total	Cov. (all)	Cov. (com.)	Cov. (rare)	*n*=1	Haps. (mean)	Haps. (med.)
12S	MiFish-U-F/MiFish-U-R	Miya et al. (2015)	2,171	0.7	0.76	0.67	0.25	1.7	1
12S	Tele02-f/Tele02-r	Taberlet et al. (2018)	2,171	0.7	0.76	0.67	0.25	1.7	1
12S	12S-V5f/12S-V5r	Riaz et al. (2011)	2,712	0.73	0.9	0.64	0.25	1.8	1
12S	L1848/H1913	Valentini et al. (2016)	1,859	0.58	0.68	0.53	0.34	1.3	1
16S	Fish16sFD/16s2R	Berry et al. (2017)	4,462	0.79	0.97	0.7	0.16	3.2	2
COI	mlCOIintF/jgHCO2198	Leray et al. (2013)	28,114	0.92	0.98	0.89	0.04	10.6	7
COI	FishF1/FishR1	Ward et al. (2005)	28,297	0.92	0.98	0.89	0.04	18.1	10
CYTB	L14912-CYB/H15149-CYB	Minamoto et al. (2012)	17,194	0.68	0.86	0.58	0.16	8.6	2

FIGURE 1 Simplified overview of bioinformatic workflow for the Meta-Fish-Lib pipeline. The pipeline runs as a series of executable R scripts for Mac and Linux. All logos and images are public domain and were obtained from https://www.wikipedia.org/ and http://www.phylopic.org/.

FIGURE 2 Examples of phylogenetic quality control output with taxonomically aware dereplication of sequences. Monophyletic species are coloured dark grey, non-monophyletic species blue, and interspecific shared haplotypes red. The first part of the branch tip label is the NCBI/BOLD database identifier for the representative sequence (mother); second part is species name; and third part is number of collapsed haplotypes, i.e. *n* dereplicated daughters belonging to that mother sequence. Panel (a) shows Clupeiformes sequences for the 12S-MiFish metabarcode (Miya et al., 2015), with two *Alosa* species sharing haplotypes, and *Sprattus sprattus* nested within *Clupea harengus*; (b) shows Ammodytidae sequences for the Leray et al. (2013) COI metabarcode, with monophyletic *Hyperoplus immaculatus* and *Ammodytes americanus*, non-monophyletic *A. marinus* and *H. lanceolatus*, and a *Conger conger* (Anguilliformes) sequence nested in *A. tobianus*; and (c) shows *Scyliorhinus* sequences for the Leray et al. (2013) COI metabarcode, with sequences of *Scyliorhinus stellaris*, *Squalus acanthias* and *Galeus melastomus* nested in *Scyliorhinus canicula*. Images are public domain and were obtained from http://www.phylopic.org/.

Figure 1

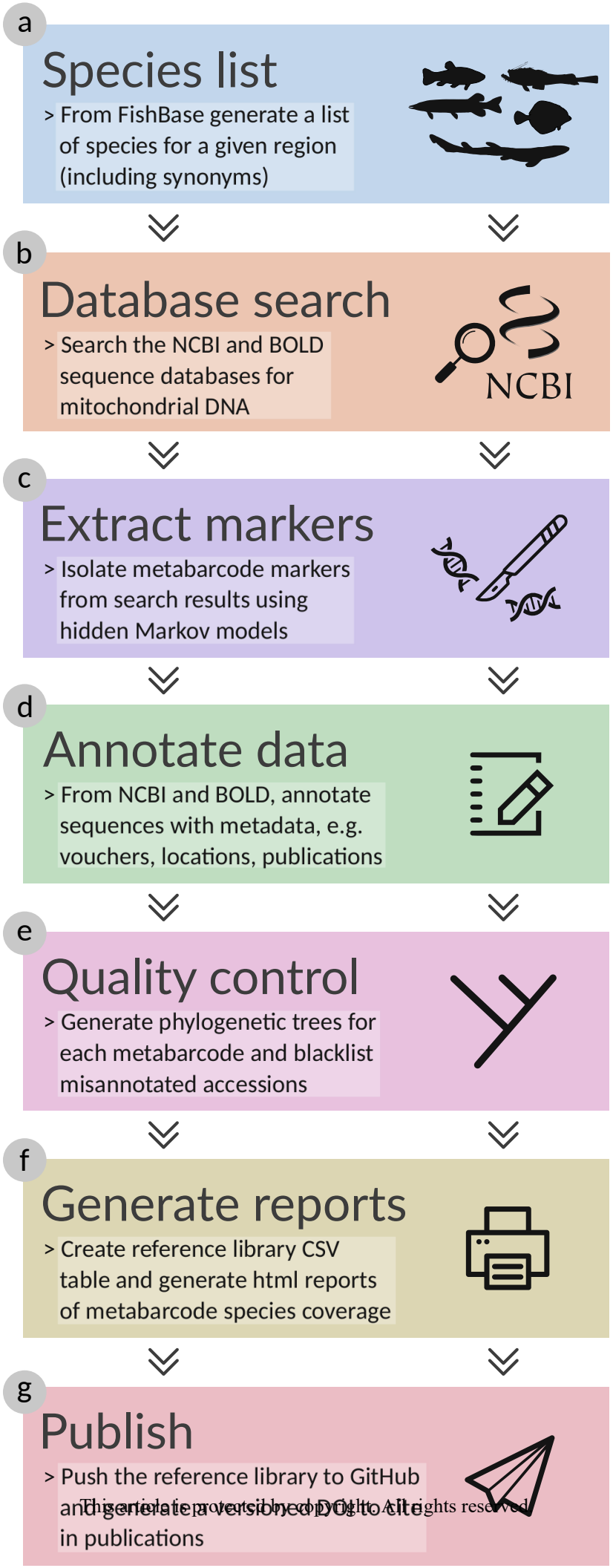

Figure 2